Optimization of the Holographic Process for Imaging and Lithography

by

José Antonio Domínguez-Caballero

M.S., Massachusetts Institute of Technology (2006)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2010

Signature of Author .....
Department of Mechanical Engineering
15 January 2010

Certified by.....
George Barbastathis
Associate Professor of Mechanical Engineering
Thesis Supervisor

Accepted by.....
David E. Hardt
Chairman, Department Committee on Graduate Students

Optimization of the Holographic Process for Imaging and Lithography

by

José Antonio Domínguez-Caballero

Submitted to the Department of Mechanical Engineering
on 15 January 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

Since their invention in 1948 by Dennis Gabor, holograms have demonstrated to be important components of a variety of optical systems and their implementation in new fields and methods is expected to continue growing. Their ability to encode 3D optical fields on a 2D plane opened the possibility of novel applications for imaging and lithography. In the traditional form, holograms are produced by the interference of a reference and object waves recording the phase and amplitude of the complex field. The holographic process has been extended to include different recording materials and methods. The increasing demand for holographic-based systems is followed by a need for efficient optimization tools designed for maximizing the performance of the optical system. In this thesis, a variety of multi-domain optimization tools designed to improve the performance of holographic optical systems are proposed. These tools are designed to be robust, computationally efficient and sufficiently general to be applied when designing various holographic systems. All the major forms of holographic elements are studied: computer generated holograms, thin and thick conventional holograms, numerically simulated holograms and digital holograms. Novel holographic optical systems for imaging and lithography are proposed.

In the case of lithography, a high-resolution system based on Fresnel domain computer generated holograms (CGHs) is presented. The holograms are numerically designed using a reduced complexity hybrid optimization algorithm (HOA) based on genetic algorithms (GAs) and the modified error reduction (MER) method. The algorithm is efficiently implemented on a graphic processing unit. Simulations as well as experimental results for CGHs fabricated using electron-beam lithography are presented. A method for extending the system's depth of focus is proposed. The HOA is extended for the design and optimization of multispectral CGHs applied for high efficiency solar concentration and spectral splitting. A second lithographic system based on optically recorded total internal reflection (TIR) holograms is studied. A comparative analysis between scalar and

vector diffraction theories for the modeling and simulation of the system is performed. A complete numerical model of the system is conducted including the photoresist response and first order models for shrinkage of the holographic emulsion. A novel block-stitching algorithm is introduced for the calculation of large diffraction patterns that allows overcoming current computational limitations of memory and processing time. The numerical model is implemented for optimizing the system's performance as well as redesigning the mask to account for potential fabrication errors. The simulation results are compared to experimentally measured data.

In the case of imaging, a segmented aperture thin imager based on holographically corrected gradient index lenses (GRIN) is proposed. The compound system is constrained to a maximum thickness of 5mm and utilizes an optically recorded hologram for correcting high-order optical aberrations of the GRIN lens array. The imager is analyzed using system and information theories. A multi-domain optimization approach is implemented based on GAs for maximizing the system's channel capacity and hence improving the information extraction or encoding process. A decoding or reconstruction strategy is implemented using the superresolution algorithm. Experimental results for the optimization of the hologram's recording process and the tomographic measurement of the system's space-variant point spread function are presented. A second imaging system for the measurement of complex fluid flows by tracking micron sized particles using digital holography is studied. A stochastic theoretical model based on a stability metric similar to the channel capacity for a Gaussian channel is presented and used to optimize the system. The theoretical model is first derived for the extreme case of point source particles using Rayleigh scattering and scalar diffraction theory formulations. The model is then extended to account for particles of variable sizes using Mie theory for the scattering of homogeneous dielectric spherical particles. The influence and statistics of the particle density dependent cross-talk noise are studied. Simulation and experimental results for finding the optimum particle density based on the stability metric are presented. For all the studied systems, a sensitivity analysis is performed to predict and assist in the correction of potential fabrication or calibration errors.

Thesis Supervisor: George Barbastathis
Title: Associate Professor of Mechanical Engineering

Acknowledgments

I dedicate this thesis to my wife Rebecca and to the Domínguez and Moon families. In particular, I would like to thank my parents, José and Elsa, my sister, Susana, and my brothers, Carlos and Javier, for their love and support throughout this journey.

I am very grateful with my advisor, Professor George Barbastathis, for his guidance, support and friendship. I also want to acknowledge all the members of the 3D Optical Systems Group, in particular to Nick Loomis, Laura Waller, Se Baek Oh, and Lei Tian for the interesting discussions and their help in some of the experimental work. Thanks to Satoshi Takahashi for his work on the fabrication of the optimized computer generated holograms.

Many thanks to Professors Jerome H. Milgram and Cabell Davis for their support during my Master of Science thesis work.

I am very thankful to our collaborators from Samsung Electronics Co. Ltd., in particular to Dr. Sung Jin Lee for his contribution to the work on the design of holographic systems applied to high resolution lithography.

Thanks to Professor Mark A. Neifeld and Dr. Michael D. Stenner for their collaboration on the work on the design and optimization of holographically corrected segmented aperture thin imagers.

Thanks to Professor Rajesh Menon for the collaboration on the design of solar concentrators based on computer generated holograms.

Finally, thanks to Dennis Gabor for inventing holography!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In 1948 Dennis Gabor invented optical holography. This invention, for which he won the Nobel Prize for Physics in 1971, was a result of an "exercise in serendipity", as Gabor explains in his autobiography. Optical holography was initially presented in the context of electron microscopy as a method to resolve the problem introduced by the spherical aberrations of electron lenses that set the limit in the resolving power [1]. Two detailed papers, [2], [3], followed his pioneering work, which explored presenting holography as a method for recording and reconstructing the amplitude and phase of a wave field. He invented the word "holography" from the Greek words "holos" meaning "whole" and "graphein" meaning "to write". Holography only attracted mild interest until the 1960s, when the concept became popular, improved predominantly by the invention of the laser.

The holographic process consists of two main steps: recording and reconstruction. In the recording step, a hologram is produced by the interference between an optical field scattered by a coherently illuminated object and a reference wave. In the conventional form, the interference pattern is recorded on a photosensitive film which is later chemically processed. This process efficiently encodes the 3D information from the object wave onto a 2D hologram plane. The interference with the reference wave allows recording both the phase and amplitude of the complex field scattered by the object. In the reconstruction step, the hologram is illuminated by the phase conjugate of the reference wave, known as

the probing wave, and diffracts the phase conjugate of the object wave that propagates through free space and forms a real image of the object. Additional diffraction orders, such as the virtual image, direct component and halo, are also produced. The virtual image is a diverging wave that appears to emanate from virtual point sources behind the hologram. It can be converted into a real image with the help of a lens. The direct component is the un-deviated diffraction order that propagates in the same direction as the probing wave. The halo is a divergent wave that depends on the intensity distribution of the object wave alone and propagates in the same direction as the probing wave.

In the original system proposed by Gabor, the reference wave is incident normal to the recording medium and a semi-transparent object is placed in its path. This is known as in-line geometry. This geometry attracted considerable interest due to its simplicity and efficient way to perform lens-less imaging. However, this geometry only yields good results when the object to be imaged is sufficiently transparent and small so that it does not significantly disturb the reference wave. Another disadvantage of this geometry is that all the diffraction orders produced during the reconstruction step co-propagate, making it difficult in some cases to recover the encoded signal. Leith and Upatnieks made significant advancements to Gabor's technique by introducing the off-axis geometry [4], [5]. In the off-axis geometry, the hologram is recorded with a reference wave (or object wave) tilted respect to the normal of the holographic medium surface. This variation introduces a high-frequency carrier signal that allows the different diffraction orders to separate during the reconstruction step. Undesirable diffraction orders can then be filtered out and the desired signal can be recovered.

A further improvement became possible when holography began to be analyzed from the viewpoint of communications theory. The holographic problem then became that of optimizing the corresponding encoding and decoding processes. This parallelism between holography, communications and systems theories allowed borrowing notions such as information transfer, channel capacity, carrier and modulating signals, impulse response and transfer function, useful in designing and analyzing holographic systems. For exam-

ple, the hologram recording process is analogous to Amplitude Modulation (AM). It was not until this analogy was found by Leith and Upatnieks that the fundamental virtual image problem present in Gabor's holograms was resolved. Other examples of systems that can be treated as communication channels include imaging and lithographic systems based holograms. In the imaging case, the channel's source is the object or scene and the receiver can be modeled as the final captured image. In the lithographic case, the information about a desired mask needs to be transferred to the final exposed pattern at the photoresist plane. In both cases, the equivalent communication channels may be subject to noise and additional system constraints, which limit its corresponding channel capacity as formulated by Shannon [6]. The addition of holography in such systems allow the manipulation of the information transfer process with great flexibility and by means of a properly designed optimization algorithm, maximal information extraction and transfer can be pursued.

In the following years, the holographic process was extended to include different recording mediums, as well as forms to implement the recording and reconstructions steps. These new methods can be broadly classified according to their type of recording and reconstruction processes as optical and numerical. In the first variation, holograms are recorded and reconstructed optically. This variation includes conventional thin and volume holograms. These holograms are recorded in a variety of materials such as silver halide holographic films, photopolymers and photosensitive crystals. The second variation consists of holograms recorded optically but reconstructed using numerical methods. This belongs to the field of digital holography. In digital holography, the holograms are recorded on a photosensitive detector such as a CCD or CMOS sensors. Numerical algorithms simulate the reconstruction process which includes probing the hologram and performing the corresponding free space propagation. In the third variation, the holograms are designed numerically and reconstructed optically. This belongs to the field of computer generated holography. Computer generated holograms provide great flexibility in their design but are limited by the chosen fabrication technique. The last modal-

ity corresponds to holograms recorded and reconstructed numerically. This modality is primarily used in the design, analysis and optimization of holographic optical systems.

Holographic methods have been adopted in a variety of optical systems used for imaging and display applications. Imaging systems include diagnostic tools designed to acquire useful information about the object or scene that would be otherwise difficult to obtain by conventional means. Examples of imaging systems include digital holographic cameras designed to study aquatic objects [7], biological cells [8] and complex flow distributions that require tracking 3D particle fields [9]; volume holographic imaging of 3D objects [10], holographic optical storage [11] and aberration correction of complex optical systems based on phase conjugation holography [12]. In contrast, display applications exploit the hologram's ability to reconstruct 3D optical fields. Examples of these applications include holographic art, 3D television based on computer generated holograms [13], and holographic lithography [14].

Over the years, holographic elements have demonstrated to be important components on a variety of optical systems and their application in new fields and methods is expected to grow. The increasing demand of holographic based systems is followed by a need for efficient optimization tools designed for maximizing the performance of the system. The different variations of holographic methods require distinct formulations that can model the system under study with sufficient accuracy. Scalar, as well as vector, diffraction theories may be implemented. In general, optical systems based on holographic elements introduce new degrees of freedom that can be exploited by the optimization algorithm. A multi-domain optimization approach is required in which variables from different domains, such as optical, geometrical, material and numerical domains are optimized in parallel to efficiently explore new regions of the optimization space. The optimization tools are required to be robust and computationally efficient. These tools should account for experimental or fabrication related error. In addition, systems analyzed using information theory can then be treated as communication channels and their performance or channel capacity can be maximized. This is equivalent of optimizing the encoding and

decoding strategies and hence maximizing the information transfer between object and image spaces.

In this thesis, a variety of multi-domain optimization tools designed to improve the performance of holographic based optical systems are presented. These tools are designed to be robust, computationally efficient and sufficiently general for the design of a variety of optical systems. All the major variations of holographic elements are studied as indicated in Table 1.1. Novel holographic based optical systems for imaging and lithography are proposed. The design, optimization and experimental implementation of these systems are conducted. An accurate model of each system is performed, as well as a sensitivity analysis to account for potential fabrication or calibration errors. The models are studied using system and information theories. The system's channel capacity and stability of the related inverse problem as a function of various control parameters are evaluated. The performance of the corresponding signal encoding and decoding processes is maximized.

Table 1.1: Holographic Processes.

Recording	**Reconstruction**	**Techniques**	**Thesis Chapter**
Numerical	*Optical*	Computer Generated Holography	2
Numerical	*Numerical*	Holographic System Design	3
Optical	*Optical*	Conventional Holography	4
Optical	*Numerical*	Digital Holography	5

1.1 Thesis Overview

In Chapter 2, the design, optimization and implementation of computer generated holograms (CGHs) that operate in the Fresnel diffraction zone is presented. In order to achieve high diffraction efficiencies, only phase CGHs are considered. Two target applications are studied: high-resolution, non-contact lithography and solar concentration. A reduced complexity hybrid optimization algorithm (HOA) is proposed. This algorithm is based on genetic algorithms and the modified error reduction method for the solution of the highly nonlinear, multivariable problem, subject to stringent constraints. The goal of

the HOA is to maximize the performance of the signal encoding process during the CGH design. This accounts for maximizing the information transfer between the hologram and reconstruction spaces. The complexity of the optimization problem is reduced by optimizing a significantly smaller set of variables by the introduction of the local diffuser phase elements (LDPE) and local negative power elliptical phase elements (LNPEPE) masks. The design of a lithographic system based on CGHs is conducted using a multi-domain optimization approach in which optical, numerical, and material parameters such as the photoresist response are considered in parallel. Three geometries are studied: in-line, off-axis and total internal reflection (TIR). The HOA is efficiently implemented on a graphics processing unit (GPU), resulting in speedups of more than 200× compared to standard central processing unit (CPU) implementations. A multiplexing method for the extension of the depth of focus is proposed to increase the tolerance of potential axial misalignments during the photoresist exposure process. A simple CGH fabrication method based on electron-beam lithography is presented, as well as an optimization scheme designed for the local correction of over and under dose errors. Experimental demonstrations of the reconstructions from the fabricated CGHs using coherent and partially coherent illuminations are presented. A sensitivity analysis is conducted to predict and assist in the correction of potential fabrication error. The presented algorithm is extended for the design and optimization of multispectral CGHs applied for high efficiency solar concentration and spectral splitting.

In Chapter 3, the design and optimization of a TIR holographic system for photoresist exposure in the Fresnel diffraction zone is presented. In contrast to the system designed in Chapter 2, this lithographic system is based on optically recorded and reconstructed holograms. The holograms are recorded on a photopolymer and operate in the near ultraviolet regime. The target application considered is a high-resolution, parallel exposure, non-contact, large area flat panel display manufacture. A comparative analysis is performed between scalar and vector diffraction theories for the modeling and simulation of the system under study. The considered methods are: Rayleigh-Sommerfeld diffrac-

tion theory, rigorous coupled wave analysis and finite-difference time-domain methods. Scalar diffraction theory is proven to be sufficiently accurate for the considered geometry and is chosen for modeling and simulating the system. The system is again designed using a multi-domain optimization approach in which optical, geometrical and material parameters are considered. First order models for simulating the material response and shrinkage of the photopolymer are presented. A novel block-stitching algorithm is introduced for the calculation of large diffraction patterns that allows overcoming current computational limitations of memory and processing time. The numerical model is implemented for optimizing the system's performance, as well as redesigning the mask to account for potential fabrication errors. The simulation results are compared to experimentally measured data. A method for extending the depth of focus of the system is presented.

In Chapter 4, the design, optimization and implementation of a high-resolution segmented aperture thin imager based on holographically corrected gradient index (GRIN) lenses is presented. The proposed compound imager utilizes a GRIN lens array to collect the light emanating from the scene, imaging it onto a photodetector effectively reducing the thickness of the system to less than 5mm. Optically recorded holographic elements based on a phase conjugation scheme are used for correcting the high-order aberrations present in the GRIN lens array. Each lens in the array is corrected for a different field angle resulting in diffraction limited performance. The new degrees of freedom introduced by the holographic elements are utilized for maximizing the information transfer from scene to measurement spaces. The optical performance of the system is evaluated using a combination of Matlab and the optical design software Zemax. The imager is model used system's theory by defining a linear operator known as the Hopkins matrix. This matrix connects the input and output signals (scene and measurement) and includes optical, geometrical and detector related parameters such as space-bandwidth product, pixel size and dynamic range. The system is analyzed using information theory treating it as a Gaussian parallel communication channel. The channel capacity as well as inver-

sion stability of the Hopkins matrix are studied. The Hopkins matrix is then treated as an alternate projection that allows extracting more scene information than conventional isomorphic methods, as well as tolerating higher levels of noise and calibration errors. A multi-domain optimization approach is implemented based on GAs for maximizing the system's channel capacity and hence improving the information extraction or encoding process. The optimized system is proven to be more efficient than conventional microlens array based compound systems. A decoding or reconstruction strategy is implemented using the superresolution algorithm. In this strategy, the captured subimages from each GRIN lens are combined to reconstruct a high-resolution output image. Experimental results for the optimization of the hologram's recording process are presented. It is found that the hologram exhibits properties of a volumetric element such as angular selectivity. The exposure and chemical processing of holograms recorded on silver halide emulsions is studied by optimizing the diffraction efficiency at the Bragg angle. A tomographic technique based on the Foucault knife-edge test is presented for the measurement of the system's space-variant point spread function. A sensitivity analysis is performed to estimate the effect of potential misalignment errors in the assembly process. A modification of geometry for polychromatic imaging is proposed.

In Chapter 5, a stochastic theoretical model based on the stability metric for the design and optimization of in-line digital holographic particle imaging velocimetry (DHPIV) systems is presented. DHPIV systems are used for the characterization of complex fluid flows by tracking micron sized tracing particles. It is found that the performance of such systems relies on the correct selection of parameters such as particle density, geometry and detector related parameters. Similar to Chapter 4, the model of the digital hologram recording process is derived based on system's theory by defining the instantaneous Hopkins matrix linear operator. This operator includes optical, geometrical and detector related parameters that connect the measured signal with the particle cloud inside the volume of interest (VOI) at a given instant in time. The theoretical model is first derived for the extreme case of point source particles using Rayleigh scattering and scalar diffrac-

tion theory formulations. The model is then extended to account for particles of variable sizes using Mie theory for the scattering of homogeneous dielectric spherical particles. The system is analyzed from an information theoretic viewpoint treating it as a Gaussian parallel communications channel. The influence and statistics of the particle density dependent cross-talk noise are studied. A metric is defined similar to the channel capacity to study the stability of the associated inverse problem. The stability metric is used for optimizing the system, maximizing the amount of 3D information from the VOI that can be encoded by a single hologram. Simulation results are presented for finding the optimum particle density based on the stability metric. An experimental evaluation is conducted to study the influence of particle density in the information extraction process by a set decoding process. The implemented decoding strategy is based on a template matching scheme designed to automatically process and count the particles present at each frame. The experimental results are compared to the predictions obtained from the stability metric.

Chapter 2

Design, Optimization and Implementation of Fresnel Domain Computer Generated Holograms: Applied to Holographic Lithography and Solar Concentration

2.1 Introduction to Computer Holography

Computer-generated holograms (CGHs) are diffraction optical elements designed to reconstruct a semi-arbitrary 2D or 3D optical field in the near, Fresnel or Fraunhofer diffraction zones as shown in Figure 2-1. The design and optimization of CGHs is performed using numerical methods allowing the use of ideal wavefronts during the encoding process, as well as choosing the most adequate encoding strategy for the given application. Typical problems that arise during the optical recording step of conventional holograms, such as aberrations of the optical components used in the system, vibrations, thermal instabilities and recording material properties (e.g. shrinkage, diffusion and lifespan),

can be avoided. CGHs are fabricated using a variety of methods such as computer driven plotter writing [15], [16], direct laser writing [17], [18], femtosecond laser micromachining [19], [20], contact or projection lithography [21], [22], [23] and electron-beam (e-beam) lithography [24], [25]. Similar to conventional holography, the reconstruction or decoding step is performed optically using a quasi-monochromatic, spatially coherent laser source. In the design of CGHs, the wave propagation between the hologram and reconstruction planes is modeled for the given system geometry with the appropriate vector or scalar diffraction theory. The inverse problem nature in the optimization of CGHs often requires field backpropagation or inverse scattering from the desired intensity or field at a given plane.

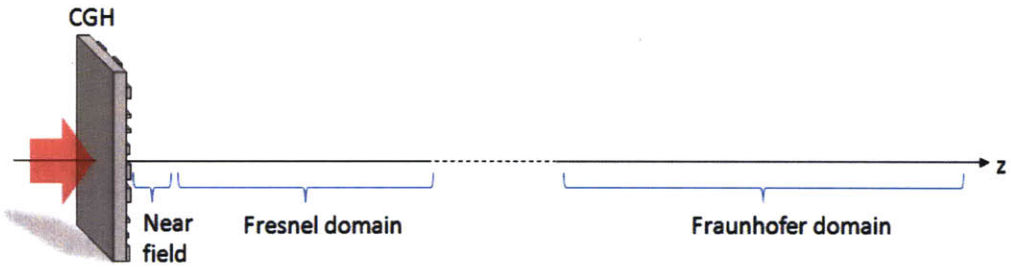


Figure 2-1: CGH diffraction zones.

The first CGH was made by Brown and Lohmann in 1966 [26]. In the following year, they presented a more rigorous paper that described in detail the working principle and required approximations for the design of a Fraunhofer CGH that is known today as the "detour phase" type [27]. As with many other researchers in those early years, Brown and Lohmann were originally motivated by the fabrication of 2D optical spatial filters similar to those used for optical signal processing [28] and imaging.

CGHs are broadly classified as point-wise or cell oriented. In the point-wise type, the transmittance function of the CGH is discretized into $N \times M$ pixels, each pixel corresponding to at least one degree of freedom (DOF). The required space-bandwidth product is typically set to match the sampling requirements and with sufficient DOF as

those from the desired reconstructed intensity. In the cell oriented type, the CGH is divided in to multiple cells, each composed of a given structure, which again corresponds to the different DOF available to transform the field transmitted by the hologram.

CGHs are also classified according to the transmittance function's type: pure phase, pure amplitude or complex (amplitude and phase) [29], [30]]. Ideal pure phase CGHs, neglecting Fresnel reflections or scattering from material defects, transmit all the power from the input field with zero absorption. These holograms are desirable to achieve high diffraction efficiency reconstructions. However, pure phase holograms rely on an efficient encoding process as the specified amplitude information (desired intensity distribution at the photoresist plane) needs to be converted into pure phase information (encoded signal) at the hologram plane. Errors in this information transfer results in noisy reconstructions with low diffraction efficiencies. The above requirement is relaxed if the hologram is of the complex type, as the desired signal can be encoded in both the amplitude and phase of the CGH's transmittance function, effectively increasing the number of DOF by a factor of 2. However, the amplitude component of the transmittance function absorbs a large portion of the input energy.

The number of DOF available on a CGH greatly depends on the method selected for its fabrication. Some fabrication techniques, such as e-beam writing, restrict the hologram's transmittance function to a binary phase or amplitude in order to avoid multi-exposure processes subject to severe misalignment errors. For a point-wise binary CGH, the total number of DOF is: $N \times M$ ($N \times M$ bits of recordable information). In contrast, multi-level CGHs quantize the amplitude or phase information using K discrete levels. The total number of DOF for a point-wise multi-level CGH is: $\log_2(K) \times N \times M$. Multi-level CGHs have been demonstrated by patterning subwavelength structures based on effective medium theory as described in [31].The effect of discrete phase or amplitude levels in the transmittance function results in quantization errors that produce poor quality reconstructions with speckle-like grainy noise [32]. Several algorithms such as error diffusion [33], pulse-density modulation [34], and phase retrieval methods [35] have

been studied to attempt to improve the signal encoding process for constrained discrete-level holograms.

An example of a cell oriented Fraunhofer CGH is the detour phase hologram. Detour phase holograms consist of many transparent dots (or apertures) on an opaque background as shown in Figure 2-2-a. To an approximation, the width of the aperture is proportional to the amplitude of the desired signal's Fourier transform, and its lateral shift respect to the center of the cell is proportional to the transform's phase. The resulting phase shift can be understood intuitively by considering the phase difference of waves originated from two separated Huygens sources [36]. The detailed derivation of the design equations for detour phase holograms is beyond the scope of this thesis and can be found at [27]. Detour phase holograms are modeled using scalar diffraction theory with the approximation of representing the amplitude reconstructed by the hologram in the form of a Fourier series which is then equated to a Fourier series representation of the image. Detour phase CGHs typically produce low diffraction efficiency reconstructions as a large fraction of the input energy gets absorbed by the opaque background. In addition, this type of holograms are very sensitive to positional errors from the plotter or alternative fabrication methods that result in undesirable phase shifts on the diffracted wavefront [37], [38]. Figure 2-2-b shows the reconstructed intensity from the detour phase hologram of Figure 2-2-a.

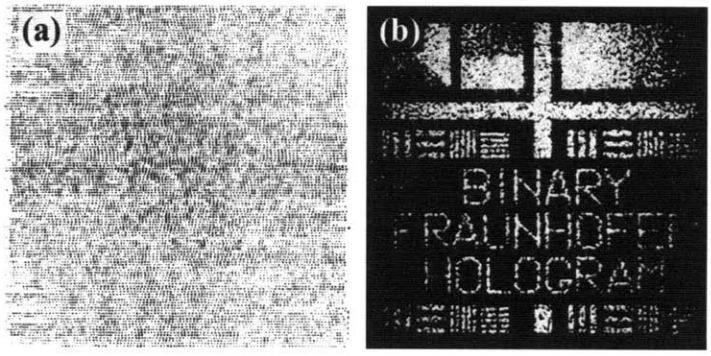


Figure 2-2: (a) Detour phase hologram; (b) Reconstructed intensity [39].

In 1977, Lee proposed one the first point-wise CGHs based on the concept of interferograms [15]. Lee realized that the apertures in a detour phase binary hologram are equivalent to the fringes of an interference pattern from conventional off-axis holographic recording. In Lee-type holograms, the positions and widths of the fringes were determined and the CGH was printed using a conventional plotter.

Kinoform CGHs were introduced by Lesem, Hirsch and Jordan in 1969 [40]. Kinoforms are point-wise pure phase Fraunhofer holograms that are designed under the assumption that most of the information on the signal can be encoded in the phase of the hologram's transmittance function. Deviations from this assumption produce low quality reconstructions as the amplitude of the signal to be encoded at the hologram plane is discarded. In contrast to the holograms described previously, Kinoforms only reconstruct a single diffraction order upon on-axis illumination. However, phase matching errors that may arise during the fabrication process resulting in noisy reconstructions - the hologram resembles an in-line conventional hologram in which undesirable diffraction orders, such as the conjugate image, direct component (DC) term and halo, co-propagate towards the reconstruction plane. To simplify the hologram design, the desired object intensity is divided into small points (or apertures) and the field at the hologram plane is calculated using scalar diffraction theory. The final design is printed using a multi-level grey scale plotter and then is photoreduced and transferred onto a photoresist which is then bleached to produce a surface relief pattern. Due to the phase nature of the transmittance function, Kinoforms can achieve much higher diffraction efficiencies than previous holograms. Figure 2-3 shows an example of a typical Kinoform CGH with its corresponding reconstruction.

Over the past several years, CGHs have been developed for a broad range of applications such as beam shaping [41], [42], [43], optical trapping [44], optical signal processing [45], [46], optical communications [47], optical testing [48], [49], 3D displays [13], [50], and lithography [51].

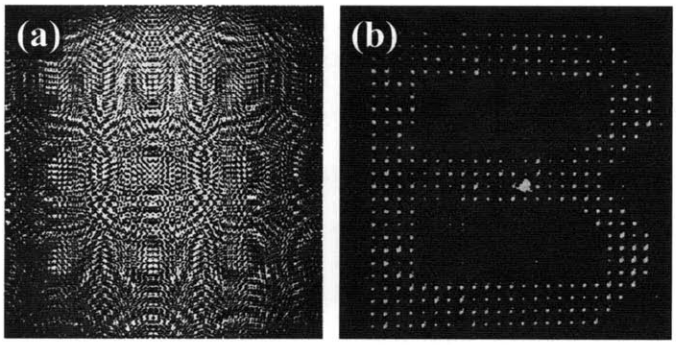


Figure 2-3: (a) Kinoform phase hologram; (b) Reconstructed intensity [40].

2.2 Motivation and Problem Definition

The increasing demand for smaller sized features, lower power and large working area semiconductor devices has led to the development of novel lithographic techniques aimed to replace conventional lithographic methods. Serial processes such as direct laser writing [52] and electron beam lithography [53] have been found to be expensive and time consuming which prevents their implementation in the mass production of semiconductor devices. In addition, serial methods require high-resolution motorized stages and are restricted to operate in small working areas to avoid stitching and positional errors. Some of these limitations are overcome by parallel processes such as Zone-Plate-Array Lithography (ZPAL) [54]. In ZPAL, an array of zone plates is used to write multiple points on the substrate in parallel. However, the required system is costly and still requires scanning. Other parallel exposure methods include contact [55] and projection lithography [56]. In contact lithography, an amplitude mask with the desired pattern is placed in direct contact with the substrate to be exposed. This method is not suitable for mass production as it is very sensitive to contaminants and degradation of the mask, due to the direct contact with the photoresist. Projection lithography is a non-contact method that involves imaging the desired pattern onto the photoresist plane. Projection lithographic systems are costly, as they are composed of multi-element optical systems

designed to demagnify the projected pattern, as well as correct for intrinsic optical aberrations. In order to achieve high-resolution exposures, expensive optical systems with large numerical apertures (NA) are required.

Holographic lithography based on computer generated holograms is a promising candidate to replace standard 2D or 3D lithographic methods. The main features provided by holographic lithography are:

- **Non-contact**: large working distances that prevent damaging the substrate or CGH. Allows high throughput, ideal for mass production
- **Parallel exposure**: fast processing as scanning requirements can be minimized or avoided
- **High-resolution**: CGHs can be designed to have a large effective NA when operating in the Fresnel diffraction zone
- **Large working area**: large area exposures are possible, ideal for applications such as large area flat panel displays (LCD) manufacture
- **Depth of focus (DOF) control**: CGHs can be multiplexed and optimized to extend the system's DOF in order to tolerate potential misalignments of the substrate during exposure
- **2D or 3D patterning**: holograms can be encoded to reconstruct 2D or 3D optical fields
- **Standard manufacture**: CGHs are manufactured using conventional 2D lithographic techniques
- **Robust design**: holograms can use their entire surface to encode the information required for the reconstruction of the desired pattern. Contaminants or manufac-

ture errors in a given section of the CGH have minimal effects on the reconstructed pattern

- **Cost effective**: simplified, compact system that doesn't require costly optical components

An implementation example of a CGH for on-axis lithography is shown in Figure 2-4-a. For 2D lithographic exposures, the CGH and substrate to be exposed are placed parallel to each other and separated by a given working distance. For Fresnel CGHs, small working distances of a few tens of microns are desirable in order to achieve large effective NAs. The CGH is then probed by a quasi-monochromatic, spatially coherent plane wave. The diffracted optical field propagates in free space and produces the desired intensity distribution at the photoresist plane where a substrate coated with photoresist is placed. Upon exposure, the substrate undergoes the standard developing and etching processes to produce the final pattern as shown in Figure 2-4-b.

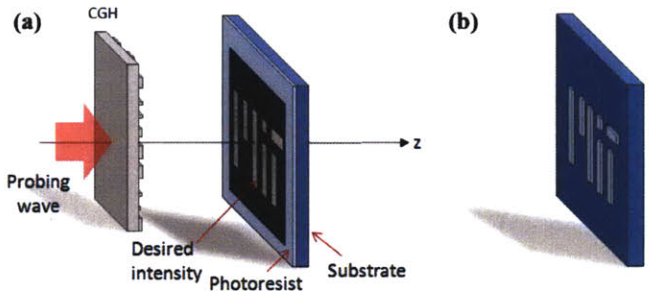


Figure 2-4: (a) In-line photoresist exposure process; (b) Final pattern after chemical post-processing.

Previous attempts in implementing CGHs for 2D lithography include the work done by Jacobsen and Howells as described in [57], [58], [59]. Their holograms were of the complex type (amplitude and phase modulation) designed to operate in the x-ray regime ($\lambda = 1 - 5$nm). Hologram designs were investigated for working distances ranging from 50μm to 200μm. They proposed to fabricate their holograms using carbon or tungsten

with at least 16 thickness levels and pixel sizes of 13nm. No experimental demonstration was conducted and the provided simulations showed low diffraction efficiency, poor quality reconstructions. Their CGH encoding strategy was based on a point-wise design optimized using a simplified version of the error reduction algorithm. Moreover, their proposed fabrication method is found to be extremely complex requiring a very small pixel size and overlay accuracy better than 16nm. Another attempt was done by Wyrowski as described in [60], [61], [62]. Their proposed CGHs were also of the complex type with 4 phase levels (0, 90, 180 and 270 degrees of phase shift) and two amplitude levels (0 and 1). Their holograms were optimized for a single wavelength (λ = 365nm) using a projection onto convex sets method designed for a gap of 50μm. For the reconstruction, they used a discrete polychromatic, partially incoherent source with wavelengths: 365nm (60%), 405nm (15%) and 436nm (25%). Their holograms were fabricated using electron beam writing as well as ion beam etching with a pixel size of 1μm. The presented results were extremely low quality, again suffering from low diffraction efficiencies. Severe chromatic aberrations were present due to their choice of illumination source. Their encoding strategy and choice of initial guess in the optimization algorithm resulted in the convergence of non-optimum solutions. Additional work on extreme ultra-violet holographic lithography includes [63]. CGHs have also been proposed for 3D lithographic exposures [64], [65], [66]. These holograms were designed using an analytical approach based on cylindrical waves with coordinate transformations for the reconstruction of line segments on a non-planar surface. The CGHs are of the complex type (binary phase and pseudo grayscale amplitude) and were fabricated using conventional photolithographic methods (pixel size of 5μm). Their presented results are low resolution (projected lines of $\sim$100μm), and suffer from non-uniformities and stitching errors. In addition, their optimization algorithm is not capable of designing holograms that project arbitrary patterns so a more general optimization algorithm is required.

From the previous discussion it can be seen that an efficient design, optimization and implementation of CGHs for lithographic applications is of upmost importance. The

choice of encoding strategy and optimization technique is crucial to achieving efficient transfer of information from the hologram to photoresist planes. The problem in hand is not simple. Its non-linear nature, large number of degrees of freedom (or decision variables) and stringent constraint result in a searching space filled with local minima threatening to trap most optimization algorithms. In addition, the selected optimization technique needs to be flexible and accommodate different CGH designs for the reconstruction of arbitrary patterns at the photoresist plane, as well as be computationally efficient. The second aspect to consider in the design of CGHs is the choice of a simple fabrication strategy, in order to reduce the required number of fabrication steps and potential manufacture errors. The selected fabrication method is closely related to the optimization algorithm by the given optimization parameters and system constraints. The development of systematic characterization techniques is also required for the evaluation of fabricated CGHs and the optimization of the fabrication procedure. Finally, a sensitivity analysis is necessary in order to estimate the effect and assist in the correction of potential manufacture errors.

In this chapter, we propose a hybrid optimization algorithm (HOA) based on genetic algorithms (GAs) and the modified error-reduction (MER) method for the efficient design and implementation of Fresnel domain point-wise, pure binary phase CGHs for 2D high-resolution lithographic exposures. To our knowledge, this is the first implementation of GAs in the optimization of large CGHs in the Fresnel region. As described in the following sections, the complexity of the optimization algorithm is reduced by the introduction of specially designed phase masks, namely the local diffuser phase elements (LDPE) and local negative-power elliptical phase elements (LNPEPE) masks, with a reduced number of degrees of freedom that improve the transfer of information between the hologram and photoresist planes. Different objective functions are studied for guiding the optimization algorithm to produce high diffraction efficiency holograms that reconstruct diffraction-limit resolution patterns, with low mean-square error of the difference between the desired and diffracted intensity distributions, at the photoresist plane. The HOA is implemented

and tested on a graphics processing unit (GPU) resulting in significant speedups (over 150 times) compared to conventional central processing unit (CPU) implementations. Three different geometries are studied and compared: in-line, off-axis, and total-internal reflection (TIR). A fabrication process based on electron-beam lithography, as well as methods for the evaluation and characterization of the fabricated CGHs, are proposed and experimentally tested. A detailed sensitivity analysis is performed to predict the effect of potential manufacture errors such as over/under dose, proximity effects, phase, position and stitching errors. The proposed algorithm is extended and applied for the optimization of multispectral CGHs to be used for solar concentrator systems.

2.3 System Geometries

Three system geometries for the implementation of CGHs for 2D high-resolution lithography are studied: in-line, off-axis, and total-internal reflection (TIR). Figure 2-5 shows the in-line geometry. In this geometry, a quasi-monochromatic spatially coherent plane wave propagating parallel to the optical axis illuminates the CGH, resulting in both desirable and undesirable diffraction orders. These orders co-propagate a given working distance towards the output plane. A substrate coated with photoresist is placed at the output plane which gets exposed by the resulting intensity distribution. This geometry is similar to the conventional in-line hologram as originally proposed by Gabor. Gabor holograms encode the desired signal as a form of interferograms (holograms) produced by the superposition of an in-line reference wave and desired signal,

$$\begin{aligned} I(x,y) &= |1+O(x,y)|^2 \\ &= 1+|O(x,y)|^2+O(x,y)+O(x,y)^*, \end{aligned} \tag{2.1}$$

where O is the signal to be encoded at the hologram plane. This choice of encoding strategy results in undesirable diffraction orders, such as the direct component (DC) term, halo and twin image that correspond to the first three terms on the right-hand-side of equation

2.1 respectively. For Fresnel domain holograms with short working distances, the effect of these undesirable orders (in particular the twin image) is very severe, resulting in noisy reconstructions. CGHs provide substantial flexibility in the choice of encoding strategy as they are designed numerically and not by an optical recording process. As discussed in the following sections, Gabor type encoding is not optimal and better encoding strategies can be implemented during the CGH design. Despite the choice of encoding strategy, it is the main task of the optimization algorithm to minimize the effect of the undesirable diffraction orders on the exposed pattern, while satisfying the stringent constraints that limit the amount of recordable information on the hologram ($N \times M$ bits for a binary phase CGH).

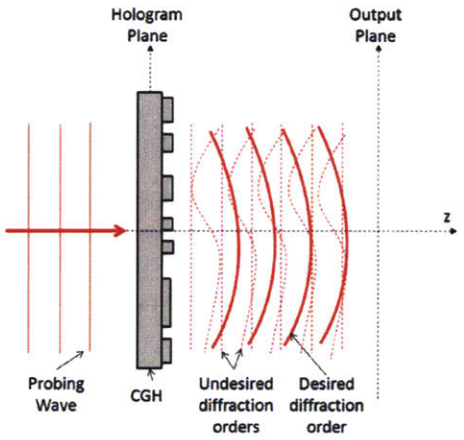


Figure 2-5: In-line geometry.

The CGH size and working distance determine the effective numerical aperture (NA) of the system,

$$NA_{eff} \approx \sin\left[\tan^{-1}\left(\frac{H_{size}}{2d}\right)\right], \tag{2.2}$$

where H_{size} is the lateral size of the CGH (a square hologram is assumed), and d is the working distance between the hologram and photoresist planes. For in-line CGHs, the choice of pixel size is relatively flexible. However, in order to prevent distortions due to aliasing, the maximum pixel size needs to be chosen to satisfy the corresponding sampling

requirements. From the Whittaker-Shannon sampling theorem [67], the maximum spatial frequency tolerated before aliasing is,

$$u_{\max} = \frac{1}{2\delta_{pix}}, \tag{2.3}$$

where δ_{pix} is the CGH pixel size. For low resolution reconstructions a large pixel size can be used; however, large pixels reduce the number of degrees-of-freedom available on the CGH, resulting in low quality reconstructions. The smallest pixel size allowed is typically determined by the selected fabrication technique. For e-beam lithography, experimental fabrication of sub-10nm features have been reported [68].However, small pixels significantly increase the fabrication time and cost. The choice of pixel size determines a second numerical aperture enforced to satisfy the sampling requirements,

$$NA_{SR} = u_{\max}\lambda, \tag{2.4}$$

where λ is the operating wavelength. As the hologram operates in the Fresnel regime, the system is ultimately limited by the evanescent cut-off,

$$u_{ev} = \frac{1}{\lambda}. \tag{2.5}$$

The system's diffraction limit resolution is given by,

$$\Lambda = 0.5\frac{\lambda}{NA}, \tag{2.6}$$

where NA is the most restrictive numerical aperture as given by equations 2.2 and 2.4. The diffraction limit resolution is independent of the type of encoding strategy used, as illustrated in the following example. Consider a point-wise multi-level phase hologram designed to reconstruct two points separated at the photoresist plane. The system parameters are: $\lambda = 500$nm, $H_{size} = 150\mu$m, $\delta_{pix} = 100$nm, and $d = 150\mu$m. From equation 2.2, the effective numerical aperture is: $NA_{eff} = 0.5$. The maximum spatial frequency

allowed as given by equation 2.3 is: $u_{\max} = 5000\text{mm}^{-1}$ (corresponds to $NA_{SR} = 2.5$, that is non-physical). The diffraction limit resolution is: $\Lambda = 500\text{nm}$. For the first encoding strategy, the phase of the CGH is given by,

$$\theta(x,y) = \arg\left\{\exp\left[ik\left(d + \frac{y^2}{2d}\right)\right]\left[\exp\left(\frac{ik}{2d}(x-x_o)^2\right) + \exp\left(\frac{ik}{2d}(x+x_o)^2\right)\right]\right\}, \tag{2.7}$$

where x_o is the point's lateral shift respect to the optical axis (10μm for this example), and $k = 2\pi/\lambda$. The CGH's phase distribution is shown in Figure 2-6-a. Figure 2-7 shows the corresponding reconstructed intensity at the photoresist plane. As can be seen, the half width of the resulting point spread function (PSF) agrees with the diffraction limit resolution as predicted by equation 2.6. It is also noted that this choice of encoding strategy failed to eliminate undesirable diffraction orders that produce additional peaks on the image plane. Figure 2-6-b shows the magnitude of the spectrum of the hologram of Figure 2-6-a. This shows that the spectrum doesn't entirely cover the system's pass-band as given by the evanescent cut-off of equation 2.5 - $u_{ev} = 2000\text{mm}^{-1}$ for this example. Other encoding strategies might be able take advantage of these unused frequency components to help eliminate the undesirable diffraction orders and reduce the mean-square error (MSE) of the reconstructed intensity.

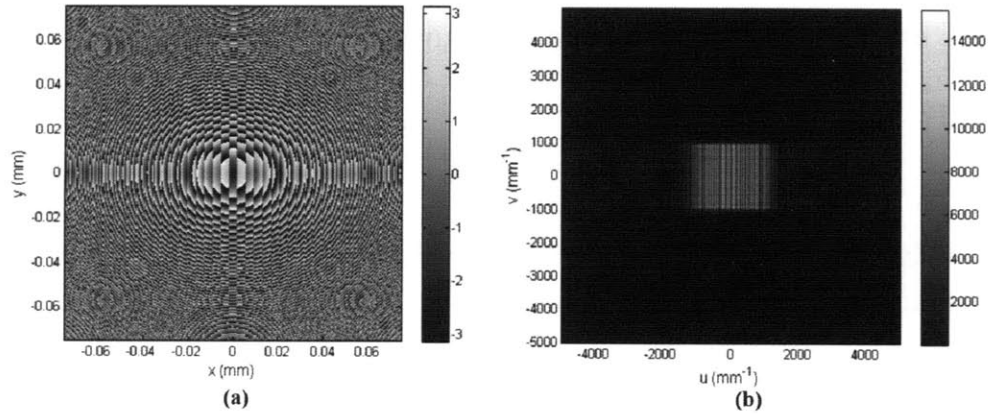


Figure 2-6: (a) CGH computed from first encoding strategy; (b) CGH spectrum.

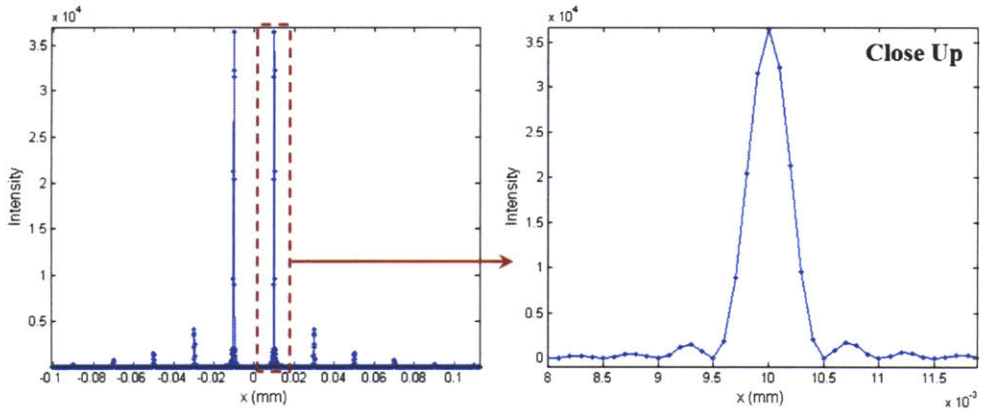


Figure 2-7: Reconstructed intensity from CGH designed using the first encoding strategy.

The second encoding strategy is based on the modified error-reduction optimization algorithm that will be discussed in the next section. Figure 2-8-a shows the resulting CGH's phase distribution. The corresponding reconstructed intensity is shown in Figure 2-9. As can be seen, the resulting PSF is still diffraction limited as predicted by equation 2.6. However, this choice of encoding strategy helped eliminate the undesirable diffraction orders at an expense of diffraction efficiency. This is the well known tradeoff between diffraction efficiency and uniformity in designing CGHs. Figure 2-8-b shows the corresponding magnitude of the CGH's spectrum. This encoding strategy better utilizes the entire system's pass-band.

The three frequency cut-offs discussed above are shown in Figure 2-10. The evanescent cut-off is fixed with the choice of operating wavelength. The geometry and sampling cut-offs are free to move according the desired hologram size, pixel size and working distance.

Additional advantages of the in-line geometry include the simplicity in the setup, making the system cost effective. Also, a compact collimation section may be used to reduce the system's overall size.

The off-axis geometry is shown in Figure 2-11. In this geometry, an off-axis plane wave illuminates the hologram, reconstructing desirable and undesirable diffraction orders that propagate in different directions. An aperture stop is placed between the hologram

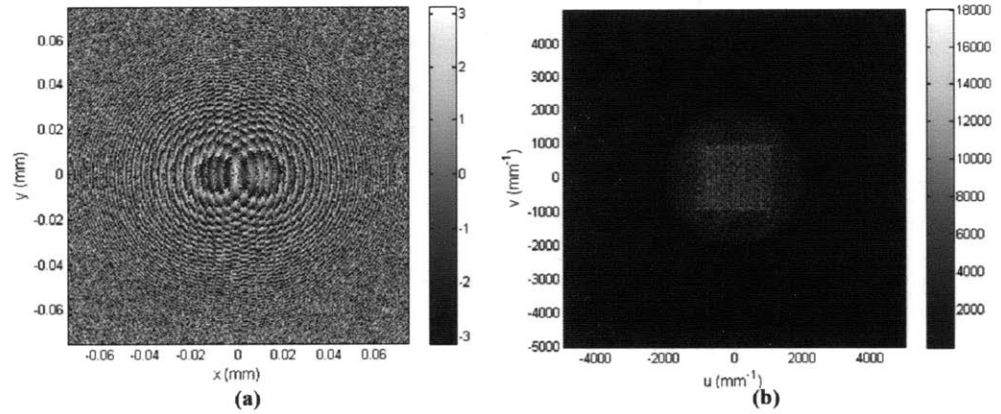


Figure 2-8: (a) CGH computed from second encoding strategy; (b) CGH spectrum.

and output planes to block the undesirable diffraction orders. However, the aperture stop can only be used in geometries with long working distances and with sufficient separation between the desirable and undesirable orders. Another technique is to place the off-axis CGH at the front focal plane of a 4f system as shown in Figure 2-12. An aperture stop is placed at the Fourier plane, one focal length behind the first lens, and is used to block the spectral components corresponding to undesirable diffraction orders.

The encoding process in off-axis CGHs is similar to amplitude modulation (AM) in which the signal is modulated by a high-frequency carrier. In conventional off-axis holography as proposed by Leith and Upatnieks [4], [5], the high-frequency carrier signal is given by the off-axis reference wave and its frequency is proportional to the off-axis angle. In this encoding method, the desired signal is again recorded by an interferogram,

$$\begin{aligned} I(x,y) &= |R(x,y)+O(x,y)|^2 \\ &= |R(x,y)|^2+|O(x,y)|^2+R^*(x,y)O(x,y)+R(x,y)O(x,y)^*, \end{aligned} \tag{2.8}$$

where O is the desired signal at the hologram plane and R is an off-axis plane wave given

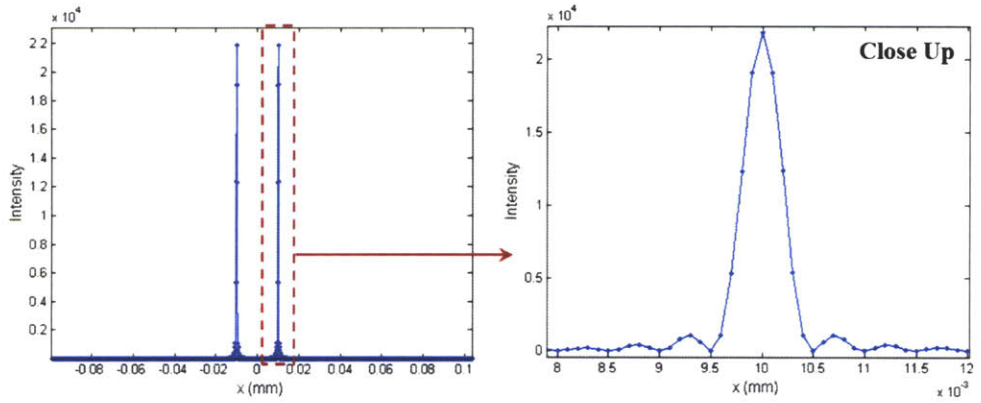


Figure 2-9: Reconstructed intensity from CGH designed using the second encoding strategy.

by,

$$R(x,y) = \exp\left[i\frac{2\pi}{\lambda}\left(\alpha x + \beta y + \gamma z\right)\right], \tag{2.9}$$

where α, β and δ are the plane wave's direction cosines referenced to the x, y and z axes respectively. Similar to the in-line case, this geometry results in undesirable diffraction orders (DC term, halo and twin image); however, these orders are now spectrally separated as shown in Figure 2-13. As discussed previously, the encoding strategy of equation 2.8 is not unique - other encoding methods can be implemented to improve the hologram's diffraction efficiency.

Off-axis CGHs are particularly useful in cases in which the designed hologram has low diffraction efficiency and the reconstruction noise, due to undesirable diffraction orders, needs to be filtered out. A less sophisticated optimization algorithm can be used as higher errors are tolerated. Two main disadvantages of off-axis CGHs are the necessity of an aperture stop or 4f system and the restrictions on the allowable CGH pixel size. In order to encode the high-frequency carrier, the pixel size is restricted to,

$$\delta_{pix} \leq \frac{\lambda}{2\sin\theta}, \tag{2.10}$$

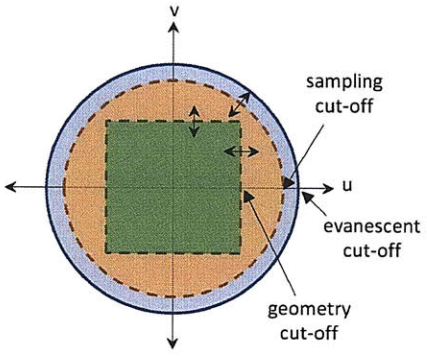


Figure 2-10: Representation of the CGH's frequency cut-offs.

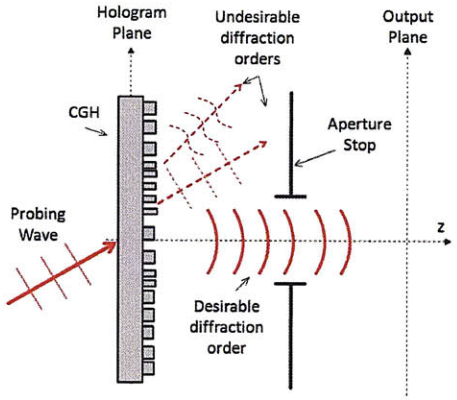


Figure 2-11: Off-axis CGH geometry.

where θ is the off-axis propagation angle (modulation in only one lateral direction is assumed). For example, consider a CGH modulated by an off-axis plane wave incident at 50 degrees with wavelength: $\lambda - 350$nm. This restricts the CGH's pixel size to: $\delta_{pix} \leq 228$nm. An additional restriction is imposed on the signal's bandwidth. For the encoding strategy of equation 2.8, the bandwidth constraints are,

$$B_x \leq \frac{\sin\theta}{3\lambda} \text{ or } B_x \leq u_{\max}^{fab} - \frac{\sin\theta}{\lambda}, \tag{2.11}$$

whichever is more restrictive. The first constraint prevents overlapping between the

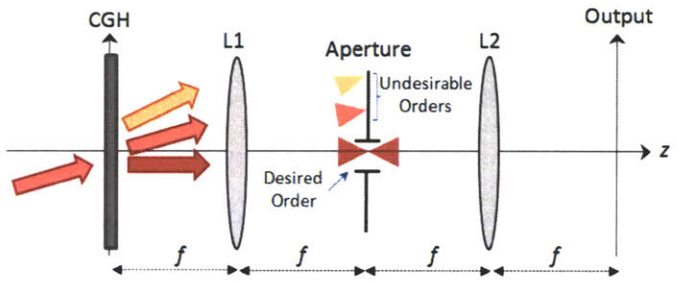


Figure 2-12: Off-axis geometry. Filtering of the undesirable diffraction orders using a 4f system.

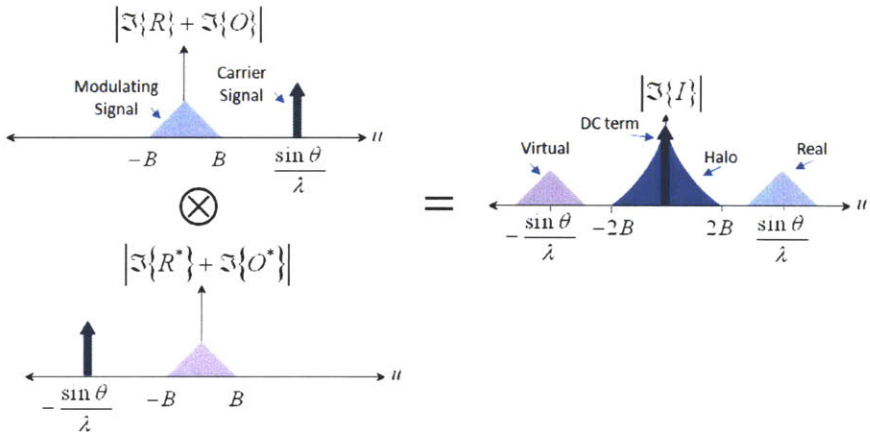


Figure 2-13: Spectral representation of off-axis geometry modulation process.

different diffraction orders (Figure 2-13). The second constraint ensures that the signal lies within the pass-band set by the maximum spatial frequency encoded on the CGH. This maximum spatial frequency is set by the chosen fabrication method. For e-beam writing, assuming a minimum pixel size of 50nm, $u_{\max}^{fab} = 10,000\text{mm}^{-1}$. For the previous example, the maximum signal's bandwidth on the x-direction is $B_x = 730\text{mm}^{-1}$, which corresponds to a minimum features size exposed of 0.7μm. Finally, off-axis CGHs are more susceptible to fabrication errors due to the small pixel size requirements.

Figure 2-14 shows the TIR geometry. In this geometry, a right angle prism is used

to couple the probing wave that propagates at an angle equal or larger than the critical angle, θ_c, required for TIR at the dielectric-air interface. The CGH and prism have approximately the same index of refraction and are held together using, for example, optical glue. The hologram is also modulated by a high-frequency carrier signal, similar to the off-axis geometry. However, no aperture stop or 4f system is required as the hologram is designed such that the undesirable diffraction orders get totally reflected and exit in a direction conjugate to the probing wave. The desired signal suffers frustrated TIR and propagates through free space towards the output plane.

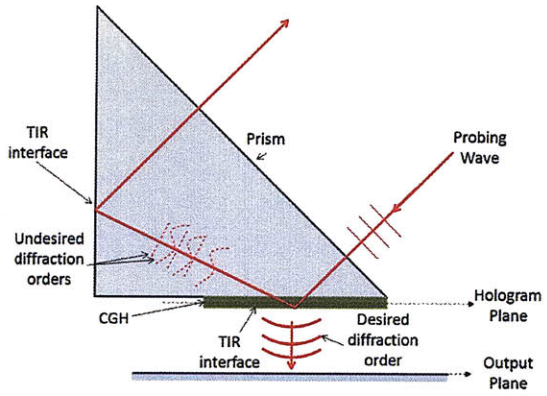


Figure 2-14: Total internal reflection geometry.

The pixel size in this geometry has the same restriction as equation 2.10; however, the wavelength is replaced by the effective wavelength inside the prism, $\lambda_{eff} = \lambda/n$, where n is the prism's refractive index, and the angle of incidence is constrained to be equal or larger than the critical angle, $\theta \geq \theta_c$, where, $\theta_c = \arcsin(1/n)$.

Figure 2-15 shows the spectral representation of the reconstruction (demodulation) process for the encoding strategy of equation 2.8. During reconstruction, the probing wave down shifts the spectrum of the desired signal to fit inside the frustrated TIR window. Only the signal within this pass-band will suffer frustrated TIR and escape

towards the photoresist plane. The bandwidth constraints for this geometry are,

$$B_x \leq \frac{\sin\theta_c}{\lambda_{eff}} \text{ or } B_x \leq \frac{\sin\theta - \sin\theta_c}{2\lambda_{eff}}, \tag{2.12}$$

where the first constraint ensures that the encoded signal fits within the frustrated TIR pass-band and the second one prevents the undesirable diffraction orders from leaking out towards the output plane. The maximum encoded frequency is also limited by the method used to fabricate the hologram.

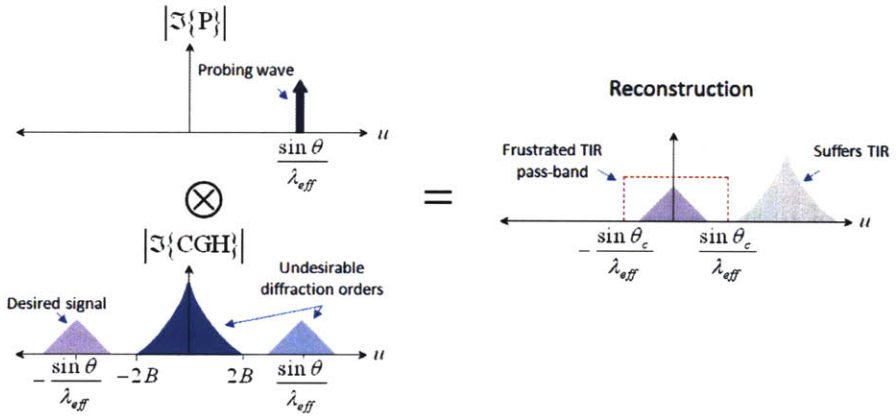


Figure 2-15: Spectral representation of TIR demodulation process.

The TIR geometry is particularly useful when the designed CGH has low diffraction efficiency and the undesirable orders need to be filtered out. In contrast to the off-axis case, this geometry allows short working distances as no additional aperture stop is required. Disadvantages of TIR CGHs include the challenging fabrication due to the small pixel size requirement and the need of a prism.

2.4 Optimization of Computer Generated Holograms

CGHs are designed and optimized using numerical methods. In this thesis, we are interested in point-wise pure phase CGHs. An optimization algorithm is required to compute

the optimum phase distribution at the hologram plane that reconstructs a desired signal at the photoresist plane. For simple reconstruction patterns, such as spots or lines, an analytic approach may be used to calculate the hologram's phase distribution [64], [65], [66].

Optimization techniques identified in the current literature can be classified into three main search classes: calculus-based, enumerative, and random [69]. Calculus-based methods subdivide into two classes: indirect and direct. In the indirect approach, a local extrema is searched by solving a nonlinear set of equations that result from setting the gradient of the objective function equal to zero. Direct methods are local search methods that seek optima by hopping on the function and moving in a direction related to the local gradient. Calculus-based methods are typically local in scope; they are used, for example, to seek a local minimum that is the best in a neighborhood of the current point. The choice of initial searching point is very important; if the algorithm starts searching too far from the global optima, it might never reach it and instead get trapped at a local extrema. Finally, calculus-based methods depend on the existence of derivatives, i.e. well-defined slope values. Enumerative schemes are designed to search the optimization space by evaluating the objective function at every point in the space, one at a time. Once all the points have been computed, the global extrema is extracted. Although enumerative methods are simple to implement and typically consider finite and discrete searching spaces, they are extremely inefficient and computationally expensive. Random search methods subdivide into two classes: fully stochastic and randomly guided. Fully stochastic methods search the optimization space randomly and save the best value found at each step of the random walk. These methods are also inefficient and in the long run are expected to perform as well as enumerative schemes. Random guided methods search the space following a deterministic set of rules that are randomly guided. They are typically robust and tolerate multimodal, noisy and discontinuous search spaces. As it will be explained later, genetic algorithms are a type of randomly guided search method.

Several iterative based algorithms have been proposed for the optimization of the

CGHs' complex transmittance function. One example is the direct binary search (DBS) method originally designed for binary amplitude Fraunhofer domain holograms [70]. This local search method consists of flipping the binary amplitude at each pixel sequentially and computing the corresponding error at the reconstruction plane after each flip. The algorithm stops when no flips are required over an entire iteration. DBS methods suffer from getting trapped at a local extrema; a combination with a random method can be used for escaping from the local minima to get closer to global or near global optimum. In addition, this method requires a large number of operations which makes it computationally inefficient with a performance close to that of an enumerative method. The error-reduction [71] and error diffusion [72] methods are other examples of local search iterative algorithms used for computer holography. In the error-reduction (ER) method, the algorithm iterates between the hologram and reconstruction planes imposing a set of constraints at each domain. In this method, the mean-square error is guaranteed to decrease or remain the same. This algorithm is also subject to becoming trapped at a local minimum and is very sensitive to the choice of the initial searching point. Some of the main advantages of this method include the fast convergence and simple implementation and enforcement of constraints. A variation of this method will be discussed in more detail later. Other variations of the ER method include the original Gerchberg-Saxton algorithm [73] and Fienup variants (input-output and hybrid input-output) [74]. Error-diffusion methods attempt to solve the quantization problem in CGHs by distributing the quantization error of a given pixel into its neighbors. Other calculus-based methods include projection onto constraint sets [75] and projection onto convex sets [64], [65]. Simulated annealing (SA) is an example of a random guided method also used for the design of CGHs [76]. This method seeks to minimize the energy of the system similar to a metallurgic annealing process, where the global minimum energy or ground state of the physical system is reached by cooling the system from a high temperature to a low temperature with a controlled cooling schedule. SA mimics this process by randomly perturbing the optimization variables, evaluating the energy function (objec-

tive function) and computing the energy difference, ΔE, between the current and past energy states (previous iteration). If $\Delta E < 0$, the perturbation is accepted unconditionally; otherwise, the perturbation is accepted based on the probability distribution $P = 1/[1 + \exp(\Delta E/T)]$, where T is the temperature parameter control based on a given schedule. This random guided scheme accepts perturbations that might increase the error helping the algorithm to avoid becoming trapped at a local extrema. However, SA still performs a local search in which the choices of initial searching point and cooling schedule significantly affect the algorithm's convergence.

2.4.1 Optimization Problem Abstraction

In order to understand this optimization problem and to propose an efficient scheme for the design and optimization of CGHs, an abstraction of the problem is performed. The problem is described in the context of holographic lithography; however, the proposed optimization scheme can be easily generalized to a broad range of applications as discussed later. When designing CGHs, the choice of encoding strategy is crucial to achieving an efficient transfer of information from the hologram to photoresist planes. The corresponding optimization problem is highly non-linear, noisy (quantization errors, etc.), with large number of optimization variables (or degrees of freedom), and is subject to stringent constraints imposed mostly by our choice of CGH fabrication method. The following are the desired features for the optimization scheme:

- Optical efficiency: optimized CGHs should have high diffraction efficiencies and be capable of reconstructing high-resolution, uniform patterns at the photoresist plane
- Computational efficiency: problem complexity should be minimized to decrease the computational time and required memory for the optimization of large CGHs
- Broad searching range: multi-point parallel optimization with broad coverage of the searching space

- Multiple constraints: allow the flexible implementation of multiple constraints and boundary conditions
- Flexible objective function: flexible choice of objective function to guide the algorithm and overcome system tradeoffs such as diffraction efficiency versus uniformity
- Avoid getting trapped at local minima: provide a way to escape from local minima and continue the search towards global minima
- Robust: algorithm independent of the choice of signal to be reconstructed at the photoresist plane

In addition, we would like to minimize the processing steps required for the fabrication of CGHs based on e-beam lithography making the system cost effective and suitable for mass production. The optimization of the fabrication procedure is beyond the scope of this thesis; however, our choice of fabrication method has a direct impact on the set of constraints and boundary conditions applied in the numerical optimization problem.

The problem geometry is shown in Figure 2-16. The hologram plane is divided by $M \times N$ pixels with pixel sizes δ_{pix_x} and δ_{pix_y}. From now on, only square CGHs with a uniform pixel size will be considered, i.e. $M = N$ and $\delta_{pix} = \delta_{pix_x} = \delta_{pix_y}$. The phase distribution of the CGH is truncated to the hologram window of size $H_{size} = H_{size_x} = H_{size_y}$. The reconstruction plane is located at a parallel plane separated by a distance d and is discretized by $M' \times N'$ pixels with pixel sizes δ'_{pix_x} and δ'_{pix_y}. For numerical convenience, only square reconstruction planes are considered and the pixel size is set to be the same as that of the hologram plane: $M' = N'$, and $\delta_{pix} = \delta'_{pix_x} = \delta'_{pix_y}$. The desired signal is restricted to an object window of size: $O_{size} = O_{size_x} = O_{size_y}$. In general, the CGH is characterized by the complex transmittance function H; however, we will restrict our analysis to pure phase holograms, $|H| = 1$ (zero absorption constraint), and neglect any losses from Fresnel reflections or other sources of scattering. The complex field at the reconstruction plane is given by R, where the amplitude and phase of the desired signal may be specified.

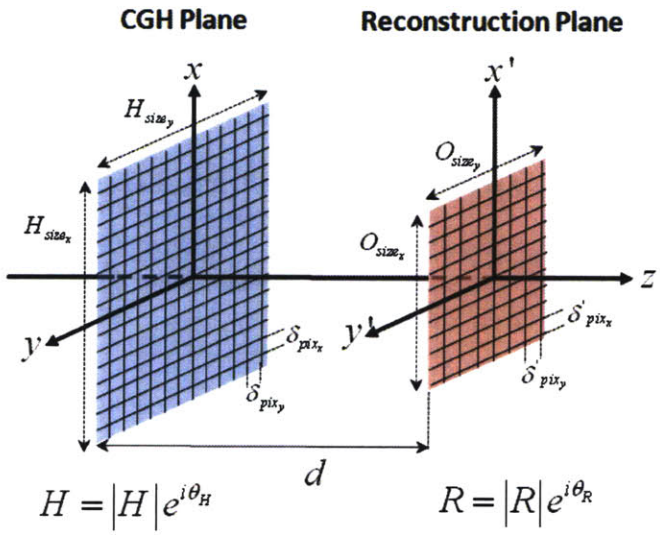


Figure 2-16: Problem geometry.

In addition to the zero absorption constraint, the hologram's phase distribution is constrained to be binary in order to avoid multi-exposure e-beam processes that would lead to severe fabrication errors. The number of degrees of freedom (DOF) resulting from the binary constraint is,

$$DOF_{binary} = M \cdot N. \tag{2.13}$$

In contrast, for a multi-level CGH the number of DOF is,

$$DOF_{multi-level} = \log_2(K_{levels}) \cdot M \cdot N, \tag{2.14}$$

where K_{levels} is the number of discrete levels used to quantize the phase – for $K_{levels} = 2$, equation 2.14 reduces to equation 2.13.

An additional constraint is imposed to the amplitude of the reconstructed field,

$$|R| = \sqrt{I_{des}}, \tag{2.15}$$

where I_{des} is the desired intensity distribution to be exposed at the photoresist plane. The phase of the field at the reconstruction plane is unconstrained and acts as a free parameter that we will exploit for the encoding of the desired signal, I_{des}, at the hologram plane.

The field propagation between the hologram and reconstruction planes is done using the Rayleigh-Sommerfeld diffraction formula valid under the scalar diffraction theory (SDT) approximation [67]. For geometries with a large working distance, d, compared to the CGH's lateral dimension, H_{size}, a Fresnel approximation of the diffraction formula is used,

$$\begin{aligned} R\left(x',y';z\right) &= \frac{e^{ikz}}{i\lambda z}\iint_{-\infty}^{\infty} D(x,y)e^{i\frac{\pi}{\lambda z}\left[\left(x'-x\right)^2+\left(y'-y\right)^2\right]}dxdy \\ &= D\left(x',y'\right)\otimes h\left(x',y';z\right), \end{aligned} \tag{2.16}$$

where $k = 2\pi/\lambda$ is the wavenumber, $\otimes$ represents a 2D convolution operation, h is the free-space propagation point spread function (PSF),

$$h\left(x',y';z\right) = \frac{e^{ikz}}{i\lambda z}e^{i\frac{\pi}{\lambda z}\left(x'^2+y'^2\right)}, \tag{2.17}$$

and D is the diffracted field at the hologram plane. For the in-line geometry, assuming a thin transparency approximation [67], the diffracted field is,

$$D(x,y) = H(x,y) = e^{i\theta_H}. \tag{2.18}$$

For the off-axis and TIR geometries,

$$D(x,y) = e^{i\frac{2\pi}{\lambda}(\alpha x+\beta y+\gamma z)}H(x,y), \tag{2.19}$$

where α, β and γ are the direction cosines of the probing plane wave as shown in Figure 2-17.

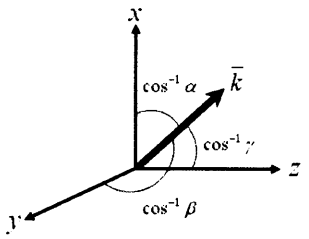


Figure 2-17: Direction cosines.

The Fresnel approximation is valid when the desired working distance satisfies the following condition,

$$z^3 >> \frac{\pi}{4\lambda}\left[\left(x^{'}-x\right)^2+\left(y^{'}-y\right)^2\right]^2_{\max}. \tag{2.20}$$

For short working distances that do not satisfy the condition of equation 2.20, the PSF of equation 2.17 is replaced with the exact form of the PSF,

$$h\left(x,y;z\right)=\frac{z}{i\lambda}\frac{e^{i\frac{2\pi}{\lambda}\sqrt{x^2+y^2+z^2}}}{x^2+y^2+z^2}. \tag{2.21}$$

The CGH optimization problem can be represented as a combination of encoding and decoding processes as shown in Figure 2-18 (for the in-line geometry case). This representation will help us understand the key aspects necessary for an efficient optimization scheme. The encoding process begins at the reconstruction plane, where the amplitude of the field is given by the square-root of the desired intensity (signal to be encoded) and the phase is a free parameter that depends on our choice of encoding strategy. Selecting a phase distribution at this plane is analogous to placing a phase mask, $P_{mask} = \exp(i\theta_{PM})$, on top of the desired amplitude mask, $\sqrt{I_{des}}$. Some examples of phase masks that result in efficient encoding strategies will be discussed later. The phase, θ_{PM}, is unconstrained (although discretized) which results in a large number of DOF. The next step of the encoding process is to propagate the field from the reconstruction to the hologram plane

by means of a "back-propagation" operator,

$$O^{-1}_{Fresnel}\{f(x,y)\} = f(x,y) \otimes h(x,y;-z). \tag{2.22}$$

In general, the resulting field at the hologram plane, $\hat{H}$, will have a non-uniform amplitude and phase that violates the zero absorption and binary constraints stated above. The last step in the encoding process consists of enforcing the hologram constraints by discarding the amplitude, $\left|\hat{H}\right|$, and binarizing the phase. The purpose of the decoding process is to retrieve the encoded signal. In contrast to the encoding process which is done numerically, the decoding process is performed optically – a method also known as optical reconstruction. The first step in the decoding process is to illuminate the CGH with the probing beam resulting in a diffracted field, $D(x,y)$, at the hologram plane. The diffracted field propagates through free space represented mathematically by the forward-propagation Fresnel operator, $O_{Fresnel}\{\}$. Finally, the modulus square of the reconstructed field, R, is computed giving rise to the estimated intensity (decoded signal) at the reconstruction plane.

A successful optimization algorithm seeks to converge to an optimum phase mask, P_{mask}, that maximizes the information transfer from amplitude (reconstruction plane) to phase (hologram plane) while satisfying the zero absorption and binary constraints,

$$\begin{aligned} \left|\hat{H}(x,y)\right| &\rightarrow 1, \\ \theta_{\hat{H}} &\rightarrow \theta_H. \end{aligned} \tag{2.23}$$

In other words, we are interested in finding a complex function pair, R and H, such that the amplitude of R is given and H is pure phase. Both functions are restricted to the object and hologram windows respectively and are related by the Fresnel operator. The solution of this problem is nonunique. The mismatch between the limited number of DOF at the hologram plane and the large number of DOF at the reconstruction plane makes the problem challenging. Errors during the encoding process in satisfying the zero absorption

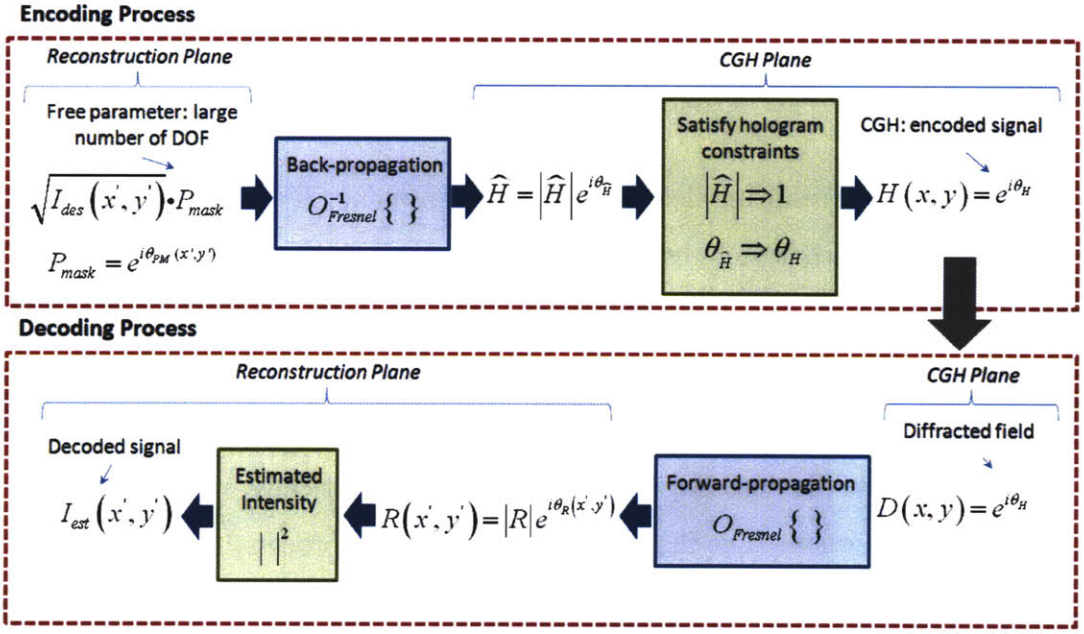


Figure 2-18: CGH encoding and decoding processes.

constraint result in noisy, low diffraction efficiency reconstructions. The violation of the binary constraint results in quantization errors that also degrade the reconstructed signal. In addition, the energy of the encoded signal must lie within the system pass-band (evanescent cut-off) to achieve high diffraction efficiency reconstructions.

2.4.2 Reduced Complexity Optimization

Holographic lithography requires the optimization of large CGHs designed for applications such as flat panel display manufacture. In addition, a large space-bandwidth product (SBP) is necessary to achieve high-resolution exposures. The number of DOF of the corresponding optimization problem is extremely large, requiring long computational times, and quickly becomes unmanageable with current computational means. Memory limitations set an upper bound on the maximum CGH size that can be handled. The optimization hypercube can only be searched locally (or with a narrow searching space)

as each searching point is composed of $M \times N$ or $M' \times N'$optimization variables, depending on whether the algorithm directly optimizes the CGH or the phase mask, P_{mask}, as shown in Figure 2-18.

These computational limitations raise the following question: can we reduce the complexity of the optimization problem and hence reduce the number of variables while still being capable of optimizing a large CGH? We found that the answer to this question is "yes", provided that we utilize some additional information about the system. The additional information that we are going to use is the type of signals encoded for holographic lithography applications. The signals of interest are binary and can be decomposed in elementary geometries such as squares, lines and circles. This is because the reconstructed intensity is used to expose a photoresist with a given contrast curve producing a binary pattern after the developing and etching processes. Our choice of signal type is not restrictive to holographic lithography. Additional applications, such as solar concentration and optical trapping, utilize similar type signals as will be explained later.

In the following sections we present two methods designed to reduce the complexity of the optimization problem, while still providing efficient encoding strategies for the design of large CGHs. These methods are based on the introduction of a specially designed phase mask, P_{mask} as in Figure 2-18, with a reduced number of DOF. The proposed phase masks are: Local Diffuser Phase Elements (LDPE) Mask and Local Negative Power Elliptical Phase Elements (LNPEPE) Mask.

Local Diffusers Phase Elements Mask

The use of deterministic or pseudorandom diffusers in holography has been studied by many researchers, including Leith and Upatnieks [77], Gerritsen [78], Gabor [79], Hirsch [80], and Katyl [81]. In conventional holography, diffused illumination of the object is used to increase the hologram's diffraction efficiency and to allow multiple viewpoints of the reconstructed image increasing the system's parallax. In addition, diffusers allow reducing the required dynamic range of the recording material by spreading the

information uniformly - each point in the object is effectively recorded over the entire hologram and thus, any subdivision of the hologram will reconstruct the complete object (restricted only by diminished viewing angles and resolution) [82], [83]. On the other hand, diffusers typically spread the energy of the signal outside the system's pass-band, resulting in speckle-like granular noise in the reconstruction. This mottled granular noise structure of the optical wavefront affects the reconstruction, making it difficult to resolve the fine details of the image close to the fundamental resolution limit of the system. For this reason, several deterministic diffusers with phase distributions that vary widely in form and/or motivation have been proposed. The diffusers implemented in conventional holography have the disadvantage that their elaborate phase distributions need to be fabricated with high accuracy. In contrast, the diffusers used in computer holography are only implemented numerically and hence sophisticated phase distributions are acceptable. The use of diffusers not only improves the hologram's diffraction efficiency but also reduces quantization errors that arise from the hologram calculation and fabrication [84], [85]. Previous diffusers proposed for Fraunhofer CGHs use uniformly distributed random phases that extend the signal's spectrum over the entire Fourier plane. The finite size of the hologram limits the amount of information that can be recorded, giving rise to errors in the reconstructed signal.

In contrast to previous work, we propose a phase mask that is composed of several small local diffusers designed to operate in the Fresnel domain. The diffusivity of each local element can be controlled independently by the optimization algorithm. We call this phase mask: "Local Diffusers Phase Elements (LDPE) Mask". During the numerical encoding process, the LDPE mask is point-wise multiplied by the desired amplitude signal at the reconstruction plane and then back-propagated to the hologram plane. The amplitude of the resulting field is discarded and the binary constraint is imposed on its phase. The encoding process is illustrated in Figure 2-19 for the case of an amplitude signal containing the letters "MIT". The modulus square of the amplitude signal is the desired intensity distribution to be exposed at the photoresist.

The LDPE mask improves the encoding process by facilitating the transfer of the amplitude information of the desired signal at the reconstruction plane to pure phase information at the hologram plane. This is done by back-propagating a field that has nearly uniform amplitude at the hologram plane as required by the zero absorption constraint. In addition, the energy of the signal is distributed evenly over the entire hologram window resulting in a smoother power spectrum that improves the hologram's diffraction efficiency and uniformity of the reconstruction. The complexity of the optimization problem is reduced by only controlling two parameters per local diffuser phase element.

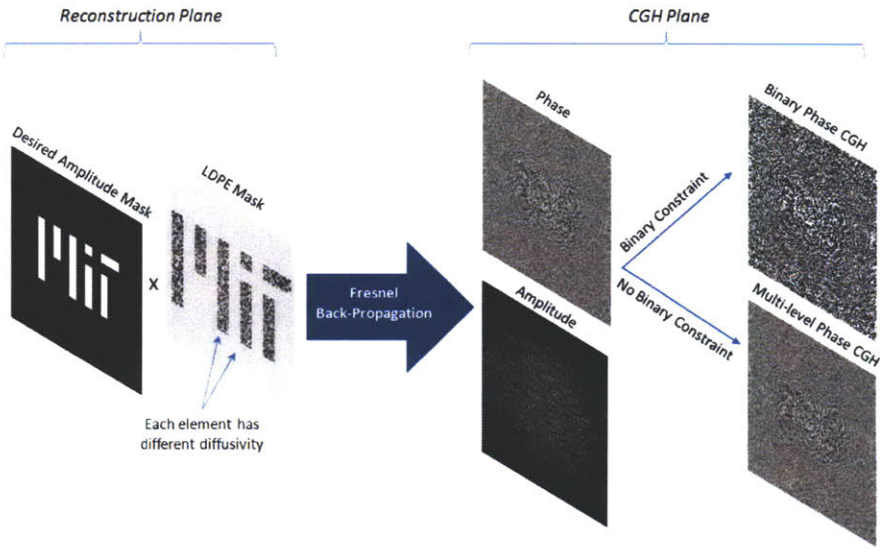


Figure 2-19: CGH encoding process based on the local diffuser phase elements mask.

The number of diffuser phase elements depends on the strategy used to decompose the mask. As mentioned before, we will restrict our design to binary signals that can be decomposed into elementary geometries, such as lines, squares and circles. Figure 2-20 shows an example of a pattern decomposition process, where the color represents the number assigned to the corresponding elementary geometry. The mask was decomposed by searching and then labeling all the connected binary structures. This decomposition strategy is not unique and different segmentations may be implemented.

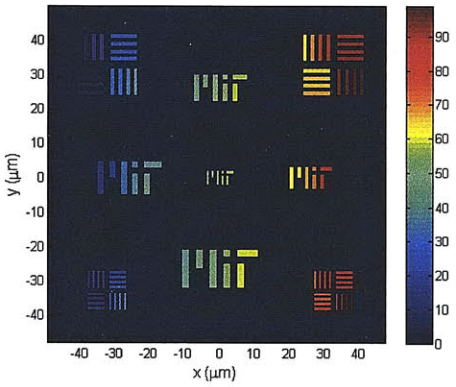


Figure 2-20: Example of mask pattern decomposition.

The LDPE mask is given by,

$$P_{LDPE}(x',y') = \sum_{q=1}^{N_{bp}} \Psi(q) \exp\left\{ i \arg\left[\exp\left(i2\pi D^{(q)}_{\text{factor}} R(x',y') - i\phi^{(q)}_{shift} \right) \right.\right. \otimes A(q) \text{Jinc}\left(\frac{2\pi r F^{(q)}_{\text{factor}}}{\rho_{ev}} \right) \left.\left.\right]\right\}, \quad (2.24)$$

where N_{bp} is the number of elementary binary geometries; $\Psi(q)$ is a binary function that describes the qth binary pattern,

$$\Psi(q) = \begin{cases} 1 & x,y \in q\text{th binary pattern} \\ 0 & \text{otherwise} \end{cases}; \quad (2.25)$$

R is a random uniformly distributed matrix, $A(q)$ is a constant factor,

$$A(q) = \pi \left(\frac{F^{(q)}_{\text{factor}}}{\rho_{ev}} \right)^2; \quad (2.26)$$

the Jinc function given by [67],

$$\text{Jinc}(r) = 2\frac{J_1(2\pi r)}{2\pi r}, \quad (2.27)$$

where J_1 is a Bessel function of the first kind, order one; r is the Jinc's radius: $r = \sqrt{x'^2 + y'^2}$; ρ_{ev} is the evanescent cut-off: $\rho_{ev} = 1/\lambda$; $D^{(q)}_{\text{factor}}$, $F^{(q)}_{\text{factor}}$ and $\phi^{(q)}_{shift}$ are the diffuser factor, frequency factor and constant phase shift of the qth binary pattern respectively – these parameters control the diffusivity of their corresponding local diffuser element.

The diffuser factor and constant phase shift specify a sector of the unit circle in the complex plane as shown in Figure 2-21-a. This sector contains all the possible complex vectors with unit magnitude and independent, uniformly distributed phase that correspond to elements of the matrix,

$$S^{(q)} = \exp\left(i2\pi D^{(q)}_{\text{factor}} R - i\phi^{(q)}_{shift}\right). \tag{2.28}$$

For the rest of the analysis we will only consider LDPE masks that have $\phi^{(q)}_{shift} = 0$. The diffuser factor is unbounded. Figure 2-21-b shows the case for $D_{\text{factor}} > 1$ – since the phase is modulo 2π, any overlapping sector result in an increased probability for the random phase elements lying inside. For $D_{\text{factor}} < 0$, the phase angle is measured in the conjugate direction.

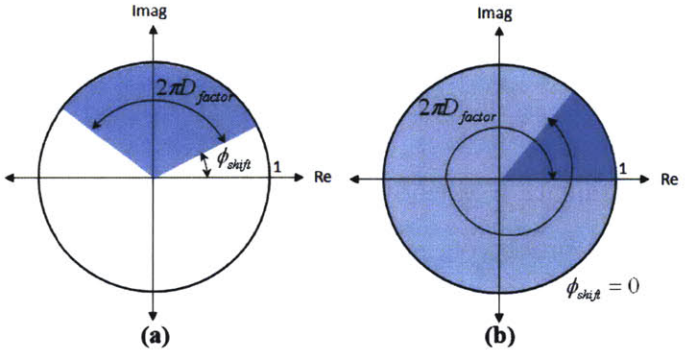


Figure 2-21: (a) Complex representation of diffuser factor; (b) $D_{\text{factor}} > 1$.

The spectrum of the matrix of equation 2.28, $\mathcal{F}\{S^{(q)}\}$, covers the entire Fourier plane

and if applied to the amplitude mask would cause the signal's spectral energy to leak outside the system pass-band, resulting in severe speckle-like noise in the reconstruction. To solve this problem, a low-pass window is used in the Fourier domain to limit the spectral bandwidth of $S^{(q)}$. We use an ideal low-pass window given by,

$$W = \text{circ}\left(\rho \cdot \frac{\rho_{ev}}{F_{\text{factor}}}\right), \tag{2.29}$$

where $\rho = \sqrt{u^2 + v^2}$, u and v are the spatial frequencies in the x and y directions and the circ function is defined as in [67]. This window gives rise to the Jinc function of equation 2.24. Other windows such as Hamming and Hann windows may be used [86], [87], [88]; however, we found that they didn't improve the performance of the LDPE mask. As indicated in equation 2.29, the frequency factor determines the diameter of the circular window effectively controlling the diffusivity of the qth local diffuser. A large diameter low-pass window causes the local diffuser to scatter the light diffracted from the amplitude mask over a large region in the hologram plane.

The optimum diffuser and frequency factors depend on the signal to be encoded. Each local phase element is coupled to the others as the diffracted fields add coherently at the hologram plane similar to a multiple slit Young interferometer. The total number of DOF required for the optimization of the LDPE mask is,

$$DOF_{LDPEmask} = 2N_{bp}. \tag{2.30}$$

This number of optimization variables is much smaller than that required for a standard optimization of a binary or multi-level CGHs (equations 2.13 and 2.14); therefore, reducing the complexity of the encoding process.

Local Negative Power Elliptical Phase Elements Mask

Next, we describe a second phase mask also designed to improve the signal encoding process while reducing the required number of variables in the optimization problem.

We call this mask "Local Negative Power Elliptical Phase Elements (LNPEPE) Mask", and is given by,

$$P_{LNPEPE}(x',y') = \sum_{q=1}^{N_{bp}} \Psi(q)\Gamma(q)\left[\exp\left\{i\frac{2\pi}{\lambda}\left[x'\sin\theta_x^{(q)} + y'\sin\theta_y^{(q)}\right]\right\}\right. \quad (2.31)$$
$$\left.\exp\left\{-i\frac{\pi}{\lambda}\left[\frac{\left(x'-x_c^{(q)}\right)^2}{f_1^{(q)}} + \frac{\left(y'-y_c^{(q)}\right)^2}{f_2^{(q)}}\right]\right\}\right],$$

where $\Psi(q)$ is also given by equation 2.25; $\Gamma(q)$ is a truncation window; $x_c^{(q)}$ and $y_c^{(q)}$ are center coordinates of the qth elemental geometry; $f_1^{(q)}$ and $f_2^{(q)}$ are the elliptical phase semi-major and semi-minor axes for the qth element or equivalently the focal lengths of two cylindrical lenses in close contact oriented along the x and y directions respectively; $\theta_x^{(q)}$ and $\theta_y^{(q)}$ are the qth off-axis angles that produce a linear phase shift reorienting the direction of the diffracted field.

The focal lengths are restricted to be negative: $f_1 < 0$ and $f_2 < 0$. This is equivalent to having two thin negative power anamorphic transforming cylindrical lenses that spread the field diffracted by the qth aperture (elemental geometry of the mask) evenly at the hologram plane. This is illustrated in Figure 2-22 for the case of an amplitude mask containing a single binary pattern. The linear phase shift term, controlled by the angles θ_x and θ_y, is equivalent to the off-axis illumination that allows redirecting the light from any binary pattern in the mask towards the center of the hologram window. Additional higher-order terms may be included to improve the mask's performance.

The truncation window of equation 2.31 is used to avoid aliasing during numerical implementation of the LNPEPE mask. The size of this window is set by the Nyquist limit,

$$u_s = 2u_{loc}, \quad (2.32)$$
$$v_s = 2v_{loc},$$

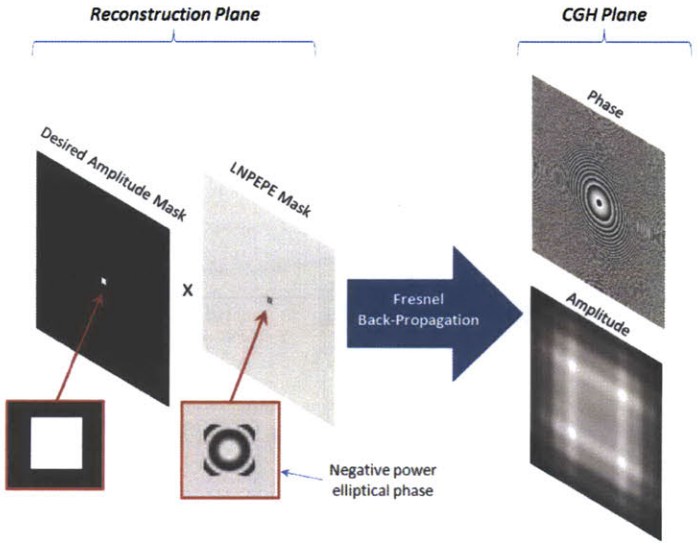


Figure 2-22: CGH encoding process based on the local negative power elliptical phase elements mask.

where u_s and v_s are the sampling frequencies: $u_s = v_s = 1/\delta_{pix}$; u_{loc} and v_{loc} are the local frequencies along the horizontal and vertical directions of the qth elliptical phase element,

$$\begin{aligned} u_{loc} &= \frac{\sin\theta_x^{(q)}}{\lambda} - \frac{(x' - x_c)}{\lambda f_1^{(q)}}, \\ v_{loc} &= \frac{\sin\theta_y^{(q)}}{\lambda} - \frac{(y' - y_c)}{\lambda f_2^{(q)}}. \end{aligned} \tag{2.33}$$

The truncation window is given by,

$$\Gamma(q) = \begin{cases} 1 & \left[\left(\frac{x'-x_c^{(q)}}{x_{\max}^{(q)}}\right)^2 + \left(\frac{y'-y_c^{(q)}}{y_{\max}^{(q)}}\right)^2\right]^{1/2} \geq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{2.34}$$

where $x_{\max}$ and $y_{\max}$ are,

$$x_{\max}^{(q)} = \left| f_1^{(q)} \sin\theta_x^{(q)} - \frac{\lambda f_1^{(q)}}{2\delta_{pix}} \right|, \tag{2.35}$$

$$y_{\max}^{(q)} = \left| f_2^{(q)} \sin\theta_y^{(q)} - \frac{\lambda f_2^{(q)}}{2\delta_{pix}} \right|.$$

The qth local phase element is ultimately limited by the most restrictive binary function: $\Psi(q)$ or $\Gamma(q)$. However, to preserve the effect of the elliptical phase, binary patterns with sizes smaller than the truncation window of equation 2.34 are desired.

The values of the focal lengths and off-axis angles depend on the signal to be encoded and are controlled by the optimization algorithm. The reduced number of DOF is,

$$DOF_{LNPEPEmask} = 4N_{bp}. \tag{2.36}$$

Similar to the LDPE mask, the optimization variables for all the binary patterns are coupled and need to be optimized together to help improve the encoding process.

2.4.3 Hybrid Optimization Algorithm

A hybrid optimization algorithm (HOA) is proposed for the robust optimization of binary and multi-level phase CGHs. The HOA is based on genetic algorithms (GAs) and a modified version of the error-reduction (MER) method. A simplified block diagram of the HOA is shown in Figure 2-23. The optimization algorithm begins by specifying the signal to be encoded and the desired encoding strategy based on the LDPE or LNPEPE phase masks. A broad multi-point parallel optimization is performed using GAs for the prescribed number of generations. This algorithm searches the nonlinear multidimensional optimization space thoroughly without becoming trapped at local minima. In addition, GAs are insensitive to the initial searching point positions as they are stochastically created. A set of constraints and boundary conditions are enforced during the optimization

process and the algorithm's performance is tracked by evaluating a fitness or objective function. GAs produce optimized solutions that lie near the global optimum and act as the initial searching point for the MER algorithm. The MER method further refines the solution and it stops when the maximum number of iterations or the minimum error tolerances is reached.

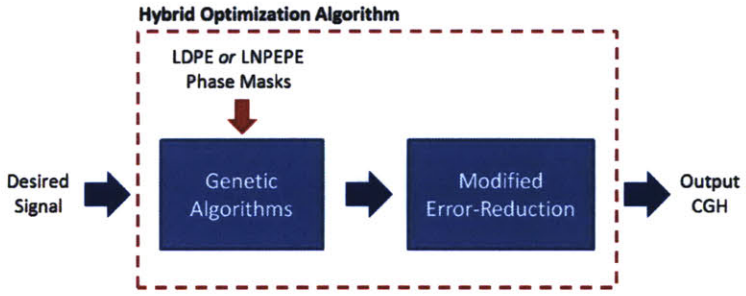


Figure 2-23: Block diagram of hybrid optimization algorithm.

The key features of the HOA are:

- Broad searching range: multi-point parallel and local searching strategies
- Robust: insensitive to initial state; avoids getting trapped at local minima; tolerates highly nonlinear, discontinuous, noisy optimization spaces
- Allows flexible choice of signal for encoding
- Optically efficient: produces high-diffraction efficiency CGHs that satisfy the system constraints and reconstruct high-resolution, uniform patterns
- Reduced complexity: small number of optimization variables by using the LDPE or LNPEPE phase masks
- Computationally efficient: algorithm implemented in parallel on a graphics processing unit (GPU) resulting in significant speedups

In the next two sections, a detailed description of the blocks comprising the HOA will be covered.

Genetic Algorithms

Genetic algorithms (GAs) are robust optimization algorithms designed based on the mechanics of natural selection and natural genetics [69]. Inspired by biological evolution, GAs combine the survival of the fittest individual from a population of multiple points used to search the optimization space. The evolution process is carried out by a structured, yet randomized, information exchange resulting in fitter offspring with stronger genes. While randomized, GAs are not simple random walks – they are randomly guided exploiting historical information to speculate on new search points with expected improved performance. The robustness of GAs enables searching in parallel highly non-linear multi-dimensional spaces without becoming trapped at local minima. GAs can be applied to nonstandard optimization problems such as those with discontinuous objective functions, non-differentiable, stochastic, highly non-linear, or with binary decision variables. In contrast to most local search optimization methods, GAs are, in theory, insensitive to the initial searching state; however in practice, if the initial searching points are too far from the solution, the algorithm might suffer premature convergence.

GAs were introduced by Holland and his colleagues at the University of Michigan, with the goal of explaining the adaptive process of natural systems and designing artificial systems software that retain the important mechanisms of natural systems [89]. Since then, GAs have been used in a broad range of applications such as biological cell simulation [90], pattern recognition [91], [92], artificial intelligence [93], finance [94], computer vision [95], and game theory [96]. In the area of computer holography, GAs have been implemented for the optimization of point-wise binary and multi-level phase Fraunhofer CGHs [97], [98], [99], [100]. In their presented work, each pixel in the hologram was treated as an independent optimization variable that controls the corresponding DOF of the individual searching point. This approach makes the optimization problem highly

complex and difficult to solve for a reasonable population size (number of individual searching points) with conventional computational means. Due to numerical limitations, only holograms with very small space-bandwidth products (32×32 and 64×64 pixels) and small number of individuals were considered. The reconstruction results from their optimized CGHs were extremely low quality due to the limited number of DOF. GAs have also been proposed for the optimization of cell-oriented holograms [101] and diffraction optical elements (DOEs) based on 1D parameterization designs [102], [103], [105].

In contrast to previous work, in this subsection we describe the implementation of GAs as part of the HOA applied to the reduced complexity problem described previously. This reduction in complexity makes it feasible to optimize efficiently large CGHs with current computational means. In addition, we present the first implementation (to our knowledge) of GAs for the optimization of Fresnel domain CGHs applied for holographic lithography.

The block diagram of the implemented version of GAs is shown in Figure 2-24. The algorithm begins with the creation function that is used to generate the initial population. A population is a collection of individuals or chromosomes. Each individual is composed of several genes that correspond to the DOF or decision variables of the problem in hand. In our implementation, the type and number of genes that comprise a given individual depends on our choice of complexity reduction strategy, i.e. encode the signal using the LDPE or LNPEPE phase masks. The form of the kth individual for the encoding strategy based on the LDPE phase mask is,

$$x_k = \left[D^{(1)}_{\text{factor}_k}, F^{(1)}_{\text{factor}_k}, D^{(2)}_{\text{factor}_k}, F^{(2)}_{\text{factor}_k}, \cdots, D^{(N_{bp})}_{\text{factor}_k}, F^{(N_{bp})}_{\text{factor}_k} \right]. \tag{2.37}$$

When the LNPEPE phase mask is used, the form of the kth individual is,

$$x_k = \left[f^{(1)}_{1_k}, f^{(1)}_{2_k}, \theta^{(1)}_{x_k}, \theta^{(1)}_{y_k}, f^{(2)}_{1_k}, f^{(2)}_{2_k}, \theta^{(2)}_{x_k}, \theta^{(2)}_{y_k} \cdots, f^{(N_{bp})}_{1_k}, f^{(N_{bp})}_{2_k}, \theta^{(N_{bp})}_{x_k}, \theta^{(N_{bp})}_{y_k} \right]. \tag{2.38}$$

The number of genes for the two individuals is given by equations 2.30 and 2.36 respec-

tively.

The creation function stochastically generates the required number of individuals for the specified population size, *Popsize*. Every generated gene is drawn out of a random uniform distribution and is forced to be within a given initial range, *Inirange*. The initial range is specified independently for each gene and its purpose is to introduce genetic diversity by controlling the initial span of the search. If the initial range is too large, the individuals of the initial population would sample the search space coarsely, potentially missing the global optimum. On the other hand, if the initial range is too small, the searching space is not sampled uniformly and the algorithm might suffer premature convergence.

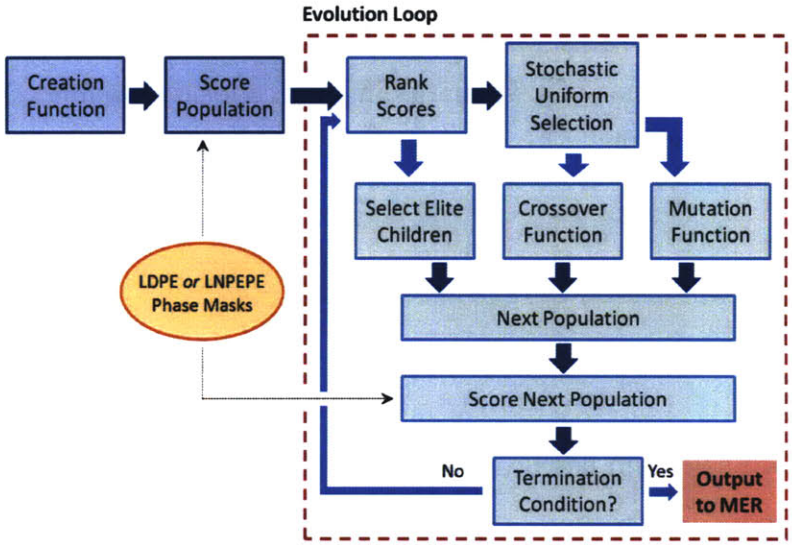


Figure 2-24: Block diagram of the GAs section.

A partial initial population can also be specified by the user and the creation function is used to generate the remaining individuals. This creation strategy is implemented for the encoding process with the LNPEPE mask to guide the algorithm to areas where optimal solutions are likely to be found. A deterministic approach based on geometrical optics is used to calculate a feasible individual which is then replicated several times to

form a partial initial population. Figure 2-25 shows the equivalent problem geometry used for the calculation of the initial focal lengths $f_1^{(q)}$ and $f_2^{(q)}$, of the qth binary pattern. The example shown consists of an amplitude mask with a single rectangular binary pattern of size $w_x \times w_y$, centered at the optical axis and represented by an aperture stop at the reconstruction plane. This amplitude mask represents the desired signal to be encoded. As mentioned before, the LNPEPE mask is equivalent to placing two negative power cylindrical lenses oriented along x and y directions (Figure 2-25 only shows the geometry for the x-direction). In the equivalent problem geometry, the complex mask (amplitude and LNPEPE masks) is illuminated by an in-line plane wave and the negative lens spreads the field which propagates towards the CGH plane. The anamorphic transformation induced by the cylindrical lens is designed to cover the entire hologram window, H_{size}, with semi-uniform amplitude to satisfy the zero absorption constraint. The resulting field is equivalent to that of a virtual point source located behind the aperture stop as shown in Figure 2-25. The initial focal lengths are given by,

$$\begin{aligned} f_1^{(q)} &= -\frac{w_x^{(q)} d}{H_{size} - w_x^{(q)}}, \\ f_2^{(q)} &= -\frac{w_y^{(q)} d}{H_{size} - w_y^{(q)}}, \end{aligned} \tag{2.39}$$

where a square hologram has been assumed. The off-axis angles, θ_x and θ_y, of the initial individual are set to zero.

The following example illustrates the deterministic creation of an initial individual. The simulation parameters are: $\lambda = 532$nm, $H_{size} = 175\mu$m, $d = 200\mu$m, $N_{bp} = 1$, $w_x = 5\mu$m and $w_y = 20\mu$m. The desired amplitude mask is shown in Figure 2-26-a. Figure 2-26-b shows the calculated LNPEPE mask. We now compare two encoding strategies: without phase mask (zero phase) and with LNPEPE mask. Figure 2-27-a shows the amplitude distribution of the diffracted field (CGH plane) when no phase mask is applied. This is the typical Fresnel diffraction pattern from a narrow slit. The non-uniform amplitude of the diffracted field violates the zero absorption constraint required

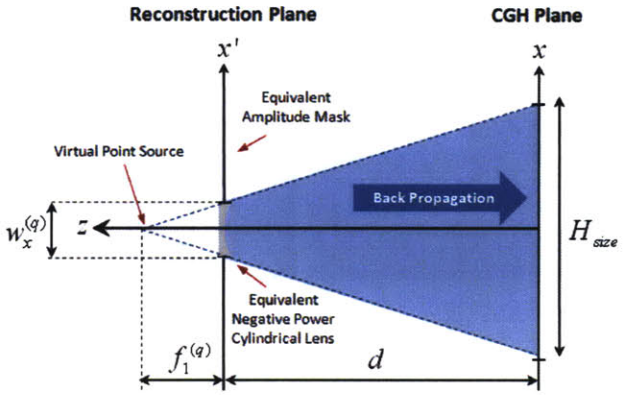


Figure 2-25: Equivalent problem geometry for creation of initial population.

for pure phase holograms. Figure2-27-b shows the corresponding amplitude distribution when the LNPEPE mask is used. This amplitude distribution uniformly covers the entire hologram window resulting in an improved encoding process. The continuous phase distributions of the corresponding in-line CGHs for the two cases are shown in Figures 2-28-a and 2-28-b (the diffracted amplitude is discarded and set to unity to satisfy the zero absorption constraint). Figures 2-29 and 2-30 show the amplitude distributions of the reconstructed field (decoded signal) for the CGHs of Figures 2-28-a and 2-28-b respectively. Despite that the parameters for the initial LNPEPE mask haven't been fully optimized by GAs, the reconstructed field is substantially higher quality than that of Figure 2-29. This proves that the LNPEPE mask improves the signal encoding process for pure phase Fresnel CGHs. The diffraction efficiencies of the reconstructed intensities for both cases are: $\eta_{nomask} = 35.85\%$ and $\eta_{LNPEPE} = 89.58\%$. The mean-square errors (MSE) of the difference between the desired and calculated intensity distributions at the reconstruction plane for both cases are: $MSE_{nomask} = 435.01$ and $MSE_{LNPEPE} = 111.25$. These metrics indicate that the improved encoding process not only produced a CGH with high diffraction efficiency, but also improved the uniformity of the reconstruction.

Following the creation function, the scores for the initial population are computed. The type of score function used depends on the choice of encoding strategy. Figure 2-

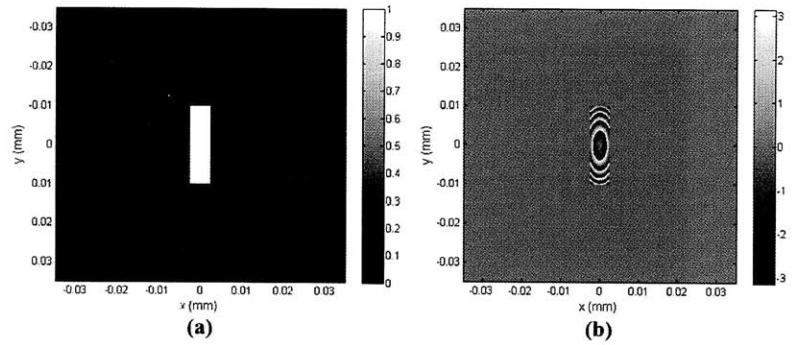


Figure 2-26: (a) Desired amplitude mask; (b) Calculated LNPEPE mask.

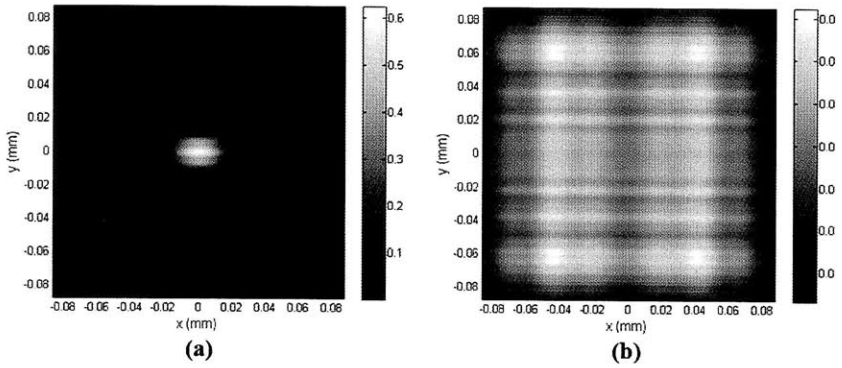


Figure 2-27: (a) Diffracted amplitude from regular mask; (b) Diffracted amplitude from mask with LNPEPE mask.

31 shows the block diagram of the score function for the encoding based on the LDPE mask. The kth individual is extracted from the population and used for the computations of a random diffuser, $S^{(q)}$, and circular window, $W^{(q)}$. A constant seed is specified for the generation of the random matrix, R, used in the diffuser for repeatable results. The random matrix is drawn out of a uniform distribution. A 2D forward Fast-Fourier Transform (FFT) operation is performed on the random diffuser and the spectrum is low-pass filtered by a circular window. The filtered spectrum is transformed back to space domain by means of an inverse 2D FFT. The phase of the resulting complex signal is extracted and point-wise multiplied by the binary truncation function, $\Psi(q)$, which

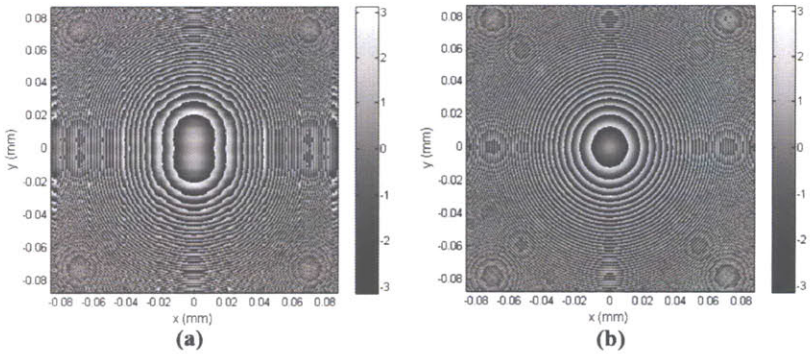


Figure 2-28: (a) Computed CGH from regular mask; (b) Computed CGH from mask with LNPEPE mask.

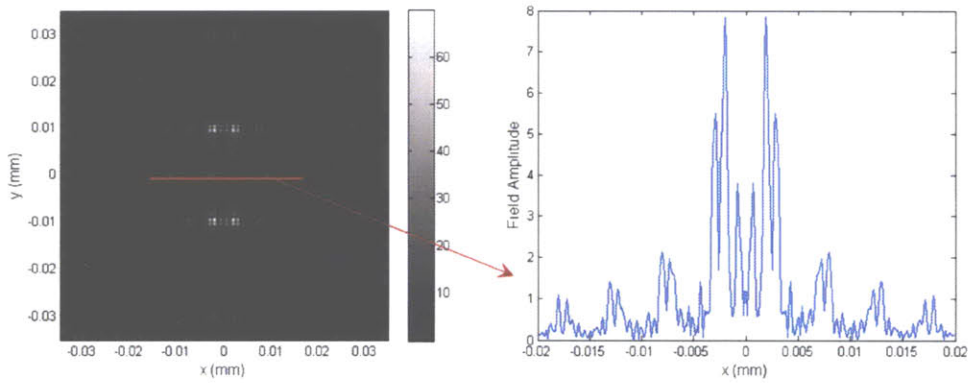


Figure 2-29: Reconstructed amplitude from regular CGH.

makes it zero everywhere except inside the corresponding binary pattern (equation 2.25). The result is then added to the total phase and the process is repeated for every binary pattern in the mask. The total phase is used to compute the LDPE mask which is then applied to the amplitude mask and back-propagated to the CGH plane resulting in the complex signal, $\hat{H}$. The CGH, H, is obtained after enforcing the zero absorption and

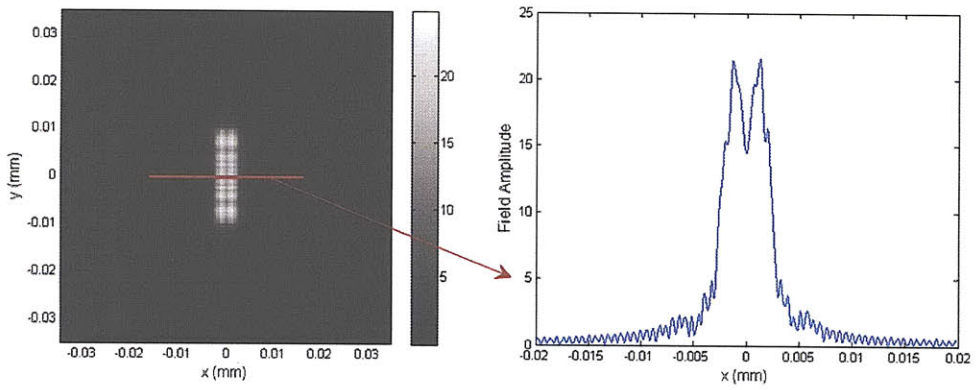


Figure 2-30: Reconstructed amplitude from CGH encoded with LNPEPE mask.

binary constraints (equation 2.23) on $\hat{H}$. The binary constraint is given by,

$$\theta_H(x,y) = \begin{cases} 0 & |\theta_{\hat{H}}(x,y)| \leq T \\ R_{phase} & |\theta_{\hat{H}}(x,y)| > T \end{cases}, \tag{2.40}$$

where R_{phase} is a specified phase delay (typically set to π), and T is a threshold value (typically set to $\pi/2$). The encoded CGH is then reconstructed by means of a forward Fresnel propagation and the intensity distribution at the reconstruction plane, I_{est}, is computed. A fitness function (objective function) is evaluated to estimate the performance of the encoding/decoding processes. The fitness functions considered in this thesis are: diffraction efficiency upper bound, effective diffraction efficiency, MSE before and after photoresist exposure, and hybrid. The details on each fitness function will be explained later. The result from the fitness function is the score, y_k, for the kth individual. High scores indicate fitter and stronger individuals with increased probability of surviving in future generations.

The block diagram of the score function for the encoding process based on the LNPEPE mask is shown in Figure 2-32. This function begins by extracting the focal lengths and off-axis angles for the kth individual, as well as the center coordinates of the

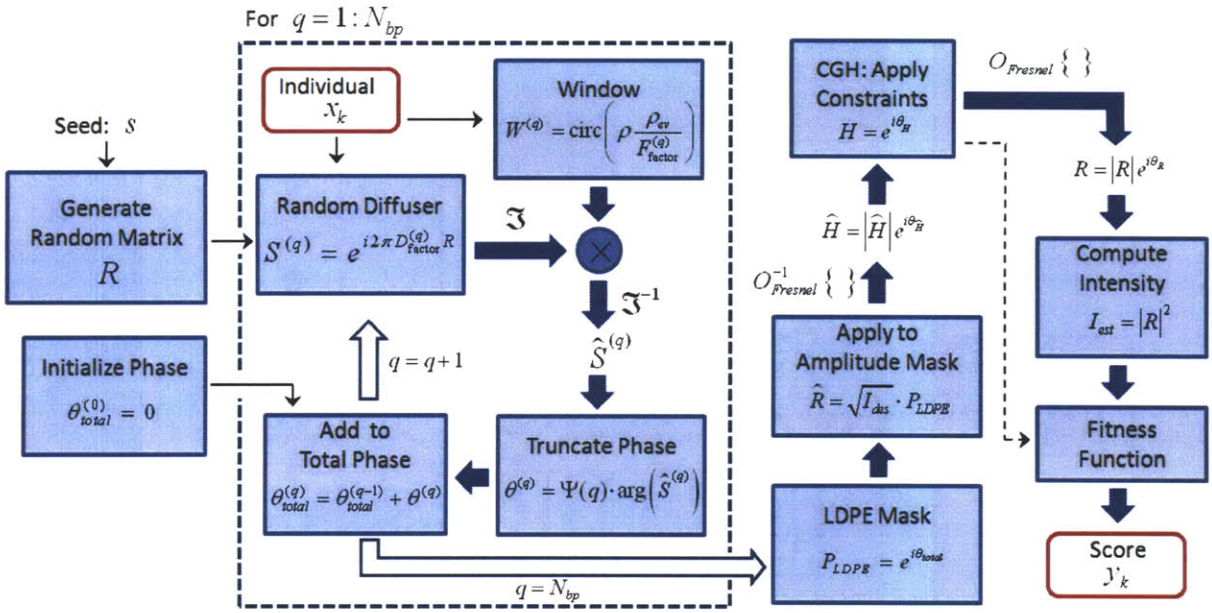


Figure 2-31: Block diagram of score function.

qth binary pattern used for the computation of the elliptical phase,

$$\theta_{elliptical}^{(q)} = \frac{2\pi}{\lambda}\left[x' \sin\theta_x^{(q)} + y' \sin\theta_y^{(q)} - \frac{\left(x' - x_c^{(q)}\right)^2}{2f_1^{(q)}} - \frac{\left(y' - y_c^{(q)}\right)^2}{2f_2^{(q)}}\right]. \tag{2.41}$$

The anti-aliasing window, $\Gamma(q)$, and truncation window, $\Psi(q)$ are applied to the elliptical phase and the result is added to the total phase. This process is repeated for all binary patterns in the mask. The final total phase is used to generate the LNPEPE mask. The following steps in the score function are similar to those described previously for the LDPE mask.

The next step in GAs is to scale the computed population scores. The scaling is done to keep appropriate levels of competition among all the individuals in the population throughout the simulation. Without scaling, GAs have the tendency to be dominated by a few superindividuals biasing the selection process. The scaling procedure implemented

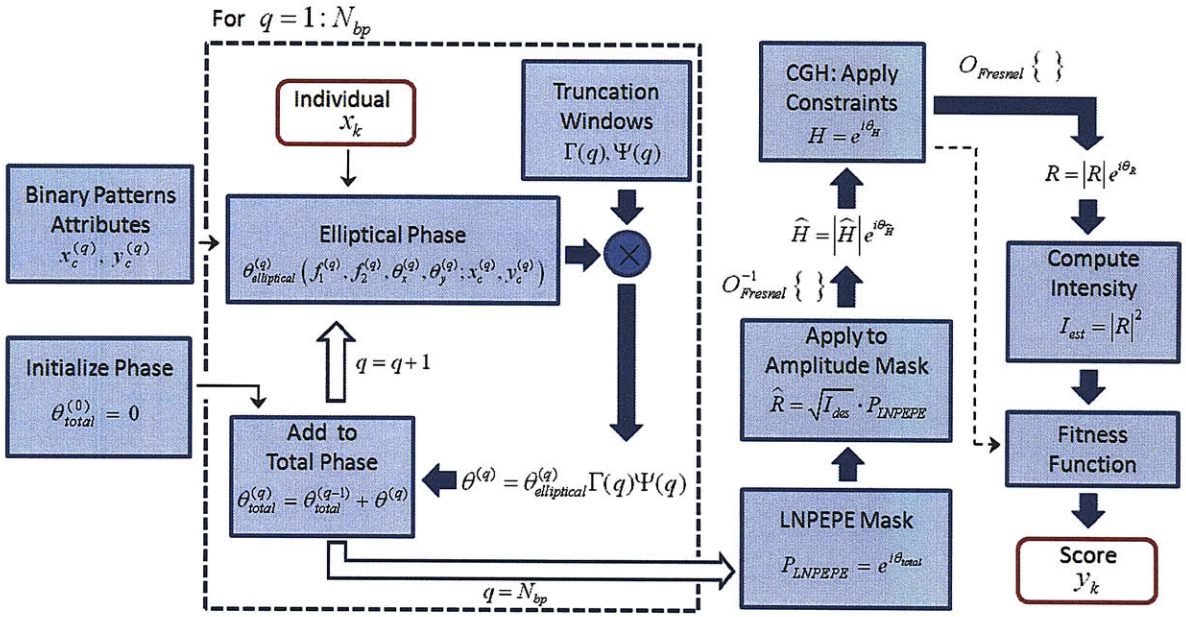


Figure 2-32: Block diagram of score function based on LNPEPE mask.

affects the performance of the algorithm. If the scaled values vary too widely, a rapid reproduction of the stronger individuals occurs, talking over the population gene pool and preventing the algorithm from searching other areas of the search space. On the other hand, if the values vary very little, the search will be progressively slower as every individual has approximately equal chance to reproduce. Several scaling methods have been studied such as linear, sigma truncation, power law scaling, top scaling and rank [69]. Every method has been found to have its strengths and weaknesses that greatly depend on the type of problem solved. In this thesis we implement the rank method, which is a nonparametric procedure that uses the rank of each individual instead of its score for scaling. The rank of the fittest individual is 1; the next most fit is 2, and so on. The implemented GA is designed to minimize the fitness function so lower raw scores have higher scaled values. The rank scaling is given by,

$$\text{rank}^{(l)} = \frac{1}{\sqrt{n}}, \tag{2.42}$$

where n is the index of corresponding individual after sorting the scores from best to worst, and l ranges from 1 to the number of individuals in the population, *Popsize*. For example, consider a population of 5 individuals with the following scores: $Scores = [3.82, 1.99, 4.85, 6.19, 0.45]$. The corresponding rank is: rank $= [0.4472, 0.7071, 1, 0.5774, 0.5]$. The rank values are then normalized to compute the expectation values of the current population that will be used during the selection, crossover and mutation processes to generate the next population. A high expectation value equals high probability of survival. The expectation values are given by,

$$\text{expectation}^{(l)} = \frac{\text{rank}^{(l)} \cdot N_{parents}}{\sum_{l} \text{rank}^{(l)}}, \tag{2.43}$$

where $N_{parents}$ is the total number of parents to be chosen during the selection process to generate the children (offspring) for the next generation. The number of parents is given by,

$$N_{parents} = 2N_{xover} + N_{mutation}, \tag{2.44}$$

where N_{xover} and $N_{mutation}$ are the number of crossover and mutation children needed for the next generation. N_{xover} an $N_{mutation}$ are specified by the user and are key parameters that control the performance of GAs as will be explained later. For the previous example consider $N_{xover} = 2$, $N_{mutation} = 1$: $N_{parents} = 5$ and expectation $= [0.6919, 1.0940, 1.5472, 0.8933, 0.7736]$.

The selection function is designed to choose parents for the next generation based on their scaled values from the fitness scaling function. The goal of the selection function is to bias the algorithm by choosing parents with strong genetic codes capable of producing even fitter children after the crossover and mutation processes. Several selection functions are available such as deterministic sampling, remainder stochastic sampling with and without replacement and stochastic uniform [69]. In this thesis, the stochastic uniform method is implemented. This method has been shown to be efficient and main-

tain diversity in the population, which prevents premature convergence. The stochastic uniform selection function performs a linear search through a roulette wheel with slots corresponding to each parent. The size of each slot on the wheel is proportional to its expectation value of equation 2.43. The search is done by moving through the wheel with steps of equal size, so as to cover the entire wheel in $N_{parents}$ steps ($stepsize = 1/N_{parents}$). At each step, a parent is selected from the slot it landed in. After the selection function, the parents' indices undergo a random permutation to prevent locality effects.

The reproduction of individuals to form the next generation is performed by the crossover and mutation genetic operators as well as the selection of elite children. Elite children are individuals with the best fitness values, guaranteed to survive intact to the next generation. The user specifies the number of elite children, N_{elite}, to be selected in every generation.

The crossover function is a genetic operator used to create new individuals (offspring) for the next generation by means of a gene exchanging process. Pairs of parents chosen by the selection function are used to generate one or more children. The fraction of the population, other than elite children, that will be created from a crossover operator is determined by the crossover fraction, F_{xover}, which is specified by the user. The number of crossover individuals generated is,

$$N_{xover} = \text{round}\left[(Pop_{size} - N_{elite}) \cdot F_{xover}\right]. \tag{2.45}$$

Several crossover strategies have been studied such as single point, double point, and scattered [69]. These functions are designed to force the algorithm to explore regions on the search space where a better solution may lie. In this thesis we implement the scattered or stochastic uniform crossover method. In this method, each gene has equal probability to be exchanged. Figure 2-33 shows an example of the scattered crossover process. The random vector, $\mathbf{r}$, is drawn out of a uniform distribution and its values are set to one if $\mathbf{r} > 0.5$ or zero otherwise. The random vector determines which genes from both parents are exchanged. This process is repeated for every parent pair in order to

generate the desired number of crossover children.

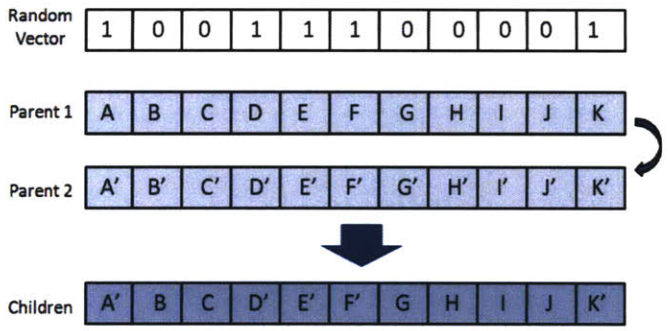


Figure 2-33: Scattered crossover process.

The mutation function generates children by applying random changes to a single individual to bring diversity to the population, preventing premature convergence and to increase the likelihood that the algorithm will generate individuals with better fitness values. In addition, it allows the algorithm to escape from local minima and continue the search towards the global optimum. The number of children to be generated from a mutation operator is,

$$N_{mutation} = Pop_{size} - N_{elite} - N_{xover}. \tag{2.46}$$

The amount of mutation performed by GAs is a critical parameter. If $N_{mutation}$ is too large, the algorithm starts becoming a random walk and the solution might never be reached. On the other hand, if $N_{mutation}$ is too small, no new information is added to the genetic pool and the algorithm might suffer premature convergence or become trapped at a local minimum. Several mutation operators have been developed such as Gaussian, uniform and adaptive feasible [69]. In this thesis we implemented the adaptive feasible mutation function. This function creates new children by perturbing genes on the remaining parents after the crossover function. The perturbations are done along feasible directions generated randomly that are adaptive with respect to the last successful or

unsuccessful generation. The feasibility of the perturbed individual is tested to ensure each gene satisfies the lower and upper bounds, LB and UB respectively. The values of LB and UB depend on our choice of encoding strategy (LDPE or LNPEPE mask). For the encoding strategy based on the LDPE mask, the bounds are: $-\infty < D_{\text{factor}} < \infty$; $0 < F_{\text{factor}} < 1$. Despite that the diffuser factor is unbounded, to control the span of the search the diffuser factor is restricted to: $0 < D_{\text{factor}} < X$, where X is a large number. For the encoding strategy based on the LNPEPE mask, the bounds are: $-\infty < f_1 < 0$; $-\infty < f_2 < 0$; $-\pi/2 < \theta_x < \pi/2$; $-\pi/2 < \theta_y < \pi/2$. Again for practical reasons, the focal lengths bounds are restricted to: $-Y < f_1 < 0$; $-Y < f_2 < 0$, where Y is a small number.

The next population is the result of combining the chosen elite children and the created crossover and mutation children. In our implementation, every population has the same size. The next step in the algorithm is to implement the score function to compute the fitness value of each individual of the new population. This score function is the same as those described previously (Figures 2-31 and 2-32). The last step in the iteration is to check if any of the termination conditions have been met: maximum number of generations or minimum error tolerance. If no termination condition has been met, the evolution loop described above repeats; otherwise, the last population is saved and the best individual is extracted and used to compute the CGH that will serve as an initial guess for the MER algorithm.

Modified Error-Reduction Algorithm

The modified error-reduction (MER) algorithm is a variation of the error-reduction (ER) method designed to refine the solution obtained by GAs as described in the previous section. The ER algorithm is a local search iterative method typically used to solve phase retrieval problems [106]. This algorithm is a generalized form of the Gerchberg-Saxton method originally designed to solve a phase retrieval problem in electron microscopy [73], [107], [108]. The algorithm iterates between two domains, such as the space and

frequency domains, imposing a set of constraints at each domain. For the original electron microscopy problem, both the modulus of a complex-valued image and the modulus of its Fourier transform were measured and the goal of the algorithm was to retrieve the phase distributions at both domains. The ER algorithm was originally proposed by Hirsh, Jordan and Lesem in 1970 to solve the synthesis problem in computer holography [80]. This method was later reinvented by Gallagher and Liu in 1973 and applied for the optimization of Kinoform type holograms [109]. In the following years, the properties of the ER method were extensively studied. It was proven that mean-square error computed at the end of each iteration is guaranteed to decrease or remain constant [110]. Also, it was shown that the algorithm tends to reduce the error rapidly within the first few iterations and then more slowly for the later iterations. Other variations of the ER method were introduced such as the input-output and hybrid methods [111], [112]. The input-output algorithm claims to speed up the convergence in certain problems, differing from the ER method only for the object domain operation.

The block diagram of the implemented MER algorithm is shown in Figure 2-34 [113]. The algorithm begins the local search from a specified initial search point. For the HOA, the initial search point is the solution computed from GAs. Two additional encoding strategies are also studied for the initialization of the MER algorithm: diffracted field (DF); simulated optically recorded hologram (SORH). The equivalent problem geometry for the DF is shown in Figure 2-35-a. This encoding strategy is used to compute the initial CGH (initial search point) for the MER algorithm. The initial CGH is computed by back-propagating the amplitude mask to the hologram plane and applying the zero absorption and binary constraints. The geometry for the encoding strategy based on the SORH is shown in Figure 2-35-b. In this case the initial CGH is computed simulating the conventional optical recording process in which an object and reference waves interfere and a phase hologram is produced. The phase hologram is given by,

$$H_{SORH} = e^{i\kappa I}, \tag{2.47}$$

where κ is a normalization constant and I is the interference pattern given by equation 2.8. The binary constraint is enforced on the phase hologram of equation 2.47.

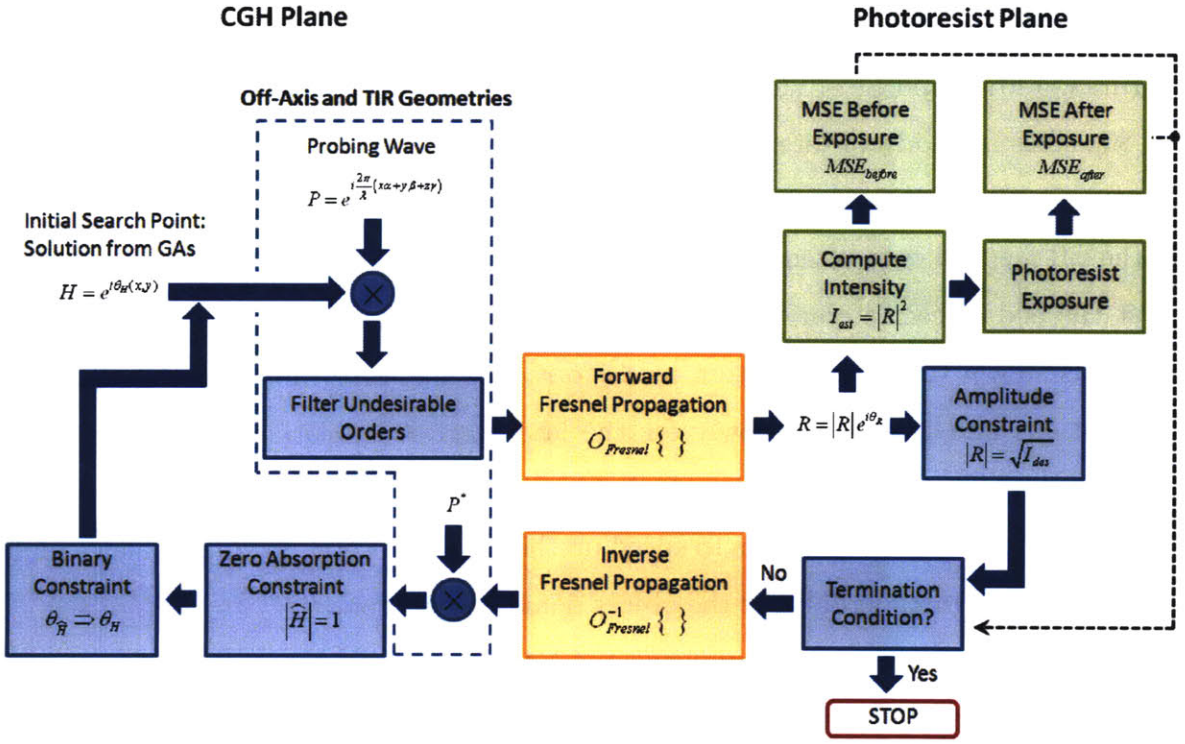


Figure 2-34: Block diagram of modified error reduction algorithm.

Once the initial CGH has been specified, it is point-wise multiplied by the probing wave. For the case of the off-axis and TIR geometries, an additional step is required to filter the undesirable diffraction orders from the modulated signal as explained before. The filtering step is performed in frequency domain by applying a low-pass circular window with the specified cut-off such as that shown in Figure 2-15.

The transition between the CGH and photoresist planes is done by means of a Fresnel transformation as explained previously. This is numerically implemented in frequency domain. The forward propagation operation is given by,

$$R(x', y') = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{D(x, y)\right\} \cdot H_{Fresnel}(u, v; d)\right\}, \tag{2.48}$$

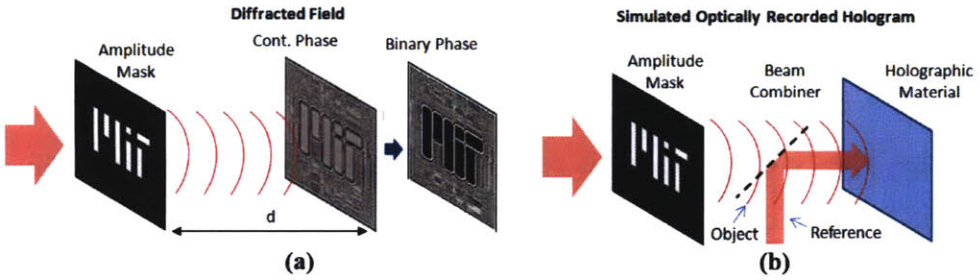


Figure 2-35: Alternative encoding strategies: (a) Diffracted field; (b) Simulated optically recorded hologram.

where D is the complex field diffracted by the hologram after probing (hologram plane), R is the reconstructed complex field after propagation (photoresist plane), and $H_{Fresnel}$ is the paraxial Fresnel transfer function,

$$H_{Fresnel}(u, v; d) = e^{-i\pi\lambda d\left(u^2+v^2\right)}. \tag{2.49}$$

The Fresnel transfer function is sampled uniformly with a pixel size δ_u. D is stored in an array of size $N_h \times N_h$, where $N_h = H_{size}/\delta_{pix}$. To prevent aliasing, the array is padded with zeros or ones forming a new array of size: $N_z \times N_z$, where $N_z = Z_{size}/\delta_{pix}$ and Z_{size} is the lateral size of the field after the padding operation. The frequency domain pixel size for a padded array is: $\delta_u = 1/Z_{size}$. A large amount of padding results in a fine frequency sampling. The Fresnel transfer function is a chirp function with a local frequency that increases linearly with u,

$$u_{loc} = \frac{1}{2\pi}\frac{d\phi(u)}{du} = \lambda du. \tag{2.50}$$

We define the Nyquist cut-off as the maximum local frequency of the chirp that can be sampled before violating the Nyquist limit,

$$\lambda du_{Nyq} = \frac{1}{2\delta_u} \rightarrow u_{Nyq} = \left(\frac{1}{\lambda}\right)\left(\frac{1}{2d\delta_u}\right) = u_{ev}\left(\frac{Z_{size}}{2d}\right). \tag{2.51}$$

In our numerical implementation, an additional circular window, $\mathrm{circ}(u/2u_{Nyq})$, is used to truncate the Fresnel transfer function to prevent it from suffering aliasing. For the in-line geometry, Z_{size} is chosen such that $u_{Nyq} \geq u_{ev}$. For the off-axis and TIR geometries, the Nyquist cut-off is required to be equal or larger than the cut-off of the corresponding demodulation window.

For small propagation distances that violate the condition of equation 2.20, an angular spectrum propagation method is adopted [69]. This propagation method is also performed in frequency domain as equation 2.48, but $H_{Fresnel}$ is replaced with the angular spectrum transfer function,

$$H_{Angular}(u, v; d) = e^{i\frac{2\pi}{\lambda}d\sqrt{1-(\lambda u)^2+(\lambda v)^2}} \tag{2.52}$$

The corresponding local frequency along the horizontal direction is,

$$u'_{loc} = \frac{1}{2\pi}\frac{d\phi'(u)}{du} = -\frac{\lambda du}{\sqrt{1-(\lambda u)^2}}, \tag{2.53}$$

and the new Nyquist cut-off is given by,

$$u'_{Nyq} = \frac{1}{\lambda\sqrt{(2d\delta_u)^2+1}} = u_{ev}\left[\left(\frac{2d}{Z_{size}}\right)^2+1\right]^{1/2}. \tag{2.54}$$

The angular spectrum transfer function is also truncated at the Nyquist cut-off to avoid numerical errors due to aliasing.

The type of padding implemented depends on the boundary conditions of the problem. The boundary conditions describe the optical transmission of the substrate surrounding the CGH. Two types of boundary conditions are studied: clear aperture and opaque aperture. A clear aperture CGH is fabricated on a transparent substrate such as fused silica in which only a small section contains the CGH structure. In contrast, the substrate surrounding a CGH with an opaque aperture is covered with an absorbing material such

as a metallic layer designed to block the incident light. The choice between these two boundary conditions is application dependent. For example, in holographic lithography an opaque aperture CGH might be desirable, in order to block the incident light outside the CGH area preventing to expose the photoresist outside the object window. On the other hand, in many applications only the signal inside the object window is of interest and thus a clear aperture CGH can be used. The clear aperture boundary condition is included in the optimization algorithm by padding the desired array with 1's. For the opaque boundary condition the array is padded with 0's.

The next step in the MER algorithm is to compute the error of the reconstructed intensity distribution at the photoresist plane. Two error metrics are considered: mean-square error (MSE) before photoresist exposure, MSE_{before}, and MSE after photoresist exposure, MSE_{after}. The MSE_{before} is given by,

$$MSE_{before} = \frac{1}{N_{o_x} N_{o_y}} \sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left[I_{est}\left(x', y'\right) - I_{des}(x', y') \right]^2 , \tag{2.55}$$

where N_{o_x} and N_{o_y} are the number of pixels along the x and y directions – for square object windows: $N_o = N_{o_x} = N_{o_y} = O_{size}/\delta_{pix}$; I_{es} is the estimated intensity: $I_{es} = |R|^2$; and I_{des} is the desired intensity at the photoresist plane. The MSE is an estimator used to quantify the amount by which the estimated intensity differs from the desired intensity distributions. The MSE as defined in equation 2.55 is the second moment (about the origin) of the error, and thus incorporates both the variance (uniformity) and the bias (diffraction efficiency) of the reconstructed intensity. An unbiased version of this estimator reduces to the variance of the estimated intensity distribution. The MSE has units of square intensity. Sometimes it is convenient to use the root-mean square error (RMSE) as it has the same units of intensity. The MSE of equation 2.55 can also be written as: $MSE = Var[I_{est}] + (Bias[I_{est}, I_{des}])^2$, where Var is the variance and $Bias$ is a function given by: $Bias[I_{est}, I_{des}] = E[I_{est} - I_{des}]$, where $E[]$ is the expectation value.

Additional error metrics can be used to evaluate the reconstruction such as the $L1$

norm,

$$L_1 = \frac{1}{N_{o_x} N_{o_y}} \sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left| I_{est}\left(x', y'\right) - I_{des}(x', y') \right|, \tag{2.56}$$

that quantifies the bias between the estimated and desired intensity distributions and the normalized cross-correlation metric,

$$NCC = \frac{\sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left(I_{est} - \bar{I}_{est}\right)\left(I_{des} - \bar{I}_{des}\right)}{\sqrt{\sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left(I_{est} - \bar{I}_{est}\right)^2} \sqrt{\sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left(I_{des} - \bar{I}_{des}\right)^2}}, \tag{2.57}$$

that quantifies the degree of correlation between both intensity distributions. NCC ranges between $[-1, 1]$, given a value of 1 if the distributions are fully correlated, 0 if they are uncorrelated and -1 if they are anti-correlated [114], [115].

The computation of MSE_{after} requires the simulation of the photoresist exposure process. Three models are considered: ideal photoresist, photoresist model and real photoresist. The performance of the photoresist is characterized by its contrast curve. The contrast curve describes the remaining resist fraction of a uniformly illuminated resist versus the logarithm of the applied exposure dose. Figure 2-36 shows the idealized contrast curves for positive and negative photoresists. For a positive photoresist, the sections exposed to light become soluble and are removed by the developer. On the other hand, for a negative photoresist, the regions exposed become insoluble and remain after the development process. The contrast of a given photoresist is characterized by the slope,

$$\gamma = \frac{1}{\log_{10}\left(\frac{D_{100}}{D_0}\right)}, \tag{2.58}$$

where D_{100} and D_0 are the dose values shown in Figure 2-36. Typical contrast values range between 2 and 3. In the ideal photoresist model, we assume a high-contrast photoresist with a contrast curve resembling a step function. The exposed pattern is the result of

binarizing the intensity distribution according to (for a positive photoresist),

$$R(x', y') = \begin{cases} 1 & I_{est} \geq K \\ 0 & \text{otherwise} \end{cases}, \tag{2.59}$$

where R is the normalized remaining photoresist thickness after development, and K is a threshold that depends on the exposure energy and time.

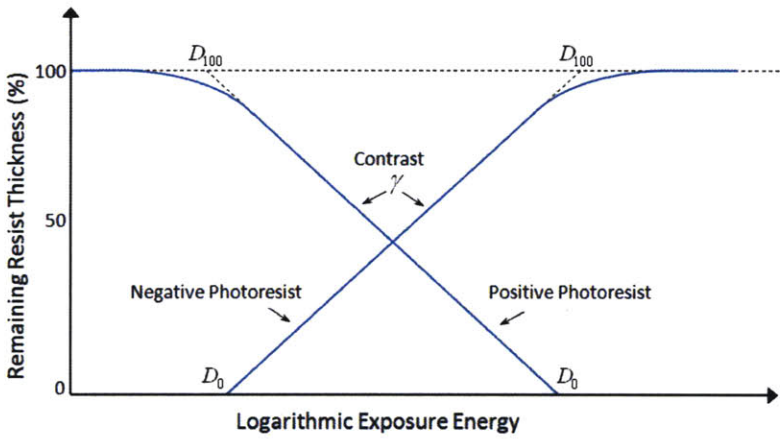


Figure 2-36: Idealized photoresist contrast curves.

The second method used is an approximated model of the photoresist in which the relative remaining photoresist thickness is given by (for a positive photoresist) [60],

$$R(x', y') = \begin{cases} 1 - \left(\frac{D(x', y')}{D_0}\right)^{\gamma} & D(x', y') \leq D_0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.60}$$

where D_0 is again the clearing dose necessary to give a complete clearing after development and D is the exposure dose: $D = I_{est}(x', y')t$, and t is the exposure time. This model is only good when the thickness of the photoresist is small compared to the imaged feature size. The third method used utilizes the real contrast curve from a given photoresist and is used to compute the exposed pattern.

The MSE after photoresist exposure is given by,

$$MSE_{after} = \frac{1}{N_{o_x} N_{o_y}} \sum_{x'=1}^{N_{o_x}} \sum_{y'=1}^{N_{o_y}} \left[R(x', y') - \hat{I}_{des}(x', y') \right]^2, \tag{2.61}$$

where $\hat{I}_{des}$ is the normalized desired intensity distribution (desired exposed pattern). For a given reconstructed intensity, the MSE_{after} is calculated for several exposure conditions and the minimum error is chosen. For example, for the ideal photoresist model, the MSE_{after} is computed for different K values within a small range to simulate the best exposure conditions.

The next step in the MER algorithm is to impose the amplitude constraint in which the amplitude of the reconstructed complex field is replaced by the square-root of the desired intensity distribution. Then the algorithm examines if any of the termination conditions have been met. The three possible termination conditions are: maximum number of iterations reached, $MSE_{before} <= tol_1$ and $MSE_{after} <= tol_2$, where tol_1 and tol_2 are user define tolerances. If any of the termination conditions has been met, the algorithm stops and the final CGH and reconstructed field are saved to disk; otherwise, the modified field gets propagated back to the CGH plane. The inverse propagation operation is done in a similar way to equation [73], but the complex conjugate of the transfer function is used. After propagation, an additional demodulation process is required for the off-axis and TIR CGHs. The last step in the iteration is to apply the zero absorption constraint ($\left|\hat{H}\right| = 1$) and binary constraint as defined in equation [67].

Objective Functions

The HOA allows the flexible minimization of one or more objective functions. The type of objective function used is application dependent. Different objective functions can guide the optimization algorithm to converge to diverse solutions. This is primarily due to the tradeoff between diffraction efficiency and uniformity. The objective function is chosen according to the desired CGH performance. For some applications, such as holographic

lithography, uniformity in the reconstruction (low MSE) is desirable. On the other hand, for other applications such as solar concentration, diffraction efficiency is more important and non-uniform reconstructions can be tolerated. In this thesis, four objective functions are investigated: diffraction efficiency upper bound, effective diffraction efficiency, MSE before photoresist exposure, and MSE after photoresist exposure. GAs allows flexibly choosing an arbitrary fitness function (objective function).

The diffraction efficiency upper bound was originally proposed by Wyrowski for Fraunhofer holograms [116], [117] and then generalized by Zhou et al. [118]. The diffraction efficiency upper bound, η_{ub}, refers to the maximum diffraction efficiency attainable in order to have a noiseless reconstruction within the object window and is given by,

$$\eta_{ub} = \frac{4\left[\left\langle \left| f_{ill}^{(0)} \right| \left| f_d^{(0)} \right| \right\rangle_{H_{size}}\right]^2 \left\langle \left| f_d^{(d)} \right|^2 \right\rangle_{O_{size}}}{\left[\left\langle \left| f_d^{(d)} \right|^2 \right\rangle_{O_{size}} + \left\langle \left| f_d^{(0)} \right|^2 \right\rangle_{H_{size}}\right]^2 \left\langle \left| f_{ill}^{(0)} \right|^2 \right\rangle_{H_{size}}}, \tag{2.62}$$

where $f_{ill}^{(0)}$ is the illumination field at the CGH plane ($f_{ill}^{(0)} = 1$ for the in-line geometry); $f_d^{(0)}(x', y')$ is the desired field at the reconstruction plane; $f_d^{(d)}(x, y)$ is the result from back-propagating the desired field to the CGH plane; the brackets,

$$\langle \rangle_{\Sigma} \equiv \iint_{\Sigma} dxdy, \tag{2.63}$$

represent a 2D integral over the surface, Σ. Equation 2.62 states that if a desired field at the reconstruction plane is "fully" specified (amplitude and phase), the diffraction efficiency upper bound can be estimated. This result is independent of encoding strategy as the desired field does not have any free parameters. However in many applications, such as holographic lithography, only the amplitude of the desired field is specified and as explained previously, the phase is a free parameter that can be used to improve the encoding process and thus maximize the diffraction efficiency upper bound. That is the purpose of the LDPE and LNPEPE phase masks. This formulation ignores any quanti-

zation effects so the phase and amplitude distributions are allowed to vary continuously and the signal's power spectrum is assumed to lie within the system's pass-band.

In order to understand the significance of the terms in equation 2.63, let us consider the case of an ideal encoding process that achieves a maximum η_{ub}. The desired field (at the reconstruction plane) is confined within the object window and is designed to have the same power as the illumination wave,

$$P_{tot} = \left\langle \left| f_{ill}^{(0)} \right|^2 \right\rangle_{H_{size}} = \left\langle \left| f_{d}^{(d)} \right|^2 \right\rangle_{O_{size}}, \tag{2.64}$$

where P_{tot} is the total power available in the system (other losses such as Fresnel reflections are ignored). Using equation 2.64, we rewrite equation 2.62 as,

$$\eta_{ub} = \frac{4\left[X_2\right]^2 P_{tot}}{\left[P_{tot} + X_1\right]^2 P_{tot}}, \tag{2.65}$$

where X_1 and X_2 are the corresponding terms of equation 2.62. The term X_1 is the amount of power of the back-propagated field that lies within the hologram window. As explained previously, different encoding masks might lead to different results. For example, for an amplitude mask containing a single rectangular binary pattern as that shown in Figure 2-26-a, a LNPEPE mask (Figure 2-26-b) can be designed to spread the field uniformly with most of its energy (close to 100%) contained inside the hologram window. The LDPE mask can also be designed to spread the field even more uniformly than the LNPEPE mask; however, it tends to leak some energy outside the hologram window. The term X_2 quantifies the uniformity of the back-propagated field amplitude over the hologram window. As discussed previously, uniform amplitude allows satisfying the zero absorption constraint for pure phase holograms and maximizes the signal's information transfer from amplitude (reconstruction plane) to phase (hologram plane). For the ideal case, let $X_1 = X_2 = P_{tot}$. This is the case when an optimum phase mask is chosen that spreads the field uniformly with all its energy lying inside the hologram window. For this ideal case, the diffraction efficiency upper bound of equation 2.65 is: $\eta_{ub} = 1$. From

the example shown in Figure 2-26 (without any optimization) it was proven that our proposed encoding strategies based on the LNPEPE and LDPE mask can significantly improve the diffraction efficiency upper bound.

Despite that η_{ub} was derived assuming a continuous phase and a band-limited signal, this metric is very useful in the evaluation of the performance of the optimization algorithm. We use this metric as one of the alternative fitness functions to guide the search from GAs. The diffraction efficiency upper bound fitness function to be minimized by GAs is given by,

$$F_1 = 1 - \eta_{ub}. \tag{2.66}$$

It is possible to design CGHs with higher diffraction efficiencies than those predicted by equation 2.62 at the expense of an increased noise inside the object window. This is another evidence of the intrinsic tradeoff between diffraction efficiency and uniformity. For holographic lithography, a certain degree of non-uniformity is tolerable as the photoresist performs some clipping of the incident energy above some threshold as explained before.

The effective diffraction efficiency objective function is used to quantify the amount of energy inside the reconstructed pattern from a CGH that satisfies the zero absorption and binary constraints. This metric takes into account the effect of quantization and the errors resulting from discarding the amplitude of the encoded signal and enforcing the zero absorption constraint. The effective diffraction efficiency fitness function is given by,

$$F_2 = 1 - \eta_{eff}, \tag{2.67}$$

and,

$$\eta_{eff} = \frac{\left\langle |R(x', y')|^2 \cdot B_{mask} \right\rangle_{O_{size}}}{\left\langle \left| f_{ill}^{(0)} \right|^2 \right\rangle_{H_{size}}}, \tag{2.68}$$

where R is the diffracted field, B_{mask} is a binary mask that is 1 inside every binary pattern and 0 elsewhere – only the energy inside the binary patterns is integrated.

The fitness functions based on the MSE are,

$$\begin{aligned} F_3 &= MSE_{before}, \\ F_4 &= MSE_{after}, \end{aligned} \tag{2.69}$$

where MSEbefore and MSEafter are defined in equations 2.55 and 2.61 respectively.

In some applications, a hybrid fitness function that compromises between diffraction efficiency and uniformity might be desirable. An example of a hybrid fitness function is,

$$F_{hybrid} = CF(\eta) + \sigma, \tag{2.70}$$

where C is a penalty parameter, σ measures the non-uniformity of the reconstruction (the unbiased version of the MSE estimator), F is,

$$F(\eta) = \begin{cases} 0 & \eta \geq \eta_d \\ (\eta_d - \eta_{eff})^2 & \eta < \eta_d \end{cases}, \tag{2.71}$$

and η_d is a desired diffraction efficiency.

The choice between the presented fitness functions depends on the application. In the next section, simulation results of CGHs designed using F_3 for holographic lithography and F_2 for solar concentration will be presented. The results will be compared to those when the other fitness functions are used to guide the search.

2.4.4 Optimization Results

In this section, examples of the optimization of in-line, off-axis and TIR CGHs based on the HOA are presented. These CGHs are designed for high-resolution holographic lithography. The optimization problem is solved using the two proposed encoding strategies: LDPE and LNPEPE masks. The resulting holograms and reconstructed fields are evaluated using the error metrics of equations 2.55 and 2.61 (for the case on an ideal

photoresist – equation 2.59). In addition, the CGHs are optimized with and without the binary constraint to show the effect of quantization errors. The optimization results are compared to those of CGHs designed using the encoding strategies based on the diffracted field (DF) and the simulated optically recorded hologram (SORH). The convergence of the HOA algorithm is studied for different control parameters of the GAs block, including crossover fraction, fitness function and population size, as well as geometrical parameters such as propagation distance. A method for extending the depth of focus (DOF) of a CGH is proposed. The HOA is implemented on a graphics processing unit and its computational performance is evaluated. Additional optimization examples are included in Appendix A.

Optimization of In-line CGHs based on the LDPE Mask

An in-line CGH is optimized to reconstruct a resolution target at the photoresist plane located at a distance, $d = 150\mu$m, from the hologram. The desired intensity distribution is shown in Figure 2-37. The smallest line in the resolution target is 700nm and the largest is 2.3μm wide. The desired intensity distribution is scaled to have an input power equal to unity,

$$P_{in} = \langle I_{des} \rangle_{H_{size}} = 1. \tag{2.72}$$

To generate the LDPE mask the pattern is decomposed as shown in Figure 2-20, resulting in 99 binary patterns. The optimization parameters are indicated in Table 2.1. The working distance and hologram size are chosen to give an effective numerical aperture: $NA_{eff} = 1$. The corresponding Nyquist (equation 2.51) and evanescent (equation 2.5) cut-offs are: $u_{Nyq} = u_{ev} = 1,879.69\text{mm}^{-1}$. The selected pixel size results in a maximum spatial frequency (equation 2.3) of: $u_{\max} = 2,500\text{mm}^{-1}$. In this geometry, the maximum pixel size tolerated to fully utilize the system's pass-band is: $\delta_{pix,\max} = 266$nm (for $u_{\max} = u_{ev}$). The diffraction limit resolution (equation 2.6) is: $\Lambda = 266$nm. The initial range and bounds for an individual chromosome are set according to the ranges of the diffuser and frequency factors. No initial population is specified – the initial population

is fully generated by stochastic creation. As the hologram is designed for lithography, the MSE metric before photoresist exposure (equation 2.55) is chosen to guide the GAs block of the HOA. For the MER block, both MSE metrics before and after photoresist exposure are used to evaluate the reconstructed signal. In this example an ideal photoresist model (equation 2.59) is used. For repeatable results an initial random is provided. The number of optimization variables in the reduced complexity problem is: $nvars = 198$ (instead of 2.25×10^6 variables required for the regular optimization problem).

Table 2.1: Optimization parameters: in-line CGH - LDPE mask.

Wavelength (λ)	532nm	**Number of Generations (GAs)**	100
Working Distance (d)	150μm	**Population Size**	100
Pixel Size (δ_{pix})	200nm	**Random Seed**	3
Hologram Size (H_{size})	300μm	D_{factor} **Range**	[0,10]
Object Window (O_{size})	180μm	F_{factor} **Range**	[0,1]
SBP (After Padding)	1500 $\times$ 1500	**Fitness Function**	MSE_{before}
Elite Children	5	**Number of Variables**	198
Crossover Fraction	0.6	**Number of Iterations (MER)**	400

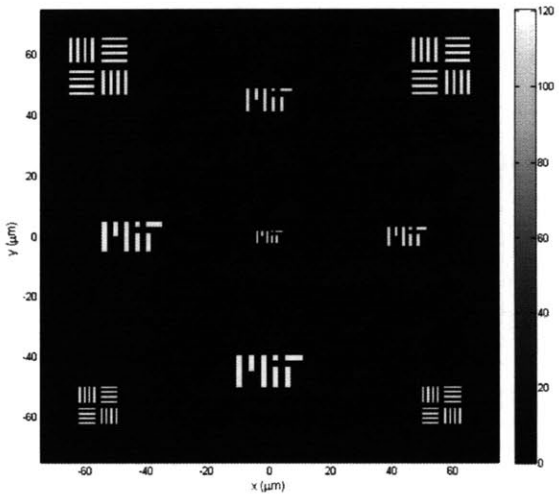


Figure 2-37: Desired intensity distribution: resolution target.

We first present the results from the GAs block. Figure 2-38-a shows the phase

distribution of the final CGH after 100 generations. In this example, no binary constraint is imposed on the GAs block. Figure 2-38-b shows the corresponding LDPE mask for the best individual in the final generation. This phase mask shows the tendency of GAs to converge to solutions in which local diffusers with small diffusivity (small frequency factor) correspond to narrow lines and diffusers with larger diffusivity correspond to wider lines. This can be physically understood as the size of the line (aperture stop in the equivalent problem) is inversely proportional to the width of the Fresnel diffraction pattern. Narrow lines suffer severe diffraction, which spreads the input energy evenly over the entire hologram window so that no additional diffuser is required. For large apertures, the diffraction effects are milder so a local diffuser is needed to spread the energy uniformly facilitating the improvement of the information transfer from amplitude (photoresist plane) to pure phase (CGH plane).

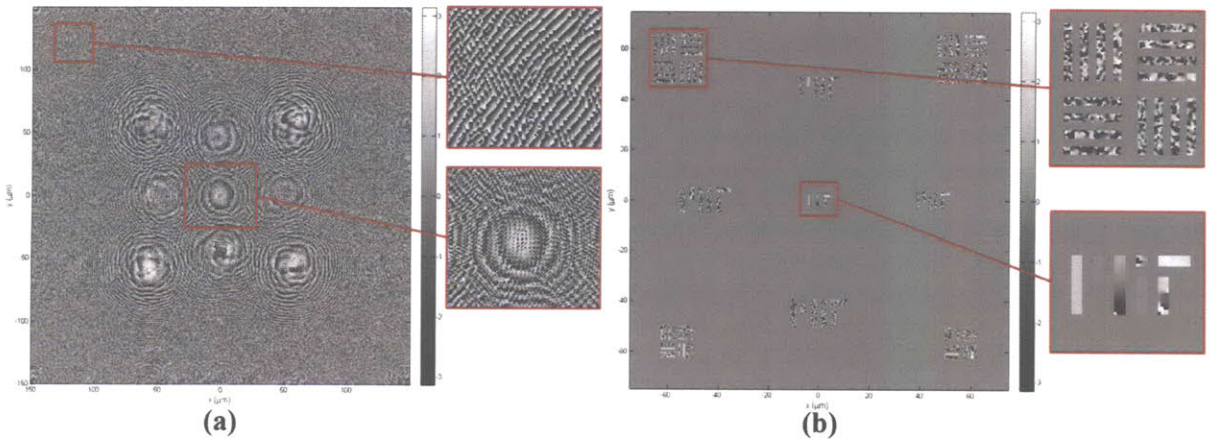


Figure 2-38: (a) Phase distribution of optimized CGH after GAs block; (b) Optimized LDPE mask.

The amplitude distribution of the reconstructed field is shown in Figure 2-39. The reconstructed amplitude of Figure 2-39 is very close to the desired distribution of Figure 2-37 which shows the importance of the LDPE phase mask in improving the signal encoding process. The grainy speckle-like noise in the reconstruction corresponds to

unresolved phase values typically encountered in diffused illumination from the inability of the local diffusers to keep the signal's power spectrum inside the pass-band of the system. This noise will be significantly reduced by the MER block. An alternative method is to use partially incoherent illumination for the reconstruction, as discussed later. The convergence of the GAs block is shown in Figure 2-40. This figure shows the fitness score of the best individual and the mean score of the population during the evolution process. The final score of the best individual is: $MSE_{before} = 89.025$. The corresponding diffraction efficiencies are: $\eta_{eub} = 73.054\%$ and $\eta_{eff} = 75.52\%$.

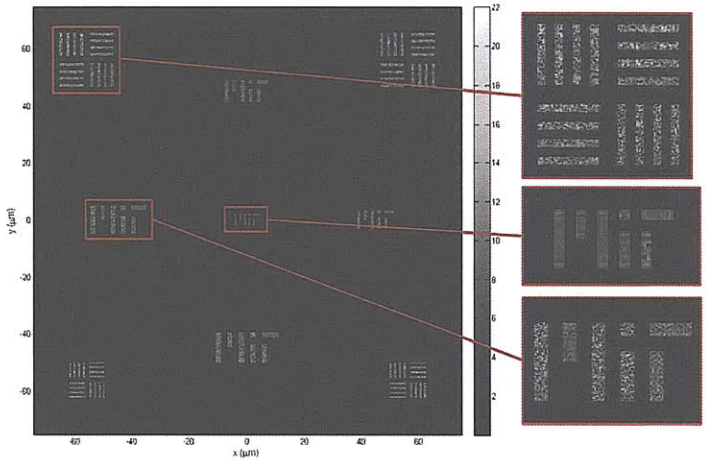


Figure 2-39: Reconstructed amplitude distribution after GAs block.

The next step in the HOA is the MER block. The binary constraint is implemented in this block. The final binary phase CGH is shown in Figure 2-41. The amplitude distribution of the reconstructed field is shown in Figure 2-42. This figure also shows a cross-section of the reconstructed intensity distribution (modulus square of the amplitude) that exposes the photoresist. The convergence of the MER block is shown in Figure 2-43.

In order to estimate the effect of the binary constraint and related quantization errors, the MER block is computed again but without binarizing the phase of the CGH (multi-level hologram). The amplitude distribution of the reconstructed field and the

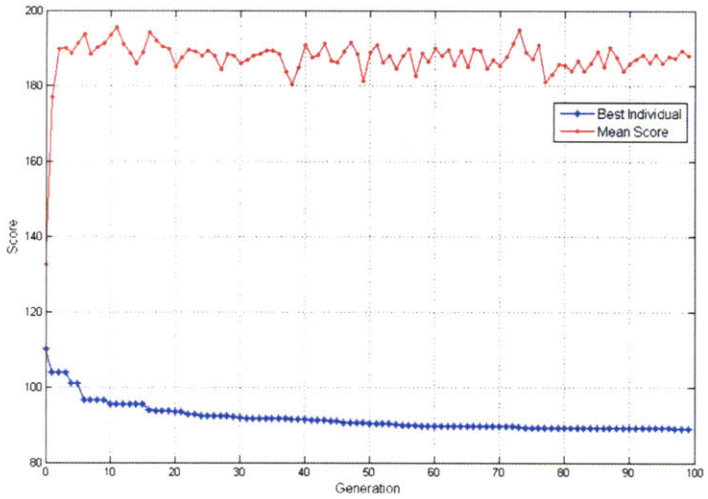


Figure 2-40: Convergence of GAs.

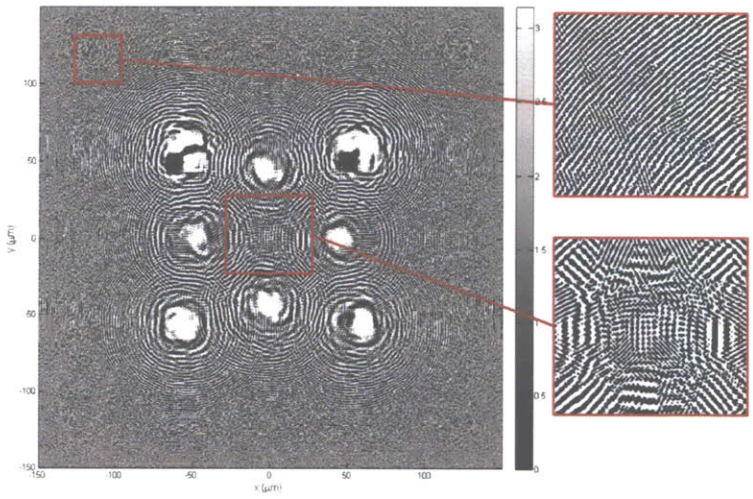


Figure 2-41: Phase distribution of final binary phase CGH.

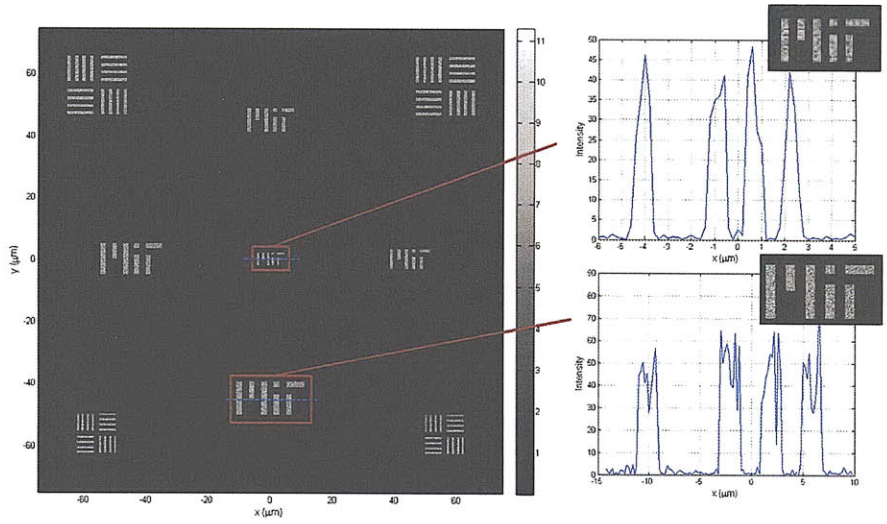


Figure 2-42: Reconstructed field from optimized binary phase CGH.

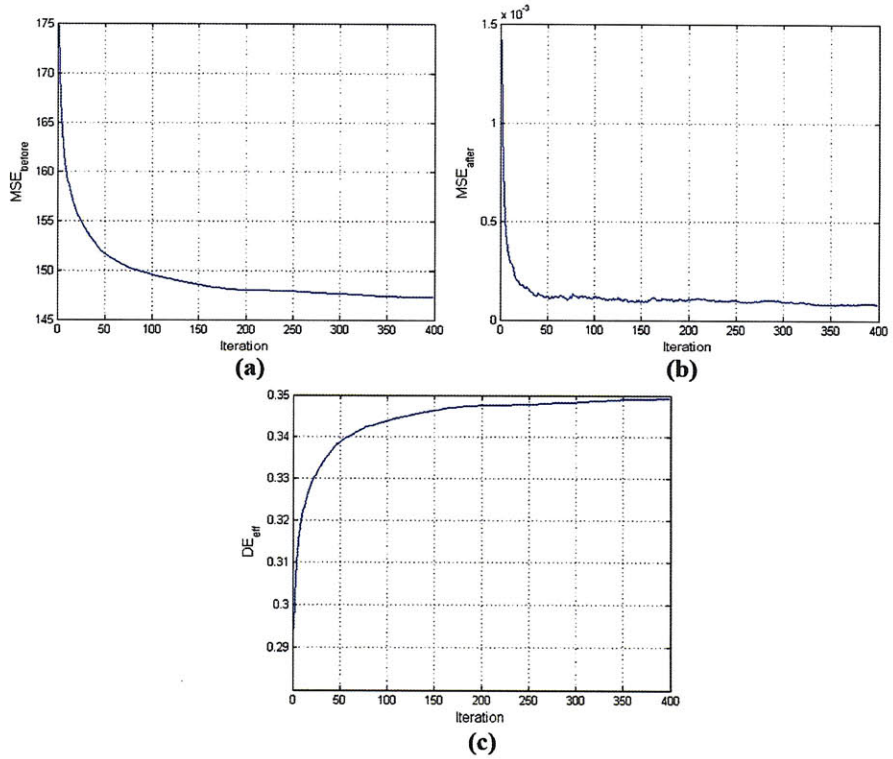


Figure 2-43: Convergence plots: (a) MSE_{before}; (b) MSE_{after}; (c) η_{eff}.

intensity cross-sections are shown in Figure 2-44. The convergence of the MER block is shown in Figure 2-45. The MSE before photoresist exposure gets substantially reduced compared to that of the binary case, and the MSE after exposure becomes zero at the iteration number 97 (this would normally terminate the algorithm, but in this example the termination condition was set to reach a maximum number of iterations).

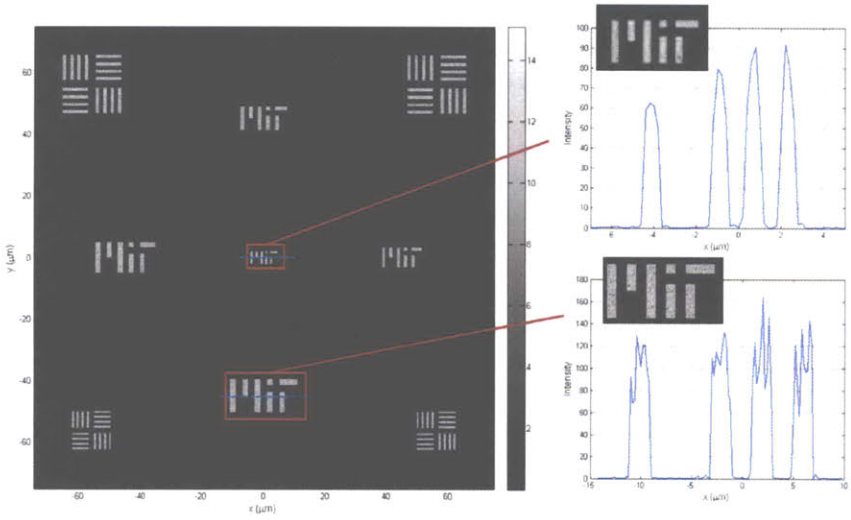


Figure 2-44: Reconstructed field from multi-level phase CGH.

Both optimized CGHs (binary and multi-level) reconstruct high quality intensity patterns that meet the resolution requirements, making the system feasible for holographic lithography. In addition, these CGH patterns can be efficiently fabricated using electron-beam lithography as will be discussed later.

Optimization of In-line CGHs based on the LNPEPE Mask

In this example, an in-line CGH is optimized based on the LNPEPE mask to also reconstruct the resolution target of Figure 2-37. The simulation parameters are included in Table 2.2. The initial range is set equal to the variables' lower and upper bounds. A partial initial population is specified using the geometrical optics solution as explained previously. The initial score for this individual is: $MSE_{before} = 97.5$. The remaining

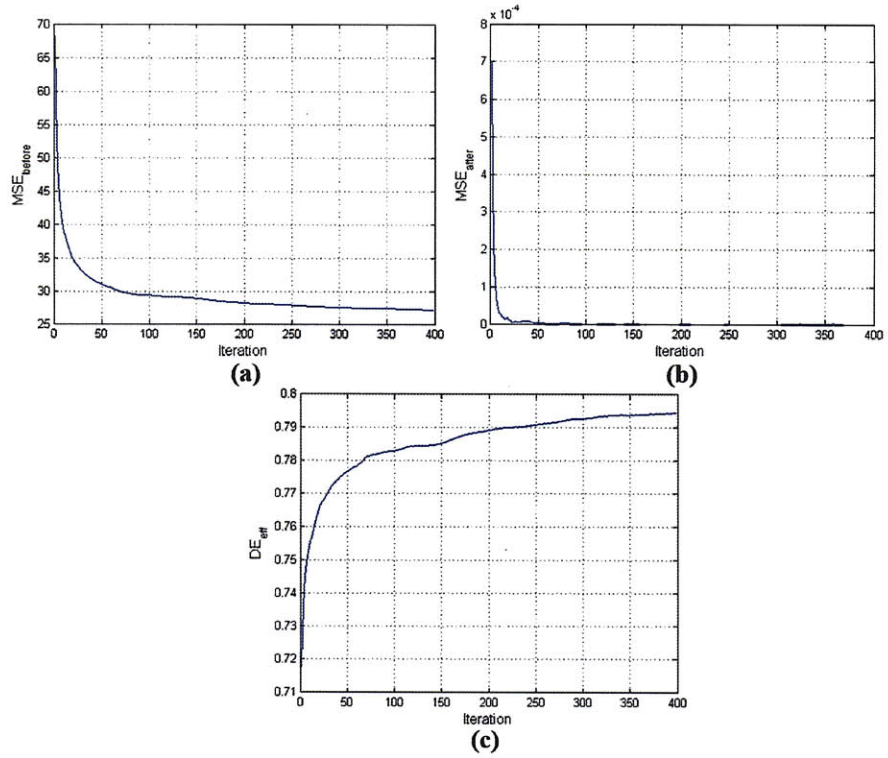


Figure 2-45: Convergence plots: (a) MSE_{before}; (b) MSE_{after}; (c) η_{eff}.

individuals in the initial population are generated by the creation function.

Again, we first present the results from the GAs block. Figure 2-46-a shows the phase distribution of the final CGH after 300 generations (no binary constraint is imposed). Figure 2-46-b shows the corresponding LNPEPE mask for the best individual in the final generation. This phase mask shows the tendency of GAs to distribute the signal evenly over the frequency spectrum in a process similar to amplitude modulation (AM). The phase mask is composed of grating-like patterns with varying periods and oriented in different directions. Each grating allocates their corresponding signal at a different section on the Fourier plane to form the final multiplexed hologram.

The reconstructed amplitude distribution is shown in Figure 2-47. This distribution is very close to the desired one (Figure 2-37) and in contrast to the reconstruction Figure

Table 2.2: Optimization parameters: in-line CGH - LNPEPE mask.

Wavelength (λ)	532nm	**Number of Generations (GAs)**	300
Working Distance (d)	150μm	**Population Size**	200
Pixel Size (δ_{pix})	200nm	**Partial Initial Population**	20%
Hologram Size (H_{size})	300μm	**Variables Lower Bound**	-1
Object Window (O_{size})	180μm	**Variables Upper Bound**	1
SBP (After Padding)	1500 $\times$ 1500	**Fitness Function**	MSE_{before}
Elite Children	5	**Number of Variables**	396
Crossover Fraction	0.7	**Number of Iterations (MER)**	400

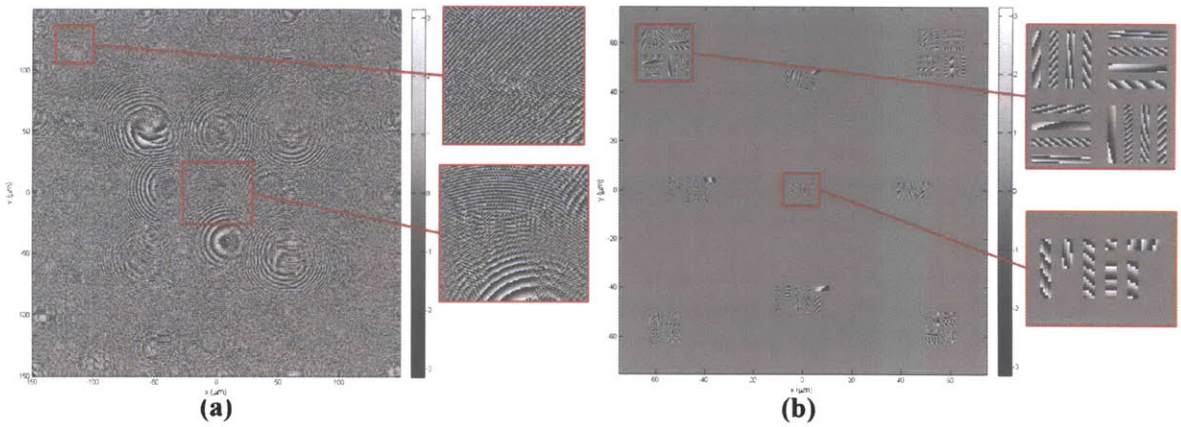


Figure 2-46: (a) Optimized CGH after GAs block; (b) Optimized LNPEPE mask.

2-39 (from the LDPE mask), no grainy speckle-like noise is present. The convergence of the GAs block is shown in Figure 2-48. The final score of the best individual is: $MSE_{before} = 47.804$. The corresponding diffraction efficiencies are: $\eta_{ub} = 76.25\%$ and $\eta_{eff} = 72.36\%$.

We now describe the results from the MER block starting with the case in which the binary constraint is imposed. The final binary phase CGH is shown in Figure 2-49. The reconstructed amplitude distribution and intensity cross-sections are shown in Figure 2-50. The convergence of the MER block is shown in Figure 2-51.

The MER block is computed again but without the binary constraint (multi-level

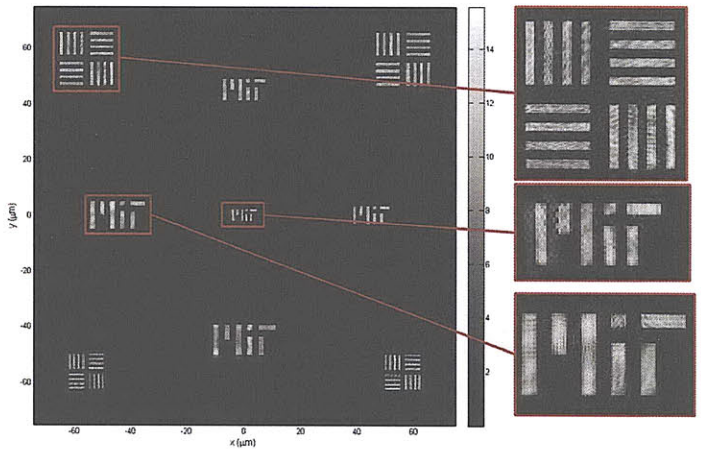


Figure 2-47: Reconstructed amplitude distribution after GAs block.

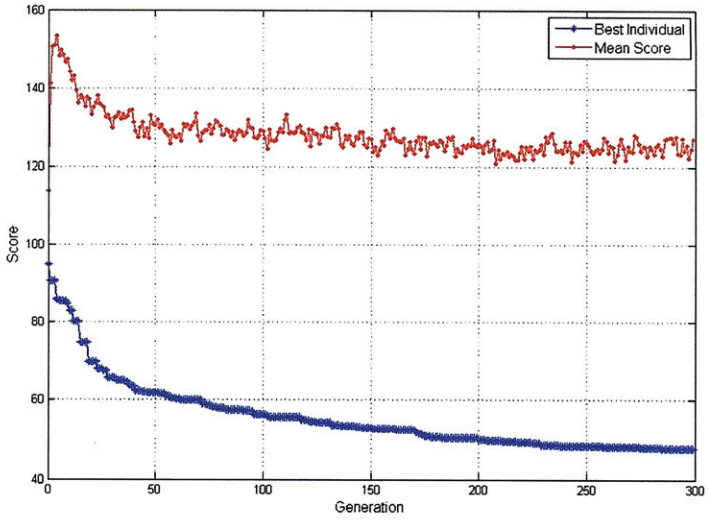


Figure 2-48: Convergence of the GAs block.

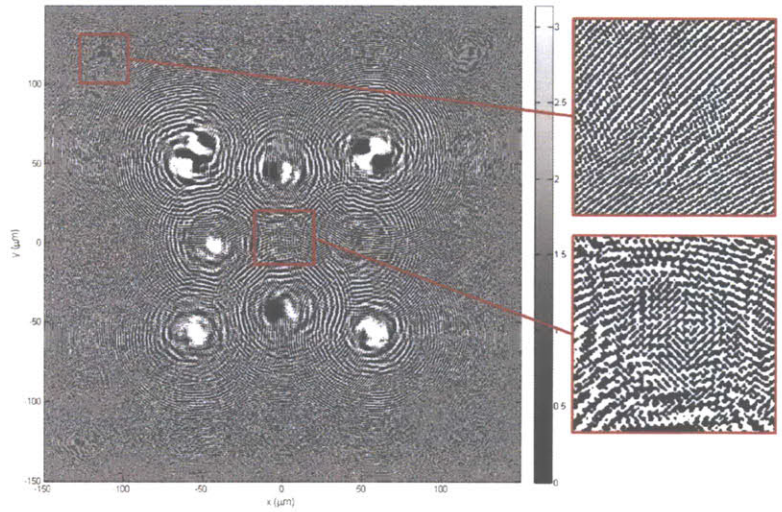


Figure 2-49: Final optimized binary CGH using LNPEPE encoding strategy.

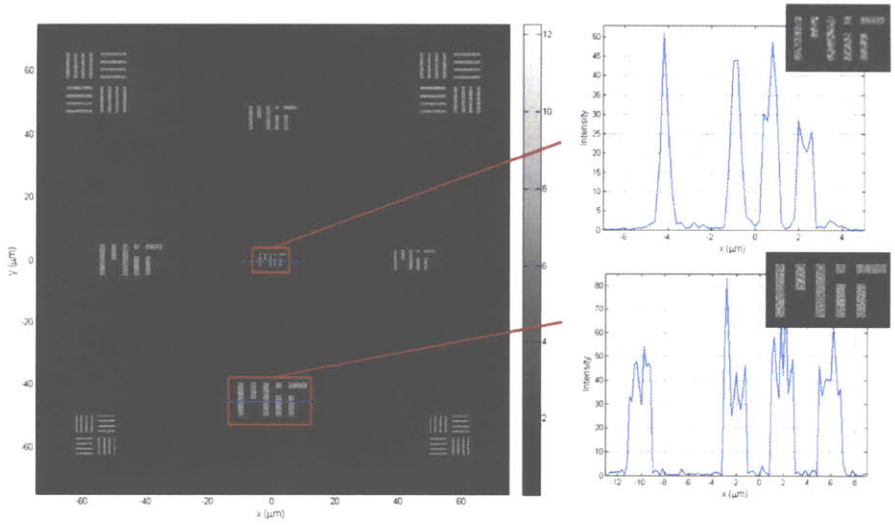


Figure 2-50: Reconstructed field from optimized binary phase CGH.

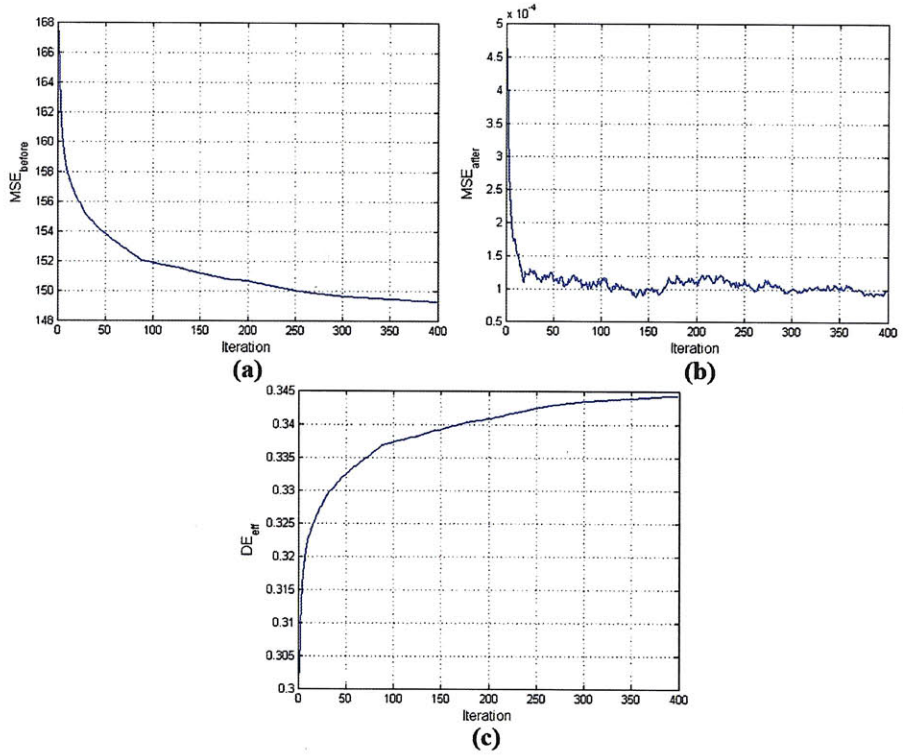


Figure 2-51: Convergence plots: (a) MSE_{before}; (b) MSE_{after}; (c) η_{eff}.

CGH) to estimate the effect of quantization errors. The reconstructed amplitude distribution and intensity cross-sections are shown in Figure 2-52. The convergence of the MER block is shown in Figure 2-53.

In-line CGH based on the Diffracted Field Encoding Strategy

We now evaluate the performance of an in-line CGH designed based on the diffracted field encoding strategy shown in Figure 2-35-a. In this encoding strategy, the desired amplitude signal is back-propagated to the hologram plane. The amplitude and phase of the resulting complex field are shown in Figure 2-54. To form a pure phase CGH, the amplitude distribution is discarded and set to unity. Also, the phase distribution may be quantized whether the CGH is binary or multi-level. The reconstructed amplitude distribution for the case of a multi-level CGH is shown in Figure 2-55. The corresponding

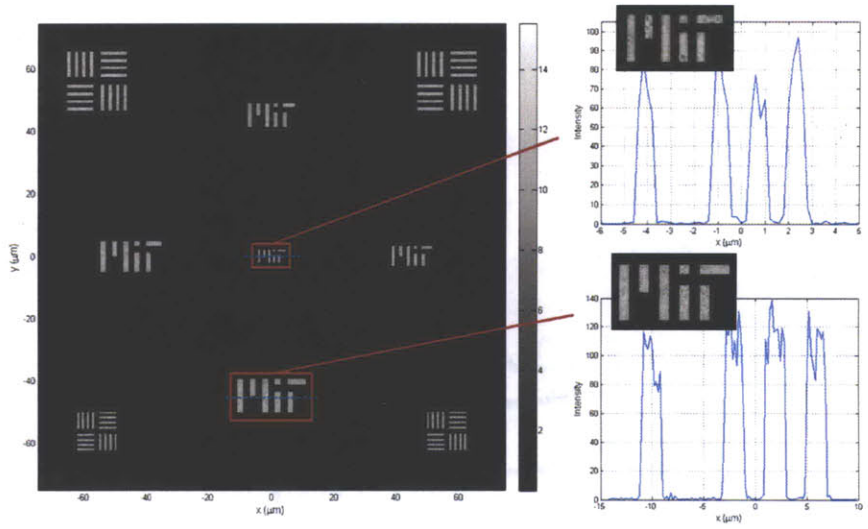


Figure 2-52: Reconstructed field from multi-level phase CGH.

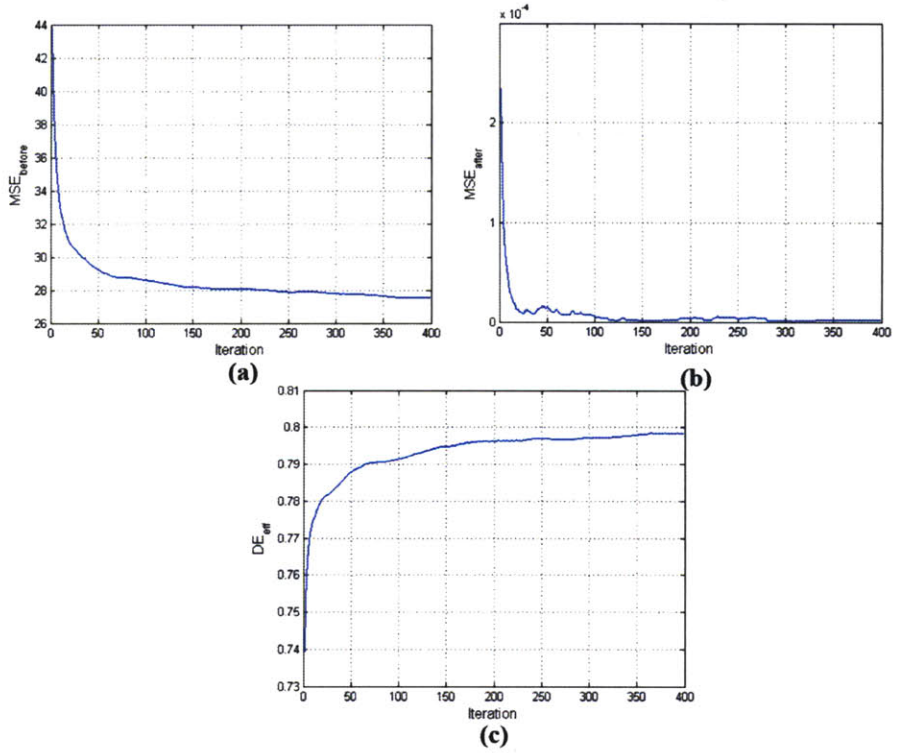


Figure 2-53: Convergence plots: (a) MSE_{before}; (b) MSE_{after}; (c) η_{eff}.

MSE before photoresist exposure and effective diffraction efficiency are: $MSE_{before} = 293.48$, and $\eta_{eff} = 49.89\%$.

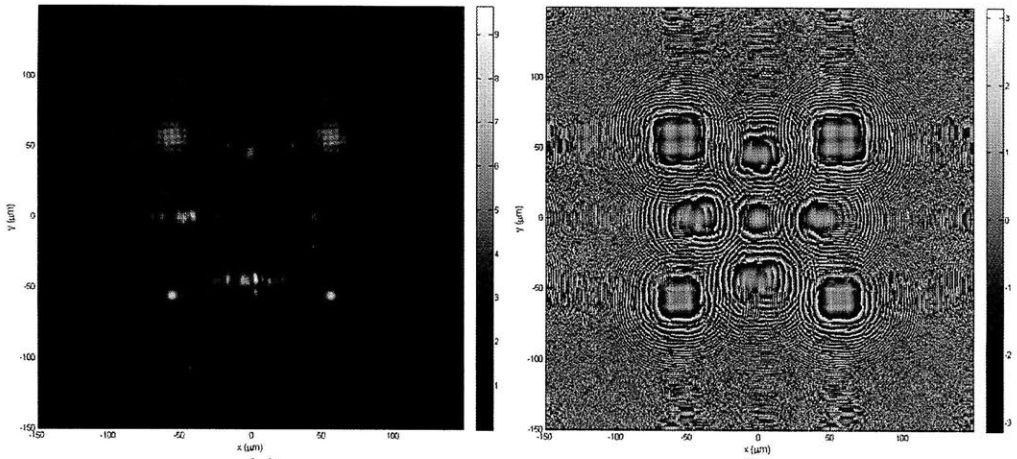


Figure 2-54: Diffracted complex field: (a) Amplitude; (b) Phase.

In-line CGH based on the Simulated Optically Recorded Hologram Encoding Strategy

An in-line CGH is designed based on the simulated optically recorded hologram (SORH) encoding strategy shown in Figure 2-35-b. The hologram's transmittance function is computed according to equation 2.47 with $\kappa = 1$. Figure 2-56-a shows the corresponding multi-level CGH. The reconstructed amplitude distribution is shown in Figure 2-56-b. The MSE before photoresist exposure and effective diffraction efficiency are: $MSE_{before} = 314.41$, and $\eta_{eff} = 3.39\%$. The CGH performance varies according to the selected normalization constant, κ.

Comparison of Encoding Strategies

Figure 2-57 shows a comparison of the MSE before photoresist exposure and the effective diffraction efficiencies for the four encoding strategies described above (before additional

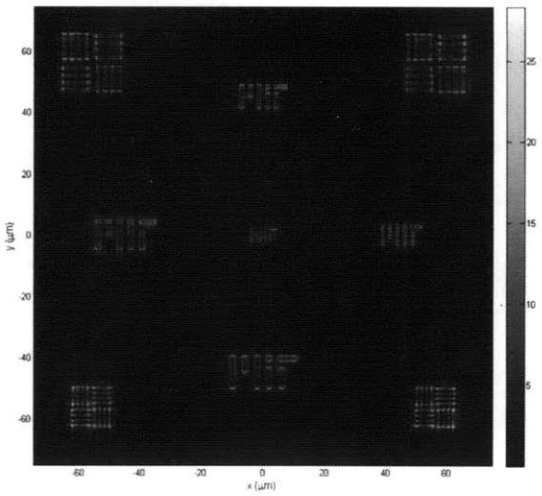


Figure 2-55: Reconstructed amplitude distribution at photoresist plane.

optimization by the MER block). As can be seen, the proposed LDPE and LNPEPE masks improve significantly the encoding process producing CGHs that reconstruct patterns very close to the desired signal (near the global optimum). The lowest MSE in this example is achieved by the LNPEPE mask as its reconstruction is free from the grainy noise that corrupts the reconstructions from holograms optimized using the LDPE mask. In addition, the CGHs optimized using the proposed encoding strategies have larger effective diffraction efficiency than that of those based on the diffracted field and SORH strategies. Higher diffraction efficiencies can be achieved by choosing a different fitness function in the GAs block (such as equations 2.66 or 2.67).

Selection of Problem Parameters

The multi-point search performed by the GAs block can be guided by our specific choice of control parameters. As explained previously, these parameters can be tuned to help the optimization algorithm search the nonlinear space more efficiently and to avoid premature convergence. The main control parameters are: crossover fraction, number of elite children, population size, initial range, number of generations, problem geometry

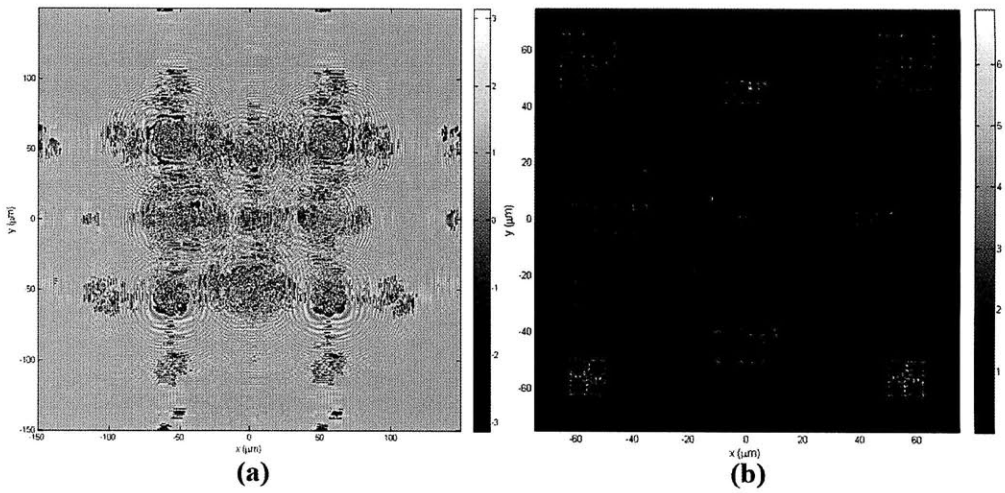


Figure 2-56: (a) Multi-level CGH optimized using the SORH encoding strategy; (b) Reconstructed amplitude.

and type of fitness function used. Four examples are presented to illustrate how the convergence of the GAs block varies as a function of crossover fraction, population size, fitness function and working distance.

In the first example, we examine the convergence of GAs for the optimization of an in-line CGH based on the LNPEPE mask for different crossover fraction values. Figure 2-58-a shows the fitness values of the best individual in the population for crossover fractions in the range [0.2, 0.9]. The optimization parameters used to generate this figure are similar to those in Table 2.2 but with a population size of 100 individuals and a pixel size of: $\delta_{pix} = 100\text{nm}$ ($SBP = 3000 \times 3000$). The crossover fraction controls how many individuals at each generation are created from the gene exchange between parents in the crossover process. If the crossover fraction is too large, the algorithm might suffer premature convergence as no new information is introduced in the genetic pool, potentially stagnating at a local minimum. In contrast, if the crossover fraction is too small, the mutation process takes over and the algorithm rapidly approaches an extremely inefficient random walk. The choice of crossover fraction is problem dependent and can be optimized by running the algorithm with different crossover fraction values for a short

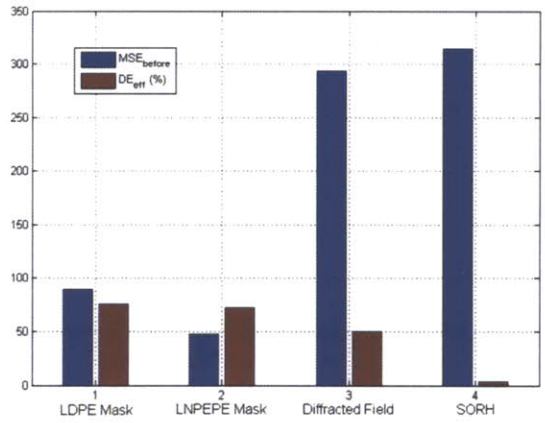


Figure 2-57: Comparison of encoding strategies.

number of generations and evaluating the results. Each case needs to be run several times to provide statistical significance due to the stochastic nature of the GAs searching process. For the case of the LNPEPE mask (Figure 2-58-a), a crossover fraction of 0.7 resulted in the lowest MSE_{before} after 20 generations. The population's mean fitness score for varying crossover fractions is shown in Figure 2-58-b. As shown, small crossover fractions generally result in large errors (except for the case of $F_{xover} = 0.3$, where the random walk landed in a lower error region), and populations with large crossover fractions do not explore the nonlinear space efficiently, displayed as small amplitude oscillations of the mean population score. For the LDPE mask it was found that a crossover fraction of 0.6 leads to the best performance of GAs.

For the second example, several population sizes are used to estimate the performance of GAs in the optimization of an in-line CGH based on the LNPEPE mask. The simulation parameters are the same as the previous example but with: $F_{xover} = 0.7$. Figure 2-59-a shows the algorithm's convergence for the best individual in the population. A larger number of individuals in the population (searching points) are expected on average to increase the probability of landing near the global optimum, as well as sample the nonlinear optimization space more efficiently. The initial range and creation strategy are also important in introducing population diversity and avoiding locality effects (cluster

of individuals) that might lead to premature convergence. For the optimization of Figure 2-59, 20% of the population was generated using the geometrical optics solution and the rest was created stochastically and distributed uniformly within the initial range. As second consideration in the choice of population size is the computational memory and time. Larger populations need to be stored by the algorithm, increasing the required memory and computational time. The reduced complexity proposed in this thesis makes it possible to compute large population sizes with conventional computational means. In addition, the algorithm is implemented in parallel on a graphics processing unit, reducing the required computational time as it will be explained later. Figure 2-59-b shows the corresponding convergence for the mean population score with varying population sizes.

In the third example we compare the GAs' performance as a function of fitness function. The GAs block provides the flexible choice of fitness function to guide the algorithm to the desired solution. This cannot be done in the MER block. As discussed previously for holographic lithography, the MSE before photoresist exposure (equation 2.69) has been found to give the best results. However, there are other applications, such as solar collection and optical trapping, in which high diffraction efficiencies are desirable and the uniformity in the reconstruction is of very little importance. For these applications, a better choice of fitness function is that of equations 2.66 or 2.67. Figure 2-60-a shows the convergence for the best individual based on the fitness functions given by the upper bound and effective diffraction efficiencies. The optimization was performed for an in-line CGH using the LNPEPE mask with the same simulation parameters as in Table 2.2, but with a population size of 100 individuals. Figure 2-60-b compares the performance metrics at the last generation for the two optimization runs.

For the final example, the performance of the GAs block is compared against 6 different working distances. The in-line CGH is optimized using the LNPEPE mask with the same parameters as in Table 2.2, but with: $\delta_{pix} = 100\text{nm}$ ($SBP = 3000 \times 3000$), $PopSize = 50$, and $Gen = 20$. Figure 2-61 shows the fitness score of the best individual

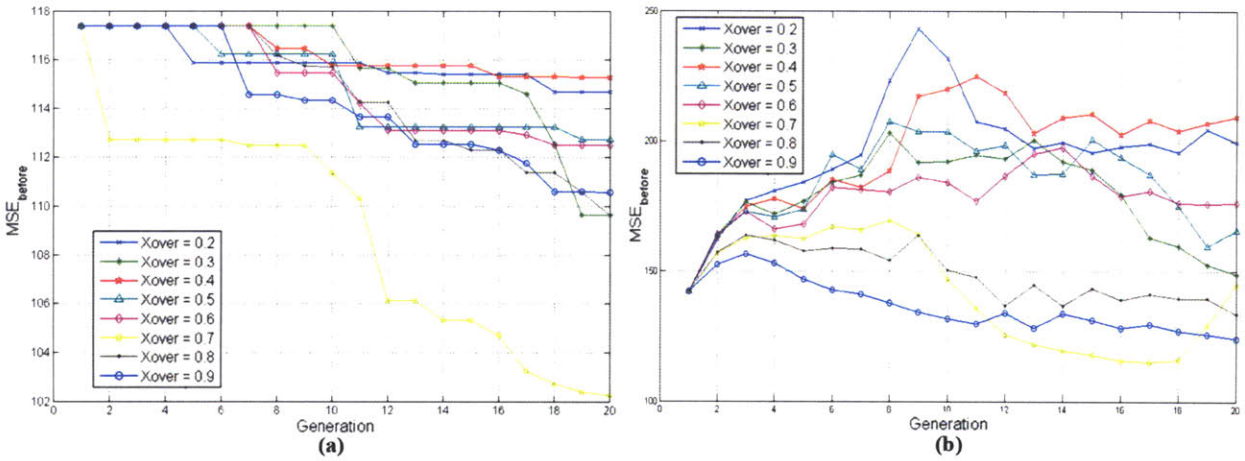


Figure 2-58: GAs convergence for different crossover fraction values: (a) Best individual; (b) Population mean.

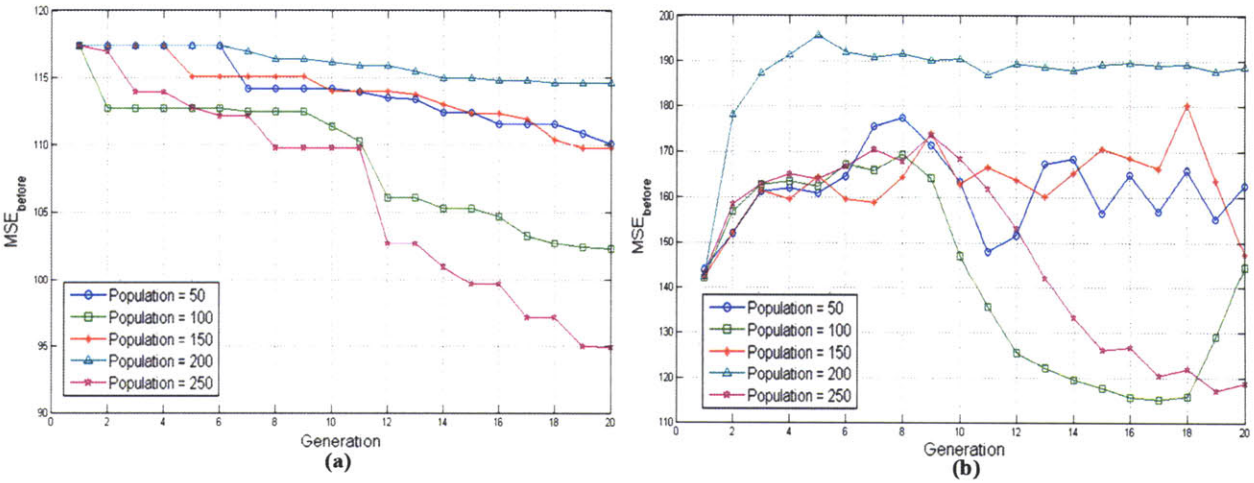


Figure 2-59: GAs convergence for different population sizes: (a) Best individual; (b) Population mean.

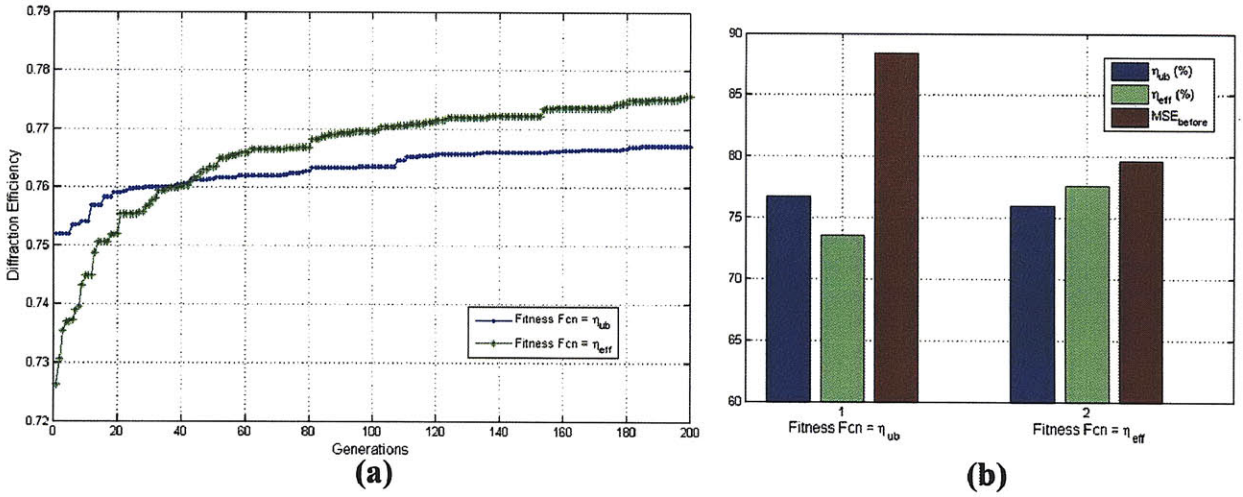


Figure 2-60: (a) Convergence of best individual for different fitness functions; (b) Comparison of fitness values at last generation.

at the final generation for the different working distances. As expected, the best performance is achieved at $d = 150\mu$m, as this results in an effective numerical aperture (equation 2.2): $NA_{eff} = 1$.

Extending the Depth of Focus

In holographic lithography, extending the system's depth of focus (DOF) is of practical importance in order to tolerate potential axial misalignments of the substrate to be exposed. A misalignment tolerance in the range of $1 - 4\mu$m is typically considered acceptable. However, high-resolution CGHs are designed to have large effective numerical apertures resulting in a small DOF. The theoretical DOF is given by,

$$\Delta z = \frac{\lambda}{2NA_{eff}^2}. \tag{2.73}$$

In order to tolerate a 4μm misalignment ($\pm\Delta z$; $\Delta z = 2\mu$m), the system needs to be designed to have a maximum effective numerical aperture of $NA_{eff} = 0.365$ (for $\lambda =$

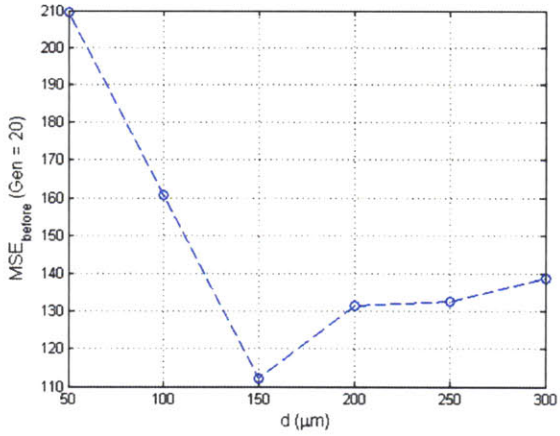


Figure 2-61: Best individual's score for different working distances.

532nm) which results in a diffraction limit resolution of $\Lambda = 0.73\mu$m. This tradeoff between DOF and resolution is characteristic of most conventional optical systems.

In this thesis we propose a method to extend the system's DOF, while maintaining a desirable resolution by performing a modification of the hologram's encoding process. The concept is similar to hologram multiplexing in which several pages of information are encoded in the same hologram and are designed to reconstruct under different decoding conditions (e.g., different wavelengths or off-axis angles of the probing waves). In the proposed method, the HOA presented previously is modified to enforce the amplitude constraint not only at the reconstruction plane, but also at additional parallel planes as shown in Figure 2-62-a. The number of planes is given by the desired DOF. However, due to the limited channel capacity of binary phase CGHs, a large number of planes result in increased noise in the reconstruction. In the results presented in this section, the modifications of the encoding process to extend the DOF were done in the MER block. The algorithm begins the search from the initial binary phase CGH of Figure 2-41. The theoretical DOF of this hologram is: $\Delta z = 266$nm. Four planes (two on each side of the reconstruction plane) are used to extend the DOF to tolerate an axial misalignment of 1.064μm. The diffracted field at the five planes (reconstruction plane

plus four additional planes) is computed by means of a forward Fresnel propagation. The amplitude constraint is enforced at each plane and the resulting fields are back propagated to the CGH plane. At the CGH plane, the fields are added coherently and the zero absorption and binary constraints are enforced. The algorithm repeats for the specified number of iterations. Figure 2-62-b shows the MSE before photoresist exposure for the regular and extended DOF CGHs for different reconstruction distances. From this figure, it is evident that the proposed method serves in extending the system's DOF. The slightly higher MSE at the reconstruction plane in the extended DOF version is the result of the limited channel capacity of binary phase holograms. The corresponding reconstruction amplitude distribution is shown in Figure 2-63 (only the smallest MIT pattern at the center of the resolution target is shown).

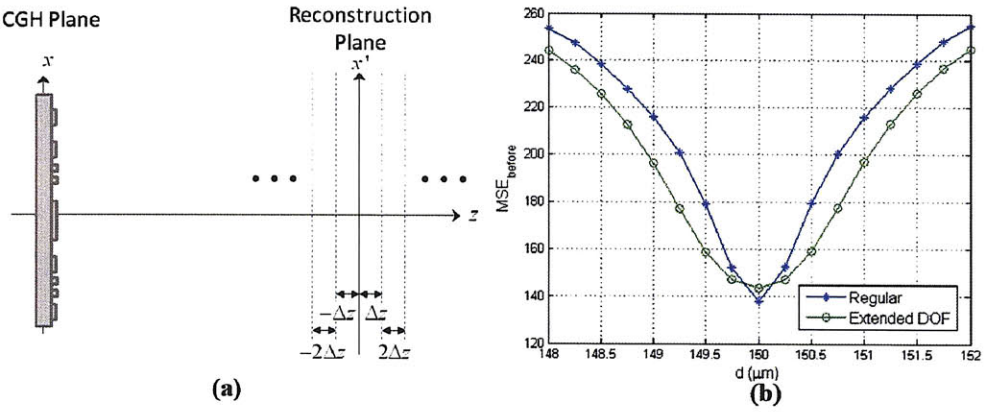


Figure 2-62: (a) Extending DOF concept; (b) MSE_{before} before and after DOF extension.

Parallel implementation on a Graphics Processing Unit

Over the past years, the development of graphic processing units (GPUs) have been driven by the insatiable market demand for real-time, high-definition 3D graphics, evolving into multithreaded, multi-core, highly parallel systems with high memory bandwidths and tremendous computational power. GPUs can be used more efficiently to solve complex problems than central processing units (CPUs) as they are designed to have more

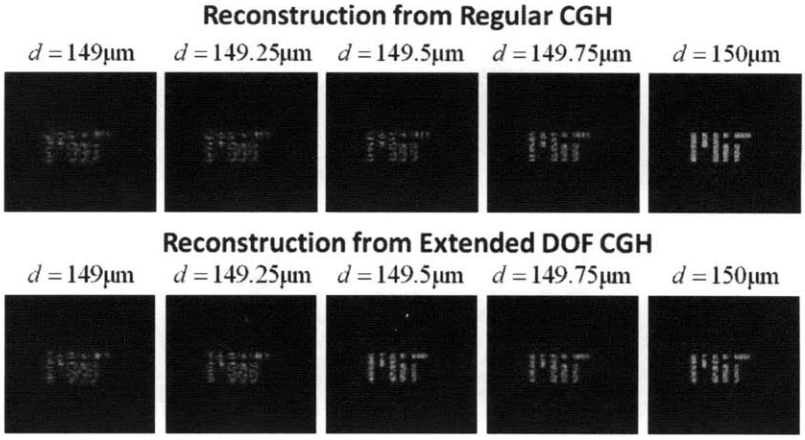


Figure 2-63: Reconstructed amplitude distributions around the focus.

transistors devoted to data processing than data caching or flow control. Figure 78 show a comparison between the computational performance of some NVIDIA GPUs and Intel CPUs [119]. In November 2006, NVIDIA introduced CUDATM (Compute Unified Device Architecture), a general purpose parallel computing architecture used to program complex computational problems in a more efficient way to be solved on NVIDIA GPUs. CUDA comes with a software environment that allows the development in C as a high-level programming language as well as supporting other platforms such as the Matlab Executable (MEX) interface [120]. In addition, codes written in CUDA are scalable and capable of operating with different compatible NVIDIA GPUs independently of the number of their multiprocessors.

The possibility of using a personal computer with a GPU to achieve the performances of an expensive CPU cluster in a simple and cost effective manner has revolutionized computational physics in a wide range of fields. GPUs have been used in a variety of applications that require high computational power such as medical imaging [121], molecular dynamics [122], fluid mechanics [123], astrophysics [124] and financial simulation [125]. In the area of holography, GPUs have been used on 3D displays [126], [127], [128], [129] and optical tweezers [130] for the near real-time computation of holograms displayed

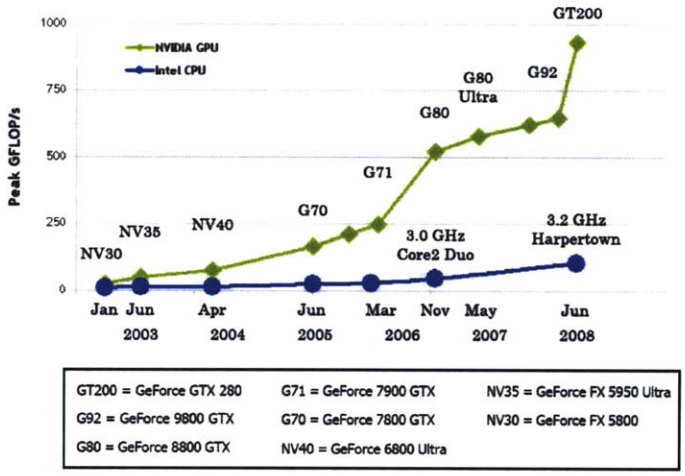


Figure 2-64: Performance comparison between NVIDIA GPUs and Intel CPUs [119]

by a spatial light modulator, as well as in digital holographic microscopy systems [131].

The HOA presented previously is particularly suitable for implementation on a GPU, due to its multi-point search nature that can be parallelized and processed simultaneously. All the optimization results presented in the previous sections were efficiently computed on a GPU. Table 2.3 outlines the main specifications of the GPU used. The GPU is hosted by a personal computer with the following specifications: Intel Pentium 4 CPU, 3.8 GHz clock rate, 3GB of memory, and Windows XP operating system.

Table 2.3: GPU Specifications.

Model	GeForce GTX 285	**Clock Rate**	1.62 GHz
Global Memory	1 GB	**Memory Bandwidth**	159 GB/sec
Multiprocessors	30	**Cores**	240

Knowledge of the GPU's architecture is necessary to understand how the HOA was implemented in parallel. A GPU is basically composed of global memory and a variable number of multiprocessors. Each multiprocessor includes eight scalar processor cores, two special function units, 8192 registers, a multithreaded instruction unit and one on-chip

shared memory. The CPU (host) controls the execution of multiple concurrent threads (independent processes) on the GPU (device) by calling the respective kernels. The kernel is a function called from the host and executed on the device. While the execution of a kernel is in progress, the host can perform other activities. When launching a kernel, threads are arranged in blocks and grids and each block is assigned to a multiprocessor and the instructions written in a kernel are executed by all the threads. Each thread has its own ID and thread cooperation is only possibly within the block – they can be synchronized and can cooperate using shared memory (block's private memory). Within a block, threads are arranged in groups of 32, referred to as warps. Threads in a warp are physically executed in parallel and are synchronized. Multiprocessors execute one warp at time, but if threads in a warp are waiting to access memory, the multiprocessor can stop executing that warp and start to process another warp eliminating memory's latency time. The management of warps is automatic and not visible by the programmer. When optimizing an algorithm, the overhead introduced from transferring memory from host to device and vice versa needs to be minimized. For this reason, in our implementation using Matlab's MEX interface, the desired mask and some preliminary calculations are done in Matlab and then transferred to the GPU at the beginning of the function. The computed data resides on the GPU for the most part and is only transferred back to the host at the end of the optimization. Most of the operations performed by the GPU are conducted using a floating-point precision (32-bit), and some operations use a double-point precision (64-bit). NVIDIA GPUs offer scalable line interconnects (SLI) that can be used to connect several graphics cards in one personal computer and to further maximize the computational power.

To evaluate the computational performance of the HOA, a comparison between CPU (Matlab function) and GPU (CUDA function) implementations of a single evaluation of the fitness function (based on the LDPE and LNPEPE masks) is conducted. The comparison is done for varying space-bandwidth products in the range of 500×500 (0.25 million) to 4000×400 (16 million) pixels. Figure 2-65-a shows the resulting computa-

tional times for a single evaluation of the fitness function based on the LDPE mask. The corresponding speedup factors are shown in Figure 2-65-b. A speedup factor of over 120 times is measured. However, a much faster performance of the GPU implementation has been observed for multiple computations of the fitness function as the memory transfer overhead is reduced. Speedup factors of over 200 times have been achieved when optimizing large CGHs. Figure 2-66-a shows the computational times for a single evaluation of the fitness function based on the LNPEPE mask. The corresponding speedup factors are shown in Figure 2-66-b. Again, higher speedup factors (over 200 times) are measured for consecutive evaluations of the fitness function in the GAs block.

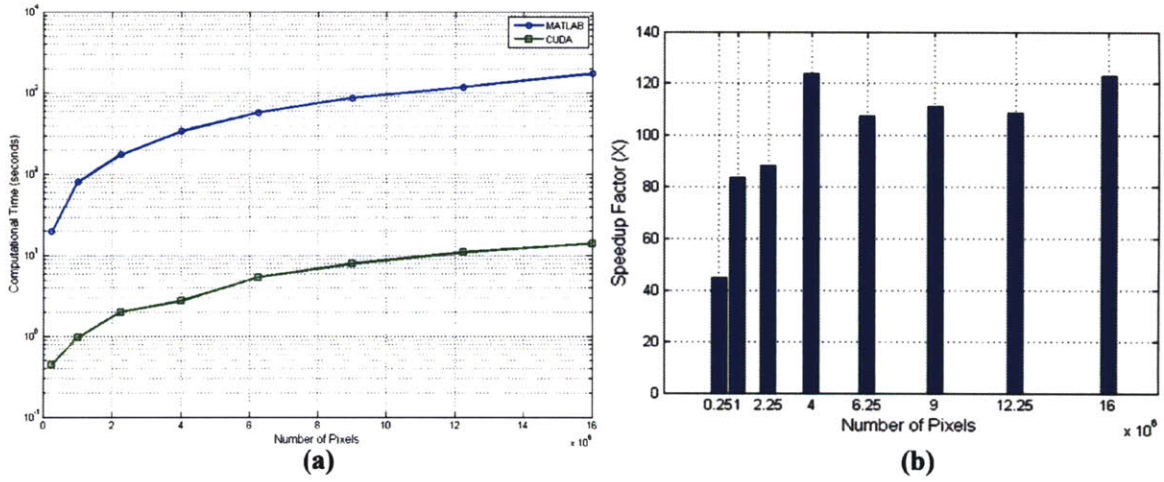


Figure 2-65: (a) Computational times: LDPE encoding strategy; (b) Relative speedup factors.

The computational time of the GAs block for optimizing the CGH based on the LDPE mask of Figure 2-38 is: $t_{GPU} = 4.47$ hours. The corresponding mean computational time per individual is: $t_{ind} = 1.611$ seconds (instead of $t_{ind} = 2$ seconds as predicted from Figure 2-65-a due to the memory transfer overhead). The estimated computational time for the same function on the CPU is: $t_{CPU} = 16.48$ days. Similarly, the computational time of the GAs block to generate the CGH based on the LNPEPE mask of Figure 2-46 is: $t_{GPU} = 13.66$ hours. The corresponding mean computational time per individual is:

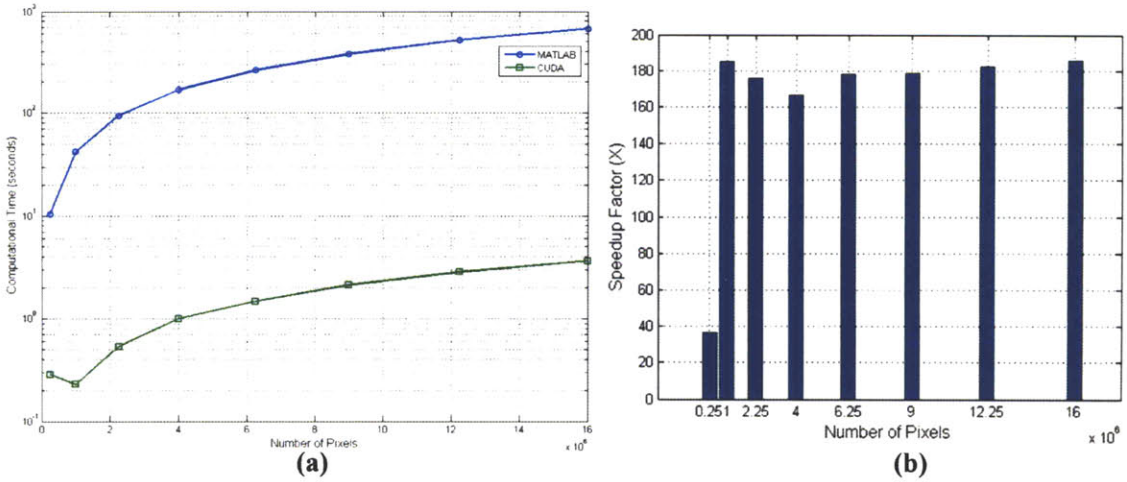


Figure 2-66: (a) Computational times: LNPEPE encoding strategy; (b) Relative speedup factors.

$t_{ind} = 0.8196$ seconds. The estimated computational time for the same function on the CPU is: $t_{CPU} = 99.96$ days. From these two examples it is clear that the implementation of the HOA on the GPU significantly improves the computational performance.

Figure 82 shows the computational time of the GAs block as a function of population size for the results corresponding to Figure 2-59 (for 20 generations of the optimization of the LNPEPE mask). As expected, the computational time increases linearly with the number of individuals in the population. It was also found that the computational time is relatively insensitive to varying crossover fractions (Figure 2-58) on the GAs block.

Optimization of Off-Axis CGHs

In this section, we present an example of the optimization of an off-axis binary phase CGH based on the geometry of Figure 2-12. The hologram is optimized using the MER block (Figure 2-34) with the additional modulation and demodulation steps as described previously. The optimization parameters are indicated in Table 2.4. During the optimization procedure, the hologram is modulated according to Figure 2-13 and demodulated in a similar form as Figure 2-15 to remove the undesirable diffraction orders. This demod-

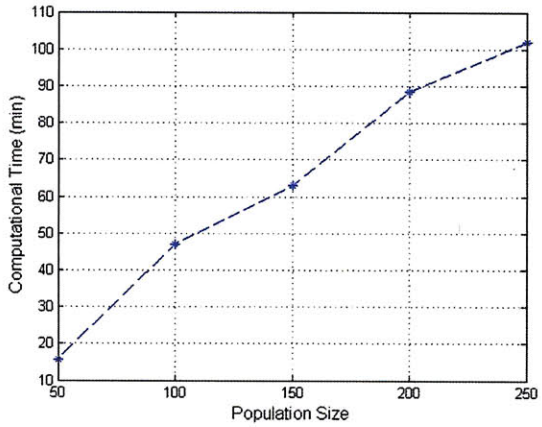


Figure 2-67: Computational time of GAs block for different population sizes.

ulation procedure can be implemented optically by probing the CGH with an off-axis wave and using a 4f system with a low-pass filter to remove the undesirable signal.

Table 2.4: Optimization Parameters: Off-Axis CGH.

Wavelength (λ)	532nm	**Object Window** (O_{size})	360μm
Working Distance (d)	250μm	**Off-Axis Angle**	45°
Pixel Size (δ_{pix})	200nm	**Initial Search Point**	Diffracted Field
Hologram Size (H_{size})	360μm	**Number of Iterations**	200

Figure 2-68-a shows the final binary phase CGH. The resulting off-axis CGH is modulated by a high-frequency carrier signal with frequency proportional to the incidence angle of the probing wave. This modulation restricts the maximum pixel size to that of equation 2.10. For the parameters of Table 2.4, the maximum pixel size is: $\delta_{pix} = 376.18$nm. In addition, the bandwidth of the encoded signal is restricted to the limits of equation 2.11. For this example, the maximum signal's bandwidth that can be encoded is: $B_x = 443.049\text{mm}^{-1}$. The corresponding reconstructed amplitude distribution is shown in Figure 2-68-b. Off-axis CGHs reconstruct high-quality patterns as the undesirable diffraction orders are filtered out. However, they are more susceptible to manufacture errors due to the small pixel size and require additional components (aperture stop and

4f system) for their implementation.

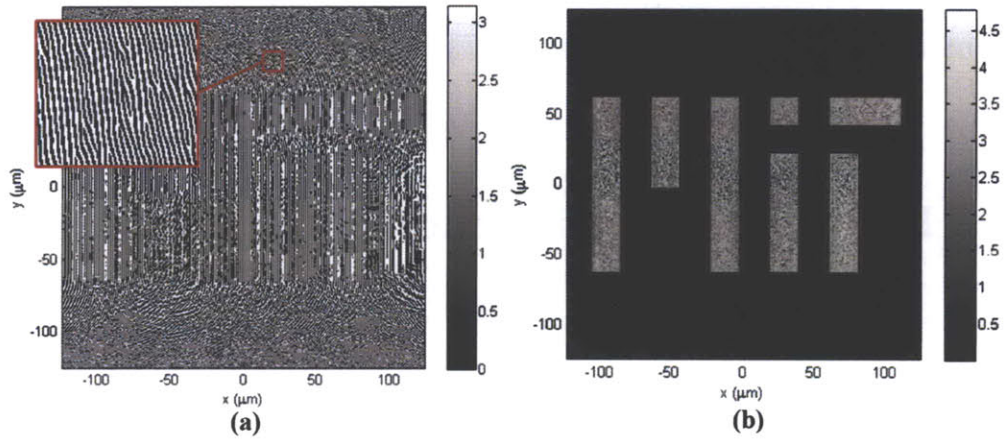


Figure 2-68: (a) Optimized off-axis CGH; (b) Reconstructed amplitude distribution.

Optimization of TIR CGHs

The TIR geometry of Figure 2-14 has the advantage, respect to the off-axis geometry, that no additional optical components (aperture stop or 4f system) are required to filter out the undesirable diffraction orders of the modulated encoded signal. Instead, the undesirable orders suffer TIR and the dielectric-air interface which can be explained by the demodulation model of Figure 2-15. In this section we present an example of a TIR CGH optimized using the MER block with the parameters indicated in Table 2.5. Figure 2-69-a shows the final optimized TIR binary phase CGH. Similar to the off-axis case, the CGH is modulated by a high-frequency carrier signal that limits the allowable pixel size (equation 2.10 but replacing λ with λ_{eff}). For this example, the maximum pixel size is: $\delta_{pix} = 165$nm. The logarithm of the magnitude of the CGH's spectrum is shown in Figure 2-69-c. The CGH is reconstructed by an off-axis wave that shifts the modulated signal back to the center of the frequency plane. The reconstructed amplitude distribution at the photoresist plane is shown in Figure 2-69-b.

Table 2.5: Optimization Parameters: TIR CGH.

Wavelength (λ)	350nm	**Object Window** (O_{size})	340μm
Working Distance (d)	200μm	**Off-Axis Angle**	45°
Pixel Size (δ_{pix})	100nm	**Initial Search Point**	SORH
Hologram Size (H_{size})	340μm	**Number of Iterations**	100
SBP (After Padding)	4000 × 4000	**Prism Refractive Index**	1.5

2.5 Experimental Fabrication and Characterization of CGHs

2.5.1 Fabrication Process

The CGHs considered in this thesis are fabricated using electron-beam (e-beam) lithography. E-beam lithography is a direct writing method in which a beam of electrons is scanned to form the desired pattern on a substrate coated with resist. After exposure the substrate commonly undergoes developing and etching processes. E-beam lithography is a high-resolution method allowing the fabrication of smaller features than alternative techniques such as direct laser writing, which is typically limited by diffraction. Fabrication of small features down to 10nm has been experimentally demonstrated [68]. For computer holography, e-beam lithography offers the possibility of fabricating CGHs with large space-bandwidth products, submicron pixel sizes and good positional accuracy within the field-of-view of the e-beam writing system [24], [25].

In this thesis, the process chosen to fabricate the optimized CGHs is restricted to single e-beam exposures to produce binary phase holograms. The presented fabrication process could potentially be extended to produce multi-level phase CGHs requiring multiple e-beam exposures or subwavelength patterns as suggested by effective medium theory. The implementation and optimization of the fabrication process are beyond the scope of this thesis; however, an understanding of the steps involved is necessary for characterizing the CGHs' performance, as well as for performing a sensitivity analysis that can be used to predict and correct potential manufacture errors. The fabrication process adopted in this

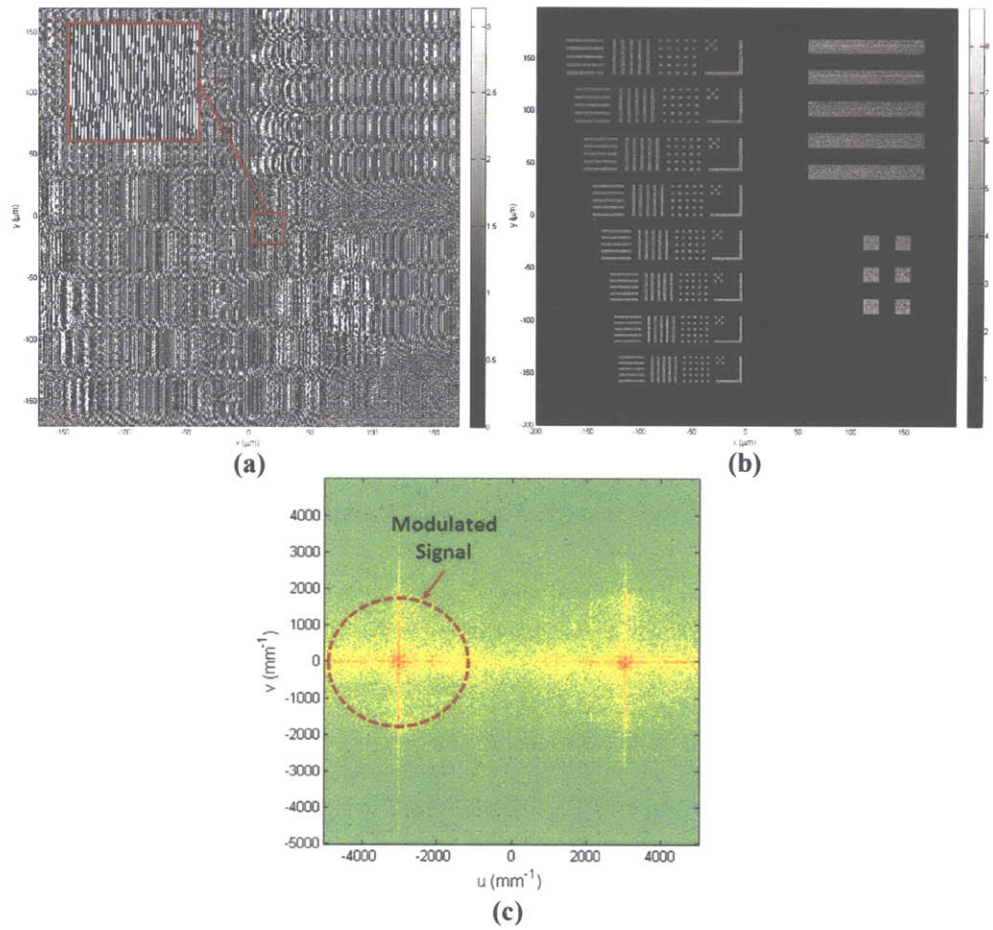


Figure 2-69: (a) Optimize TIR CGH; (b) Reconstructed amplitude; (c) CGH spectrum.

thesis is shown in Figure 2-70. The first step consists of spinning a layer of Hydrogen Silsesquioxane (HSQ Dow Corning Fox-series) onto a fused silica substrate. HSQ is a negative resist sensitive to e-beam and X-ray radiations. The thickness of the HSQ layer is proportional to the desired phase delayed of the CGH,

$$t = \frac{\phi\lambda}{2\pi\left(n_2 - n_1\right)}, \tag{2.74}$$

where ϕ is the required phase delay (corresponding to R_{phase} in equation 2.40; typically set to $\phi = \pi$i), n_2 is the refractive index of the HSQ layer at the operating wavelength,

and n_1 is the refractive index of the surrounding medium (typically $n_1 = 1$). The spin speed is determined from the spin (spin speed-thickness) curve of the HSQ with the particular dilution. After spinning, the HSQ is baked at 150 degrees for 2 minutes, then at 220 degrees for another 2 minutes using hotplates. Next, a thin (5-6nm) metal layer is deposited on the HSQ by e-beam evaporation. This metal layer is required to avoid charging during the e-beam exposure as all the layers underneath are non-conductive. Any type of conductive material that does not react with HSQ and can be easily deposited and removed without significantly damaging the layers underneath is appropriate for this layer. The next step is the e-beam exposure process. The phase distribution of the CGH optimized by the HOA is converted into a file compatible with the e-beam writing machine which indicates the locations on the substrate that need to be exposed by the electron beam. The CGHs presented in the following section were fabricated using the Raith 150 e-beam exposure system located at the NanoStructures Laboratory (NSL) facilities in MIT. The time for e-beam write varies according to the hologram's size, pixel size and dose, taking approximately 30 minutes to 1.5 hours to write an entire structure. After the e-beam writing step, the metal layer is removed by wet chemical etching, and the HSQ is developed for 2 hours with Shipley Microdeposit MF CD26, a TMAH-based developer.

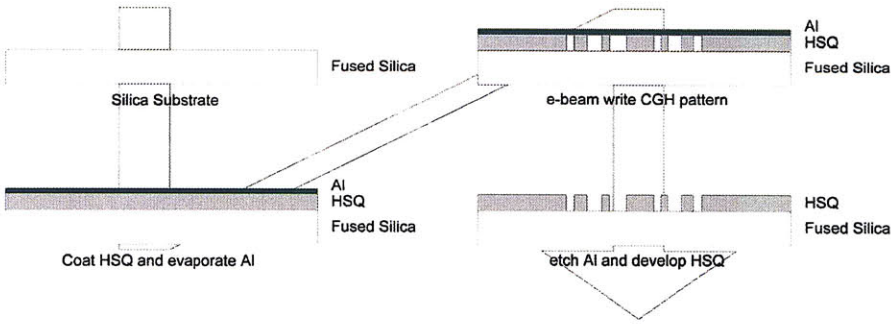


Figure 2-70: CGH fabrication process.

At the beginning of the fabrication run, a small section of the CGH is written several times using different uniform e-beam doses to determine the correct exposure energy. The

written structure is then evaluated using a scanning electron microscope (SEM) and the dose that results in the highest quality pattern is chosen. A more quantitative method to characterize the fabricated samples based on a 2D error map will be introduced later. In general, e-beam writing is known to suffer from proximity effects that distort the resulting pattern in the resist. In order to minimize this effect, the dose of the beam must be properly chosen according to the beam energy, pattern and the stack of layers that the e-beam writes on. Due to the complex structure of the CGH's phase distribution, an additional local dose correction is employed. In the local dose correction, the phase pattern is divided into several regions and the optimum dose (above or below the uniform dose obtained from the dose matrix) is estimated using the 2D error map. An alternative method for correcting these errors is by the implementation of specialized proximity effect correction (PEC) software [132].

The proposed fabrication process of Figure 2-70 is simple and does not require an additional etching step as HSQ, once exposed and developed, acquires mechanical and optical properties similar to glass [133]. In addition, HSQ is a high-resolution material with a long lifespan. HSQ can be spin coated with thickness accuracies better than 1%, avoiding phase errors in the fabricated CGH.

2.5.2 Examples of Fabricated In-line CGHs

The first fabrication example is of an in-line binary phase CGH optimized by the MER block with the initial search point given by the diffracted field encoding strategy. The optimization parameters are given in Table 2.6. Figure 2-71-a shows the phase distribution of the optimized CGH and Figure 2-71-b shows the converge plot of the optimization algorithm for the error metrics of equations 2.55, 2.56 and 2.57. The corresponding simulated reconstructed amplitude at the photoresist plane is shown in Figure 2-72-a. Figure 2-72-b shows a SEM image of a fabricated CGH. The fabricated CGH has not been fully corrected for errors due to proximity effects, as shown in the close up section of Figure 2-73 in which the red circles indicate examples of good pattern replication and the yellow

circles show examples of missing structures.

Table 2.6: Optimization Parameters: Fabricated In-line CGH.

Wavelength (λ)	532nm	**Object Window** (O_{size})	350μm
Working Distance (d)	250μm	**SBP (After Padding)**	2500 $\times$ 2500
Pixel Size (δ_{pix})	200nm	**Initial Search Point**	Diffracted Field
Hologram Size (H_{size})	358μm$\times$353μm	**Number of Iterations**	200

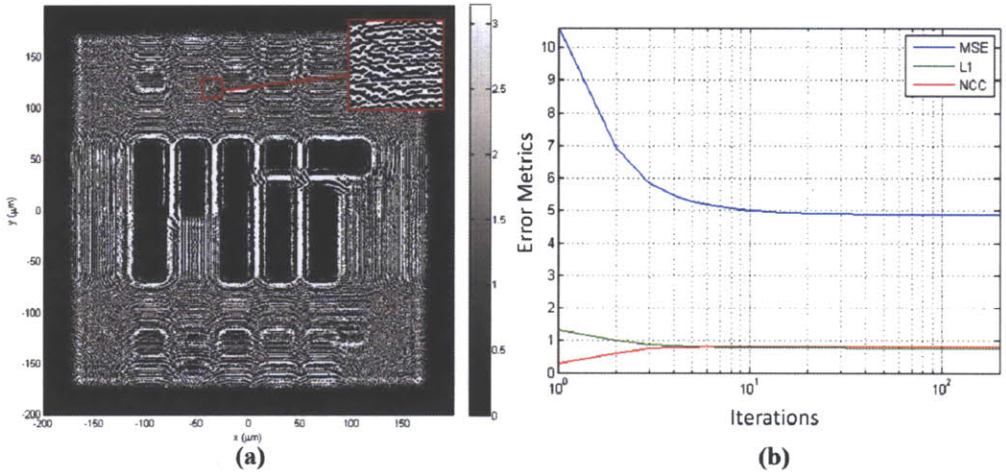
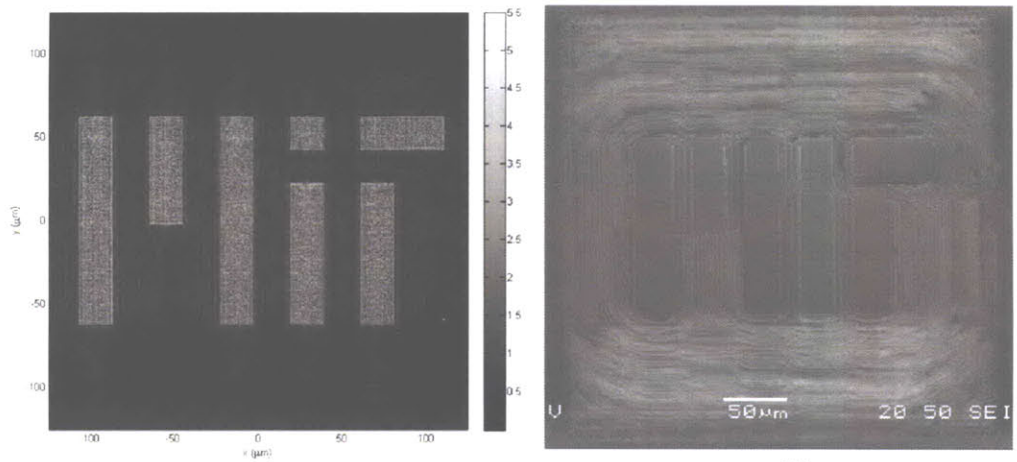


Figure 2-71: (a) Optimized in-line CGH; (b) Convergence plot.

The second example consists of the fabrication of an in-line binary phase CGH optimized based on the LDPE mask. Figure 2-74-a shows the computed CGH and Figure 2-74-b shows the corresponding convergence plot. The optimization parameters are the same as those of Table 2.6. The simulated reconstructed amplitude distribution is shown in Figure 2-74-c. As discussed previously, the LDPE mask helps to significantly improve the encoding strategy, producing high-quality reconstructions with lower errors than that of the previous example. Figure 2-75 shows an SEM image of the fabricated CGH. Similar to the previous case, the fabricated sample has not been fully corrected for proximity effects. In the next section, a quantitative evaluation of this sample will be performed.



(a) (b)

Figure 2-72: (a) Simulated reconstructed amplitude; (b) SEM of fabricated CGH.

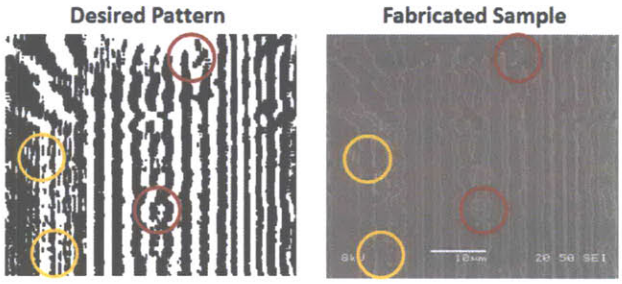


Figure 2-73: Comparison between designed and fabricated CGHs.

2.5.3 Experimental Characterization of Fabricated CGHs

The fabricated CGHs contain geometrical distortions due to fabrication errors such as non-ideal exposure doses, proximity effects, stitching errors and beam positioning errors. Three methods are implemented for a quantitative characterization of the fabricated samples: evaluation algorithm (2D error map), optical characterization and photoresist exposure test.

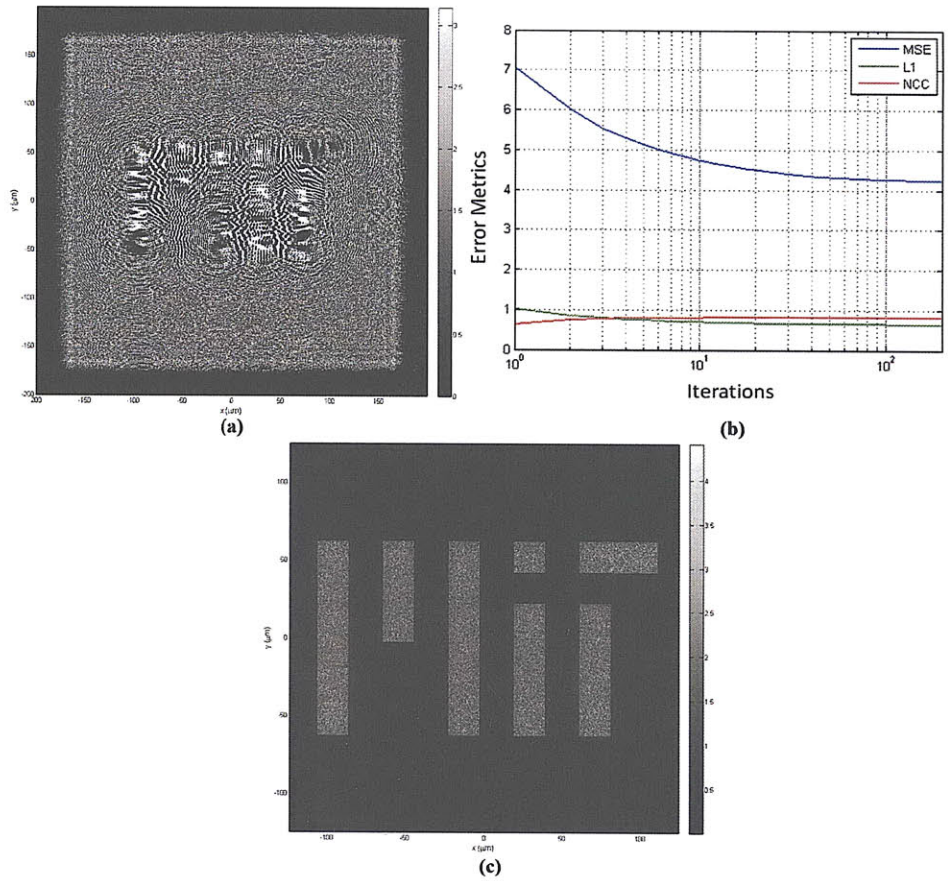


Figure 2-74: (a) Optimized in-line CGH; (b) Convergence plot; (c) Simulated reconstructed amplitude.

Evaluation Algorithm: 2D Error Map

A 2D error map of the difference between the desired and fabricated CGH patterns is computed and used to quantitatively evaluate the fabricated sample. The block diagram of the evaluation algorithm used to calculate the 2D error map is shown in Figure 2-76. The first step is to acquire N high-resolution images of the fabricated CGH using a confocal or scanning electron microscopes. Due to the microscope's limited field of view when operating at high-resolution, multiple images of different sections of the sample are captured. An auto-stitching algorithm is used to produce a complete, high-resolution image of the entire fabricated CGH. The auto-stitching algorithm begins by performing a

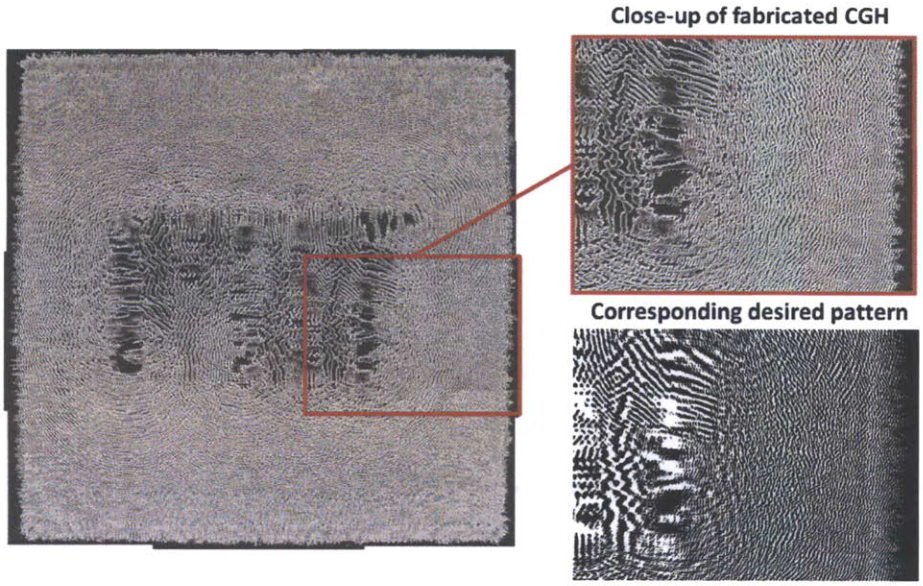


Figure 2-75: SEM of fabricated CGH.

noise reduction and 2D interpolation (for sub-pixel stitching accuracy) of each captured image. In order to stitch an image to its neighbors, horizontal and vertical template patches are extracted and a spectral cross-correlation algorithm is performed. For the nth image, a 2D cross-correlation map is computed,

$$xC^{(n)}(x,y) = \mathcal{F}^{-1}\left\{\left(\mathcal{F}\left\{P_{neighbor}(x,y)\right\}\right)\left(\mathcal{F}\left\{P^{(n)}_{template}(x,y)\right\}\right)^{*}\right\}, \tag{2.75}$$

where $P^{(n)}_{template}$ is the extracted horizontal or vertical template patch (zero padded to match the size of $P_{neighbor}$),$P_{neighbor}$ is the corresponding neighboring image, and $*$ represents the complex conjugate operation. The coordinates of the correlation peak ($\max[xC^{(n)}]$) are computed and used for placing the image at the corresponding location on a high-resolution grid. The final stitched image is then binarized using Otsu's method, which chooses the threshold to minimize the intraclass variance of the black and white pixels [134]. Figure 2-77-a shows an example of the resulting high-resolution binary image generated from 12 SEM images after the auto-stitching and binarization processes. This

figure corresponds to the fabricated sample of Figure 2-75. The high-resolution binary image is then rotated and scaled to match the desired CGH pattern as closely as possible. The optimum rotation and scale factors are found by genetic algorithms by minimizing the fitness function,

$$F = \mathbf{W}\left[\mathcal{F}^{-1}\left\{\left(\mathcal{F}\left\{P_{des}\right\}\right)\left(\mathcal{F}\left\{\mathbf{R}\left\{\mathbf{S}\left\{P_{fab}\right\}\right\}\right\}\right)^{*}\right\}\right]_{\max}^{-1}, \tag{2.76}$$

where P_{des} and P_{fab} are the desired and fabricated CGH binary patterns, $\mathbf{R}$ and $\mathbf{S}$ are the rotation and scale operators, and $\mathbf{W}$ is a weighting factor. The next step in the evaluation algorithm is to compensate for lateral shift again using a spectral cross-correlation algorithm. The final step consists of computing the 2D error map, which is the result of the difference between the desired (P_{des}) and fabricated (P_{fab} – after lateral shift compensation) binary patterns. Figure 2-77-b shows the 2D error map computed for the fabricated hologram of Figure 2-75. The blue and red patches (error values of -1 and 1) correspond to sections on the fabricated sample that need to have a local e-beam dose correction. In addition to the 2D error map, a global MSE metric is computed and used for comparing different fabricated samples. The 2D error map is used to assist in the local correction of e-beam doses for future fabrication runs.

Optical Characterization

An optical characterization test is implemented to evaluate the quality of the reconstructed intensity distributions from the fabricated in-line CGHs, presented in the previous section. The CGH optical characterization setup is shown in Figure 2-78-a. A green laser ($\lambda = 532$nm) is spatially filtered and collimated to produce a plane wave that probes the fabricated in-line CGH. The reconstructed intensity distribution at the photoresist plane is imaged by a microscope objective ($100\times$, $NA = 0.75$) onto a 16 megapixel CCD camera (pixel size = 9μm$\times 9\mu$m). The microscope objective and CCD are placed on top of a motorized linear stage with 1μm axial resolution designed to axially scan the recon-

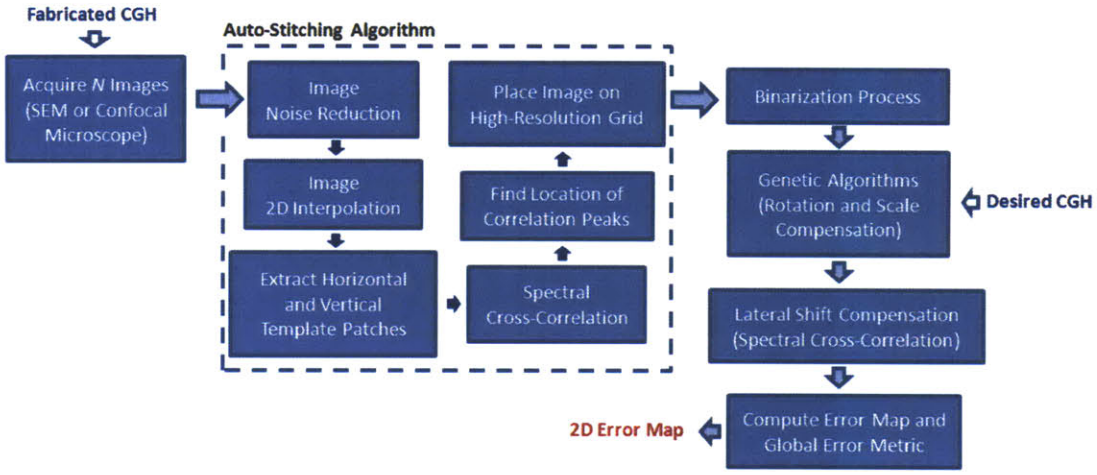


Figure 2-76: Block diagram of the evaluation algorithm.

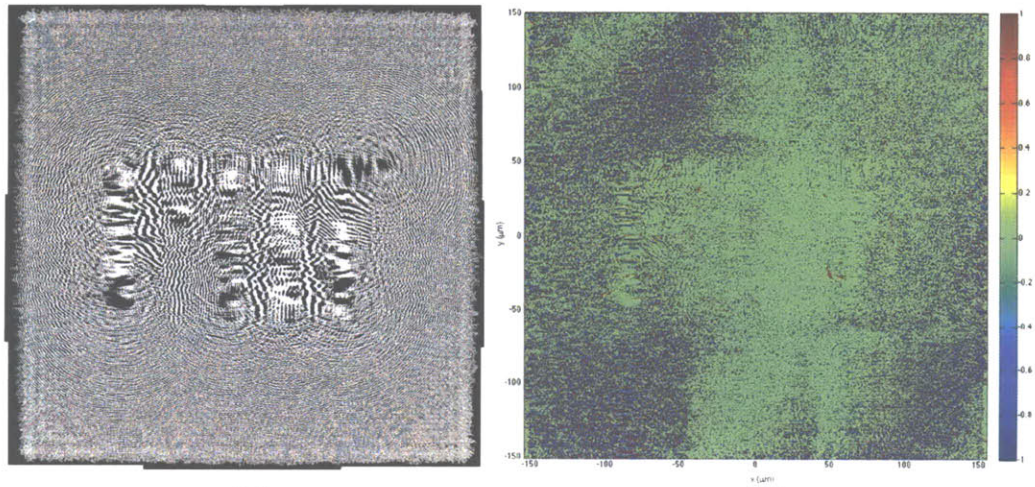


Figure 2-77: (a) Example of high-resolution image produced by the auto-stitching and binarization processes; (b) 2D error map.

struction space to find the location of the focal plane (photoresist plane). A LabView interface was developed to automatically perform the characterization of the fabricated CGH samples. A screenshot of this interface is shown in Figure 2-78-b. The CGH characterization interface allows controlling the camera and motorized stage parameters, as well as performing automatic scanning cycles of the reconstruction space (linear motion plus image acquisition). In addition, the captured images are evaluated comparing them with a target pattern and the correct focus (intensity distribution at photoresist plane) is automatically extracted.

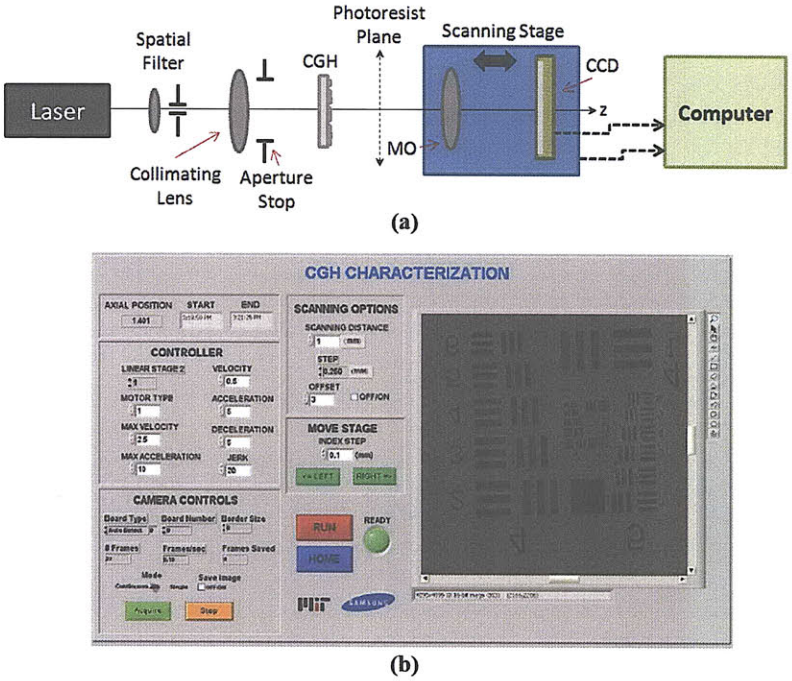


Figure 2-78: (a) Optical characterization setup; (b) Measuring station GUI.

Figure 2-79 shows the experimental reconstructed intensity distribution from the fabricated hologram of Figure 2-72-b. The reconstructed pattern shows good contrast and well-defined edges, which indicates that a significant amount of high-frequency components survived the fabrication process and produced a high-resolution pattern. An additional local dose correction is necessary to improve the diffraction efficiency and

uniformity of the reconstructed pattern.

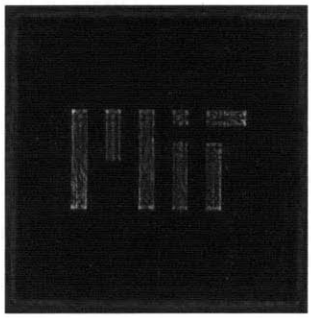


Figure 2-79: Measured reconstructed intensity distribution.

Figure 2-80 show the experimental reconstructed intensity distribution from the fabricated hologram of Figure 2-75. Despite the fabrication errors indicated on the 2D error map of Figure 2-77-b, this CGH still reconstructs the MIT logo at the photoresist plane. Also, this experiment shows how the resulting CGH is robust against potential contaminants as the LDPE encoding strategy spreads the encoded signal over the entire hologram window. The grainy speckle-line noise present in the reconstruction is the result from lost spatial frequencies after the fabrication process. An additional optimization of the fabrication process for this sample is required.

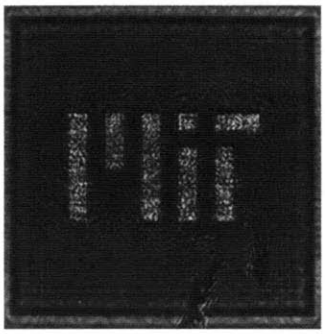


Figure 2-80: Measured reconstructed intensity distribution.

Reconstructing with Partially Spatially Incoherent Illumination

The speckle-like nose present in the reconstructed intensity distribution of Figure 2-80 corrupts the photoresist exposure process and hence needs to be eliminated. Two potential techniques to eliminate this noise are: improve fabrication process (reduce error in 2D error map following the optimized CGH as closely as possible), and reconstruct the hologram with partially spatially incoherent illumination. The effects of reconstructing the CGH with partially spatially incoherent illumination are simulated by probing the hologram with several plane waves oriented at slightly different off-axis angles (equation 2.9). The angles are drawn from a random uniform distribution with a range given by the desired degree of coherence. The reconstructed field from each plane wave is calculated using the Fresnel operator and the fields at the photoresist plane are added incoherently: $I_{tot} = I_1 + I_2 + I_3 + \cdots I_N$, where I_n is the reconstructed intensity distribution at the photoresist plane corresponding to the nth probing wave.

The experimental implementation is conducted with the optical setup of Figure 2-81. The illumination source is a tungsten halogen white lamp with variable output power. The white light source delivers the light by means of an optical fiber bundle. A narrow band-pass filter (center wavelength: $\lambda = 532$nm) is used to produce a quasi-monochromatic spatially incoherent illumination. A variable aperture is used to manually control the effective degree of coherence. Larger apertures results in a more incoherent illumination. The rest of the optical components are similar to those of the coherent illumination case of Figure 2-78-a.Figure 2-82 shows the captured reconstructed images using partially incoherent illumination for the hologram of Figure 2-75. The size of the variable aperture is progressively increased from images 1 to 6 (from high to low effective degree of coherence). This technique reduces the effect of the speckle-like noise at the expense of spatial resolution.

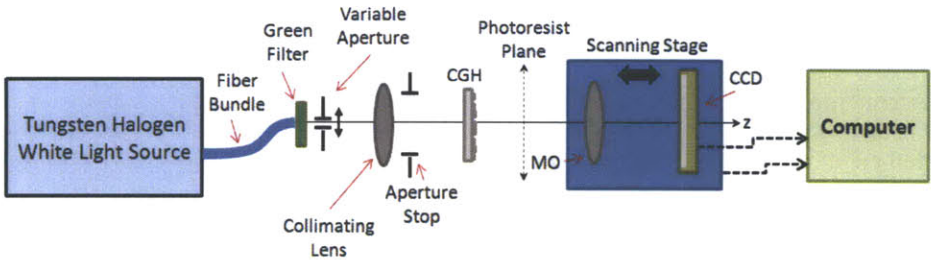


Figure 2-81: Optical characterization setup for partially coherent illumination.

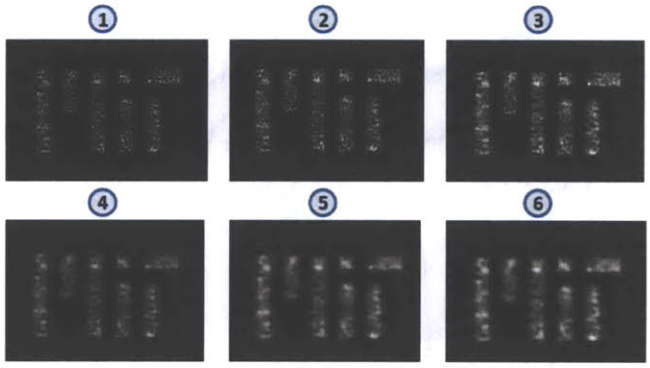


Figure 2-82: Measured intensity distribution for different degrees of coherence.

Photoresist Exposure Test

In the photoresist exposure test, the imaging section of the optical characterization setup (microscope objective and CCD) is replaced with a substrate coated with photoresist. Several exposures are conducted with different exposure energies and axial positions to evaluate the optimum printing conditions. The exposed substrate is then developed and the fabricated sample is inspected using a confocal or scanning electron microscope.

The test results presented in this section were performed at the Mechatronics Center facilities of Samsung Electronics in Suwon, South Korea. The CGH used in the experiments was designed to reconstruct a resolution target consisting of a 5×5 grating array with periods ranging from 700nm to 1400nm. The CGH was optimized using the MER

block with the optimization parameters indicated in Table 2.7. Eight CGH samples with different e-beam doses were fabricated and analyzed using the evaluation algorithm. The global error metric was used to choose the best sample. Figure 2-83 shows the binary phase distribution of the desired (optimized) and chosen fabricated CGHs as well as the corresponding 2D error map. The fabrication errors indicated in the 2D error map are primarily from e-beam proximity effects. An additional local dose correction is required. Figure 108 shows the desired reconstructed intensity and the expected reconstructed intensity from the fabricated sample (simulated using the phase distribution of Figure 2-83-b). The CGH fabrication errors result in non-uniformities on the reconstructed intensity distribution. Figure 109 shows a confocal microscope image of the printed pattern after the photoresist exposure test. The non-uniformities present on the printed pattern very closely match those predicted in the simulations (Figure 2-83-b).

Table 2.7: Optimization Parameters: In-line CGH for Photoresist Exposure Test.

Wavelength (λ)	364nm	**Object Window** (O_{size})	50μm
Working Distance (d)	50μm	**Hologram Size** (H_{size})	62.5μm$\times$120μm
Pixel Size (δ_{pix})	100nm	**Number of Iterations**	100

2.6 Sensitivity Analysis

A sensitivity analysis is performed to estimate and assist in the correction of potential manufacture errors that may occur during the e-beam writing and development processes. Five different types of errors are studied: over and under dose, proximity effect, phase, stitching and positional errors. Simulation algorithms are presented to model these errors and quantitatively evaluate their influence on the reconstructed patterns.

For uniform featured size patterns, such as gratings, a uniform dose can yield accu-

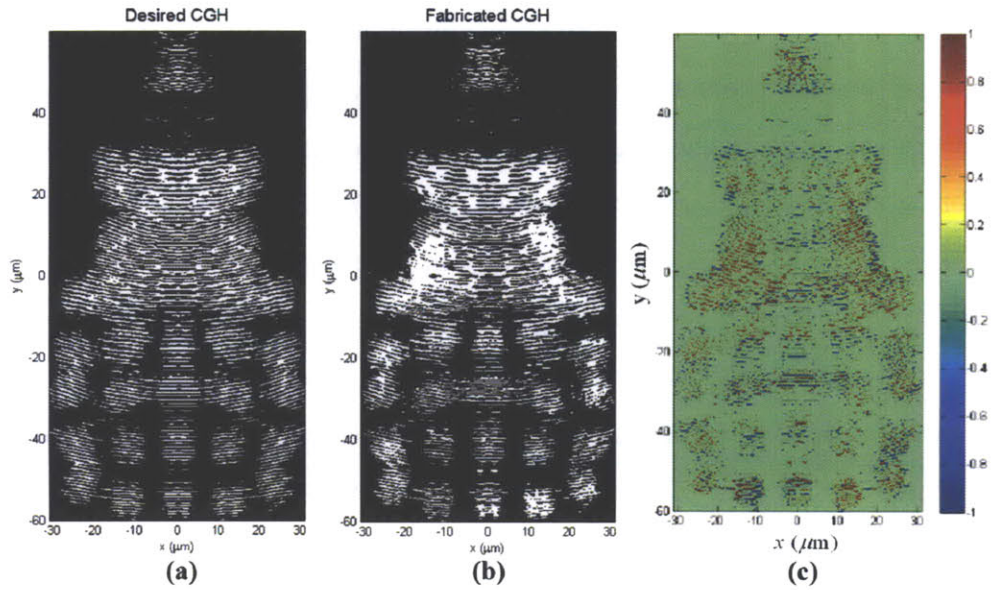


Figure 2-83: (a) Optimized CGH; (b) Fabricated CGH; (c) 2D error map.

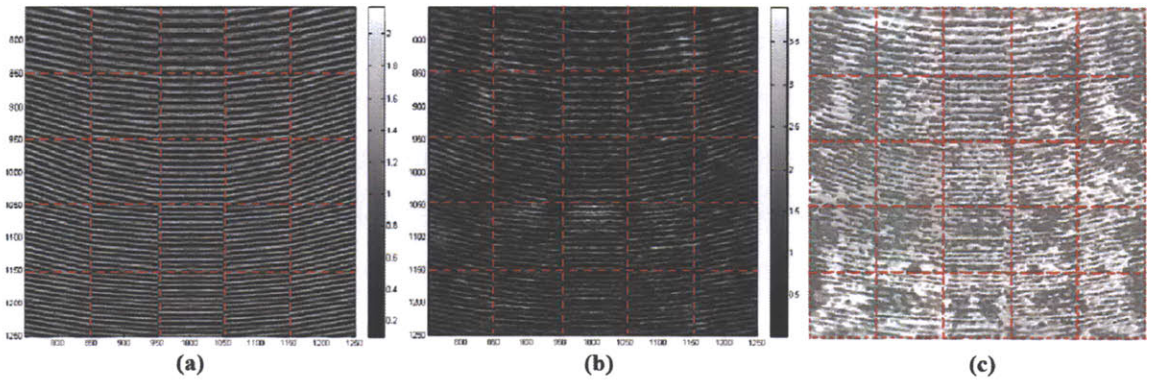


Figure 2-84: (a) Simulated reconstruction from optimized CGH; (b) Simulated reconstruction from fabricated CGH; (c) Confocal microscope image of reconstructed pattern.

rate results. Finding the correct dose can be done experimentally by writing a matrix of patterns with different dose energies and performing a post evaluation of the processed structures. Since HSQ is a negative resist, an underexposed pattern washes out most of the features leaving the substrate behind after development. For overexposed patterns, the features are also washed out and a substantial amount of HSQ is left behind. For the fabrication of CGHs, over and under doses result in the loss of spatial frequencies (or fringes in the CGH pattern) that contain information of the encoded signal. Small features (or high frequency components) are washed out first, decreasing the hologram's diffraction efficiency and producing non-uniform reconstructions. To model the over and under dose fabrication errors, an algorithm based on dilation and erosion structures is implemented. In the case of overexposure, a dilation pattern of a specified shape and size is defined and applied to the binary phase pattern of the desired (optimized) CGH. Disk structures are typically used and their diameter is given by the desired level of dilation. Dilating the CGH's phase pattern causes the fringes to become thicker and progressively disappear by merging together. A similar effect happens for the under dose case in which the erosion process causes the fringes to become thinner and progressively disappear. For each dilation or erosion state, the hologram is reconstructed and the MSE before photoresist exposure is computed. Figure 2-85-a shows an example of the CGH's phase distribution and reconstructed amplitude for three dilation states, designed to model a potential overexposure error of the hologram of Figure 2-71-a. The corresponding 1D cross-section of the reconstructed intensity distributions is shown in Figure 2-85-b. As can be seen, the simulated error introduces undesirable diffraction orders such as the virtual image and direct component. CGHs designed with the diffracted field encoding strategy do not spread the signal's information uniformly over the entire hologram window. The encoded signal is only spread over a very small path with the high-frequency components located near the edges of the hologram and larger features towards the center. After the dilation or erosion processes, the fringes near the edges disappear first, significantly affecting the reconstructed pattern. In contrast, the information encoded in

CGHs optimized using the LDPE or LNPEPE masks is distributed uniformly over the entire hologram window introducing redundancy and making them less sensitive to over and under dose errors.

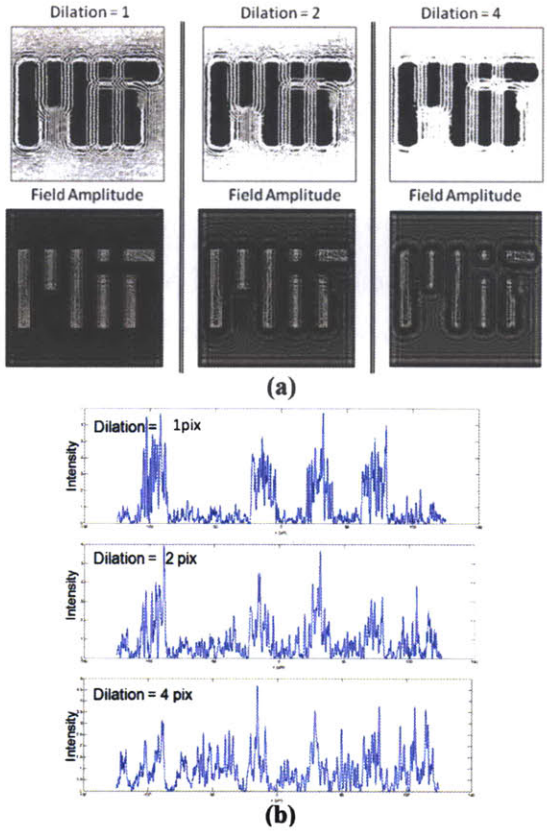


Figure 2-85: (a) Example of dilation analysis; (b) Intensity cross-sections.

Proximity effects occur when high-energy electrons interact via electromagnetic fields with resident electrons continuously losing energy through the excitation of secondary electrons. These electrons undergo multiple small angle scatterings producing a gradual spread of the well-defined incident beam. In addition, some of the incident electrons undergo large scattering angles, producing backscattering electrons that emerge at points that are remote to the original entry point. To a first degree approximation, this error is also modeled using the algorithm based on the dilation and erosion structures.

The phase error consists of variations in the height of the pattern which translates to errors in the phase delay produced by the diffraction element. Two types of binary phase errors are modeled: uniform and non-uniform. In the uniform case, a constant phase shift is added to the range R of equation 2.40 and the hologram is reconstructed and evaluated. For the non-uniform case, the pattern's phase delay is perturbed by an arbitrary-shaped distribution. For the fabrication process considered in this thesis, the uniform case is more relevant as the binary phase error is directly proportional to errors in the thickness of the HSQ after spin coating. The spin coating process is very accurate and can be controlled within a few nanometers by optimizing the spin speed and viscosity of the HSQ. For this reason, a small degree of phase error is expected on the fabricated samples. Figure 2-86-a shows the reconstructed amplitude distribution for three different degrees of uniform phase error for the hologram of Figure 2-71-a. The corresponding 1D cross-section of the reconstructed intensity is shown in Figure 2-86-b. This analysis reveals that even relatively large uniform phase errors do not significantly degrade the reconstructed signal.

E-beam stitching errors are primarily the result of calibration errors which produce overlapping or discontinuous fields. Stitching errors can be mitigated by the implementation of techniques such as spatial phase-locked e-beam lithography. To model this error, the CGH's phase distribution is divided into multiple e-beam fields and each block is laterally translated according to the desired degree of calibration error. An example of the e-beam stitching error analysis is shown in Figure 2-87. The CGH pattern is divided into four blocks of $200\mu\text{m}\times200\mu\text{m}$ which correspond to the typical e-beam field size. A positive or negative offset is applied to simulate cases when the fields overlap or become discontinuous. For each state, the hologram is reconstructed and the MSE before photoresist exposure is computed.

E-beam positional errors are predominantly caused by stray fields that deflect the incident beam, causing field distortions. Additional causes include thermal expansion, charging and laser interferometer quantization errors. Positional errors are quantified by

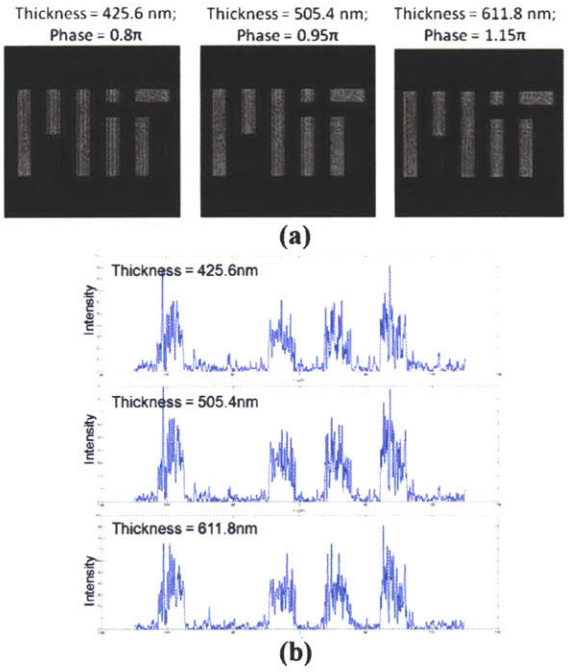


Figure 2-86: (a) Example of phase error analysis; (b) Intensity cross-sections.

computing the 2D error map as explained above.

By using a combination of the error modeling algorithms discussed above, it is possible to decouple the errors present in a given fabricated sample and help in the correction of future fabrication runs. This can be done by analyzing the reconstructed intensity using the optical characterization procedure and matching the parameters of the sensitivity variables to simulate a hologram that contains similar distortions. Figure 2-88 shows an example of a simulated reconstruction pattern from a hologram perturbed to match the optical reconstruction of Figure 2-79. The simulated pattern very closely matches the experimental result, enabling the estimation of the different degrees of errors present in the fabricated sample.

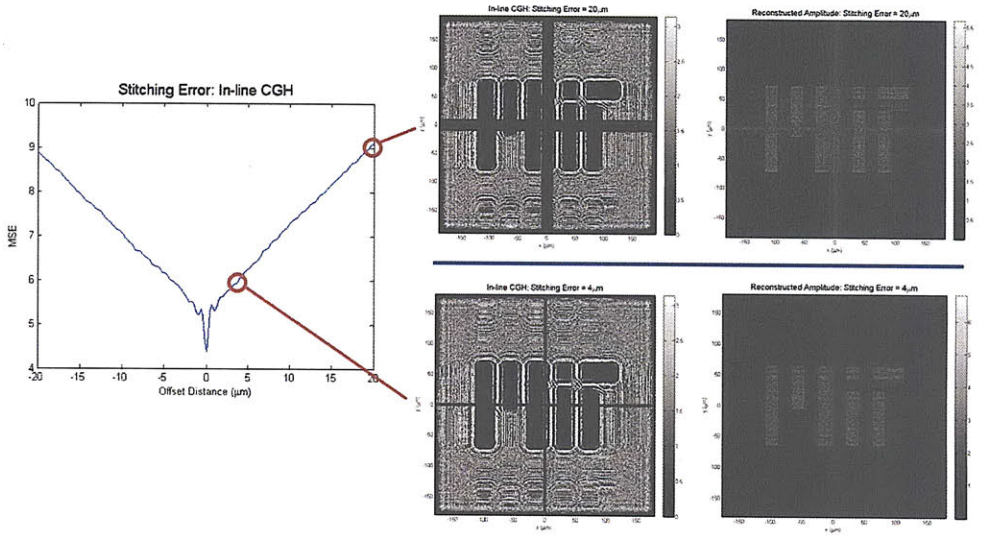


Figure 2-87: Example of stitching error analysis.

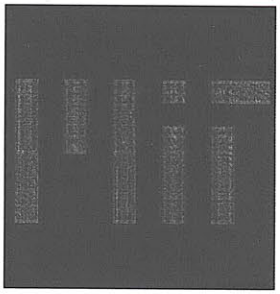


Figure 2-88: Simulated reconstructed pattern from perturbed CGH.

2.7 Optimization of Multispectral CGHs for High-Efficiency Solar Concentration

In recent years, a great effort has been placed on the development of new technologies that exploit solar energy for the production of electricity to satisfy current demands. The sun provides two orders of magnitude more resource availability over wind and more over biomass, geothermal, and waves [135]. Some of the challenges faced by new emerging solar technologies today include the development of efficient and cost effective systems. One of the main factors driving the cost of solar systems is the expensive photovoltaic (PV) cells required. Solar concentrators allow the replacement of large PV devices with smaller cells reducing the overall systems' cost and improving the collection efficiency. These systems are characterized by the concentration ratio,

$$C = \frac{D_{input}}{D_{output}}, \tag{2.77}$$

where D_{input} and D_{output} are the input (entrance pupil) and output (field stop, e.g. PV cell size) diameters of the solar concentrator system [136]. The collection angle is related to the concentration ratio by the Lagrange invariant of the optical system (given by the characteristic étendue of the optical system),

$$\begin{aligned} n_1 D_1 \sin\theta_1 &= n_2 D_2 \sin\theta_2, \\ \rightarrow\ C &= \frac{n_2 \sin\theta_2}{n_1 \sin\theta_1}, \end{aligned} \tag{2.78}$$

where n_1 and n_2 are the refractive indices of the surrounding media and PV cell respectively, D_1 and D_2 are the ray's heights (corresponding to the input and output diameters), and θ_1 and θ_2 are the acceptance half angles of the system. Solar concentrator systems are broadly classified as 2D and 3D [137]. In the 2D case, the system concentrates the incident solar energy into a line. The concentration ratio for the maximum collection

angle ($\theta_2 = 90$degrees and $n_1 = 1$) simplifies to [138],

$$C_{2D} = \frac{n_2}{\sin\theta_1}. \tag{2.79}$$

Small concentration factors lead to large acceptance angles, making 2D systems suitable for passive operation (no tracking required – tolerate daily or seasonal variations of the sun). For the 3D case, the solar energy is concentrated into a tight spot and the concentration ratio for the maximum collection angle is given by,

$$C_{3D} = \left(\frac{n_2}{\sin\theta_1}\right)^2. \tag{2.80}$$

Large concentration ratios can be achieved by 3D systems, provided that an efficient optical system is used for the collection of sunlight. The effective concentration ratio is: $C_{eff} = C\eta_{eff}$, where η_{eff} is the effective efficiency of the optical system. Only relatively narrow collection angles can be achieved in 3D concentrator systems, requiring the implementation sun tracking devices.

Solar concentrator systems based on CGHs are strong candidates, promising to deliver high concentration efficiencies in a cost effective manner. As demonstrated in the previous sections, CGHs can be designed to reconstruct arbitrary patterns at a parallel plane with high diffraction efficiencies (over 80% for the case of multi-level phase CGHs). In solar concentrator systems, CGHs can be used to redirect the incidence light to a narrow spot where an optimized PV cell is placed, replacing conventional systems such as those based on lenses, Fresnel zone plates or parabolic mirrors. CGHs can be designed to operate with extremely short working distances (distance from the CGH to the PV cell) reducing the overall size of the system. In addition, a multiplexing based encoding strategy can be implemented to design holograms that perform a spectral splitting operation on the incident light. The CGHs not only concentrate the incident energy, but can also redirect particular spectral bands to a desired location where a PV cell optimized for that particular spectral band is located. Undesirable spectral bands can

be redirected outside the region of interest, improving the overall efficiency of the PV cell. Alternatively, bands not used for the PV conversion process can be deflected and used for thermal solar heating, effectively using a greater portion of the solar spectrum compared to most conventional techniques. In addition, CGHs can be multiplexed to perform passive tracking of the sun at different points during the day.

Conventional holographic elements have been developed and applied for solar collection [138], [139], [140]. Most of these systems perform a 2D collection and are based on transmission or reflection gratings designed to deflect the incident beam to a desired direction. These holograms are recorded optically, limiting the encoding process, resulting in low diffraction efficiencies over the entire solar spectrum. These holograms are typically unstable over long periods and have a short lifespan.

Figure 2-89 shows the CGH based solar concentrator system studied in this section. The optimized CGH concentrates the incident sunlight and spectrally separates different bands deflecting them to single junction solar cells with matching bandgaps.

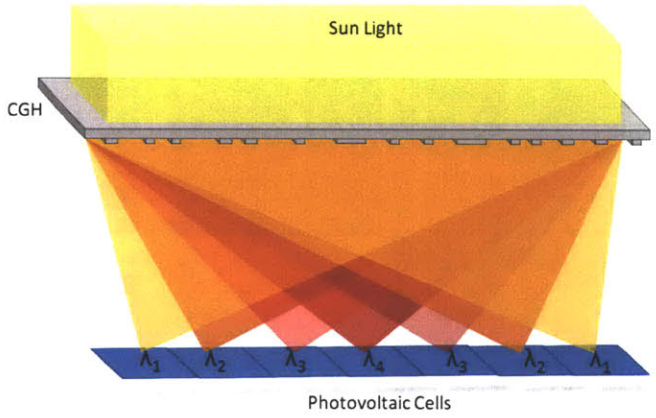


Figure 2-89: CGH based solar concentrator.

We present some preliminary results on the optimization of multispectral CGHs. In the presented example, the optimization is done for four discrete wavelengths. The optimization parameters are indicated in Table 2.8. The corresponding concentration

ratio and maximum full acceptance angle are: $C_{3D} = 10, \theta_1 = 56.63$degrees. Figure 2-90-a shows the optimized multispectral CGH. To demonstrate the spectral splitting capability of the designed CGH, the hologram is reconstructed for each wavelength independently. The reconstructed amplitude is shown in Figure 2-90-b. The effective optical diffraction efficiencies are plotted on Figure 2-91. Diffraction efficiencies over 50% are achieved. We predict that an even higher performance can be obtained by optimizing the multispectral CGH using the HOA discussed above. An additional analysis is required to take into account dispersion effects and to include extended spectral bands.

Table 2.8: Optimization Parameters: Multispectral CGH.

λ_1	731nm	**Working Distance (d)**	25mm
λ_2	887nm	**Hologram Size (H_{size})**	4mm×4mm
λ_3	1.13μm	**Pixel Size (δ_{pix})**	1μm
λ_4	2.48μm	**PV Cell Size**	0.4mm×0.4mm

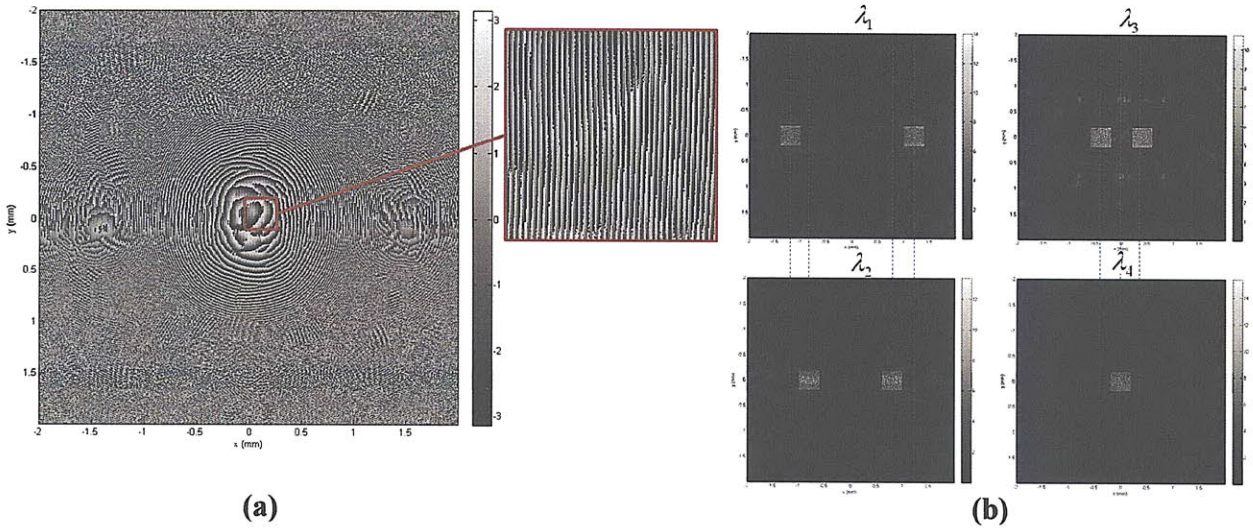


Figure 2-90: (a) Optimized multispectral CGH; (b) Reconstructed amplitude for different operating wavelengths.

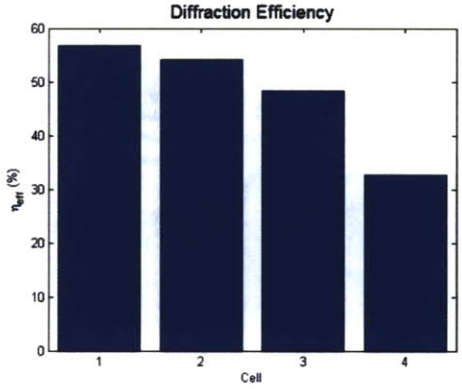


Figure 2-91: Computed diffraction efficiencies per solar cell.

Chapter 3

Design and Optimization of Total Internal Reflection (TIR) Holographic System for Photoresist Exposure in the Fresnel Diffraction Zone

In the previous chapter, the design, optimization and implementation of computer generated holograms (CGHs) and their application to high-resolution lithography was discussed. CGHs were proven to reconstruct high-quality patterns that satisfy the high diffraction efficiency and uniformity demands when optimized using the proposed hybrid optimization algorithm with the reduced complexity formulation that allowed improving significantly the signal encoding process in a computationally efficient manner.

In this chapter, the design and optimization of a total internal reflection (TIR) holographic system is presented. This system is also applied for high-resolution lithography. In contrast to the previous method based on CGHs, the holographic elements used in the TIR system are optically recorded on a photosensitive material. The hologram record-

ing process fixes the type of strategy used to encode the desired signal. This encoding strategy can be optimized by the manipulation of several geometrical, material and fabrication related parameters. Similar to the previous system, the reconstruction or decoding process is performed optically and is applied for the exposure of a substrate coated with photoresist in the Fresnel diffraction zone. The correct modeling and simulation of the recording and reconstruction processes is very important to assist in optimizing the system parameters, as well as designing an optimum binary mask (desired signal) that minimized potential fabrication errors. Scalar and vector diffraction formulations are implemented and their performance under the studied geometry is compared. The results obtained from simulations are compared to those from experiments verifying the validity of the implemented algorithm. The response of the photosensitive material used for the recording of TIR holograms is studied. A method for extending the system's depth of focus is proposed. A novel block-stitching algorithm is introduced for the calculation of large diffraction patterns that overcomes current computational limitations of memory and processing time. Finally, the advantages and disadvantages of the TIR holographic system and the system based on CGHs are compared.

3.1 Background and Problem Definition

The optical recording of holograms using the TIR geometry was originally proposed by Stetson in 1967 [14], [141]. This new form of holography made it possible to record the complex field scattered by an object located in close proximity to the holographic recording material in the near or Fresnel regimes. High spatial frequency components including evanescent waves can be used during the signal encoding process. In addition, the TIR geometry provides an effective way to filter out the undesirable diffraction orders that result from the hologram reconstruction process without the need for additional optical components, such as an aperture stop or 4f system, as required in the off-axis geometry. Over the following years, several researchers experimentally studied the phys-

ical properties of TIR holograms such as Bragg selectivity (over varying incident angle and wavelength), polarization dependence [142], emulsion thickness, diffraction efficiency, and signal-to-noise ratio. Simplified theoretical models based on Kogelnik's coupled wave theory [143] and vector diffraction theory [144] for the extreme cases of thick and thin emulsions were developed to try to explain the experimentally measured data.

Recently, TIR holography has been found to be a promising technique for high-resolution lithography [145], [146]. This geometry allows exposing high-resolution patterns with low background noise while having the photoresist in close proximity to the holographic emulsion. As a result, high effective numerical apertures ($\tilde{}1$) can be easily achieved. In addition, the required system is lensless, inexpensive and allows the exposure of large areas - ideal for applications such as liquid crystal display (LCD) panels [147] and field emission displays [148]. The photoresist exposure process can be integrated with scanning systems for the manufacture of large area semiconductor devices. Subwavelength exposures of periodic patterns have been demonstrated using off-axis illumination of the mask during the TIR recording process [149].

Despite the increasing popularity of TIR holographic lithographic systems, currently there is no multi-domain optimization algorithm designed to improve the overall system's performance. This chapter will present a numerical model that allows simulating the TIR recording and reconstruction processes, including the material response of the holographic emulsion used. The optical, geometrical and material parameters of the simulation algorithm can be used to optimize the photoresist exposure for arbitrary mask patterns.

3.2 System Geometry

The TIR recording and reconstruction geometries are shown in Figure 3-1. In the recording step, the field diffracted by an amplitude mask (such as a chromium pattern on a fused silica substrate) located at a working distance, d, from the holographic material in-

terferes with the incident and reflected components of the reference wave. The reference wave gets coupled into the system with the help of a right angle prism and illuminates the hologram at an angle, θ, larger than the critical angle, $\theta_c = \arcsin(n_1/n_2)$. As a result, the reference wave suffers TIR at the prism-air and hologram-air interfaces and exits the system in the conjugate direction. The holographic emulsion has a refractive index approximately equal to that of the prism and is held together using, for example, optical glue or oil. Three effective holograms are recorded: transmission, reflection, and Lippmann [150]. The transmission hologram results from the interference of the reflected component of the reference wave and the object wave and has relatively low spatial frequencies. The reflection hologram is produced by the interference of the incident reference wave and the object wave and has higher spatial frequencies. The Lippmann hologram is produced by the self-interference of the incident and reflected reference waves. The influence of this hologram can be reduced to some extent by controlling the polarization of the reference wave and the incident angle [151], [152]. Upon total reflection, the wave suffers a phase shift known as Goos-Hänchen shift [153], which changes the polarization state of the reflected wave. It is sometimes desirable to eliminate the additional recorded holograms as much as possible, as they utilize the finite dynamic range provided by the photosensitive holographic material lowering the resulting diffraction efficiency. A replication of the recorded hologram to form a surface relief hologram has been proven to reduce the effect of these additional holograms at the expense of a more elaborate fabrication process [154]. Surface relief holograms have a higher index contrast that results in higher diffraction efficiency and longer lifespan compared to that of conventional photopolymers, as they are typically etched into a glass substrate. After recording, the hologram undergoes developing and bleaching processes or alternatively a fixation process using UV uniform illumination.

In the reconstruction step, the hologram is illuminated by a probing wave that is phase conjugate of the reference wave. This results in desirable and undesirable diffraction orders. The desirable diffraction orders suffer frustrated TIR and propagate towards the

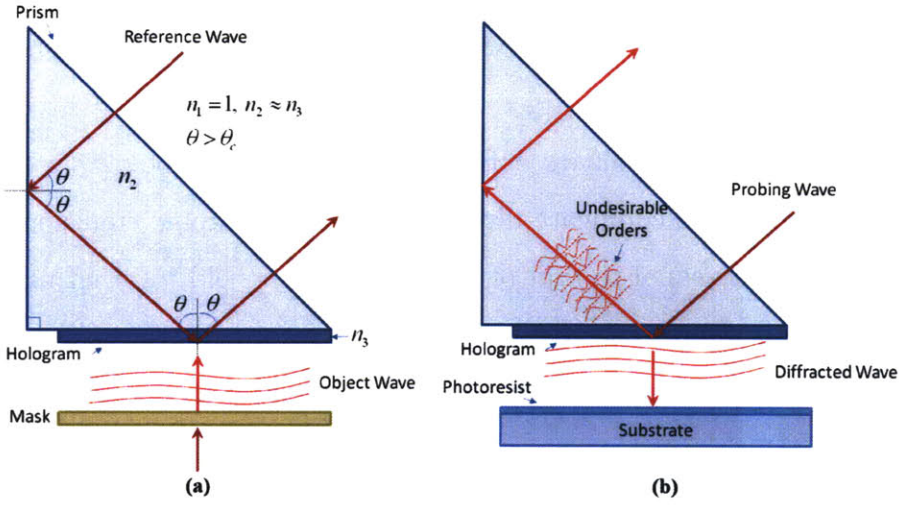


Figure 3-1: (a) TIR recording geometry; (b) TIR reconstruction geometry.

photoresist plane located at the same working distance, d. The undesirable diffraction orders suffer TIR at the hologram-air interface and exit the system in the conjugate direction as the probing wave. After exposure, the photoresist undergoes the standard developing and etching processes.

3.3 Comparison of Vector and Scalar Diffraction Formulations for the Simulation of TIR Holographic Systems

We begin our analysis by comparing vector and scalar formulations for calculating the optical field diffracted by the mask at the hologram plane. Three methods are compared: rigorous coupled wave analysis, finite-difference time-domain and scalar diffraction theory (Fresnel approximation and exact formulation). The modeled masks consist of phase and amplitude binary gratings with periods: $\Lambda = 2\mu$m. The phase mask is used to simulate the diffraction from the hologram in the reconstruction step. The operation wavelength

and working distance are: $\lambda = 350$nm and $d = 200\mu$m. Two polarization states are considered: TE and TM. After the analysis, the best method (in terms of accuracy and computational performance) will be selected and used for modeling the entire TIR holographic system.

3.3.1 Rigorous Coupled Wave Analysis

Rigorous coupled wave analysis (RCWA) is one of the most widely used methods for the accurate analysis of the diffraction of electromagnetic waves by periodic structures such as gratings, holograms and surface relief structures. It provides a relatively straightforward technique for obtaining the exact solution of Maxwell's equations for the diffraction of grating structures. It is a noniterative, computationally efficient, deterministic method based on state-variable space representation that converges without numerical instabilities. The accuracy of the solution depends solely on the number of terms retained in the field space-harmonic expansion.

RCWA was originally proposed by Moharam and Gaylord in 1981 [155]. Since then, this method has been successfully used to analyze a variety of structures, including transmission and reflection planar dielectric-absorption holographic gratings [156], binary phase gratings [157], arbitrary profiled dielectric-metallic surface relief gratings [158], [159], multiplexed holographic gratings [160], and anisotropic gratings for both planar and conical diffraction [161].

Diffraction from Binary Phase Grating

We begin by describing the RCWA formulation for the case of a binary phase grating mask under TE polarized incident illumination (where the electric field oscillates perpendicular to the plane of incidence). The general geometry of the grating structure is shown in Figure 3-2. The grating is assumed to be of infinite extent. The problem parameters are indicated in Table 3.1. We will consider the case of normal incidence; however, the general formulation will be presented.

Table 3.1: Problem Parameters: Binary Phase Grating.

Wavelength (λ)	350nm	n_I	1
Period (Λ)	2μm	n_{II}	2.04
Duty Cycle (f)	0.5	**Thickness** (t)	200nm
Incidence Angle (θ)	0	**Polarization**	TE

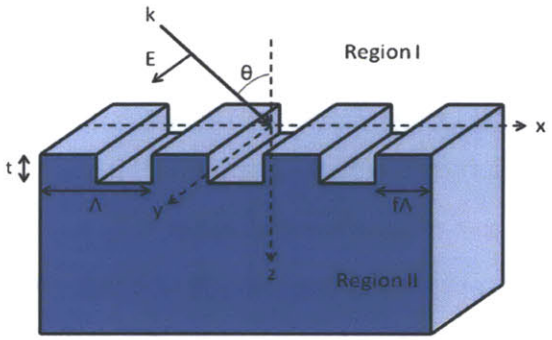


Figure 3-2: Grating geometry: TE polarization.

Three regions are defined: region I (air), region II (dielectric) and grating region ($0 < z < t$). In the grating region, the periodic relative permittivity can be represented using Fourier series,

$$\varepsilon(x) = \sum\nolimits_h \varepsilon_h \exp\left(i\frac{2\pi hx}{\Lambda}\right), \tag{3.1}$$

where the Fourier coefficient is given by,

$$\varepsilon_h = \frac{1}{\Lambda}\int_{-\frac{\Lambda}{2}}^{\frac{\Lambda}{2}} \varepsilon(x)\exp\left(-i\frac{2\pi hx}{\Lambda}\right)dx. \tag{3.2}$$

The refractive index and relative permittivity are related by: $n = \sqrt{\varepsilon}$ (where we assumed

$\mu_{relative} \approx 1$). The relative permittivity within one grating period is,

$$\varepsilon(x) = \begin{cases} n_I^2 & -\frac{\Lambda}{2} < x < -\frac{f\Lambda}{2} \\ n_{II}^2 & -\frac{f\Lambda}{2} < x < \frac{f\Lambda}{2} \\ n_I^2 & \frac{f\Lambda}{2} < x < \frac{\Lambda}{2} \end{cases} . \tag{3.3}$$

Substituting equation 3.3 into equation 3.2 we find the expression for the Fourier harmonics,

$$\begin{aligned} \varepsilon_0 &= f n_{II}^2 + n_I^2 (1 - f), \\ \varepsilon_h &= \left(n_{II}^2 - n_I^2\right) \frac{\sin(\pi h f)}{\pi h}, \end{aligned} \tag{3.4}$$

where ε_0 is the average value of the relative permittivity. Figure 3-3 shows the relative permittivity modulation for the problem parameters of Table 3.1. The Gibbs phenomena shown in this figure is related to the number of retained Fourier harmonics.

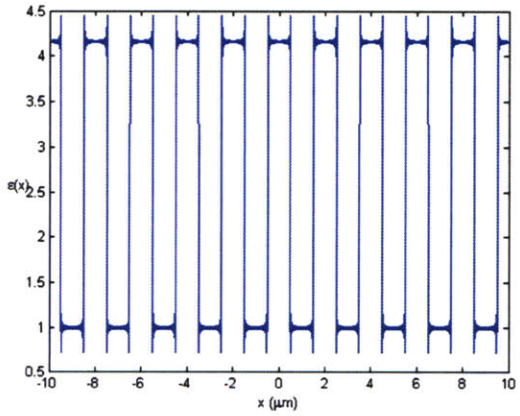


Figure 3-3: Relative permittivity modulation of phase binary grating.

The binary grating structure is illuminated by a TE polarized plane wave given by,

$$E_{inc,y} = \exp\left[-ikn_I\left(\sin\theta x + \cos\theta z\right)\right], \tag{3.5}$$

where $k = 2\pi/\lambda$. The total electric field in region I is,

$$E_{I,y} = E_{inc,y} + \sum_q R^{(q)} \exp\left[-i\left(k_x^{(q)}x - k_{I,z}^{(q)}z\right)\right], \tag{3.6}$$

where $R^{(q)}$ is the qth reflection coefficient, and $k_x^{(q)}$ and $k_{I,z}^{(q)}z$ are the components of the qth reflected wave vector along the x and z directions respectively. Similarly, the total electrical field in region II is,

$$E_{II,y} = \sum_q T^{(q)} \exp\left[-i\left(k_x^{(q)}x - k_{II,z}^{(q)}(z-t)\right)\right], \tag{3.7}$$

where $T^{(q)}$ is the qth transmission coefficient, and $k_x^{(q)}$ and $k_{II,z}^{(q)}$ are the components of the qth transmitted wave vector. The x-component of the reflected and transmitted wave vectors, $k_x^{(q)}$, is required to satisfy the Floquet condition (phase matching condition),

$$k_x^{(q)} = k\left[n_I \sin\theta - q\frac{\lambda}{\Lambda}\right]. \tag{3.8}$$

The z-component of the wave vectors in both regions are derived from the dispersion relation for the cases of external and internal reflection,

$$k_{\mathcal{L},z}^{(q)} = \begin{cases} k\left[n_{\mathcal{L}}^2 - \left(\frac{k_x^{(q)}}{k}\right)^2\right]^{1/2} & n_{\mathcal{L}}k > k_x^{(q)} \\ -ik\left[\left(\frac{k_x^{(q)}}{k}\right)^2 - n_{\mathcal{L}}^2\right]^{1/2} & n_{\mathcal{L}}k < k_x^{(q)} \end{cases}, \tag{3.9}$$

where $\mathcal{L} = I,\ II$.

Next, we express the tangential electric field in the grating region as a Fourier expansion,

$$E_{g,y} = \sum_q S_y^{(q)}(z) \exp\left(-ik_x^{(q)}x\right), \tag{3.10}$$

where $S_y^{(q)}$ is the normalized amplitude of the qth space-harmonic electric field. The Fourier expansion of the x-component of the magnetic field is found using equation 3.10

and Maxwell's equation: $\bar{H} = (i/\omega\mu) \bigtriangledown \times \bar{E}$,

$$H_{g,x} = -i \left(\frac{\epsilon_0}{\mu_0} \right)^{1/2} \sum\nolimits_q U_x^{(q)}(z) \exp\left(-ik_x^{(q)} x\right), \tag{3.11}$$

where $U_x^{(q)}$ is the normalized amplitude of the qth space-harmonic magnetic field, and ϵ_0 and μ_0 are the permittivity and permeability of free space. Equations 3.10 and 3.11 satisfy Maxwell's equations,

$$\begin{aligned} \frac{\partial E_{g,y}}{\partial z} &= i\omega\mu_0 H_{g,x}, \\ \frac{\partial H_{g,x}}{\partial z} &= i\omega\varepsilon_0\varepsilon(x) E_{g,y} + \frac{\partial H_{g,z}}{\partial x}. \end{aligned} \tag{3.12}$$

Substituting equations 3.10 and 3.11 into equation 3.12 we obtained the set of coupled-wave differential equations,

$$\begin{aligned} \frac{\partial S_y^{(q)}}{\partial z} &= k U_x^{(q)}, \\ \frac{\partial U_x^{(q)}}{\partial z} &= \left[\frac{\left(k_x^{(q)}\right)^2}{k} \right] S_y^{(q)} - k \sum\nolimits_p \varepsilon\,(q-p)\, S_y^{(p)} \end{aligned} \tag{3.13}$$

We can rewrite equation 3.13 into the reduced matrix form,

$$\left[\frac{\partial^2 \mathbf{S}_y}{\partial\,(z')^2} \right] = [\mathbf{A}]\,[\mathbf{S}_y], \tag{3.14}$$

where $z' = kz$; $\mathbf{A} = \mathbf{K}_x^2 - \mathbf{E}$ ($\mathbf{A}$ is an $n \times n$ matrix, where n is the number of harmonics retained after the expansion and is symmetric for dielectric gratings and Hermitian for lossy gratings); $\mathbf{E}$ is the matrix formed by the permittivity harmonic components, with the q, p element being equal to $\epsilon(q-p)$; and $\mathbf{K}_x$ is a diagonal matrix with the q, q element being equal to $k_x^{(q)}/k$.

The set of coupled differential equations is solved using the eigenvalues-eigenvector

method. The general solution is given by,

$$\begin{aligned} S_y^{(q)}(z) &= \sum_{m=1}^{n} v_m^{(q)} \left[c_m^+ \exp\left(-k\zeta_m z\right) + c_m^- \exp\left(k\zeta_m (z-t)\right) \right], \\ U_x^{(q)}(z) &= \sum_{m=1}^{n} \kappa_m^{(q)} \left[-c_m^+ \exp\left(-k\zeta_m z\right) + c_m^- \exp\left(k\zeta_m (z-t)\right) \right], \end{aligned} \tag{3.15}$$

where $v_m^{(q)}$ are the elements of the eigenvector matrix $\mathbf{V}$; ζ_m are the positive square root eigenvalues of the matrix $\mathbf{A}$ (elements of diagonal matrix $\boldsymbol{\zeta}$); c_m^+ and c_m^- are unknown constants to be determined using the boundary conditions at $z = 0$ and $z = t$; and $\kappa_m^{(q)} = \zeta_m v_m^{(q)}$ are elements of the matrix $\boldsymbol{\kappa} = \mathbf{V}\boldsymbol{\zeta}$.

Next we use the boundary conditions to match the amplitudes of the electric and magnetic components to $R^{(q)}$ and $T^{(q)}$ in regions I and II. The resulting set of coupled differential equations (in matrix form) is,

$$\begin{aligned} \begin{bmatrix} \delta^{(q)} \\ in_I \cos\theta \delta^{(q)} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ -i\mathbf{Y}_I \end{bmatrix} [\mathbf{R}] &= \begin{bmatrix} \mathbf{V} & \mathbf{VX} \\ \boldsymbol{\kappa} & -\boldsymbol{\kappa}\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{c}^+ \\ \mathbf{c}^- \end{bmatrix}, \\ \begin{bmatrix} \mathbf{VX} & \mathbf{V} \\ \boldsymbol{\kappa}\mathbf{X} & -\boldsymbol{\kappa} \end{bmatrix} \begin{bmatrix} \mathbf{c}^+ \\ \mathbf{c}^- \end{bmatrix} &= \begin{bmatrix} \mathbf{I} \\ i\mathbf{Y}_{II} \end{bmatrix} [\mathbf{T}], \end{aligned} \tag{3.16}$$

where $\delta^{(q)} = 1$ for $q = 0$ and $\delta^{(q)} = 0$ for $q \neq 0$; $\mathbf{X}$, $\mathbf{Y}_I$ and $\mathbf{Y}_{II}$ are diagonal matrices with the diagonal elements $e(ik\zeta_m t)$, $(k_{I,z}^{(q)}/k)$ and $(k_{II,z}^{(q)}/k)$; and $\mathbf{I}$ is the identity matrix.

Equation 3.16 is solved simultaneously to find the values of $R^{(q)}$ and $T^{(q)}$. Figure 3-4 shows the magnitude and phase of the total optical fields in regions I and II.

The diffraction efficiencies for the reflected and transmitted waves are defined as,

$$\begin{aligned} DE_R^{(q)} &= R^{(q)} R^{(q)^*} \operatorname{Re}\left(\frac{k_{I,z}^{(q)}}{k n_I \cos\theta} \right), \\ DE_T^{(q)} &= T^{(q)} T^{(q)^*} \operatorname{Re}\left(\frac{k_{II,z}^{(q)}}{k n_I \cos\theta} \right). \end{aligned} \tag{3.17}$$

Energy conservation requires that,

$$\sum_q DE_R^{(q)} + DE_T^{(q)} = 1. \tag{3.18}$$

Figure 3-5 shows the corresponding diffraction efficiencies for the binary phase grating structure.

Diffraction from Binary Amplitude Grating

We now use RCWA to calculate the diffracted field from a binary amplitude grating. The grating consists of a quartz substrate and a thin layer of chromium, patterned to form the grating lines in a similar geometry as that of Figure 3-2. The results presented in this section are for TE polarized illumination. The problem parameters are indicated in Table 3.2. The complex index of refraction of chromium was obtained from [162].

Table 3.2: Problem Parameters: Binary Amplitude Grating.

Wavelength (λ)	350nm	n_I	1
Period (Λ)	2μm	n_{II}	1.5655
Duty Cycle (f)	0.5	$n_{chromium}$	$1.812 - i2.619$
Incidence Angle (θ)	0	**Thickness** (t)	200nm

The problem formulation is similar to that described in the previous section except that the complex permittivity function is used to characterize the dielectric modulation within the grating region. Figure 3-6 shows the complex relative permittivity within the grating region. The phase and amplitude of the total fields in regions I and II are shown in Figure 3-7. The corresponding diffraction efficiencies for the reflected and transmitted waves are shown in Figure 3-8.

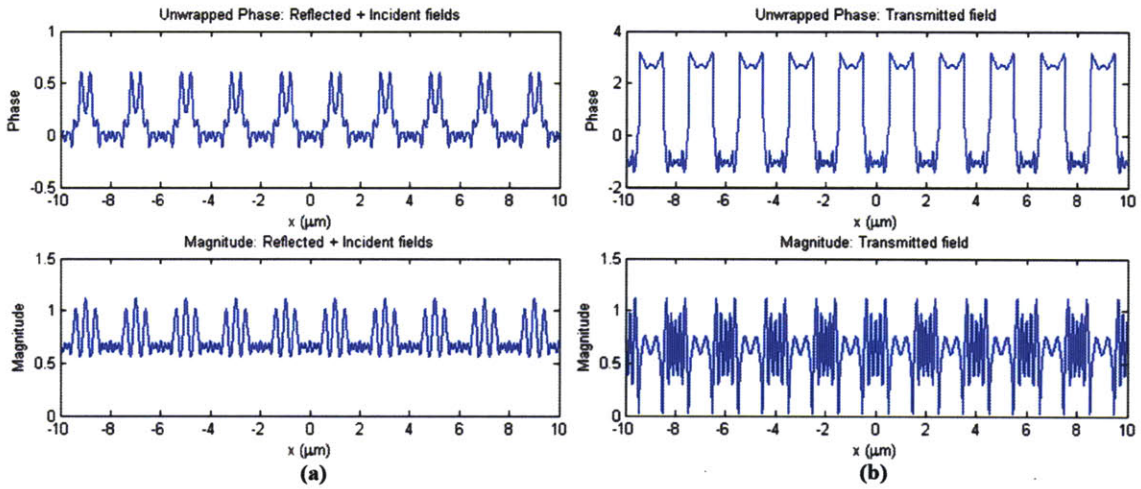


Figure 3-4: (a) Complex field in Region I; (b) Complex field in Region II.

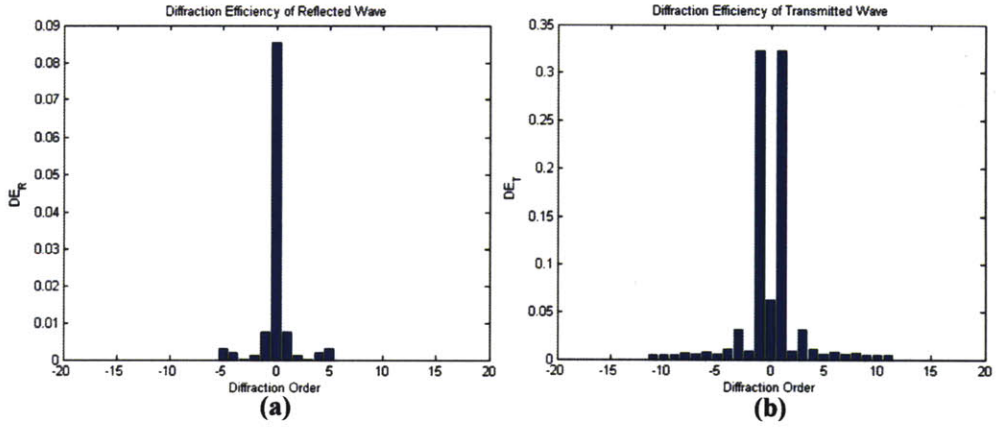


Figure 3-5: (a) Diffraction efficiency of reflected wave; (b) Diffraction efficiency of transmitted wave.

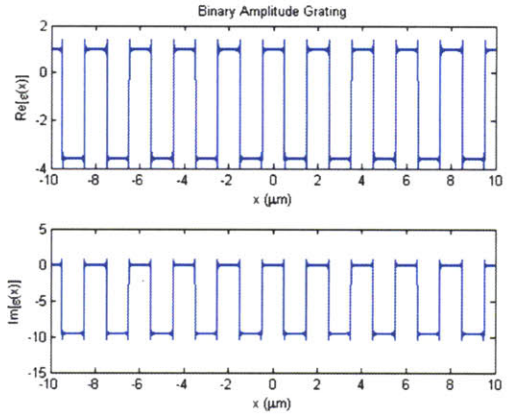


Figure 3-6: Complex permittivity of amplitude binary grating.

3.3.2 Finite-Difference Time-Domain (FDTD) Method

In the FDTD method, the time-dependent Maxwell equations are discretized using central-difference approximations and the resulting set of finite-difference equations is solved. The solution of the problem is done in a leapfrog manner: the electric field vector components are solved in a volume of space (Yee cell [163]) at a given instant in time; then the magnetic field vector components at the same volume of space are solved at the next instant in time. The process is repeated until the desired steady-state electromagnetic field behavior is reached. The main strengths of FDTD modeling are its versatility and ease of use. In addition, animations of the electrodynamics of the problem can be produced as the algorithm calculates the electric and magnetic fields everywhere in the computational domain as they evolve in time. Some of the disadvantages are that the entire computational domain must be gridded, and the grid spatial discretization must be sufficiently fine to resolve both the smallest geometrical feature and operation wavelength. This results in long computational times and memory problems. For this reason, the computational domain must be truncated by inserting artificial boundaries, such as a perfectly matched layer (PML). The FDTD method has been used in a variety of problems in computational electrodynamics such as photonic crystals, nanoplasmonics,

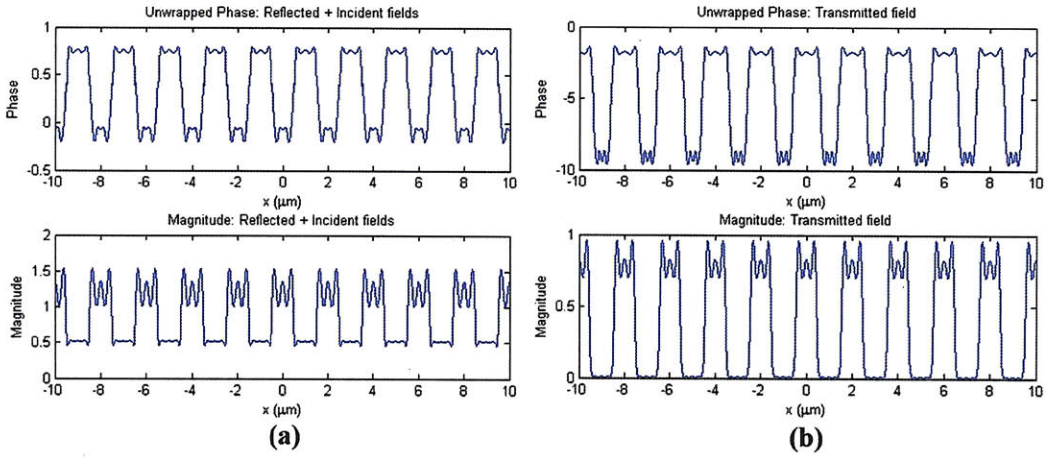


Figure 3-7: (a) Complex field in Region I; (b) Complex field in Region II.

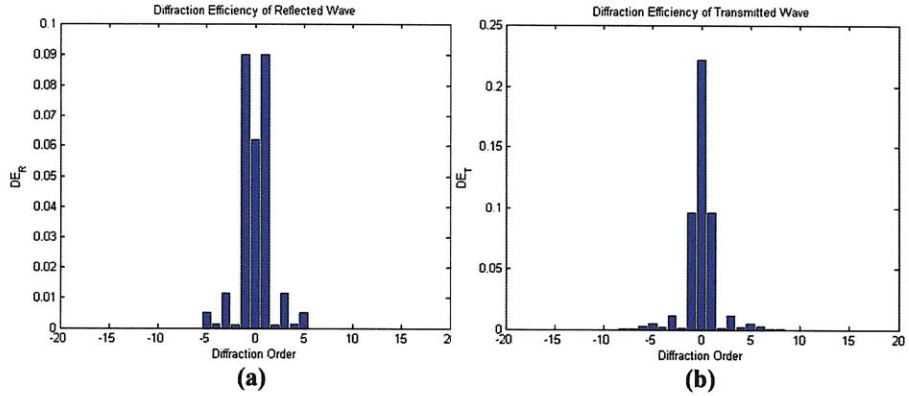


Figure 3-8: (a) Diffraction efficiency of reflected wave; (b) Diffraction efficiency of transmitted wave.

solitons, antennas and radar [164].

The numerical results presented in this section were performed by Satoshi Takahashi using the open source software Meep [165]. Figure 3-9-a shows the evolution of the electric field for the case of the binary phase grating discussed previously. The simulation parameters are the same as those in Table 3.1. The simulation results for the case of the binary amplitude grating are shown in Figure 3-9-b. In this simulation, the chromium layer was replaced by a perfect conductor.

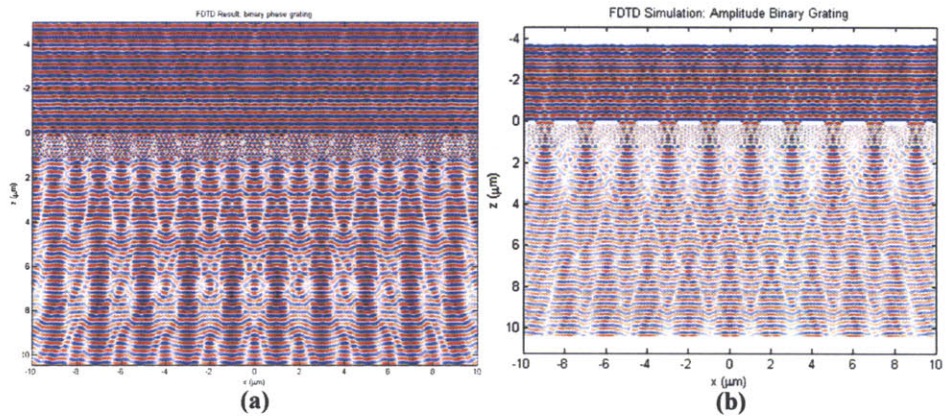


Figure 3-9: (a) FDTD results for binary phase grating; (b) FDTD results for binary amplitude grating.

3.3.3 Scalar Diffraction Theory Analysis

The behavior of the electric and magnetic field components inside a dielectric medium that is linear, isotropic, homogeneous and nondispersive is identical and it is fully described by the scalar wave equation. In the treatment of light using scalar diffraction theory analysis (SDTA), phenomena associated with the vectorial nature of the electric and magnetic fields, such as polarization, is ignored. For the analysis of TIR holography, this formulation is accurate under the following assumptions: thin hologram (Raman-Nath regime [67]), to avoid coupling effects due to the variations of the refractive index

inside the volume; neglect polarization; and diffracting structures large compared to the operation wavelength.

SDTA offers a variety of formulations including Huygens' Principle [166], Rayleigh-Sommerfeld theory [167], Kirchhoff formulation [67] and angular spectrum propagation [168]. In the previous chapter, the formulations based on the first solution of the Rayleigh-Sommerfeld and angular spectrum theories were discussed for the optical field propagation in the Fresnel regime. The free space propagation impulse response for the Fresnel approximation and exact formulation are given by equations 2.17 and 2.21 respectively. As discussed previously, the computation of the diffracted field is numerically implemented in frequency domain (equation 2.48) using the Fresnel approximation (equation 2.49) or angular spectrum propagation (equation 2.52) transfer functions depending on the geometry of the system (equation 2.20).

Block and Stitching Method for the Calculation of Large Diffraction Patterns

Holographic lithography often requires the calculation of a field diffracted from a large pattern (transmission function) such as a mask. The large size of the transmission function makes the computational implementation particularly challenging. To numerically compute the diffracted field, the transmission function needs to be discretized using $N \times N$ pixels. The resulting large space-bandwidth product makes it impossible to process this function with conventional computational means. Memory and computational time limitations promptly arise.

To solve this problem we propose an algorithm designed for the efficient calculation of large diffraction patterns. We call this algorithm block-stitching (BS) method. The BS method is based on SDTA and it computes the diffracted field by splitting the problem into manageable portions (blocks), perform field propagation operations on each block and stitching them afterwards. This algorithm leverages on the spatial locality of most Fresnel diffraction patterns. For binary masks that are composed, for example, of squares, lines and circles such as those used in lithography, the field diffracted by each el-

ement is confined within a small patch of size given by the corresponding Fresnel number.

The geometry of the studied diffraction problem is shown in Figure 3-10. The corresponding block diagram of the BS method is shown in Figure 3-11. A transmission function, $t(x, y)$, of size X_{size} is defined on plane 1 and is divided into $M_x \times M_y$ blocks of size $B_{size,x} = X_{size}/M_x$ (for the x-direction). The total number of blocks depends on: 1. Computational time (each processed block adds memory transfer overhead); 2. Spatial frequency content of the signal inside the block. The second point is related to the required block size after zero padding. Zero padding the block before performing the field propagation operation is necessary, as the diffracted wave diverges at a rate proportional to the signal's spatial frequency content. In addition, a zero padded signal results in better sampling in frequency domain and avoids aliasing in the evaluation of the propagation transfer function.

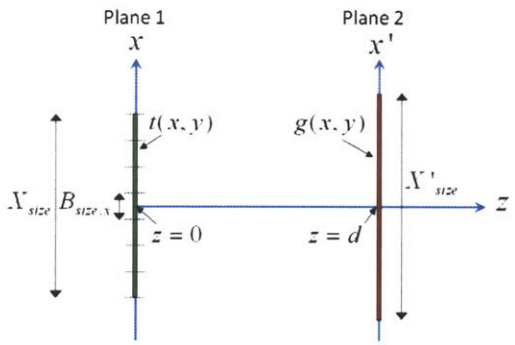


Figure 3-10: Geometry of diffraction problem.

To determine the block size after zero padding two methods are employed: Fresnel number and signal's power spectrum. The first method based on the Fresnel number is particularly suitable for transmission functions composed of binary patterns, such as traditional masks used in lithography. Depending on their signal content, different blocks might lead to different block sizes after zero padding. The size of the mth zero padded

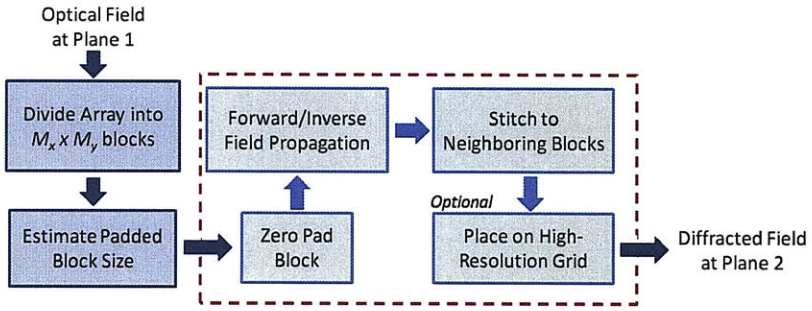


Figure 3-11: Block diagram of BS method.

block along the x and y directions is given by,

$$\hat{B}^{(m)}_{size,\mathcal{L}} = w^{(m)}_{h,,\mathcal{L}} \left(\frac{\kappa}{F^{(m)}_{\mathcal{L}}} + 2 \right) + l^{(m)}_{offset,\mathcal{L}}, \tag{3.19}$$

where $\mathcal{L} = x, y$ (specifies the direction); w_h is the half width of the smallest aperture inside the mth block (equals to the smallest radius for the case of a circular aperture); $\kappa = 1$ for rectangular apertures and $\kappa = 1.22$ for circular apertures (from the location of the first ring of the Airy pattern); l_{offset} is the length offset of the center of the smallest aperture respect to the center of the block; and F is the Fresnel number given by,

$$F^{(m)}_{\mathcal{L}} = \frac{\left(w^{(m)}_{h,,\mathcal{L}} \right)^2}{\lambda d}, \tag{3.20}$$

where λ is the operation wavelength and d is the propagation distance. The zero padded block size of equation 3.19 guarantees that most of the energy of the diffracted signal at plane 2 is contained within the computed padded patch in order to avoid potential stitching errors.

For the second method, the power spectrum of the mth block is computed. Next, the cut-off frequencies along the horizontal and vertical directions for an idea low-pass window that contains more than 90% of the energy after being applied to the signal are found.

The cut-off frequencies correspond to the direction cosines of the highest off-axis plane wave (highest spatial frequency of interest) emanating from plane 1: $\alpha_{cutoff} = u_{cutoff}\lambda$, $\beta_{cutoff} = u_{cutoff}\lambda$. The resulting off-axis angle sets the effective numerical aperture of the system: $NA_{eff,x} = \arccos(\alpha_{cutoff})$, $NA_{eff,y} = \arccos(\beta_{cutoff})$. The zero padded block size is given by,

$$\hat{B}^{(m)}_{size,\mathcal{L}} = 2dNA_{eff,\mathcal{L}} + B^{(m)}_{size,\mathcal{L}}. \tag{3.21}$$

The second method for determining the zero padded block size is useful for general signals other than the binary patterns used in lithography. An alternative method can be implemented based on the Wigner distribution function [169]. Despite the chosen method, the maximum block size after padding is ultimately limited by the memory available in the processing unit. In the presented work, a maximum array size of 4000×4000 elements ($\hat{B}_{size,x}/\delta_{pix} \times \hat{B}_{size,y}/\delta_{pix}$) is implemented.

The following example illustrates the calculation of the zero padded block size using the two methods described above. Consider a finite size transmission function given by (1D case),

$$\begin{aligned} t(x) &= \left[\frac{1}{2} + \frac{1}{2}\cos\left(2\pi f_o x\right)\right] \text{rect}\left(\frac{x}{X_{size}}\right) \\ &= g(x)\text{rect}\left(\frac{x}{X_{size}}\right), \end{aligned} \tag{3.22}$$

where f_o is the grating's frequency. The transmission function is divided into three blocks ($M = 3$),

$$\begin{aligned} t_{-1}(x) &= g(x)\text{rect}\left(\frac{x - B_{size}}{B_{size}}\right), \\ t_0(x) &= g(x)\text{rect}\left(\frac{x}{B_{size}}\right), \\ t_1(x) &= g(x)\text{rect}\left(\frac{x + B_{size}}{B_{size}}\right), \end{aligned} \tag{3.23}$$

where $B_{size} = X_{size}/M$. It can be easily proved that by adding the equations 3.23 we

arrive at equation 3.22. The Fourier transform of each block's transmission function is given by,

$$\begin{aligned} T_0(u) &= \frac{B_{size}}{2}\text{sinc}\left(B_{size}u\right) + \frac{B_{size}}{4}\left[\text{sinc}\left(B_{size}(u-f_o)\right) + \text{sinc}\left(B_{size}(u+f_o)\right)\right] \quad (3.24) \\ T_{-1}(u) &= \frac{B_{size}}{2}\text{sinc}\left(B_{size}u\right)e^{-i2\pi B_{size}u} + \frac{B_{size}}{4}\left[\text{sinc}\left(B_{size}(u-f_o)\right)e^{-i2\pi B_{size}(u-f_o)}\right. \\ &\quad \left.+\text{sinc}\left(B_{size}(u+f_o)\right)e^{-i2\pi B_{size}(u+f_o)}\right], \\ T_1(u) &= \frac{B_{size}}{2}\text{sinc}\left(B_{size}u\right)e^{i2\pi B_{size}u} + \frac{B_{size}}{4}\left[\text{sinc}\left(B_{size}(u-f_o)\right)e^{i2\pi B_{size}(u-f_o)}\right. \\ &\quad \left.+\text{sinc}\left(B_{size}(u+f_o)\right)e^{i2\pi B_{size}(u+f_o)}\right]. \end{aligned}$$

The summation of the frequency spectrums of equation 3.24 leads to the spectrum of the original transmission function,

$$\begin{aligned} T(u) &= \sum\nolimits_m T_m(u) = \frac{X_{size}}{2}\text{sinc}\left(X_{size}u\right) \quad (3.25) \\ &\quad + \frac{X_{size}}{4}\left[\text{sinc}\left(X_{size}(u-f_o)\right) + \text{sinc}\left(X_{size}(u+f_o)\right)\right]. \end{aligned}$$

To estimate the zero padded block size using equation 3.21, consider the following parameters: $\lambda = 500\text{nm}$, $\delta_{pix} = 100\text{nm}$, $X_{size} = 600\mu\text{m}$, $\Lambda_{grating} = 1\mu\text{m}$ ($\Lambda_{grating} = 1/f$), and $d = 100\mu\text{m}$. The corresponding space-bandwidth product is, $SBP = 6000 \times 6000$, which is too large to process conventionally. The block size is: $B_{size} = 200\mu\text{m}$ ($SBP_{block} = 2000 \times 2000$). From equation 3.24, we see that the block's spectrum is composed of a sinc function centered on the frequency axis and two additional sinc functions shifted by f_o. The cut-off frequency is given by the location of the first zero of the shifted sinc functions: $u_{cutoff} = f_o + 1/B_{size} = 1005\text{mm}^{-1}$. The corresponding effective numerical aperture is: $NA_{eff} = 0.5025$. The zero padded block size given by equation 3.21 is: $\hat{B}_{size} = 300.5\mu\text{m}$ ($SBP = 3005 \times 3005$). Dividing the function into blocks allowed us to reduce the SBP to an amount that is manageable with current computational means. We now compare the zero padded block size computed using equation 3.19. To use this equation, the grating pattern is binarized: $g(x) = 1$ for $g(x) \geq 1/2$ and $g(x) = 0$ for $g(x) < 1/2$. Each grating

line becomes a rectangular aperture ($\kappa = 1$) with a half size: $w_h = 0.25\mu$m. For the center block, the offset distance of the furthest aperture is: $l_{offset} = B_{size}/2 = 100\mu$m. The corresponding Fresnel number is: $F = 1.25 \times 10^{-3}$. The resulting zero padded block size is: $\hat{B}_{size} = 300.5\mu$m.

The next step in the BS method is to zero pad the mth block to the calculated size. Then, the block's diffraction pattern is computed by performing forward or inverse field propagations. Next, the diffracted field at plane 2 is stitched to the neighboring diffracted blocks coherently adding the overlapping regions. As an optional step, the computed diffraction block is placed on a high-resolution grid (provided there is enough memory); otherwise, the result is saved to hard drive. These steps repeat until all the blocks have been processed.

Figure 3-12-a shows an example of the segmentation of a large binary mask into 16 blocks to be processed by the BS method. The zero padding and diffraction steps for the first block are shown in Figure 3-12-b. The stitching procedure for the first four blocks is illustrated in Figure 3-13-a. Figure 3-13-b shows the final diffracted pattern. Notice how the stitching procedure did not introduce any boundary errors.

3.3.4 Comparison of Diffraction Theories

The three methods described above (RCWA, FDTD and SDTA) are used to simulate the field diffracted from an amplitude binary grating mask with a working distance: $d = 200\mu$m. Figure 3-14 shows a comparison of the resulting field amplitude along the x-direction. As can be seen, the results from SDTA using both the Fresnel approximation and exact formulations very closely follow the solution computed from RCWA. The discrepancies between FDTD and RCWA are attributed to the type of material used and the selected computational region. For the simulation results based on RCWA, the pattern of the amplitude mask is composed of chromium. On the other hand, a perfect conductor is used for the simulation based on FDTD. In addition, spurious reflections from the perfectly matched layer are observed that contributed to the resulting field

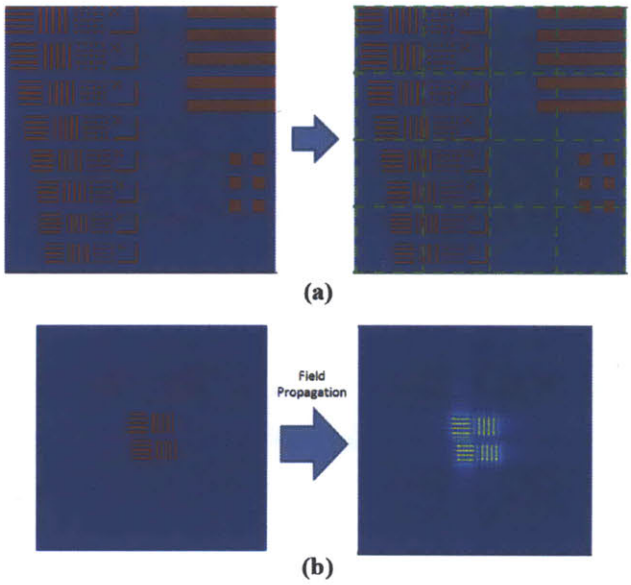


Figure 3-12: (a) Mask segmentation process; (b) Zero padding and diffraction calculation of first block.

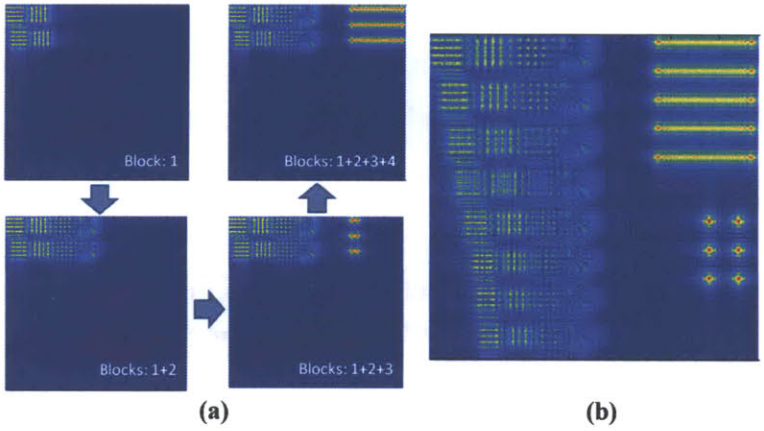


Figure 3-13: (a) Stitching process; (b) Final diffracted field.

amplitude.

From the results of Figure 3-14, we conclude that the SDTA formulation is feasible for modeling the recording and reconstruction processes of TIR holograms. SDTA is computationally efficient, accurate and allows simulating different mask designs. In contrast, RCWA can only be used to simulate diffraction from periodic structures.

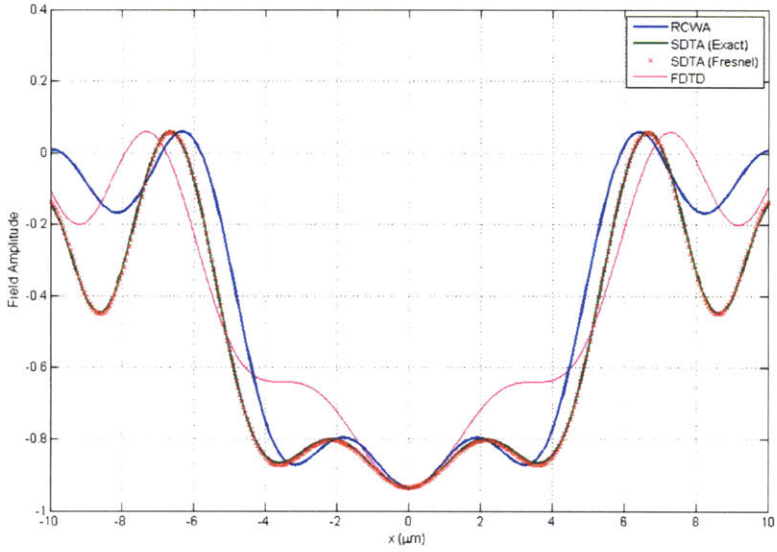


Figure 3-14: Comparison of diffraction theories.

3.4 Simulation and Optimization of TIR Holographic Systems

3.4.1 Optical Recording Process

The TIR hologram recording geometry is shown in Figure 3-15. To simplify the analysis, the incidence angle is set to: $\theta = \pi/4$. This angle is slightly larger than the critical angle: $\theta_c \approx 41.8$degrees (for $n_2 \approx n_3 = 1.5$). The reference wave suffers TIR at the dielectric-air

interface, giving rise to two plane waves with wave vectors,

$$\bar{k}_1 = \frac{2\pi}{\lambda_{eff}} \begin{bmatrix} -\sin\theta \\ -\cos\theta \end{bmatrix}, \quad \bar{k}_2 = \frac{2\pi}{\lambda_{eff}} \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}, \tag{3.26}$$

where λ_{eff} is the effective wavelength inside the dielectric media: $\lambda_{eff} = \lambda/n$. At the hologram plane ($z = d$), the reference waves are given by,

$$r_1 = \exp\left[-i\frac{2\pi}{\lambda_{eff}\sqrt{2}}(x' + d)\right], \quad r_2 = \exp\left[i\frac{2\pi}{\lambda_{eff}\sqrt{2}}(-x' + d)\right]. \tag{3.27}$$

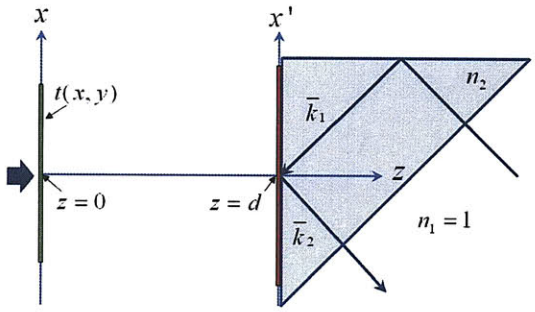


Figure 3-15: TIR recording geometry.

The holographic emulsion is exposed by the intensity distribution,

$$I = |r_1 + r_2 + O|^2 = |r_1|^2 + |r_2|^2 + |O|^2 + (r_1 + r_2)\,O^* + (r_1^* + r_2^*)\,O + r_1^* r_2 + r_1 r_2^*, \tag{3.28}$$

where O is the field diffracted by the mask: $O = t(x, y) \otimes h(x, y; d)$. The spectral representation of the exposed intensity distribution is shown in Figure 3-16-a. The encoded signal is modulated by a high frequency carrier with frequency proportional to incident angle of the reference wave. Similar bandwidth constraints, as those described in the previous chapter (equation 2.12), apply.

Depending on the type of holographic material used, two types of holograms can be

recorded: amplitude and phase. The amplitude hologram is recorded, for example, on a silver halide emulsion [170]. Upon exposure, the silver halide grains are converted into metallic silver at a rate proportional to the incident illumination. Following the exposure, the hologram undergoes developing and fixation processes. The resulting amplitude transmission function is: $H \propto I$. Figure 3-16-b shows the recorded intensity distribution for the resolution target mask. The case of phase holograms will be discussed later.

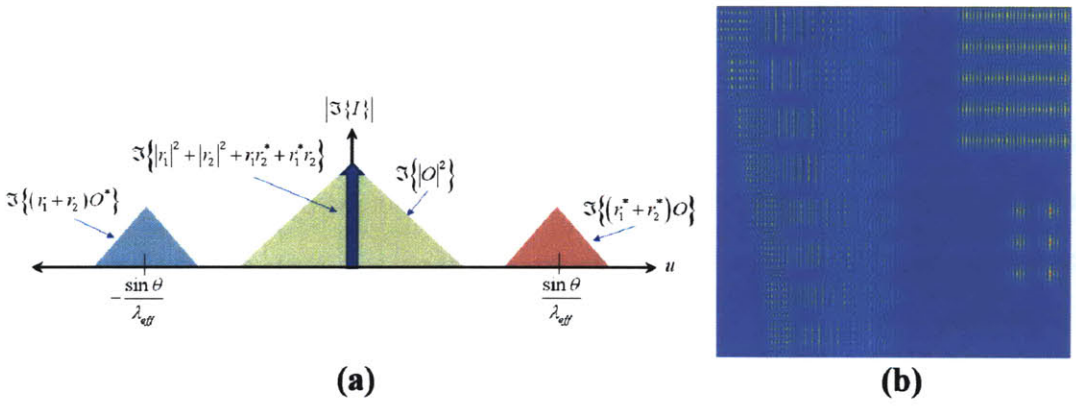


Figure 3-16: (a) Spectral representation of exposed intensity; (b) Recorded instensity distribution.

3.4.2 Optical Reconstruction Process

In the reconstruction step, the phase conjugate of the reference wave is used to illuminate the hologram. The diffracted field at the hologram plane is given by (for the amplitude hologram case),

$$R(x', y'; d) = (r_1^* + r_2^*)\, H \propto U_1 + U_2 + U_3 + U_4 + U_5 + U_6, \tag{3.29}$$

where,

$$\begin{aligned} U_1 &= r_1^* \left(|r_1|^2 + 2|r_2|^2 + |O|^2\right); \quad U_2 = r_2^* \left(2|r_1|^2 + |r_2^2| + |O|^2\right); \\ U_3 &= O\left[(r_1^* + r_2^*)^2\right]; \quad U_4 = r_2 r_1^{*2}; \\ U_5 &= r_1 r_2^{*2}; U_6 = O^* \left(|r_1|^2 + |r_2|^2 + r_1 r_2^* + r_1^* r_2\right). \end{aligned} \tag{3.30}$$

The effect of the probing wave is to shift the spectrum of the encoded signal (Figure 3-16-a) in a similar demodulation process as that of Figure 2-15. An ideal circular low-pass window with cut-off frequency, $\rho_{TIR} = \sin(\theta_c)/\lambda$, is applied to the shifted spectrum simulating the TIR process and eliminating the undesirable diffraction orders. Provided all the diffraction orders do not overlap in frequency domain, the filtering process is equivalent to,

$$r(x', y'; d) = R(x', y'; d) - U_1 + U_2 + U_3 + U_4 + U_5 = \left[2 + 2\cos\left(\frac{4\pi}{\lambda_{eff}\sqrt{2}} d\right)\right] O^*. \tag{3.31}$$

The resulting field of equation 3.31 is the reconstructed signal that suffers frustrated TIR and propagates towards the photoresist plane. The output intensity distribution is given by,

$$I_{photo} = |r(x', y'; d) * h(x', y'; d)|^2 = |r(x, y; 0)|^2. \tag{3.32}$$

Figure 3-17 shows the reconstructed intensity distribution at the photoresist plane. This simulation was processed using the BS method described previously. The reconstructed intensity distribution exposes the photoresist and this process is simulated using the methods described in the previous chapter.

3.4.3 Modeling Material Response

The previous model for the recording and reconstruction processes assumes ideal experimental conditions and does not consider the response of the holographic material. In addition, the analysis is restricted for the case of amplitude holograms. We now extend

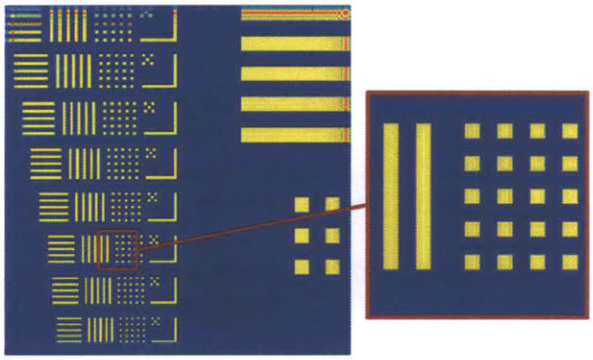


Figure 3-17: Reconstructed intensity distribution at the photoresist plane.

the model for the case of phase holograms, as well as include the nonlinear response of the holographic material.

Phase holograms can be produced by directly modulating the refractive index of the holographic material or by an additional bleaching step that produces a surface relief hologram. For silver halide emulsions, bleaching returns metallic silver grains back to a transparent silver halide compound. The variation of concentration of silver grains produces variations in refractive index and a surface relief. The transmission function of a thin hologram is given by,

$$t = \exp(-\alpha p) \exp\left(i\frac{2\pi}{\lambda} pn\right) = |t| \exp(i\phi), \tag{3.33}$$

where p and n are the emulsion's thickness and refractive index. For a lossless hologram: $\alpha = 0 \rightarrow |t| \approx 1$. As indicated in the Hurter-Driffield curve (Figure 2-36), the phase distribution (in the form of index modulation or thickness variation) is proportional to the logarithm of the exposed energy when operating in the linear regime: $\phi \propto \log E$, where $E = I \cdot T$ and T is the exposure time. The exposed intensity distribution of

equation 3.28 can be rewritten as,

$$\begin{aligned} I &= |r_1 + r_2 + O|^2 = |R + O|^2 \\ &= \left(|R|^2 + |O|^2\right)\left[1 + \frac{2\,|R|\,|O|}{|R|^2 + |O|^2}\cos\left(\phi_O - \phi_R\right)\right], \end{aligned} \tag{3.34}$$

where ϕ_O and ϕ_R are the phases of the object and reference waves respectively. The modulation of the emulsion's refractive index is proportional to the exposed intensity distribution and is given by,

$$n(x, y) = n_0 + \Delta n(x, y), \tag{3.35}$$

where n_0 is the average refractive index. Under the weak reference condition ($|R| << |O|$) [171], the phase hologram's transmission function reduces to,

$$\begin{aligned} t &= \exp\left[ikp\left(n_0 + \Delta n_R + \Delta n_O\right)\right]\exp\left[i2kp\sqrt{\Delta n_R \Delta n_O}\cos\left(\phi_O - \phi_R\right)\right] \\ &= \exp\left[ikp\left(n_0 + \Delta n_R + \Delta n_O\right)\right]\sum_{q=-\infty}^{\infty}(i)^q J_q\left(2kp\sqrt{\Delta n_R \Delta n_O}\right)\exp\left[iq\left(\phi_O - \phi_R\right)\right], \end{aligned} \tag{3.36}$$

where $k = 2\pi/\lambda$; Δn_O and Δn_R are the refractive index modulations corresponding to the exposures with $|O|^2$ and $|R|^2$; and J_q is Bessel function of the first kind, order qth. The multiple diffraction orders present is characteristic of phase holograms. We are interested to recover the -1 order as all the other terms suffer TIR. The reconstructed diffraction order at the hologram plane is,

$$t_{-1} \approx \exp\left[ikp\left(n_0 + \Delta n_R + \Delta n_O\right)\right]\left(2kp\sqrt{\Delta n_R \Delta n_O}\right)\exp(-i\phi_O). \tag{3.37}$$

The following example illustrates how the reconstructed pattern at the photoresist plane is affected by the holographic material's response. Consider a hologram recorded on a Dupont photopolymer: OmniDex613 [172]. Figure 3-18 shows the material's photo-response. Figure 3-19-a shows the material's response curve for the following simulation

parameters: $T = 100\text{sec}$, $P_{tot} = 0.240\text{W/cm}^2$, $n_0 = 1.5$, and $d = 200\mu\text{m}$. The reconstructed intensity distribution is shown in Figure 3-19-b.

If we set $T = 400\text{sec}$ and all other parameters remain the same, we find the material's response curve shown in Figure 3-20-a. The corresponding reconstructed intensity distribution is shown in Figure 3-20-b.

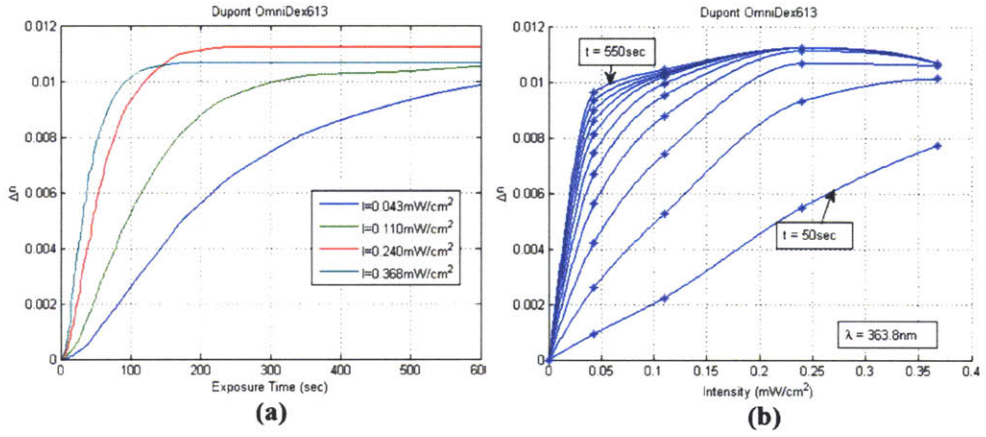


Figure 3-18: OmniDex613 photo-response: (a) Exposure time; (b) Intensity.

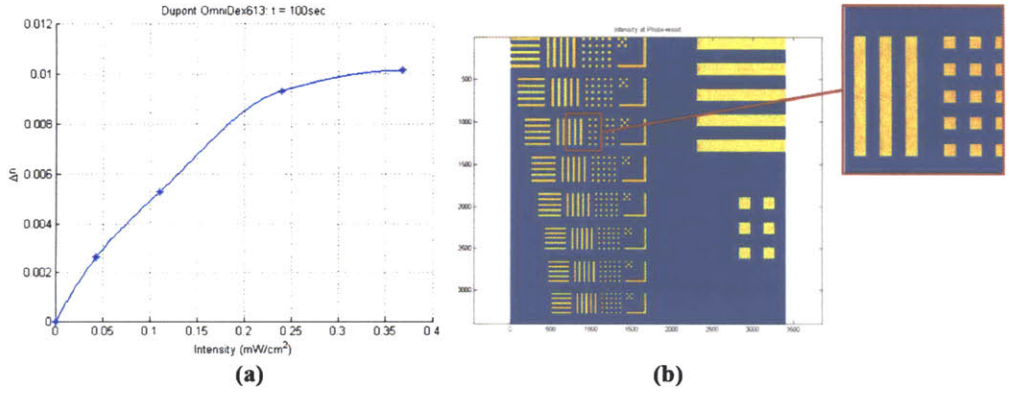


Figure 3-19: (a) Material response curve; (b) Reconstructed intensity distribution.

The second aspect in modeling the material response is to analyze the effects of shrinkage of the holographic emulsion after the recording step and additional post processes.

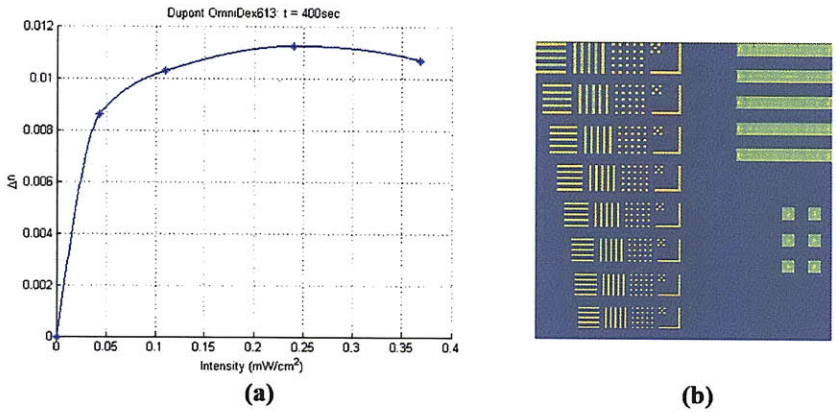


Figure 3-20: (a) Material response curve; (b) Reconstructed intensity distribution.

This analysis departs from the assumption of thin holograms, as shrinkage modifies the modulated refractive index within the material's volume introducing errors in the reconstruction. The presented shrinkage model can be applied to a variety of recording materials; however, specific examples for Dupont photopolymers will be presented.

The properties of Dupont photopolymers have been studied theoretically and experimentally by several research groups [173]. These photopolymers are mainly composed of polymeric binders, free monomers, photoinitiators, and sensitizing dyes [174]. Upon exposure, the activated dye in the material photoinitiates polymerization and the monomer molecules start moving toward regions of higher polymer concentrations in a diffusion-based process. To model this process, local and non-local diffusion models have been developed [175], [176]. In the diffusion process, monomers are polymerized forming polymer chains that grow away from the point of origin, removing active monomers as they grow and smearing the exposed pattern. This smearing limits the highest spatial frequency recordable on a given photopolymer. For Dupont photopolymers, polymer chains of around 50-110nm have been predicted [175]. The main advantages of Dupont photopolymers are: relatively large index modulation ($\Delta n \approx 1 \times 10^{-2}$), dry processing, long lifetime, good photo-speed, wide spectral sensitivity, high-resolution and low cost. To study the dynamics of the recording process, real-time monitoring systems based on

tracking the Bragg diffraction angle have been implemented [174], [177]. The shrinkage of the holographic emulsion is mainly attributed to the recording and fixation processes. For the recording process, it was found experimentally that the DC term is the dominant factor contributing to the material's shrinkage [177]. The fixation process consists of exposing the recorded hologram with a uniform high intensity light to polymerize the residual monomers. This process also causes shrinkage due to an increase of volume density in the material. It was found experimentally that post baking the hologram after the fixation process can reduce the effect of shrinkage due to the irreversible thermal expansion property, the decrease of the average refractive index and the loss of volatile components of the recording material [174], [178].

To model the effects of shrinkage of the holographic material, a model based on the weak diffraction approximation (1st-order Born approximation) is implemented [179]. In this model, multiple diffractions are neglected and the reconstruction probing wave is assumed to be unaffected while propagating through the hologram (in practice, there is a depth-dependent absorption of this wave). The total response is given by the superposition of all the contributing point sources. In addition, only uniform shrinkage along the axial direction is assumed. From experimental measurements, shrinkage factors of around 1-2% have been estimated for the Dupont photopolymer: OmniDex 613. The implemented shrinkage model is shown in Figure 3-21. In this model, the recorded phase distributions before shrinkage at several parallel planes separated by a distance, Δz, within the hologram's volume are calculated using the method described above (only two planes are shown). The computed planes are then shifted axially by a shrinkage factor, Δt. The resulting hologram is reconstructed and the fields diffracted by each plane are added coherently at the photoresist plane and the output intensity distribution is computed. Due to shrinkage, the optical path traversed by each plane will be different than that before shrinkage.

Figure 3-22 shows an example of the reconstructed intensity distributions before and after shrinkage for a hologram recorded on an OmniDex 613 photopolymer with the fol-

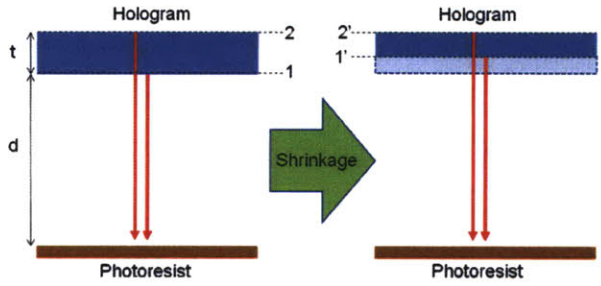


Figure 3-21: Geometry of shrinkage model.

lowing simulation parameters: $\lambda = 350\text{nm}$, $d = 200\mu\text{m}$, $n_0 = 1.5$, $I_{tot} = 0.240\text{mW/cm}^2$, $T = 100\text{sec}$ and 10% shrinkage.

3.4.4 Extension of the Depth of Focus

Similar to the previous chapter, extending the system's depth of focus (DOF) is critical to tolerate potential axial misalignment of the substrate to be exposed. The theoretical DOF is given by equation 2.73. In this section, we propose two methods to extend the DOF of TIR holographic systems: 4f relay system and finite beam scanning. In the first method, a 4f system is placed between the mask and the hologram and is used to relay the field diffracted by the mask to the hologram plane during the recording step. A variable circular aperture is placed at the Fourier plane and is used to limit the frequency extent of the signal. This results in a reduction of the effective numerical aperture (NA) and thus an increase in DOF at the expense of some resolution loss. In the second method, the reference wave and mask are illuminated by a finite size beam which is raster scanned to sequentially expose the entire hologram. The finite beam size also reduces the effective NA extending the system's DOF.

For the resolution target example presented in the previous sections, the simulation parameters are: $\lambda = 350\text{nm}$, $H_{size} = 400\mu\text{m}$, $d = 200\mu\text{m}$, and $\delta_{pix} = 100\text{nm}$. The corresponding Nyquist and evanescent cut-offs (equations 2.51 and 2.5) are: $u_{Nyq} = u_{ev} =$

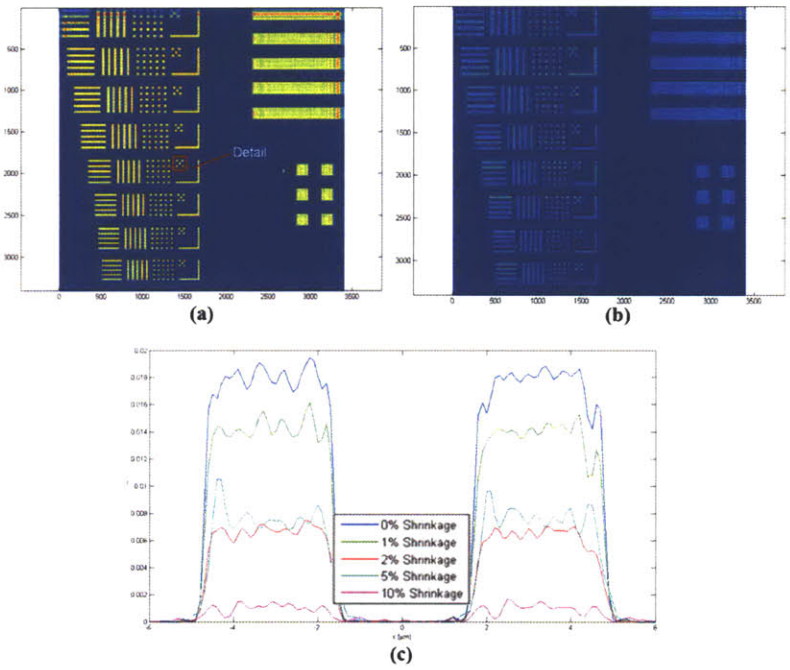


Figure 3-22: (a) Reconstructed intensity with no shrinkage; (b) Reconstructed intensity with 10% shrinkage; (c) Intensity cross-sections.

$2,857.2\text{mm}^{-2}$. The effective NA (equation 2.2) is: $NA_{eff} = 0.7071$. The diffraction limit resolution and theoretical DOF are: $\Lambda = 247.5\text{nm}$, and $DOF = \pm 350\text{nm}$. In order to extend the DOF, the effective NA of the system is reduced to: $NA_{eff} = 0.21$. This corresponds to a cut-off frequency of the low-pass filter (aperture stop in the 4f system) of: $u_{cutoff} = 600\text{mm}^{-1}$. The diffraction limit resolution and extended DOF are: $\Lambda = 1.02\mu\text{m}$ (feasible for LCD panel manufacture), and $DOF = \pm 3.97\mu\text{m}$.

Figure 3-23 shows the normalized intensity at the center of the PSF as a function of axial distance for the system before and after extending the DOF. The DOF is estimated as the region where more than 80% of the irradiance is contained. Figure 3-23-c show an example of the reconstructed defocused image (at $d = 203\mu\text{m}$) for the regular and extended DOF holograms, using the same simulation parameters as those from the previous example. The holograms are designed to reconstruct the desired resolution pattern

at a working distance: $d = 200\mu$m.

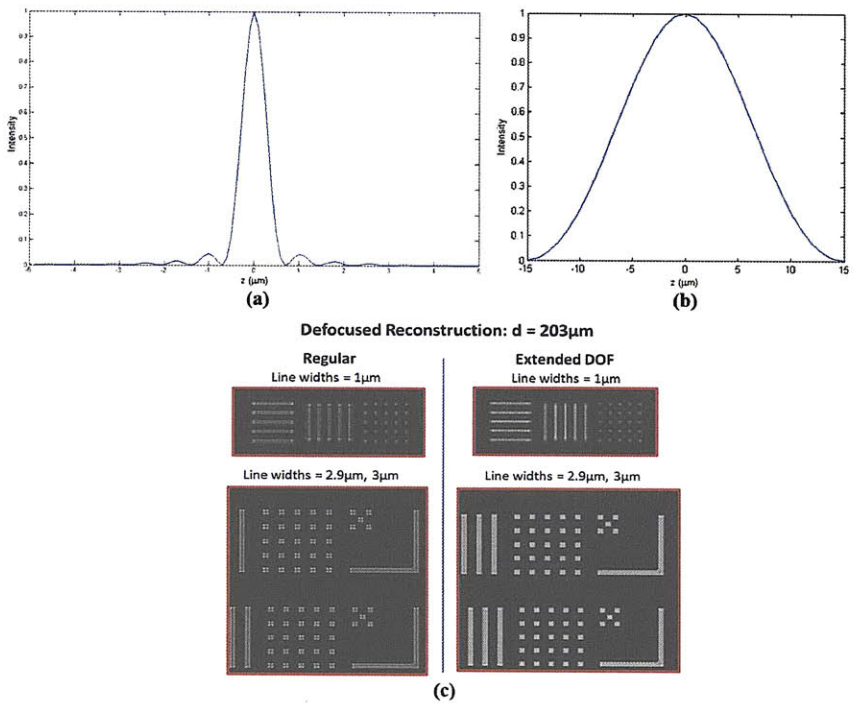


Figure 3-23: Normalized intensity PSF's center: (a) Before extension; (b) After Extension; (c) Reconstructed intensity before and after extending DOF.

3.4.5 Optimizing the Mask Design

The presented algorithm allows simulating the recording process and predicting the optical reconstruction under a variety of optical, geometrical, material and experimental conditions. The corresponding controlling parameters can be optimized to produce high quality reconstruction patterns. An additional degree of freedom is added by redesigning the mask in a way that allows for compensating potential reconstruction errors. This can be particularly useful in systems with low effective NA such as aperture limited systems or those designed to extend the DOF. The reconstructions from this type of system are characterized by low-pass filtering the desired mask. When a line pattern in the mask

gets low-pass filtered, the sharp edges become blurry and the reconstructed line becomes narrower. This is shown in Figure 3-24-a.

To solve this problem, the mask design is modified by the introduction of dilations, serifs and mousebites. This optical proximity correction technique has been widely used to enhance the resolution in conventional optical lithographic systems [180], [181]. The amount of dilation, location and size of the serifs and mousebites is found by an optimization algorithm with an error metric given by the MSE of the difference between the desired and reconstructed intensity distributions at the photoresist plane. Figure 3-24-b shows the optimized mask design for a single line pattern.

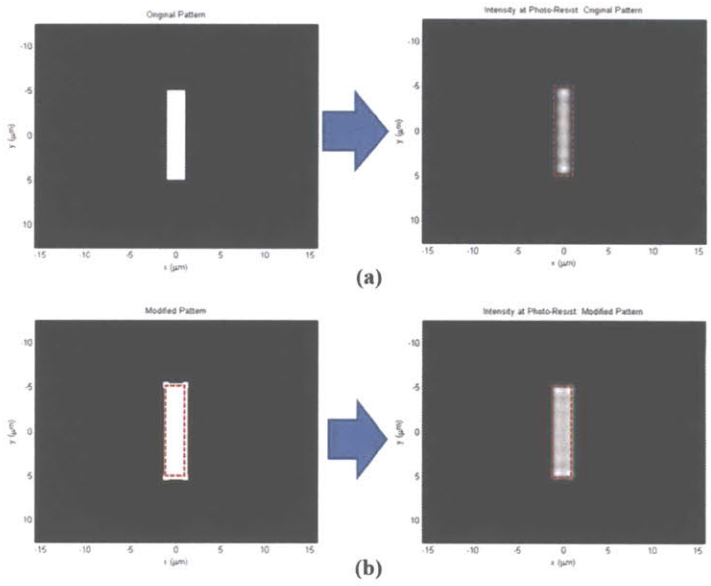


Figure 3-24: (a) Low-pass filtering of line pattern; (b) Mask correction process.

3.5 Experimental Validation

A graphical unit interface (GUI) is designed for the flexible implementation of the presented simulation algorithm. The GUI is used for the design and optimization of TIR

holographic lithographic systems for LCD panel manufacture. A screen shot of the GUI is shown in Figure 3-25.

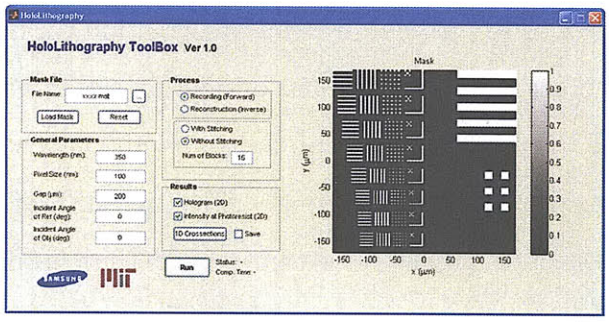


Figure 3-25: HoloLithography ToolBox.

Experiments conducted at the Mechatronics Center of Samsung Electronics Corporation in Suwon, South Korea, are used to validate the accuracy of the presented simulation algorithm. Figure 3-26-a shows a section of the desired gate pattern mask. Figure 3-26-b shows a close up section of the simulated intensity distribution at the hologram plane. Figure 3-26-c shows the corresponding microscopic image of the real fabricated surface relief hologram. Figure 3-26-d shows a microscopic image of the pattern exposed onto a photoresist after developing and etching. It is found that a critical factor in the design of TIR surface relief holograms is to avoid nonlinear shape distortion of the resulting grating [182]. These distortions can be avoided by designing holograms with small aspect ratios of width versus depth.

3.5.1 Comparison of Holographic Lithographic Methods

We now compare the holographic lithographic systems based on CGHs presented in the previous chapter and optically recorded TIR holograms presented in this chapter. The comparison is performed against the in-line CGH geometry. Similar arguments can be applied to the off-axis and TIR geometries. The main advantages and disadvantages of both systems are described in Table 3.3.

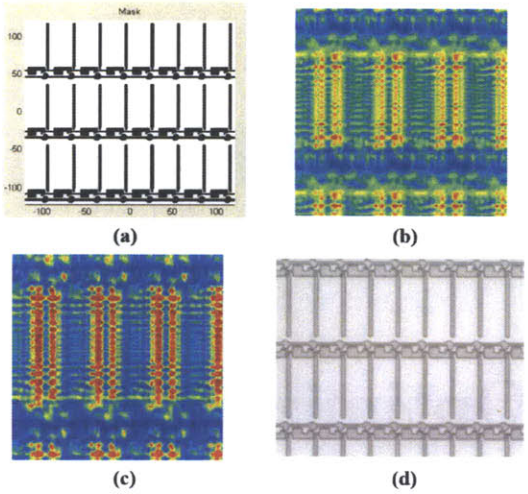


Figure 3-26: (a) Desired gate pattern mask; (b) Simulated intensity at hologram plane; (c) Image of fabricated hologram; (d) Image of exposed pattern.

Table 3.3: Comparison of CGH and TIR Holographic Lithographic Systems.

Conventional TIR System	**In-line CGH System**
Optical recording	Numerical design
Optical reconstruction	Optical reconstruction
Difficult recording: sensitive to vibrations, temperature, etc.	Simple fabrication process using e-beam lithography
Material problems (photopolymers): shrinkage and lifespan	HSQ: robust material, no degradation and long lifespan
Complex reconstruction system: large prism and robust motion stage	Simple optical set up
Undesirable diffraction orders suffer TIR	Optimization algorithm need to eliminate undesirable orders
Fixed encoding strategy	Flexible encoding strategy

Chapter 4

Design, Optimization and Implementation of High-Resolution Segmented Aperture Thin Imager Based on Holographically Corrected GRIN Lenses

The optimization problems discussed in the previous chapters are formulated to maximize the information transfer between object and hologram spaces during the signal encoding and decoding processes. Several geometrical, optical, material and fabrication related parameters are used to control the system's performance in a multi-domain optimization approach. For the case of holographic lithography based on CGHs, it is found that the choices of encoding strategy and optimization algorithms when designing a hologram subject to stringent constraints are critical to produce high-quality reconstruction patterns at the photoresist plane. In the case of lithographic systems based on conventional TIR holograms, the encoding strategy is fixed by the optical recording process and the optimization is done by controlling the corresponding optical (wavelength, exposure con-

ditions), geometrical (working distance, numerical aperture), and material related (photo response, post-processing) parameters. In both cases, the signal decoding is performed by an optical reconstruction process. This process can also be optimized by implementing techniques such as the extension of the system's depth of focus. In addition, a sensitivity analysis allowed us to actively predict and assist in the correction of potential fabrication errors. Finally, both systems presented previously utilize holograms as active components in the system exploiting their ability to encode complex 2D or 3D information efficiently on a 2D plane.

In this chapter, another optimization example designed to maximize the information transfer between object and detector spaces is presented - but this time it is applied to imaging. A novel segmented aperture thin imager based on holographically corrected gradient refractive index (GRIN) lenses is proposed and studied. This non-conventional imager is designed to be high-resolution and have a maximum thickness of 5mm. The system is modeled using system's theory and simulated by a combination of Matlab and Zemax [183], [184] interfaces. An analysis of the system is conducted using information theory. A multi-domain optimization is performed using GAs. An example of decoding strategy is presented based on the superresolution method for post-processing of the captured image. Experimental results for the optimization of the hologram's recording process as well as the measurement and characterization of the system's point-spread function are presented. A sensitivity analysis is performed to estimate the effect of potential misalignment errors in the assembly process. A modification of the system's geometry is proposed for polychromatic imaging.

4.1 Motivation and Problem Definition

The rapid evolution of electronics witnessed over the past few decades has drastically changed the design of imaging systems [185]. In particular, the introduction of digital photodetectors and numerical post-processing methods have allowed the development of

non-conventional imaging systems designed to extract scene information more efficiently than conventional systems. In a similar way, the revolution in mobile technology and the increasing trend to develop smaller pixel-size photodetectors has led to the development of small-size imaging systems. Thin imaging systems have been a popular area of research due to their wide range of potential applications. Emerging military applications envision large-scale deployment of compact imaging systems mounted on a variety of platforms such as soldier's helmets, ground vehicles and unmanned aerial vehicles (UAVs) for navigation, guidance, target localization and recognition. These systems are required to be low power, lightweight, high-resolution and with a large field-of-view, as well as overcome stringent challenges in cost and fabrication volume. To this end, the Defense Advanced Research Projects Agency (DARPA) initiated a program called "Multiple Optical Non-redundant Aperture Generalized Sensors (MONTAGE)" [186] under which the presented work was developed. In the medical industry, minimally invasive procedures require the implementation of small-size imagers in devices such as endoscopes. Mobile devices such as cell phones, portable computers and video cameras also rely on these types of imagers with a market growing at a pronounced rate. Compact, low-power sensors are also required for collecting sensitive oceanographic data in a variety of platforms such as drifters. Finally, more sophisticated surveillance systems have currently been developed.

The design of high-resolution thin imaging systems is challenging and requires the development of non-conventional optical and numerical methods. To illustrate the problem, consider a conventional imaging system based on a single lens. From the imaging condition we find that an object located at a distance, s_o, is imaged by the lens at a distance: $s_i = s_o f/(s_o - f)$, where f is the lens' focal length. The minimum imaging distance occurs when $s_o \to \infty$ and $s_i \to f$. To make the system thin, the focal length is reduced while maintaining constant f/ #. Because the f/# is being held constant, a reduction in focal length is followed by a reduction in aperture diameter. This results in a loss of photon count that in many cases will produce an unacceptable loss of image quality. The collected optical power is a quadratic function of the focal length and in

the presence of additive white Gaussian detector noise the corresponding signal-to-noise ratio is: $SNR \propto 1/f^2$. In addition, the system would suffer a loss of resolution and field-of-view. If the f/# is allowed to vary, a decrease in focal length would lead to complex curved surfaces, physical constraints, optical aberrations and cost-related limitations. In order to reduce the resulting optical aberrations a multi-element design would normally be required, increasing the size of the imaging system. A similar argument applies when using Fourier optics. In this case, diffraction also plays a role in the degradation of the captured image in aperture-limited systems. In addition, the limited space-bandwidth product and pixel-size of current CCD and CMOS sensors limits the sampling of the captured image.

Several methods have been proposed for the design of non-conventional thin imaging systems. Examples of such systems include: thin folded imagers [187], [188], systems based on compressive sensing applied to visible and infrared imaging [189], [190], task-specific imaging systems [191], and compound systems using planar, spherical and rotational based geometries [192], [193], [194]. In addition, methods that rely on alternate projections of the collected field have been investigated such as wavefront coded (using, for example, a cubic phase mask) [195] and coded aperture systems [196].

Compound imaging systems have been demonstrated to be a good candidate for ultra-thin imaging configurations. They are segmented aperture systems typically composed of a lenslet or pinhole array and a digital detector. Compound imagers try to mimic natural apposition compound eyes found in, for example, flies or moths, where an array of lenses is used to capture the incident light and each lens has a small number of associated photoreceptors forming a channel called "ommatidium". For small invertebrates that have external skeletons, eyes are very expensive in terms of weight and consumption of metabolic energy. Compound eyes simplify their visual system in comparison to single refractive lens based eyes and provide a larger field-of-view at the cost of comparatively lower spatial resolution. The concept of compound imaging was originally introduced by Lippmann in 1908 [197]. An example of an artificial compound imaging system is the

"thin observation module by bound optics (TOMBO)" [198]. This system is composed of a microlens array, signal separator (to prevent cross-talk from adjacent lenses) and a digital detector. Each lens produces a low-resolution elementary image of the scene. The captured elementary images are later combined to retrieve a high-resolution image by means of a post-processing algorithm. Several reconstruction algorithms have been proposed, such as the sampling method [198], pixel rearrange method [199] and back projection method [200]. A variation of the pixel rearrange method known as the superresolution algorithm will be presented later. Additional applications of compound systems include the measurement of 4D light fields [201], [202], Shack-Hartmann sensors [203] and task-specific imaging systems [204].

Thin segmented aperture systems based on refractive microlens arrays suffer from severe optical aberrations due to the short focal length requirement. Refractive lenses bend the light at the dielectric-air interfaces so a short focal length (increase in optical power) requires highly curved lens surfaces that are difficult to make. Aberrations and diffraction artifacts degrade the quality of the elementary images and fail to extract useful information from the scene. This results in low quality reconstructed images after the post-processing algorithm. Moreover, a proper optical characterization in terms of, for example, the system's point-spread function (PSF) or modulated transfer function (MTF), has not been conducted on previous work for compound systems such as TOMBO. No methods for the correction of such aberrations applied to thin segmented aperture systems have been proposed.

4.2 Description of Proposed System

In this thesis, a novel segmented aperture thin imager based on holographically corrected GRIN lenses is proposed. This system is high-resolution and has a relatively large field-of-view. In contrast to conventional systems based on refractive lenses, the refractive index of GRIN lenses is modulated quadratically allowing them to utilize their entire volume to

smoothly guide the incident light to the focal plane. As a result, shorter focal lengths can be achieved with a better optical performance (lower aberrations). To reduce the cost of the system, commercially available GRIN lenses are utilized. These lenses are optimized for on-axis illumination, requiring an additional mechanism for the correction of optical aberrations. In the proposed system, the GRIN lenses are corrected by a specially designed holographic element. Figure 4-1 shows the geometry of the proposed system. The light scattered by the scene is collected by the hologram-GRIN lens array combo and is focused onto a photodetector such as a CCD or CMOS sensor. The captured frame is sent to the computer for image post-processing. Each lens in the array forms a low resolution elementary image of the scene. The aberration correction for each lens is performed for different field angles (correction points in the image plane) as shown in Figure 4-2. This alternate projection scheme allows recording of additional information from the scene that otherwise would be lost, as will be proven using an information theoretic analysis. The location of the correction points introduces a new degree of freedom that can be exploited to optimize the system to extract scene information much more efficiently than can be achieved using conventional isomorphic methods. This optimization is performed using a multi-domain approach based on genetic algorithms with a fitness function given by the system's channel capacity. The encoding process is then defined by the acquisition of a frame composed of N elementary images. The proposed encoding process is proven to be more efficient than that of conventional systems such as TOMBO. A reconstruction algorithm is used to combine the captured information from all the elementary images and produce a restored high-resolution image. Several decoding strategies can be implemented. In the presented work, the superresolution algorithm is used.

The aberration correction process is conducted using phase conjugation holography (PCH). The holograms used in PCH are recorded and reconstructed optically. The recording geometry for PCH is shown in Figure 4-3-a. A diffraction limited point source is generated at the front focal plane of the GRIN lens using, for example, a pinhole or microscope objective. An ideal GRIN lens would produce a collimated output plane

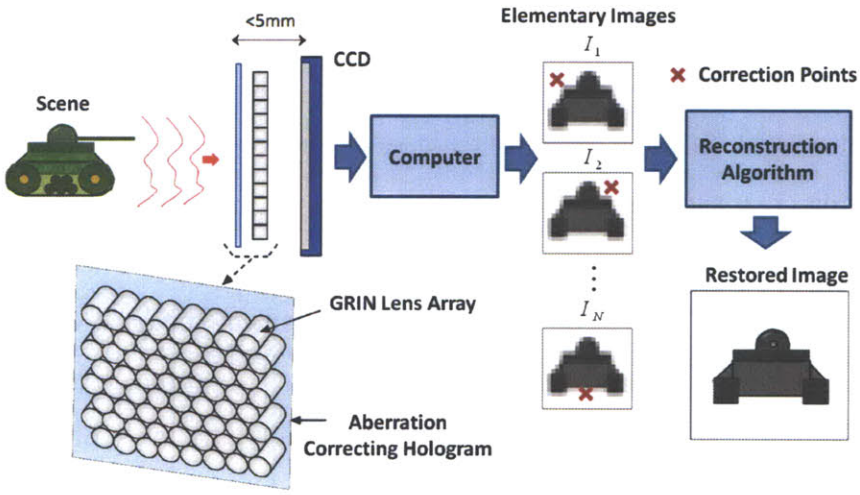


Figure 4-1: Geometry of holographically corrected segmented aperture thin imager.

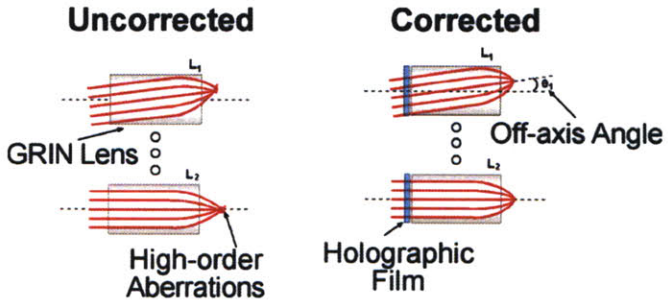


Figure 4-2: Holographic correction of GRIN lens array for different field angles.

wave. Instead, the aberrated GRIN modifies the wavefront of the output field generating a phase distorted signal wave. A beam splitter is used for coupling the plane reference wave in transmission geometry. A photosensitive holographic material is used to record the interference pattern produced by the superposition of the signal and reference waves. The hologram then undergoes the standard development, bleaching and fixation steps and is placed back into the system for reconstruction. The reconstruction geometry is shown in Figure 4-3-b. An incident plane wave probes the hologram reconstructing the phase conjugate of the signal wave that propagates back through the GRIN lens, cancelling the intrinsic aberrations and producing a corrected point source at the camera's back focal plane. To reduce the thickness of the system, the beam splitter is removed and the recorded hologram is optically glued to the front surface of the GRIN lens. The effects of removing the beam splitter on the system's PSF are analyzed by means of a sensitivity analysis.

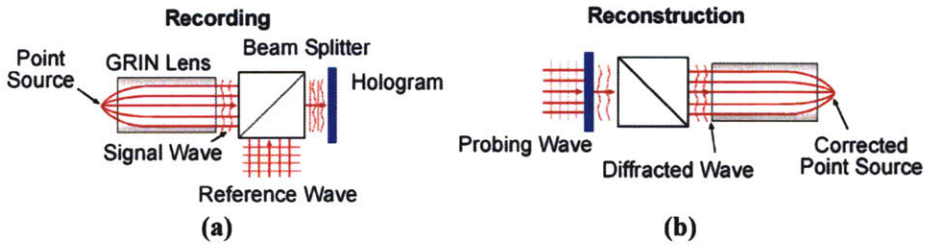


Figure 4-3: Phase conjugation holographic process: (a) Recording; (b) Reconstruction.

The PCH aberration correction method shown in Figure 4-3 is performed for a single field angle (point in the image plane) per lens. Each lens in the array is corrected for a different field angle by laterally translating the point source and rotating the reference wave to the corresponding angle during the hologram recording step. An alternative technique, not explored in this work, involves recording multiple holograms by means of a multiplexing technique. During the image acquisition process, the light scattered by an object located in the far-field arrives at the nth Hologram-GRIN lens combo as a fan of plane waves propagating in different directions with each direction corresponding to a

point on the object. Plane waves with angles close to the correction angle are efficiently focused by the system forming a close to diffraction limit spot at the image plane. Only the plane wave travelling at the same angle as that used during the recording process will produce a diffraction limited spot. Plane waves propagating at angles significantly different to the correction angle produce aberrated spots at the image plane. In other words, each elementary image has an aberration-free patch centered at a particular location which is different from that of the other lenses. The correction positions are found by the optimization algorithm to maximize the information exchange from object to detector planes.

Correction of aberrations by means of PCH dates back to 1966 with the work of Leith and Upatnieks [205]. Since then, PCH has been extensively used in applications such as aberration correction in telescopes [206], [207], [208], microscopes [209], [210], refractive lenses and zone plates [211] and imaging through diffusing media [212], [213].

An alternative geometry is proposed to increase the system's field-of-view and is shown in Figure 4-4. In this geometry all the lenses in the array are corrected for an on-axis point source but the incidence angle of the reference wave is varied. The off-axis field coming from the scene is rotated by a holographic element, allowing it to be imaged by the lens and hence increasing the system's field-of-view. A hybrid combination of the two proposed geometries (Figures 4-2 and 4-3) may be implemented. The holographic element can also be replaced by a CGH designed to correct the optical aberrations for different points on the image plane. In the following analysis, we will concentrate on the system geometry of Figure 4-2.

4.3 Optical Performance of Uncorrected and Corrected GRIN Lenses

We begin our analysis by studying the optical performance of uncorrected and corrected GRIN lenses using the PCH method described above. The evaluation is done using the

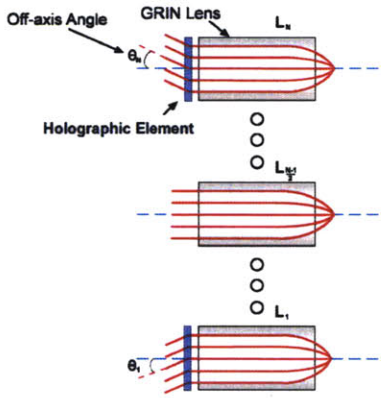


Figure 4-4: Geometry for increasing the system's field-of-view.

optical design software Zemax. The specification parameters of the selected GRIN lens are indicated in Table 4.1. This lens is chosen such that the total thickness of the optical system is less than 5mm. In addition, the pitch of the lens allows the back focal plane to be away from the lens surface at a working distance (back focal length), d. This helps bypassing the glass cover that protects the photodetector's pixels, ensuring a sharp focus of the incoming plane waves. The imager is designed for infinite conjugates to avoid complex focusing mechanisms. The GRIN lens is made out of SELFOC® material provided by NSG America, Inc. [214]. These GRIN lenses are manufactured using a high-temperature ion exchange process within the glass and the host material that results in a controlled variation of the refractive index. However, this fabrication process introduces variations on the index gradient constant up to $\pm 3\%$ between ion exchange of different batches and $\pm 1\%$ for the same batch that modifies the optical performance of each lens. The PCH approach allows correcting for these aberrations independently for each lens.

The GRIN lens has a radial parabolic modulation of its refractive index given by,

$$n(r) = n_0 \left(1 - \frac{A}{2} r^2\right), \tag{4.1}$$

Table 4.1: Specification Parameters of Selected GRIN Lens.

Company	Newport	**Transmission**	$\geq 89\%$, $320 - 2000$nm
Model	LGI630-1	**Clear Aperture**	70%
Operation Wavelength (λ)	630nm	**Coating**	Single-layer MgF_2 AR
Pitch (p)	0.23	**Material**	SELFOC
NA	0.46	**Diameter** (ϕ)	1.8mm
Working Distance (d)	0.21mm	**Length** (L)	4.26mm

where n_0 is the refractive index at the center of the lens ($n_0 = 1.6075$), and A is the gradient index constant given by,

$$\sqrt{A} = \frac{2\pi p}{L}. \tag{4.2}$$

For the selected lens: $A = 0.1148$.The modulated index of refraction is shown in Figure 4-5 (for $\lambda = 630$nm).

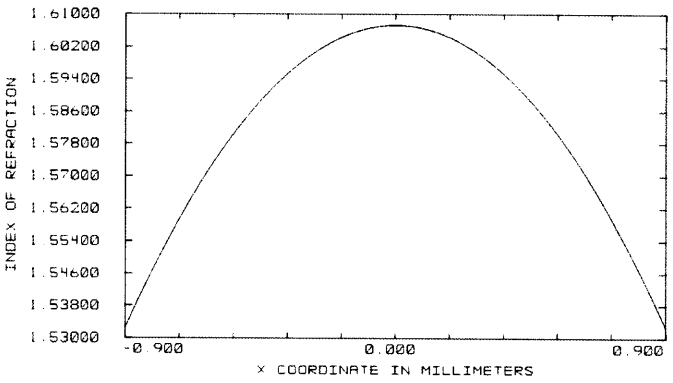


Figure 4-5: Refractive index profile of GRIN lens: LGI630-1.

Inside the GRIN lens, a given ray follows a sinusoidal trajectory with period: $P = 2\pi/\sqrt{A} = L/p$. To find the unaberrated back focal length or working distance (from the last lens surface to the focal plane), we consider an input marginal ray propagating parallel to the optical axis and evaluate its trajectory inside the lens and outside towards the focal plane. The resulting unaberrated working distance (under the paraxial

approximation) is given by,

$$d_{ideal} = \frac{1}{n_0\sqrt{A}} \cot\left(L\sqrt{A}\right). \tag{4.3}$$

Using equation 4.3, we find the position of the second principal plane and use it to compute the effective focal length,

$$EFL = \frac{1}{n_0\sqrt{A}\sin\left(L\sqrt{A}\right)}. \tag{4.4}$$

The corresponding location of the principal planes is given by,

$$l_{principal} = \frac{1}{n_0\sqrt{A}} \tan\left(\frac{L\sqrt{A}}{2}\right). \tag{4.5}$$

Figure 4-6 shows geometry of the GRIN lens. For the selected GRIN lens: $d_{ideal} = 0.232$mm, $EFL = 1.85$mm, and $l_{principal} = 1.618$mm.

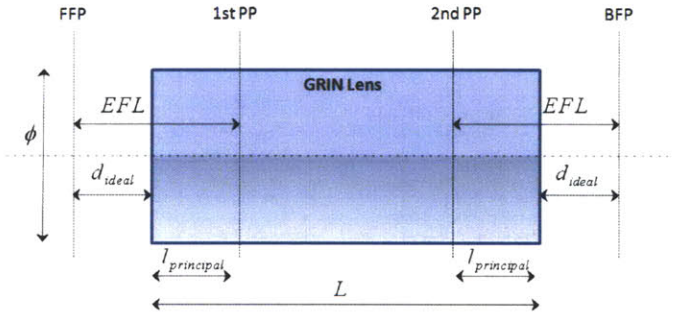


Figure 4-6: GRIN lens geometry.

This sinusoidal trajectory followed by a ray inside the lens also specifies the maximum acceptance angle of an incoming off-axis ray,

$$\theta_{\max} = \sin^{-1}\left(\frac{n_0\phi\sqrt{A}}{2}\right). \tag{4.6}$$

The corresponding field-of-view is: $FOV = 2\theta_{\max}$. For the selected lens: $FOV = 40.14$ degrees (only using 70% of the lens aperture).

Ray tracing is performed to evaluate the optical performance of the uncorrected GRIN lens using geometrical optics. Figure 4-7-a shows the ray tracing results for two field angles. The corresponding spot diagram is shown in Figure 4-7-b. As can be seen, the GRIN lens suffers from aberrations deviating from the perfect Gaussian image. To quantify these aberrations, the optical path difference (OPD) for the two fields along the meridional (tangential) and sagital planes is plotted, as shown in Figure 4-8-a. The OPD represents the difference between the optical path length of the ray under consideration and the chief ray in traveling from the point object to the reference Gaussian sphere at the exit pupil [215]. From the geometrical optics point of view, for an unaberrated system all the rays launched from the object traverse equal optical paths and arrive at the image plane forming a perfect spot. The corresponding wavefront map error (from the difference between the wavefront at the exit pupil and the ideal Gaussian reference sphere) is shown in Figure 4-8-b. A calculation of the Seidel aberration coefficients reveals that the dominant aberrations present in the system are astigmatism and distortion.

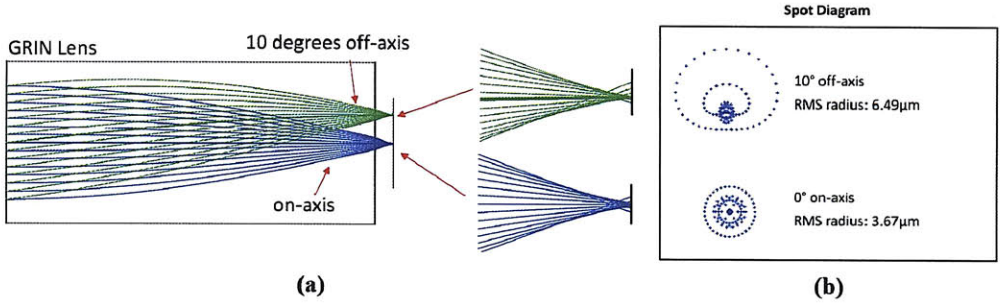


Figure 4-7: Uncorrected GRIN lens: (a) Ray tracing; (b) Spot diagram.

The effects of diffraction and evaluate the optical performance using the system's modulated transfer function (MTF) are now considered. The MTF of the uncorrected GRIN lens is shown in Figure 4-9 for the aberrated and diffraction limited cases. The

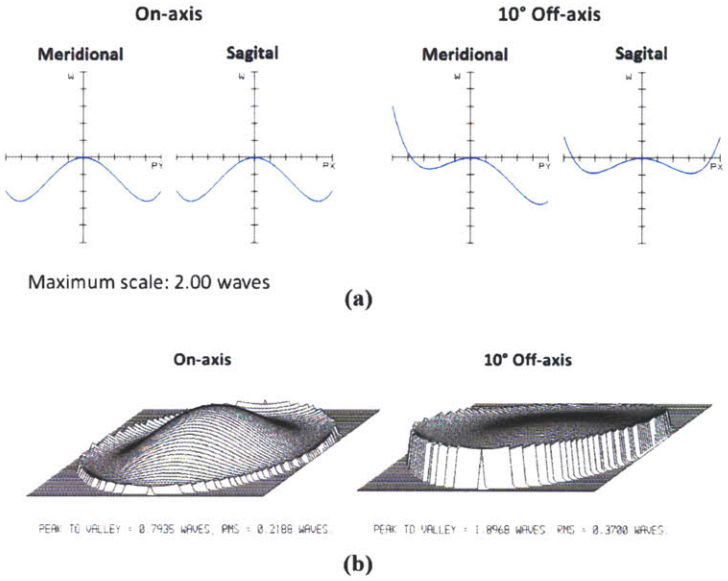


Figure 4-8: Uncorrected GRIN lens: (a) Optical path difference; (b) Wavefront map.

aberrated lens performs unsatisfactory for both on-axis and off-axis configurations, attenuating most high spatial frequencies.

The numerical model of the holographically corrected GRIN lens is performed using Zemax's optically fabricated hologram surface. A recording geometry is specified similar to that shown in Figure 4-3-a for a given correction point and field angle. A hologram with high diffraction efficiency is assumed in the simulations. The performance of the corrected lens is first evaluated using geometrical optics. Figure 4-10-a shows the ray tracing results for a GRIN lens corrected for an on-axis plane wave. The corresponding spot diagram is shown in Figure 4-10-b. According to geometrical optics, PCH results in perfect aberration correction for the designed field angle. These results can be verified from the optical path difference plots of Figure 4-11-a (for $\lambda = 658$nm). The MTF of the corrected GRIN lens is shown in Figure 4-11-b. This curve shows that a diffraction limited performance is achieved for the designed correction angle.

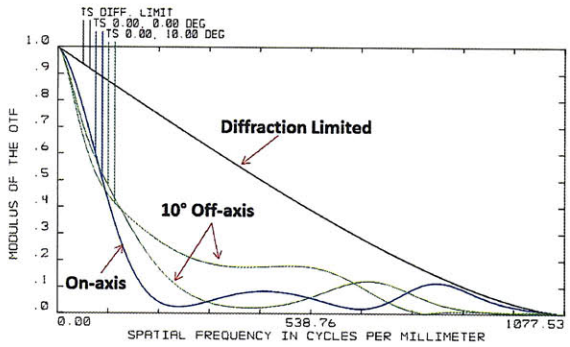


Figure 4-9: Modulated transfer function of uncorrected GRIN lens.

4.4 System Model

A model of the segmented aperture thin imager is used to characterize the image acquisition or encoding process. Using 1D linear operator formalism, the model of the system is given by,

$$\mathbf{g} = H\mathbf{f} + \mathbf{n}, \tag{4.7}$$

where $\mathbf{g}$ is a $M \times 1$ vector with elements given by raster scanning (or rearranging in lexicographic order) the measured intensity distribution; $\mathbf{f}$ is a $N \times 1$ vector corresponding to raster scanned scattered intensity distribution by the object (or scene) sampled at the desired resolution; H is a $M \times N$ tensor that functions as a linear operator also known as the Hopkins matrix; and $\mathbf{n}$ is a $M \times 1$ vector of additive white Gaussian noise (AWGN). The Hopkins matrix, H, includes the responses of both, the optical components (GRIN lens array and holographic film) and digital photodetector. Figure 4-12 shows the graphical interpretation of the system model for the case of a thin imager based on a 2×2 GRIN lens array. The object space is sampled into M discrete point sources. As the object space is located in the far-field, each object point source corresponds to a plane wave arriving at the camera with a different field angle. For the nth hologram-GRIN lens unit, the field before the photodetector is given by the superposition of the space-variant incoherent PSF responses. The resulting field is then subsampled by the

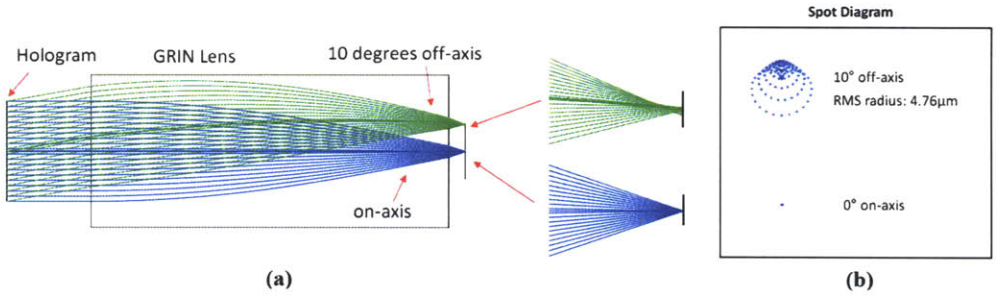


Figure 4-10: Corrected GRIN lens: (a) Ray tracing; (b) Spot diagram.

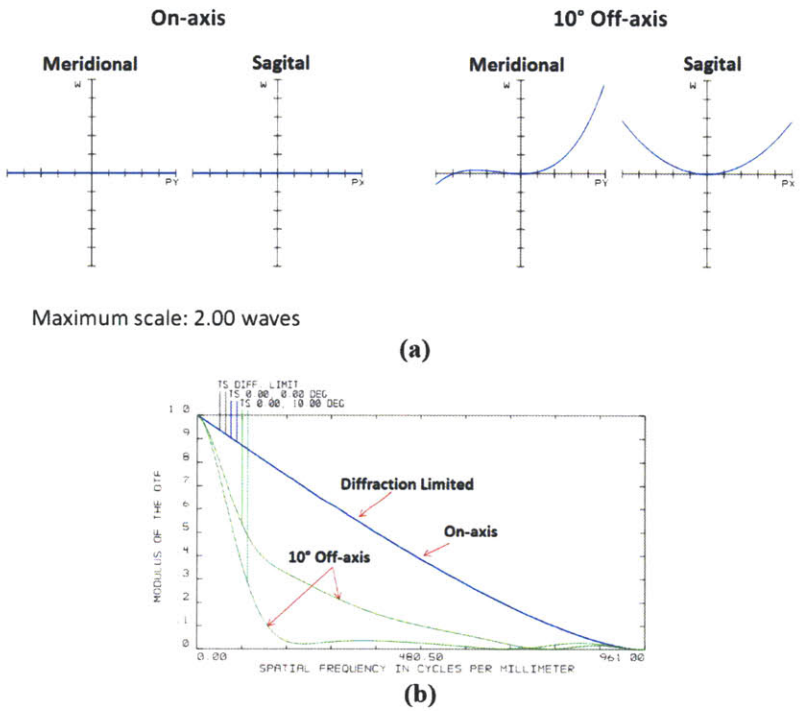


Figure 4-11: Corrected GRIN lens: (a) Optical path difference; (b) Modulated transfer function.

photodetector with a given space-bandwidth product. In the example shown, the output signal is composed of 2×2 low resolution images of the object space, each one with a different modulation (holographic correction), subpixel shift, and added noise.

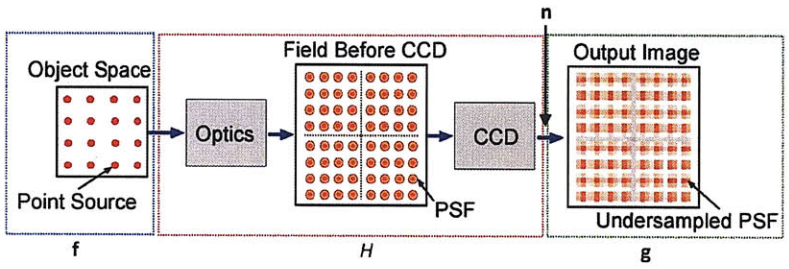


Figure 4-12: Graphical interpretation of system model.

The structure of the Hopkins matrix for the case of a holographically corrected thin imager based on a 2x2 GRIN lens array is shown in Figure 4-13-a. To compute the Hopkins matrix, the space-variant impulse response of the system (subsampled incoherent PSF) corresponding to the nth GRIN lens is computed for each field angle sequentially. The impulse response is then placed at the corresponding location on the image plane and the resulting image is raster scanned. Figure 4-13-b shows an example of the images captured by the four GRIN lenses corresponding to unraster scanning the first column of the Hopkins matrix. In general, the Hopkins matrix is block-space-variant. For a diffraction limited system, the Hopkins matrix is block-Toeplitz as the system becomes space invariant.

Using the model of the system presented above, the optical performance of an uncorrected and corrected GRIN lenses based on their space-variant PSFs is compared. The case of a single uncorrected GRIN lens used to image an array of 63×63 point sources in the object space is first examined. Each object point source corresponds to an incident plane wave that is imaged by the lens. Figure 4-14 shows the resulting intensity distribution at the image plane before the photodetector (before subsampling). The incoherent PSFs are normalized based on their Strehl ratio. The Strehl ratio is the ratio between the

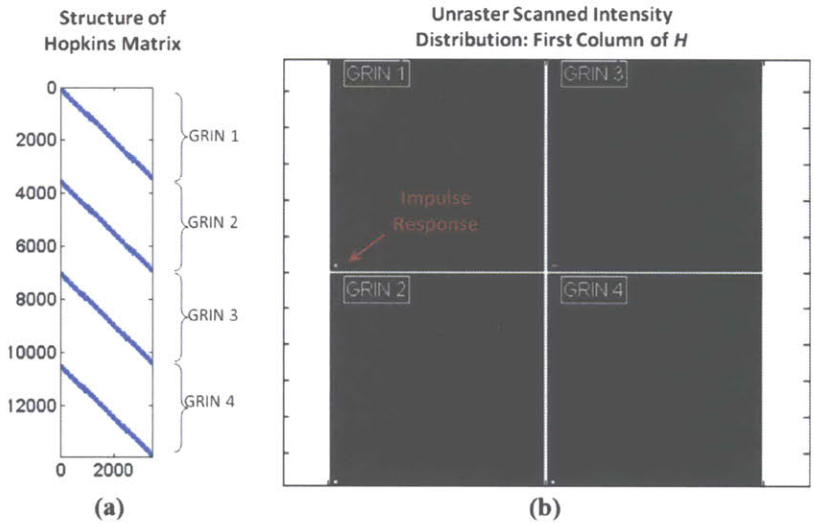


Figure 4-13: (a) Structure of Hopkins matrix; (b) Unraster scanned image from first column of Hopkins matrix.

aberrated and diffraction limit response observed at the peak intensity. The imaged point sources correspond to a FOV of approximately 15.5 degrees. The aberrated PSFs have large sidelobes smearing the image and spreading the optical energy. Next the intensity distribution before the photodetector for the case of a GRIN lens corrected for an on-axis field (center of the image plane) is computed. The result is shown in Figure 4-15. At the center of the image plane, the incoherent PSF is diffraction limited with: Strehl ratio = 1. As the hologram used for correcting the aberrations is relatively thin, the PSFs within a region centered at the image plane have near diffraction limited performance. Overall the optical performance of the corrected GRIN lens is much better than the case without correction, as will be shown in the next section.

4.5 System Analysis Based on Information Theory

Information theory [6] offers a powerful framework for characterizing the quality of an imaging system in terms of its information content. The result from this characterization

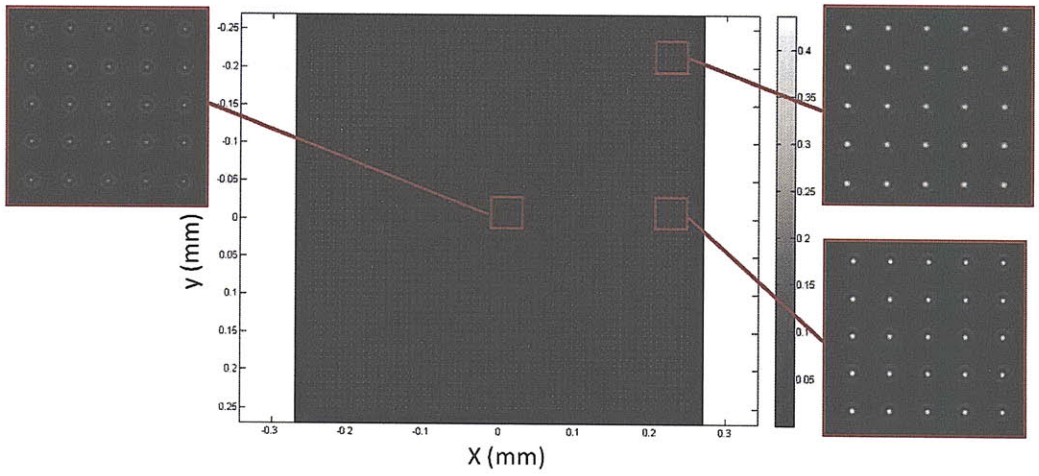


Figure 4-14: PSFs from uncorrected GRIN lens.

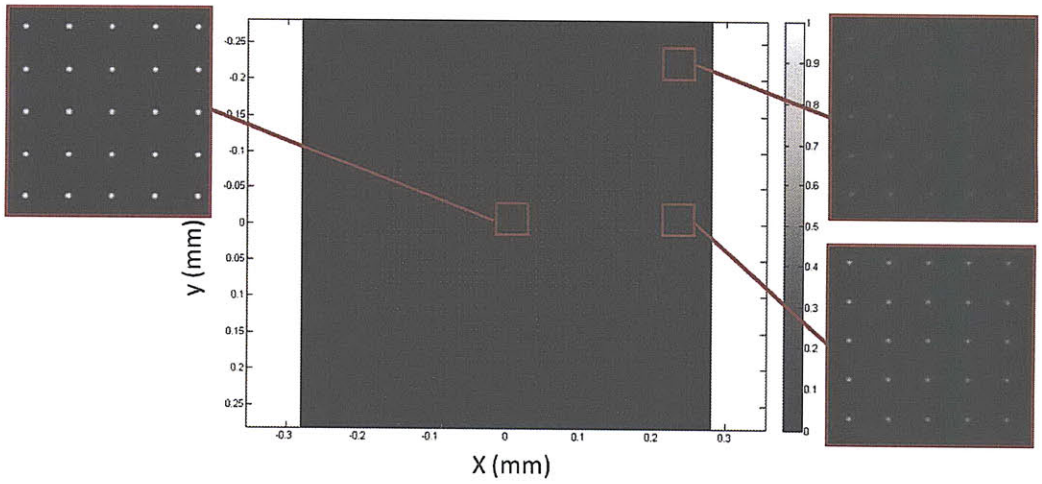


Figure 4-15: PSFs from holographically corrected GRIN lens.

can then be used to assist in the design of the optical system to maximize information exchange process. In this formulation, the optical system is treated as a communication channel and the captured image is treated as a received message, which gives information about the object or scene. The model of the channel includes factors such as power loss, aberrations, limited bandwidths (optical and detector) and noise.

It is found that standard metrics used for the characterization of optical systems, such as the two-point Rayleigh resolution criterion or the detector's space-bandwidth product (SBP), do not offer significant insight into the true reliability or fidelity of the system [216]. For example, an imaging system subject to severe blur might suggest low fidelity in the standard sense; however, by implementing a deconvolution procedure, the high-resolution pattern might be retrieved. A better metric for the evaluation of the system's performance is the Shannon or channel capacity [217]. The channel capacity specifies the highest information transfer rate that can be achieved at a given channel with an arbitrary low probability of error. It provides a useful bound that allows determining the ultimate performance limitations of the system subject to a set of design variables. The SBP metric is commonly taken to represent the amount of information content of an image; however, this measure fails to estimate the system's channel capacity and it does not include the effects of noise. In contrast, the signal-to-noise ratio (SNR) provides a better description of the system performance and is included in the description of channel capacity as will be derived later. For the proposed system model, the Hopkins matrix represents an alternate projection that can be optimized using this information-based metric. This allows maximizing the extracted scene information and relies on the existence of an inversion procedure that can be implemented to retrieve the encoded signal.

In the analysis considered in this section, the optical system is modeled as a channel subject to additive white Gaussian noise (AWGN). In this formulation, the quantum fluctuations that lead to Poisson noise are ignored. Instead, the system is assumed to be corrupted by shot noise produced by ambient light, as well as electronic noise from the

photodetector [218]. To simplify the analysis, the object space is discretized and each point is mapped to another discrete point inside the working area of the image plane. Each pixel in the detector represents a Gaussian channel of the form: $Y_i = X_i + Z_i$, where X_i is the input, Y_i is the output, and Z_i is the noise. The noise is assumed to be independent of the input signal. The power of the input signal is constrained to,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P, \tag{4.8}$$

where P is the total available power. The amount of information on average required to describe the input and output signals is given by the entropy. The discrete entropy of the random input signal is,

$$h(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x), \tag{4.9}$$

where $p(x) = \Pr\{X = x\}, x \in \mathcal{X}$, is the probability mass function and the entropy has units of bits. For example, the entropy for an input signal that only contains two possible states with equal probability is: $h(X) = 1\text{bit}$. Similarly, the joint and conditional entropies for the input and output signals (random variables X and Y) are,

$$\begin{aligned} h(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}} p(x,y)\log_2 p(x,y), \\ h(Y|X) &= -\sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}} p(x,y)\log_2 p(y|x). \end{aligned} \tag{4.10}$$

The expressions of equations 4.9 and 4.10 are related by the chain rule as: $h(X,Y) = h(X) + h(Y|X)$. We now define the mutual information that allows quantifying the amount of information than the output contains about the input. The mutual information can be viewed as the intersection in the Venn diagram of two regions defined by the

entropies of the input and output signals. The mutual information is given by,

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \\ &= h(Y) - h(Y|X). \end{aligned} \tag{4.11}$$

Equation 4.11 can be understood as the reduction of uncertainty or entropy of the measured signal, Y, by some shared knowledge about the input. The channel capacity of a discrete channel as the maximum mutual information between input and output signals is defined,

$$C = \max_{p(x)} I(X;Y), \tag{4.12}$$

where the maximum is taken over all possible input distributions $p(x)$. The definitions of entropy, mutual information and channel capacity can be extended for the case of continuous random variables. In doing so, the summations are replaced by integrals over the support set of the random variable, and the probability mass function is replaced by the probability density function. This notion is used to compute the differential entropy of the noise, Z, that is assumed to have a zero mean normal distribution,

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{4.13}$$

The corresponding differential entropy is: $h(Z) = (1/2)\ln 2\pi eN$ nats (natural bits) or by changing the base of the logarithm, $h(Z) = (1/2)\log_2 2\pi eN$ bits, where $N = \sigma^2$ is the noise power.

The mutual information for the considered Gaussian channel is now calculated,

$$\begin{aligned} I(X;Y) &= h(Y) - h(X+Z|X) \\ &= h(Y) - h(Z), \end{aligned} \tag{4.14}$$

as the noise, Z, is independent of X. The entropy of the output signal is bounded by [217]: $h(Y) \leq (1/2)\log_2(2\pi e)E[Y^2]$, where $E[Y^2] = E\left[(X+Z)^2\right] = E[X^2] + E[2XZ] + E[Z^2] = P + N$, and $E[\,]$ is the expectation value. Substituting these results to equation 4.14 we get,

$$\begin{aligned} I(X;Y) &\leq \frac{1}{2}\log_2 2\pi e(P+N) + \frac{1}{2}\log_2 2\pi eN \\ &= \frac{1}{2}\log_2\left(1+\frac{P}{N}\right) = \frac{1}{2}\log_2(1+SNR). \end{aligned} \tag{4.15}$$

The mutual information of equation 4.15 is maximized (equality) when the probability distribution of the output is also Gaussian [219]. The entropy is greatest when the parameters are statistically independent and from the central limit theorem the probability distribution is Gaussian. Using equation 4.15 the channel capacity for a Gaussian channel is found,

$$C = \frac{1}{2}\log_2\left(1+\frac{P}{N}\right). \tag{4.16}$$

Now the formulation for a set of parallel Gaussian channels (pixels on the photodetector) is extended. The total information capacity of a 2D pixilated optical imaging system is determined by the capacity of a single pixel times the number of pixels [216]: $C_T = C_{pix}M$, where M is the total number of pixels. The power constraint requires the total power to be distributed among the channels,

$$E\left[\sum_{i=1}^{M} X_i^2\right] \leq P. \tag{4.17}$$

Using the same methods as above, the channel capacity for parallel Gaussian channels is derived,

$$C = \sum_i \frac{1}{2}\log_2\left(1+\frac{P_i}{N_i}\right), \tag{4.18}$$

where $P_i = E[X_i^2]$, and $\sum_i P_i = P$. Finally, we use these results on our system model

of equation 4.7. The resulting channel capacity is given by,

$$C = \sum_i \ln\left(1 + \frac{\mu_i}{\sigma^2}\right), \tag{4.19}$$

where μ_i is the ith singular value of the Hopkins matrix and σ^2 is the noise variance (assumed to be constant over every channel). Singular value decomposition (SVD) is performed on the Hopkins: $H = U\Sigma V^T$, where U and V are unitary matrices and Σ is a diagonal matrix with elements equal to the singular values. The units of the channel capacity are nats due to the change of base in equation 4.19. Also, power normalization is assumed. The term μ_i/σ^2 can also be identified as the SNR of the system.

The singular values and channel capacities for two thin imagers, one based on an uncorrected 2x2 GRIN lens array (similar to the TOMBO system) and the second one based on a corrected GRIN lens array, are now compared. For the imager based on the holographically corrected array, two correction configurations are simulated. The correction points are chosen arbitrarily and are indicated in Table 4.2. The simulation parameters are: $\lambda = 658$nm, detector $SBP = 120 \times 120$pixels (pixel size= 9μm), Hopkins matrix computed for a grid of 63×63 point sources. The singular values for the three cases are shown in Figure 4-16-a. Notice how the overall magnitude of the singular values is increased for the both cases of the corrected imager. This increase is substantial despite that no optimization of the correction position has been conducted. To understand the significance of the improvement the system's SNR is first considered. As indicated in Figure 4-16-a, in the presence of noise the low magnitude singular values are lost making it difficult to recover the original Hopkins matrix that characterizes the system. This concept is analogous to that used in matrix approximation theory [220]. For optical systems, high magnitude singular values correspond to low frequency components of the imaged object or scene. Similarly, low magnitude singular values correspond to high frequency components of the scene. If these singular values are lost, the system is unable to resolve small features. The dependence of spatial frequency can be seen in Figure 4-16-b when computing the Karhunen-Loeve modes corresponding to the singular values:

μ_1, μ_{550}, and μ_{900}.

Table 4.2: Correction Positions.

	Correction 1		Correction 2	
Lens	x (mm)	y (mm)	x (mm)	y (mm)
GRIN 1	-0.13	-0.13	0	0
GRIN 2	-0.13	0.13	0	0.2
GRIN 3	0.13	-0.13	0.2	0
GRIN 4	0.13	0.13	0.2	0.2

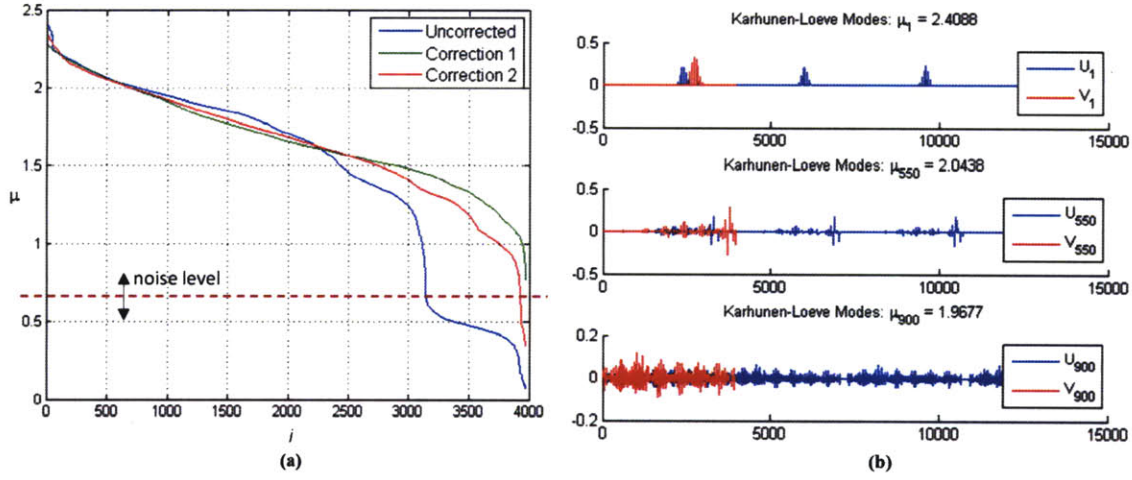


Figure 4-16: (a) Singular values of uncorrected and corrected 2×2 GRIN lens arrays; (b) Karhunen-Loeve modes.

A second viewpoint for quantifying the improvement of the holographically corrected imager is based on matrix stability. In general, the associated inverse problem is ill-posed in the sense of Hadamard [221]. For the problem to be well-posed, the corresponding solution must be unique and exists for arbitrary data as well as depend continuously on the data. If any of these conditions is not satisfied, the problem is considered ill-posed. The non-existence of the solution and the lack of uniqueness in imaging systems are commonly attributed to the fact that the system does not transmit complete information

about the Fourier transform of the object at certain frequencies. This information loss cannot be recovered by any mathematical trickery or decoding algorithm. The alternate projection, introduced by the Hopkins matrix of the holographically corrected system, allows recovering some of the information lost and hence improves the stability of the system. The third condition that requires continuous dependence on the input data relates to numerical instabilities that may arise during the inversion process. If a small error in the initial data results in a large error in the solution, the associated problem is called ill-conditioned and is quantified by the condition number. If the condition number is large, the problem is ill-conditioned. On the other hand, if the condition number is close to one, the problem is well-conditioned. The condition number controls the error propagation from the data to the solution. The condition number is given by,

$$cond = \frac{\mu_{\max}}{\mu_{\min}}, \tag{4.20}$$

where $\mu_{\max}$ and $\mu_{\min}$ are the maximum and minimum singular values of the Hopkins matrix. Figure 4-17-a shows a comparison of the condition numbers of the uncorrected and corrected systems using the first configuration indicated in Table 4.2. Finally, the channel capacity as defined in equation 4.19 for both cases with and without holographic correction is compared. Figure 4-17-b shows the computed channel capacity for a noise level of 0.08. This figure proves that the holographically corrected system is more efficient in extracting information from the scene than the TOMBO-like system. This result will be further improved in the next section by performing a multi-domain optimization.

In the last example, the system is evaluated for different numbers of lenses in the array. Figure 4-18-a shows the computed channel capacity. Figure 4-18-b shows the corresponding singular values. This result is expected, as adding more lenses increases the effective aperture of the system. The larger effective aperture allows collecting more power from the object signal. An opposite result would be expected if the total power available is constrained and a comparison is performed between a single lens and a segmented aperture system, where both systems have the same effective aperture. Ignoring

the effect of aberrations, the single lens system yields to a larger channel capacity compared with the segmented aperture system, as the power in this system is distributed over the lenses. In addition, there is a tradeoff between SNR and number of pixels or pixel size that can be exploited to find an optimum SBP or pixel size in the sense of information theory [216].

4.6 Multi-Domain Optimization based on Genetic Algorithms

In a traditional imaging system, the optical components are optimized to perform as close as possible a linear mapping between the object and measured signals. This formulation then requires a Hopkins matrix: $H \rightarrow I$, where I is the identity matrix. Such formulation is suboptimal, as the optical system is ultimately limited by diffraction and by the sampling properties of the photodetector, limiting the amount of information that can be extracted from the scene. In contrast, we are interested in finding an alternate projection as represented by H not equal to I and $\mathbf{g}$ (intensity measurement) being the input of an image post-processing algorithm. The output of the post-processing algorithm is,

$$\begin{aligned} \mathbf{a} &= P\left(H\mathbf{f} + \mathbf{n}\right) \\ &= PH\mathbf{f} + \hat{\mathbf{n}}, \end{aligned} \tag{4.21}$$

where P is a potentially nonlinear operator representing the post-processing (decoding) algorithm. The equivalent operator, $W = PH$, combines the actions of the optics, detector and post-processing, and can be optimized simultaneously using a multi-domain optimization (MDO) approach [222]. If the output, $\mathbf{a}$, is required to be an isomorphic or geometrical representation of the source, $\mathbf{f}$, the operator P needs to be proportional to the inverse of the Hopkins matrix. Other applications, such as machine vision or pattern identification, might not require an isomorphic representation of the object. An

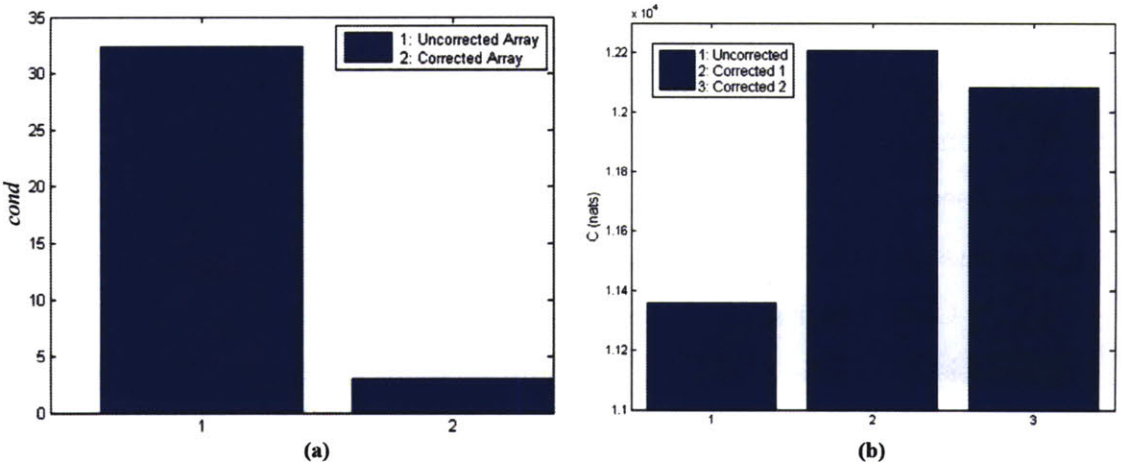


Figure 4-17: Comparison between uncorrected and corrected systems: (a) Condition number; (b) Channel capacity.

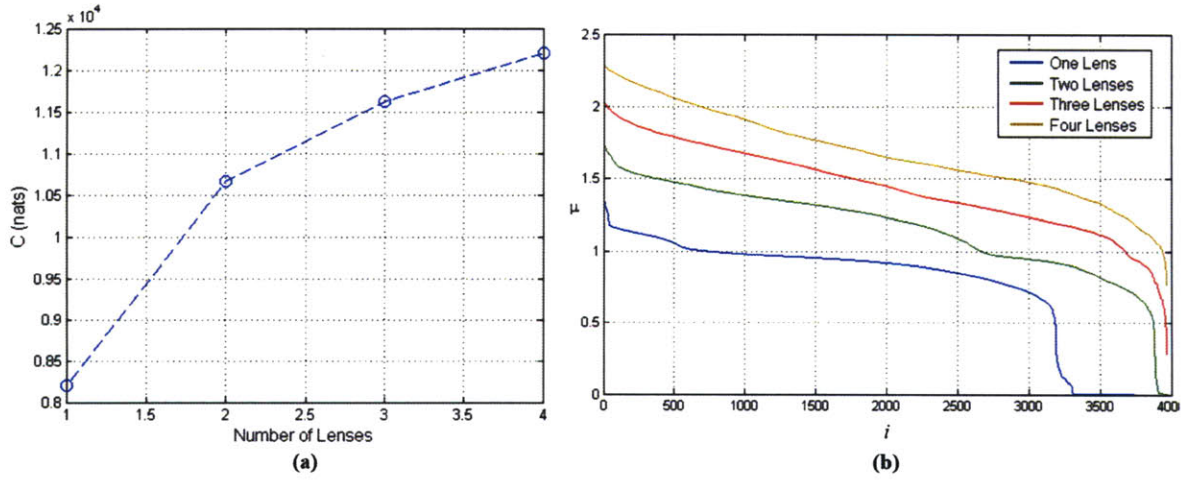


Figure 4-18: Comparison of systems with different number of lenses: (a) Channel capacity; (b) Singular values.

alternative MDO relies on maximizing the channel capacity of the Hopkins matrix to extract more scene information and to improve the stability of the combined operator, W, respect to noise perturbations in the measurement or imperfect calibrations of the optical system. This is the approach implemented in this section.

The configuration (positions and field angles) of the holographic correction for the proposed segmented aperture imager is optimized using genetic algorithms (GAs) based on the fitness function: $f = 1/C$, where C is the channel capacity as defined in equation 4.19. The GAs optimization is conducted in the same way as in Chapter 2. To simplify the analysis, a smaller grid of 5×5 source points is used to construct the Hopkins matrix. This model assumes that the resulting PSF is locally shift-invariant within a small isoplanatic patch [219]. A grid of 6×6 possible image plane correction points per GRIN lens is considered. The system evaluation and correction points are indicated in Figure 4-19-a. The optimization parameters are indicated in Table 4.3. Figure 4-19-b shows the computed optimum correction positions. The singular values of the optimized Hopkins matrix for the holographically corrected system and those from the uncorrected TOMBO-like imager are shown in Figure 4-20-a. The corresponding channel capacities are (assuming a noise level $N = 0.8$): $C_{corrected} = 20.78$nats, $C_{uncorrected} = 17.74$nats. Figure 4-20-b shows the convergence plot (fitness function score) for the best individual and the population mean.

Table 4.3: MDO Parameters.

Array Size	2×2	**Number of Generations (GAs)**	25
Operation Wavelength (λ)	658nm	**Population Size**	50
Photodetector Pixel Size	2μm	**Crossover Fraction**	0.5
SBP (pixels)	233×233	**Elite Children**	1
FOV	15.6°	**Number of Variables**	8

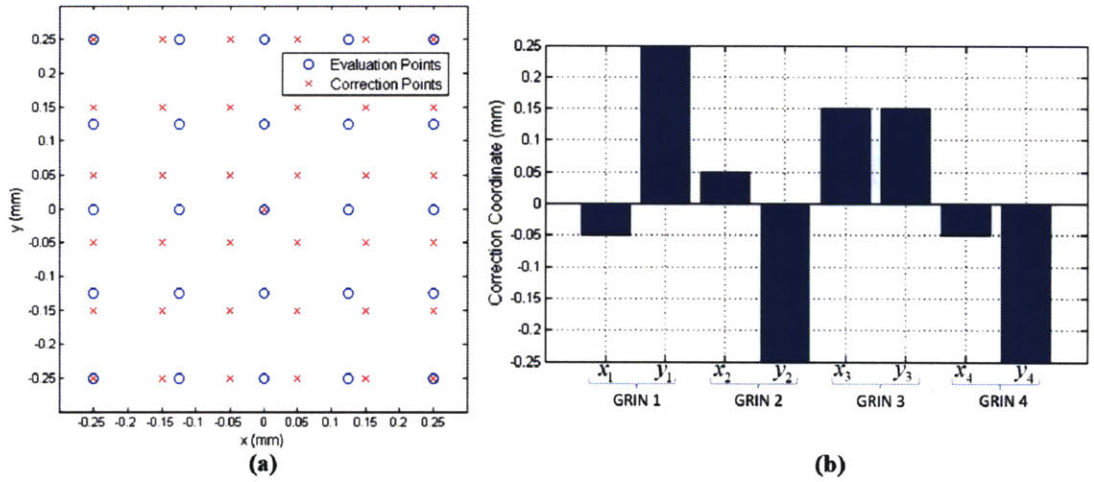


Figure 4-19: (a) Evaluation and correction points; (b) Optimized correction points.

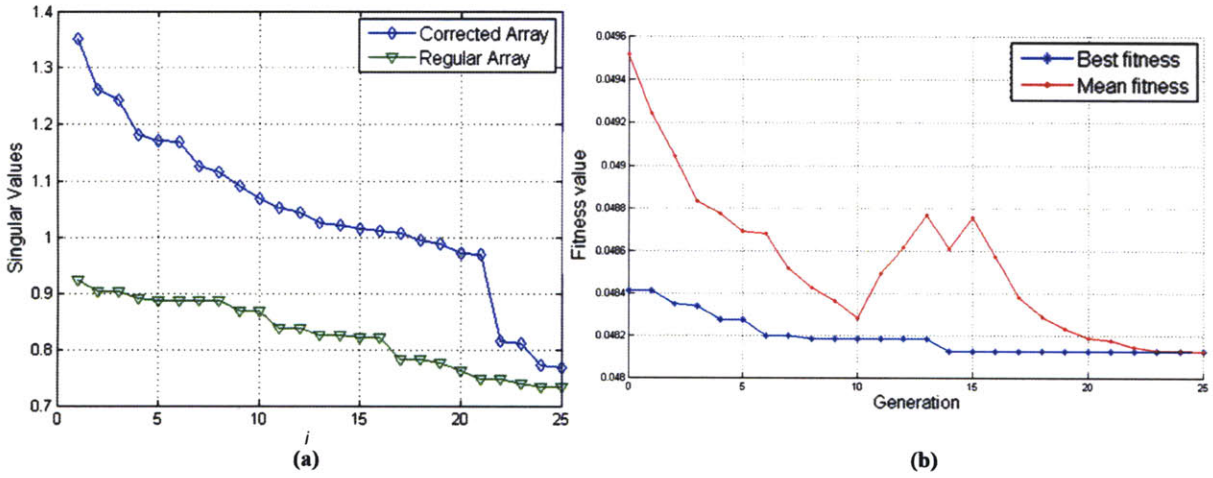


Figure 4-20: (a) Singular values comparison; (b) GAs convergence plot.

4.7 Decoding Algorithm: Image Post-Processing

The decoding algorithm involves solving the inverse problem to retrieve an estimate of the object, **a**, from the measured intensity, **g**. In the simplest form, the decoding operator takes the form: $P = H^{-1}$. The computation of such inverse might not be stable, requiring the implementation of the pseudoinverse based, for example, on the SVD of the Hopkins matrix: $P = V\Sigma^{\dagger}U^{T}$, where $\dagger$ represents the inverse of the diagonal singular values matrix. This technique is also known as the back projection method [198] and is a common way to compute the best fit in the least squares sense of a solution to a system of linear equations that lacks of unique solution. However, in practical implementation, the calculation of the pseudoinverse might not be possible, due to the large size of the Hopkins matrix resulting in memory and computational time problems. An alternative method is then required. Several decoding algorithms have been investigated for segmented aperture systems. Examples include the linear minimum mean-square error method [222], sampling method [199] and pixel rearrange method [200], also known as the superresolution algorithm. In this section the implementation of the superresolution algorithm for the post-processing of the captured images is described.

In the superresolution algorithm, the low-resolution captured images from each GRIN lens in the array are combined to produce a high-resolution image [223], [224]. The measured low-resolution images suffer aliasing where high-frequency components are folded into the low frequency components of the image and, as a result, the subtle or detailed information is lost. Various superresolution algorithms have been reported in the literature and can be broadly classified as iterative and noniterative. Examples of iterative methods include maximum a posteriori [225], and expectation maximization algorithms [226]. Noniterative methods include the k-nearest weighted neighbor approach [227].

The superresolution method is a form of de-aliasing algorithm that can be used to recover this information in systems with PSFs smaller than the pixel size of the detector. By utilizing k images produced by an array of k lenses, the expected resolution improvement from the information theory viewpoint is: $\delta_{pix}/\sqrt{k} \times \delta_{pix}/\sqrt{k}$, where δ_{pix} is the

detector's pixel size. For example, for the simulation results presented in the previous section based on a 2×2 GRIN lens array and a pixel size, $\delta_{pix} = 2\mu$m, the resulting improvement is $\delta_{pix,HR} = 1\mu$m, which results in an increase of the bandwidth of the recovered signal by a factor of two. The performance of the superresolution algorithm is ultimately limited by diffraction (limited pass-band of the optical system).

There are three mayor steps in the superresolution image reconstruction method: 1. Acquisition of multiple images with subpixel shifts; 2. Estimation of subpixel shift; 3. Reconstruction of high-resolution image. The first step is done in a single shot on a segmented aperture imager. For imagers with a small number of lenses, the perspective variations of the scene can be neglected. Otherwise, a homomorphic transformation may be applied. This perspective variation is exploited in some applications such as 4D light field imaging [201]. For the example presented here, any perspective variations of the scene are neglected. In the second step, an algorithm is implemented to estimate the subpixel shifts relative to a reference frame. Here we implement an estimation algorithm based on spectral cross correlation to explore the translational differences of Fourier transforms of the analyzed images. The block diagram of this algorithm is shown in Figure 4-21. Before estimating the subpixel shift, the algorithm of Figure 4-21 is implemented without the upsampling and low-pass windowing sections to estimate integer pixel shifts. Then, the algorithm is used to compute the $k-1$ subpixel shifts, x_{shift} and y_{shift}, using the first image, $k = 1$, as a reference frame. The upsampling is done by zero padding the cross correlation spectrum, $\hat{G}$, to a size larger that the optical system's pass-band given by the highest spatial frequency of the MTF for the diffraction limited system. In order to avoid ripples in space domain, the spectrum is low-pass filtered using the power window [87] given by,

$$W(u,v) = \text{circ}\left(\frac{\rho}{\rho_{cutoff}}\right) \exp\left(-\alpha\rho^n\right), \tag{4.22}$$

where α and n are constants, ρ_{cutoff} is the desired cut-off, and $\rho = \sqrt{u^2+v^2}$. The power

window is a Fourier-based smoothing window that preserves most of the spatial frequency components within the pass-band and attenuates quickly at the transition band. It is differentiable at the transition point which gives the desired smooth property and limits the ripple effect. After cross correlation, the location in the spatial domain of the peak identifies the relative shifts.

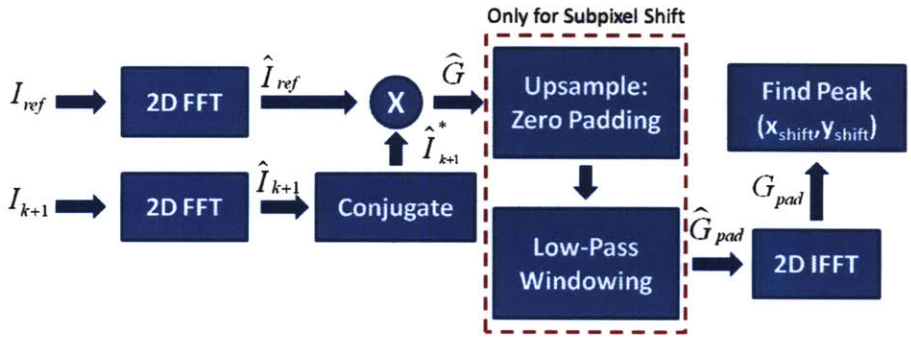


Figure 4-21: Block diagram of subpixel shift estimation algorithm.

In the third step, the low-resolution images are placed on a high-resolution grid using the subpixel shift estimates. The samples that do not correspond to any of the captured low resolution images are set to zero. This high-resolution grid is called the processing array. Several algorithms have been developed to retrieve a high-resolution image from the processing array. These techniques include the nonlinear interpolation method [228], regularized inverse processing [229], and the iterative error-energy reduction (EER) method [230]. In this section the EER method is implemented. This method is a variation of the modified error reduction algorithm implemented in Chapter 2. The block diagram of this method is shown in Figure4-22. The algorithm iterates between space and frequency domains imposing constrains at each domain. A bandwidth constraint is imposed in frequency domain by low-pass filtering the signal spectrum using the power window with a cut-off frequency given by the diffraction limited optical system pass-band. For the space domain constraint, the samples of the space domain signal at the measured locations are replaced by those from processing array. The algorithm repeats until the

maximum number of iterations is reached.

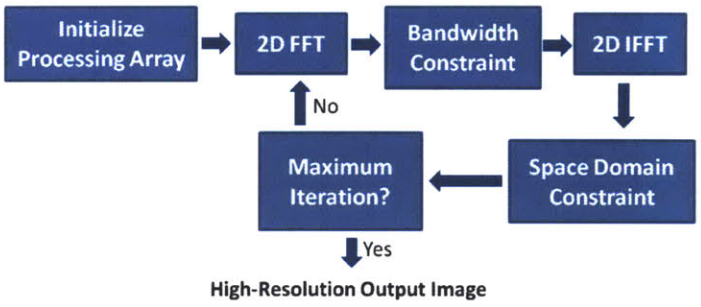


Figure 4-22: Block diagram of error-energy reduction algorithm.

The model presented in the previous sections is used to simulate the subimages captured by a 2×2 holographically corrected segmented aperture thin imager. The correction positions are those found by GAs indicated in Figure 4-19-b. The object or scene is formed by an array of point sources forming a resolution target pattern. The object is placed in the front focal plane of an ideal lens that collimates the point sources simulating an object in the far field. The space-variant PSFs are computed at the image plane before and after being sampled by the photodetector. Figure 4-23 shows the intensity distribution at the image plane by GRIN len 3 before the photodetector. The intensity distribution is normalized to the Strehl ratio. The image plane area corresponds to a FOV of approximately 9 degrees. Figure 4-24 shows the subsampled images captured by the four GRIN lenses in the array. Each low-resolution image has a different subpixel shift and distribution, due to the different holographic correction points. Despite that the detector's pixel size is small ($\delta_{pix} = 2\mu$m), the PSFs are averaged by the pixel photosensitive area (a fill factor of 100% has been assumed) and the small details about the object are lost.

The superresolution algorithm is then used to combine the information contained on the four captured images of Figure 4-24 to produce a higher resolution image. After estimating the subpixel shifts and initializing the processing array, the EER algorithm is run for 300 iterations. Figure 4-25 shows the reconstructed high-resolution image. As

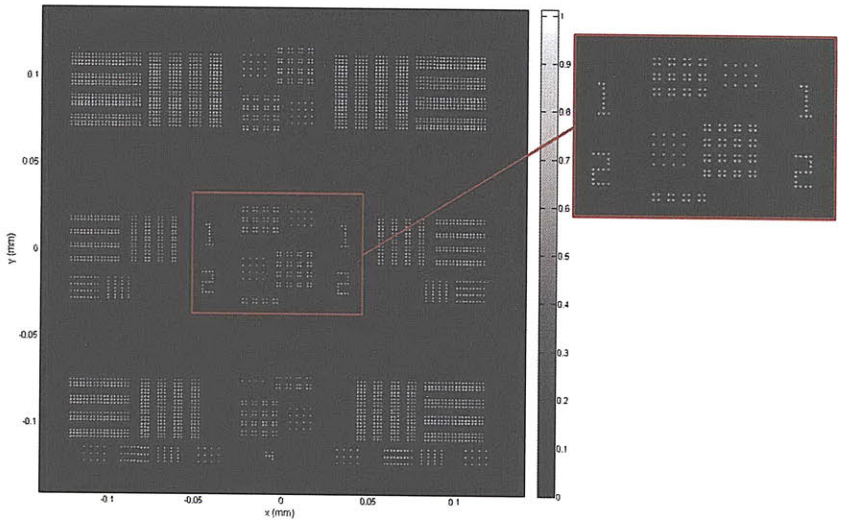


Figure 4-23: Intensity distribution before photodetector imaged by GRIN lens 3.

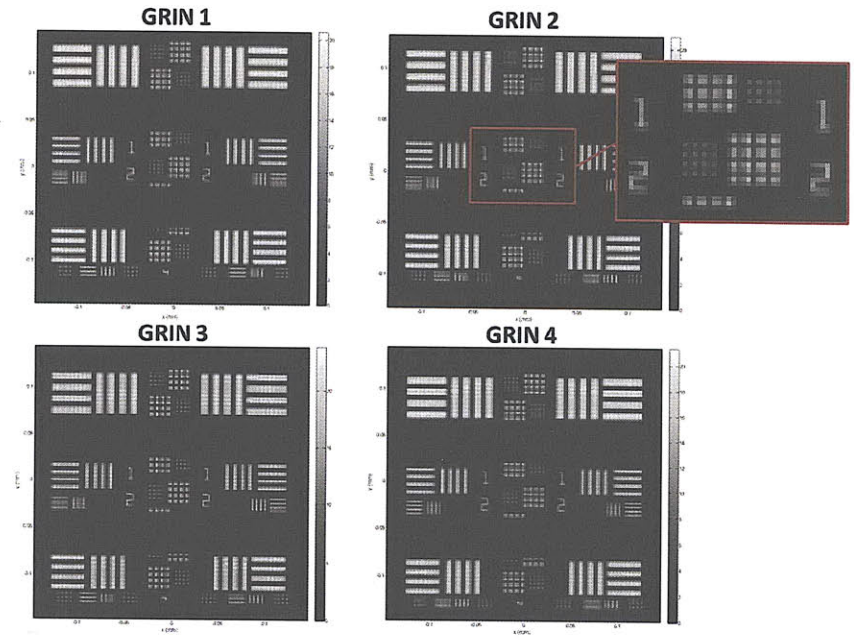


Figure 4-24: Simulated captured images from holographically corrected segmented aperture system.

can be seen, some of the modulation due to the array of point sources is retrieved. A better result is expected for a larger number of lenses. In addition, the information about the model (configuration, correction position, hologram/GRIN response) can be used to improve the reconstruction. For example, a decoding strategy based on the pseudo inverse of the Hopkins matrix can be applied to each subimage, where the Hopkins matrix only contains the response for that single lens. In this way, distortions due to the hologram can be compensated. The processed images are then used by the superresolution algorithm to generate a high-resolution output image.

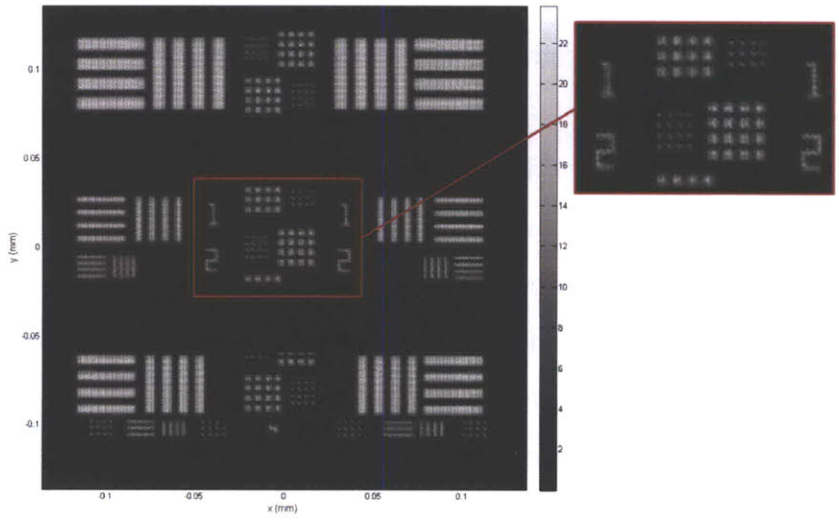


Figure 4-25: Reconstructed high-resolution image.

4.8 Experimental Implementation

In this section we present some experimental results on the design and optimization of the proposed holographically corrected segmented aperture thin imager. The holographic recording process is optimized to achieve high diffraction efficiency and the experimental results are compared to those predicted by Kogelnik's coupled wave theory. A method for characterizing the system's PSF based on the Foucault knife edge test is presented.

A comparison between an uncorrected and holographically corrected GRIN lenses is performed. A sensitivity analysis is conducted for estimating potential axial and lateral misalignment errors.

4.8.1 Optimization of Hologram Optical Recording Process

The optimization results presented in the previous sections rely on high diffraction efficiency holograms to minimize the optical aberrations produced by undesirable diffraction orders on the reconstructed field. The hologram's diffraction efficiency is closely related the recording conditions, processing steps and holographic material used, as well as its properties such as thickness, refractive index modulation and absorption coefficient. Over the last few decades, several materials have been studied for the optimum recording of holograms such as dichromated gelatins [231], photopolymers [232] and photographic materials [233]. Among these materials, silver halide (AgH) photographic emulsions are the most widely used, due to their high energy and spectrum sensitivities, repeatability of results and commercial availability. Silver halide emulsions are composed of silver halide grains suspended in a gelatin support. In addition, sensitizing agents are added to the gelatin to influence the introduction of dislocation centers within the crystals. Most commercially available emulsions are coated on a base such as a glass plate, mylar or acetate (in the case of films). Light incident on the emulsion initiates a complex physical process consisting of the following major steps: 1. The incident photon absorbed by a grain releases an electron-hole pair within the grain; 2. The resulting mobile electron in the conduction band is eventually trapped at a crystal dislocation; 3. A silver ion is attracted by the trapped electron and combines to form a metallic silver atom; 4. The collection of silver atoms forms a development speck, which comprises the latent image. The exposed emulsion then undergoes a prescribed chemical post-processing.

The holograms studied in this section are recorded on a silver halide emulsion with the specifications indicated in Table 4.4. To optimize the diffraction efficiency of the hologram, an unslated transmission hologram is recorded under varying exposure ener-

gies using the geometry of Figure 4-26. A Helium Neon laser ($\lambda = 632.8$nm) is used as the illumination source. A computer activated shutter allows the accurate control of the exposure time and hence the exposure energy. The beam expansion and collimation is conducted using a spatial filter and a plano-convex lens. A polarized beam splitter divides the input beam into the reference and object beams. The relative optical power of these beams is controlled by properly adjusting the $\lambda/2$ waveplates. The light reflected by two mirrors interferes at the hologram plane, forming an unslated hologram within the volume of the emulsion. The fixed angle between the object and reference beams is set to 30 degrees. The polarization of the interference beams is perpendicular to the plane of incidence. The holographic film is sandwiched between two microscope slides and placed on a holder on top of a motorized rotation stage. The rotation stage is a Newport URS Series with angular resolution of 0.0005 degrees. It is used to calibrate the recording geometry to the desired incidence angle, as well as for characterizing the recorded angular selectivity of the hologram. Two chemical processing methods are studied: fixation free rehalogenating bleaching and silver halide sensitized gelatin methods. For each processing method, several holograms are recorded with varying exposure energies. After processing, the holograms are characterized by measuring the diffraction efficiency of the first order at the Bragg angle, angular selectivity of the diffracted (1st order) and transmitted (0th order) beams, and total diffracted, transmitted and scattered powers. The measurements are conducted with Newport's dual-channel power meter model 2832-C, which communicates with the computer using a LabView interface. The second photodetector is placed on top of a linear motorized stage (Newport UTM50CC1DD with 1μm position accuracy) to track the diffracted beam during the angular selectivity test. The tracking position is calculated by estimating the output diffraction angle based on the grating's equation: $\Lambda(\sin\theta_m - \sin\theta_i) = m\lambda$, where Λ is the grating period, θ_m is the output angle of the mth order, $\theta_i$is the incident angle, and $m = 1$ for the first diffraction order. The rotation and linear motorized stages are operated using the universal motor drive controller UniDrive 6000 from Newport.

Table 4.4: Silver Halide Emulsion Specifications.

Manufacturer	Slavich	**Emulsion Thickness** (d)	7.32μm
Model	PFG-01	**Substrate Thickness**	180μm
Type	Film	**Spectral Sensitivity**	600-680nm
Substrate	Triacetate ($n_{sub} = 1.478$)	**Emulsion Resolution**	~3000 lines/mm
Grain Size	~40nm	**Extinction Coefficient** (κ)	8×10^{-5}

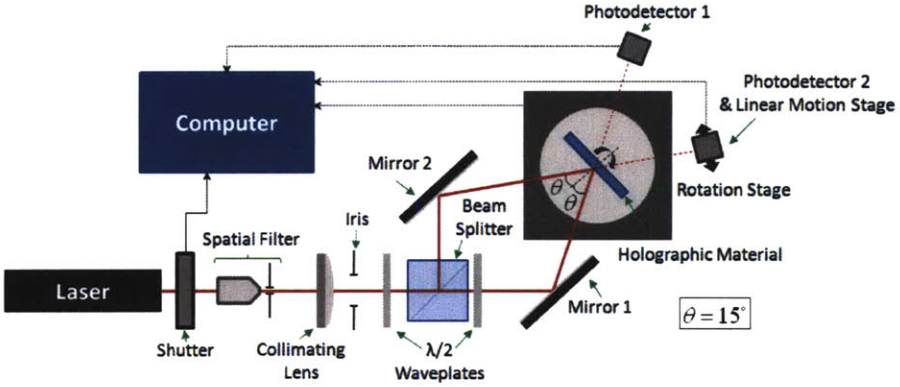


Figure 4-26: Hologram recording geometry.

In the chemical post-process the exposed holograms are first developed to transform the silver specks into metallic silver, amplifying the effect of exposure. Common developers include one-part solutions, such as Kodak's D-19 which tends to produce low contrast results, and two-part developers that are mixed before use. Before development the holographic film is transparent, turning black after the development bath. Volume holograms recorded on silver halide emulsions are commonly bleached to obtain phase holograms with increased diffraction efficiencies. The fixation free rehalogenating (FFR) bleaching process has been studied by several researchers using a variety of commercially available holographic emulsions [234], [235]. It is assumed that diffusion of the material from exposed to unexposed regions occurs during the bleaching bath. The metallic silver grains are converted back to silver halide grains by an oxidation process. As a result,

a refractive index modulation is produced between the exposed and non-exposed zones and is proportional to the size difference of the silver halide grains and concentration. The implemented fixation free rehalogenating process contains the following steps: 1. Immerse hologram in two-part developer for 2 minutes; 2. Rinse in water for 3 minutes; 3. Immerse in bleaching solution for 2 minutes; 4. Second rinse for 3 minutes; 5. Wetting agent bath for 1 minute to promote uniform drying of the processed film. The developer is prepared according to the following recipe: part 1 – catechol (20g), ascorbic acid (10g), sodium sulfite anhydrous (10g), urea (75g), distilled water (1000ml); part 2 – sodium carbonate (60g), distilled water (1000ml). The two parts are mixed before processing with a ratio 1/1. The bleaching solution is prepared according to the following recipe: potassium dichromate (5g), sodium bisulfate (80g), distilled water (1000ml). The developer and bleaching solutions are kept at room temperature. The wetting solution is Kodak's PhotoFlo and is diluted in distilled water (1ml/1l). After the chemical process, the holograms are hung to air dry.

In the silver halide sensitized gelatin (SHSG) process, the variations of optical density between exposed and unexposed regions in the emulsion are converted to variation in the degree of hardening of the emulsion gelatin [236]. As opposed to the previous method, this method includes a fixation step in which the silver halide grains are removed, resulting in an increased diffraction efficiency and lower scattering loss. This method has proven to be a good alternative to dichromated gelatin in the production of transmission holograms [237]. The implemented SHSG consists of the following steps: 1. Developer bath for 5 minutes; 2. Rinse in water for 2 minutes; 3. Bleaching bath for 3 minutes; 4. Rinse for 2 minutes; 5. Fixation bath for 2 minutes; 6. Rinse for 10 minutes; 7. Three step dehydration process. Kodak's D-19 developer is used and is composed of: metol (2g), sodium sulfite anhydrous (45g), hydroquinone (8g), sodium carbonate anhydrous (50g), potassium bromide (5g), distilled water (1l). The bleaching solution is composed of two parts and is prepared according to the following recipe: part 1 – ammonium dichromate (20g), sulfuric acid (14ml), distilled water (1l); part 2 – potassium bromide (92g), distilled

water (1l). The two parts are mixed on a ratio of 1/30. Two temperatures are considered for the bleaching solution: 20°C and 30°C. Kodak's nontanning F-24 is used as a fixer solution. The dehydration steps consist of immersing the hologram in baths of 50%, 90%, and 100% isopropanol, each bath for 3 minutes. The processed holograms are then hung for air drying.

We first present the diffraction efficiency results from the holograms processed using the fixation free rehalogenating bleaching method. The diffraction efficiency is calculated as the ratio of the measured powers from the diffracted beam (1st order) and input power (illumination beam). Figure 4-27-a shows the measured diffraction efficiencies as a function of exposure energy. The highest diffraction efficiency without taking into account reflection losses is 53% (corresponding to an exposure energy of $213.12\mu J/cm^2$). If the effect of Fresnel reflections from all the interfaces involved (microscope slides, substrate and emulsion) is subtracted, the corresponding diffraction efficiency is 64%. The calculation of the reflection losses is conducted using the Fresnel equations for TE polarized light [153]. The reflection coefficients at the five interfaces (air-glass-emulsion-triacetate-glass-air) are computed using the following refractive indices: $n_{glass} = 1.5$, $n_{emulsion} = 1.609$, $n_{triacetate} = 1.478$. The power of the additional transmitted and reflected diffraction orders is measured and subtracted to the input power to estimate the scattering and absorption losses. These losses account for 15.5% of the incident energy as no fixation step is used. Figure 4-27-b shows the measured powers of the diffracted and transmitted (zeroth order) beams. The corresponding input power is: $P_{in} = 3.18 mW/cm^2$. These results are consistent to those presented in [235].

The holographic emulsion used in the experiments is close to the boundary between the Raman-Nath and Bragg regimes. This boundary is described in terms of the Q factor,

$$Q = \frac{2\pi d\lambda}{\Lambda^2 n_{emul}}. \tag{4.23}$$

If $Q < 2\pi$, the operation is on the Raman-Nath regime, while if $Q > 2\pi$, operation is in the Bragg regime. The recorded hologram has a period: $\Lambda = (\lambda/2)\sin(15) = 1.22\mu m$.

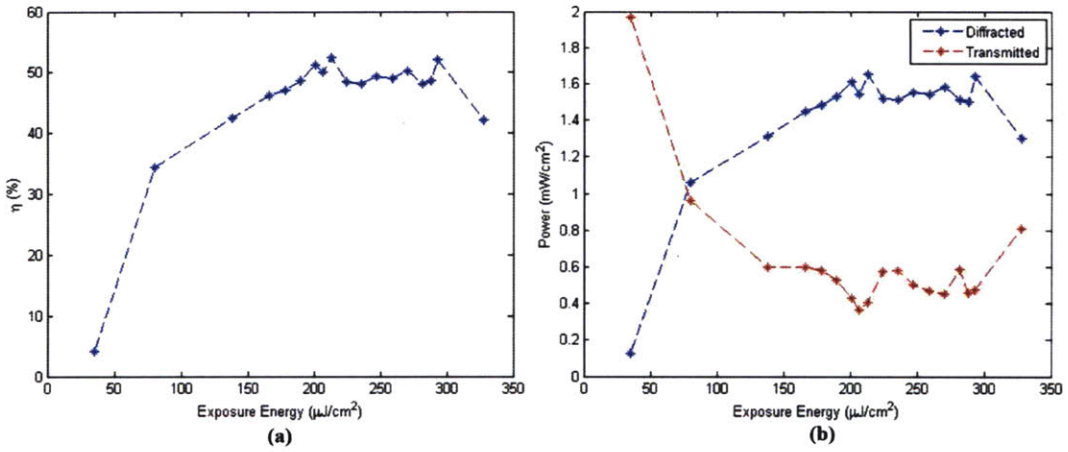


Figure 4-27: (a) Measured diffraction efficiency; (b) Measured powers of diffracted and transmitted orders.

From Table 4.4 and with $n_{emul} = 1.609$ and $\lambda = 632.8$nm, the quality factor is: $Q = 12.103$. The resulting Q factor is slightly larger than 2π so the hologram is expected to exhibit properties from volumetric materials such as Bragg selectivity.

The hologram's angular selectivity is measured by a controlled rotation and automatic measurement of the power from the first diffraction order. The measurements are then compared to the predictions from Kogelnik's coupled wave theory [143]. Kogelnik derived an approximated formulation to describe the behavior of thick holographic gratings. The main assumptions of Kogelnik's formulation are: sinusoidal modulation of the refractive index and absorption coefficient; small index modulation, $\Delta n << 1$; small absorption loss per wavelength; slow energy interchange between two coupled waves; light incident at or near the Bragg angle; neglect other diffraction orders. Kogelnik considered the cases of reflection and transmission holograms, lossless and lossy phase (dielectric) and amplitude (absorption) holograms, slanted and unslanted gratings, TE and TM polarizations, and mixed gratings (amplitude and phase). We are interested in the solution for a lossy, unslanted, phase, transmission hologram. The field of the first order diffracted wave at the output of the hologram is given by (a detailed derivation of this equation can be

found in [143]),

$$S = -i \exp\left(-\frac{\alpha d}{\cos\hat{\theta}}\right) \exp\left(-i\xi\right) \frac{\sin\left(\sqrt{\nu^2+\xi^2}\right)}{\sqrt{1+\frac{\xi^2}{\nu^2}}}, \tag{4.24}$$

where α is the mean absorption coefficient; d is the hologram thickness; $\hat{\theta} = \arcsin(\sin(\theta_h)/n_0)$, is the half angle of the probing or reconstruction wave inside the medium, where θ_h is the corresponding half angle in air and n_0 is the hologram's average index of refraction. Also, $\nu = (\pi \Delta n d)/(\lambda \cos\hat{\theta})$, where Δn is the refractive index modulation, and ξ is proportional to the dephasing measure that estimates deviations from the Bragg angle and is given by,

$$\xi = \frac{\pi d}{\Lambda \cos\hat{\theta}} \left(\left|\sin\hat{\theta}\right| - \frac{\lambda}{2n_0\Lambda}\right). \tag{4.25}$$

The diffraction efficiency is given by,

$$\eta = SS^* = \exp\left(\frac{-2\alpha d}{\cos\hat{\theta}}\right) \frac{\sin^2\left(\sqrt{\nu^2+\xi^2}\right)}{\left(1+\frac{\xi^2}{\nu^2}\right)}. \tag{4.26}$$

At the Bragg angle, $\xi = 0$, and the diffraction efficiency of equation 4.26 reduces to,

$$\eta_{Bragg} = \exp\left(\frac{-2\alpha d}{\cos\hat{\theta}_{Bragg}}\right) \sin^2\left[\nu\left(\hat{\theta}_{Bragg}\right)\right], \tag{4.27}$$

and it will be maximum when $\nu = \pi/2 \rightarrow \Delta n = (\lambda \cos\hat{\theta}_{Bragg})/(2d)$. The maximum diffraction efficiency depends on the attainable refractive index modulation from a given chemical processing. For the considered geometry and optical parameters, the required index modulation is: $\Delta n = 0.0446$. If the refractive index of the emulsion is under or over modulated, the hologram will reconstruct with low diffraction efficiencies. Both considered chemical processes (FFR and SHSG) can be controlled to achieve index modulations close to the ideal. However, as we will see later, the absorption losses resulting from the FFR method are larger to those from the SHSG technique. The hologram's extinction

coefficient is related to the absorption coefficient by: $\alpha = (2\omega\kappa)/c$, where $\omega = 2\pi f$, $f = c/\lambda$, and c is the speed of light. For the parameters of Table 4.4, the absorption coefficient is: $\alpha = 1.5886x10^3 m^{-1}$.

Figure 4-28-a shows a comparison of the measured and theoretical diffraction efficiencies with different probing angles (without considering Fresnel reflections) for the hologram with the highest diffraction efficiency of the batch processed using the FFR bleaching method. As can be seen, the couple wave theory model of equation 4.26 fits very closely the measured data. The thickness of the emulsion limits the system's FOV to approximately 20 degrees (relative to the Bragg angle). Because the expression for the theoretical diffraction efficiency depends on the refractive index modulation, thickness and absorption coefficient, information about these parameters can be obtained by fitting the theoretical function to the experimental data. However, these parameters can also be determined experimentally using, for example, ellipsometry or polarization based methods [238]. Figure 4-28-b, shows the corresponding measured angular selectivity for the transmitted wave. The small shift from the Bragg angle is a direct result of shrinkage of the emulsion after chemical processing.

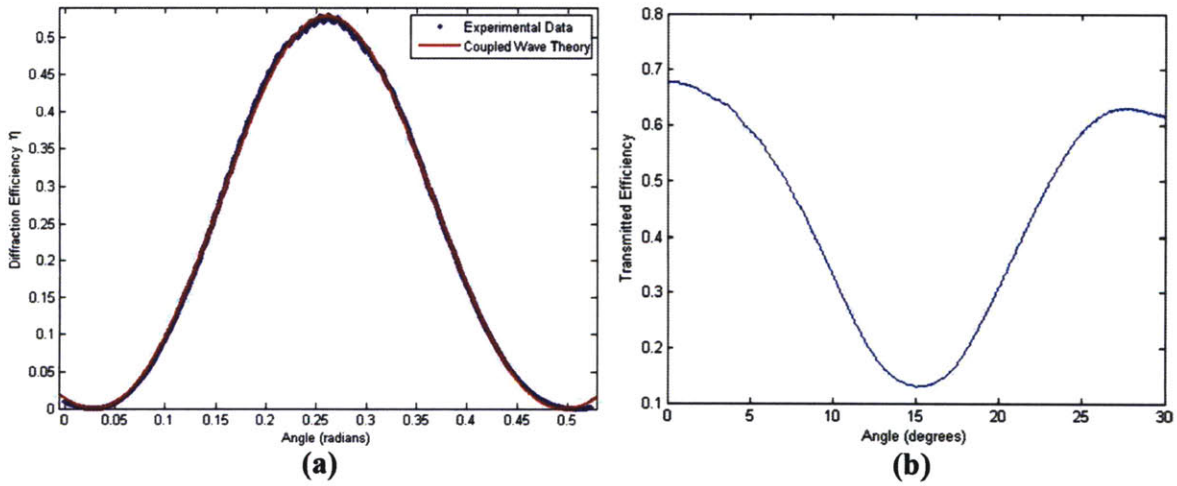


Figure 4-28: (a) Comparison of measured and theoretical diffraction efficiencies; (b) Measured angular selectivity of the transmitted wave.

A similar study is performed on holograms processed using the SHSG method. Taking into account Fresnel reflections, the highest measured diffraction efficiency is 82%. However, it is found that the results from this process are not fully repeatable due to the unstable chemical reaction. It is also found that the processed holograms suffer from higher shrinkage than those processed using the FFR method. This is due to the material removal that occurs during the fixation process. A further characterization and optimization of the chemical recipe is required. Control parameters include the influence of time spent at each bath as well as the temperature of the solution.

4.8.2 Measurement and Evaluation of the System Point Spread Function

The performance of the optical system can be fully characterized by the measurement of the space-variant PSF of the optical system. An automatic test station for the measurement and evaluation of the PSF of an uncorrected and corrected GRIN lenses is implemented. The measurement technique used is based on the knife-edge test, originally proposed by Foucault in 1859 for testing concave mirrors [239]. Since then, this method has been extended for the characterization of optical aberrations in a variety of optical systems [240], [241]. In the traditional Foucault knife-edge test, the optical system under study is illuminated by a collimated beam and a razor blade is used to scan the output beam on the image space. The razor blade is positioned in front, at, and after the focal plane. A screen or photodetector is used to capture the shadowgram corresponding to each position of the razor blade. An example of this test is shown in Figure 4-29-a for the evaluation of a single ideal and aberrated lens. In the case of an ideal lens, the shadowgram produced when the razor blade is positioned before or after the focal plane has an opposite orientation than the razor blade. If the razor blade is placed at the focal plane and scanned laterally, the corresponding shadowgram is either all bright or all dark. In the case of an aberrated lens, the shadowgram can be used to infer the degree of aberration present. Figure 4-29-b shows the shadowgrams corresponding to an

uncorrected GRIN lens subject to on-axis and off-axis (10 degrees) illumination.

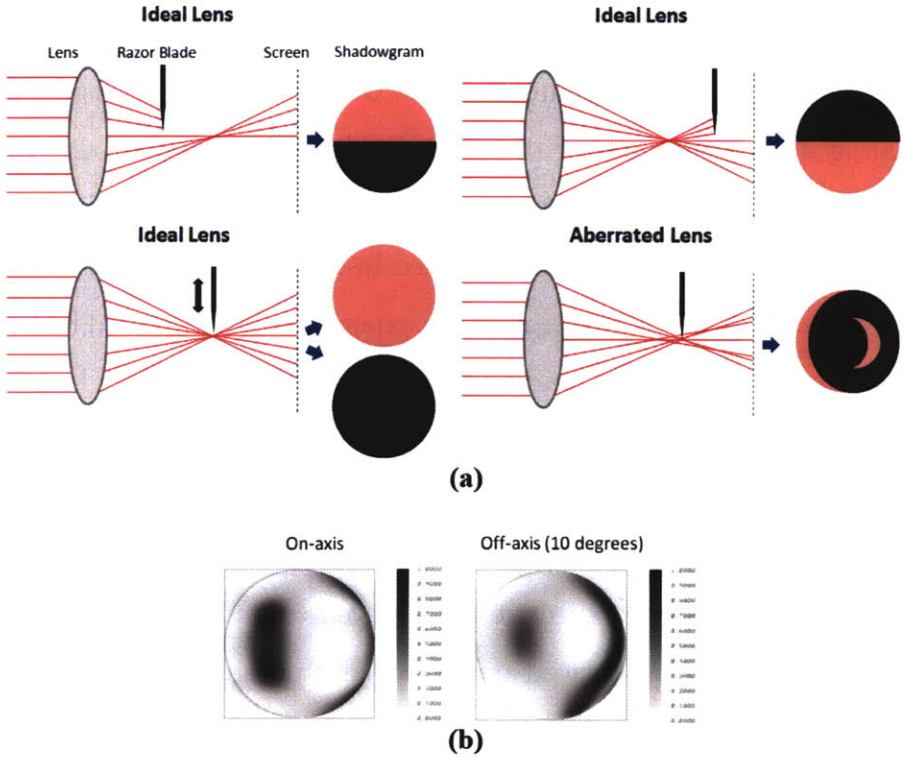


Figure 4-29: (a) Foucault knife-edge test; (b) Shadowgrams of uncorrected and corrected GRIN lenses.

To measure the PSF of the GRIN lens we used a variation of the knife-edge test explained above. Instead of capturing the shadowgram, we measure the optical power integrated by a single photodetector while scanning the razor blade laterally and axially. The axial scan allows us to find the location of the focal plane. The measured power during a single lateral scan corresponds to the cumulative sum (or integral) of the projected intensity distribution along a line perpendicular to the scanning direction. This can be expressed mathematically as,

$$P_{meas}(l) = \int_{-\infty}^{l} h^{(\theta)} dl, \tag{4.28}$$

where $h^{(\theta)}$ is the incoherent PSF projected along a line L with angle θ. For example,

if the edge of the razor blade is oriented parallel to the y-axis, the incoherent PSF is projected to the x-axis and the cumulative sum is measured. To reconstruct the system's PSF, the first derivative of the measured power is computed and used to populate an $M \times N$ tensor where M is the number of discrete power measurements obtained from a single scan and N is the number of projection directions measured. An inverse Radon transform [221] is performed on the resulting tensor to recover the 2D PSF, h. This reconstruction process is similar to that used in X-ray tomography.

Figure 4-30-a shows a photograph of the experimental setup used for the measurement and evaluation of the corrected and uncorrected GRIN lens PSFs. Two linear motorized stages (Newport UTM50CC1DD) with positional accuracy of 1μm are used for the axial and coarse later scans. A 3-axis piezo stage (Thorlabs MDT630A) is used to perform the fine later scan. The GRIN lens is illuminated with a collimated beam and the output light is collected by a microscope objective and relayed towards the photodetector (Newport 2832-C). The GRIN lens is placed on a specially designed mount that also holds the holographic film and is secured on top of a motorized rotation stage. The rotation stage allows measuring the projected PSF at different angles that can be then reconstructed using the inverse Radon transform. A graphical unit interface implemented in LabView is designed to automate the PSF evaluation process and coordinate the execution of all the motion stages and photodetector. Figure 4-30-b shows a screenshot of the implemented Foucault knife edge test interface.

We first present experimental results on the characterization of an uncorrected GRIN lens subject to on-axis illumination. Figure 4-31-a shows the raw measured power for an axial and lateral scans with steps: $\Delta z = 1\mu$m, $\Delta x = 20$nm. The raw data is then filtered using the power window of equation 4.22 to reduce the measurement noise. A 2D bilinear interpolation of the filtered data is performed and the result is shown in Figure 4-31-b. The data processing scheme allows minimizing the amplification of noise when taking the derivative. A finite difference approximation is implemented to compute the space derivative of the processed measured data along the x-direction. Figure 4-32-a shows

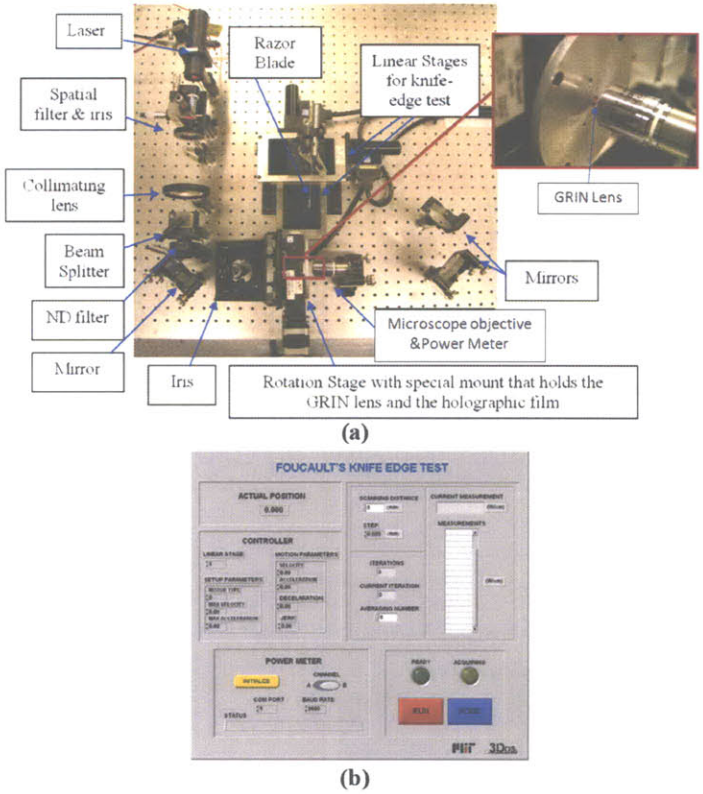


Figure 4-30: PSF evaluation station: (a) Optical setup; (b) GUI.

the resulting projected PSF. To estimate the location of the focal plane, the maximum gradient value along the x-direction is computed for every axial position. The peak of the gradient cue indicates the position of the focal plane as shown in Figure 4-32-b. The gradient cue can also be used to estimate the effective depth-of-focus (DOF). The effective DOF is defined as the range of the gradient cue with values larger than 0.9. For the example shown: $DOF_{eff} = \pm 2.48\mu$m. The corresponding theoretical DOF is: $DOF = \pm 2.37\mu$m. As the GRIN lens is illuminated on-axis, the PSF is expected to be axially symmetrical. The inverse Radon transform is computed using the projected PSF data at the focal plane. Figure 4-32-c shows the reconstructed 2D normalized PSF and a comparison between its cross-section and the simulated PSF using Zemax. The measured PSF matches very closely the expected response from the model.

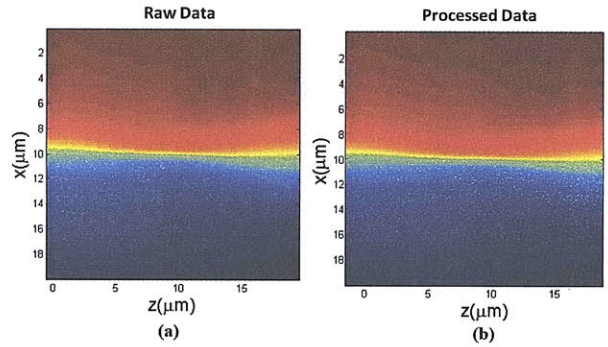


Figure 4-31: Measured intensity: (a) Raw data; (b) Processed data.

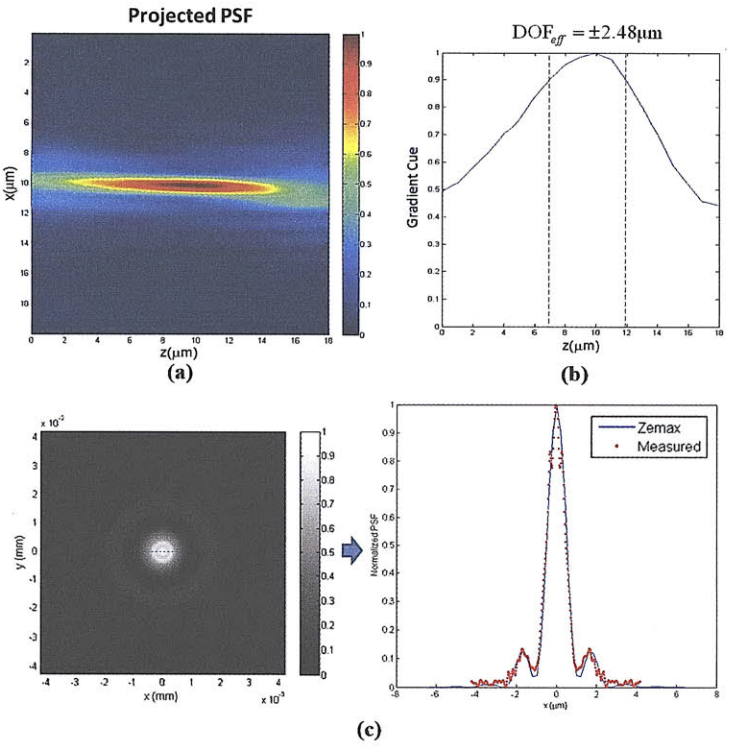


Figure 4-32: (a) Computed projected PSF; (b) Focal plane estimation process; (c) Reconstructed PSF.

In the second set of experiments, the GRIN lens is holographically corrected for on-axis illumination. A photograph of the experimental setup is shown in Figure 4-33-a. Figure 4-33-b shows a photograph of a small mount designed to hold the GRIN lens, holographic film and prism for the easy recording and realigning with micron resolution of the processed hologram. The hologram is recorded on a silver halide film and is chemically processed with the optimized technique presented in the previous section. The lens response is measured using the same procedure as described for the uncorrected GRIN lens. Figure 4-34 shows the reconstructed 2D normalized PSF and a comparison between its cross-section and the simulated PSF using Zemax. Again, the measured PSF closely matches the expected response from the model. The corrected PSF has decreased sidelobes and has a higher energy contained within the main lobe as quantified by the Strehl ratio. The effect of holographic correction is even more pronounced for off-axis angles.

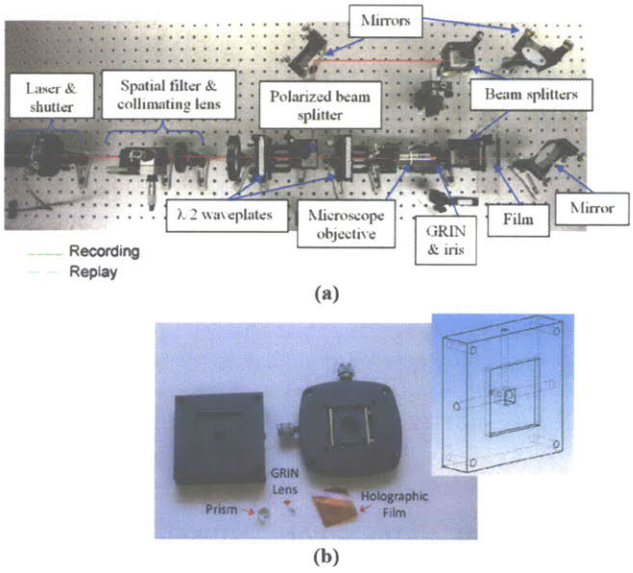


Figure 4-33: (a) Experimental setup; (b) Hologram-GRIN lens mount.

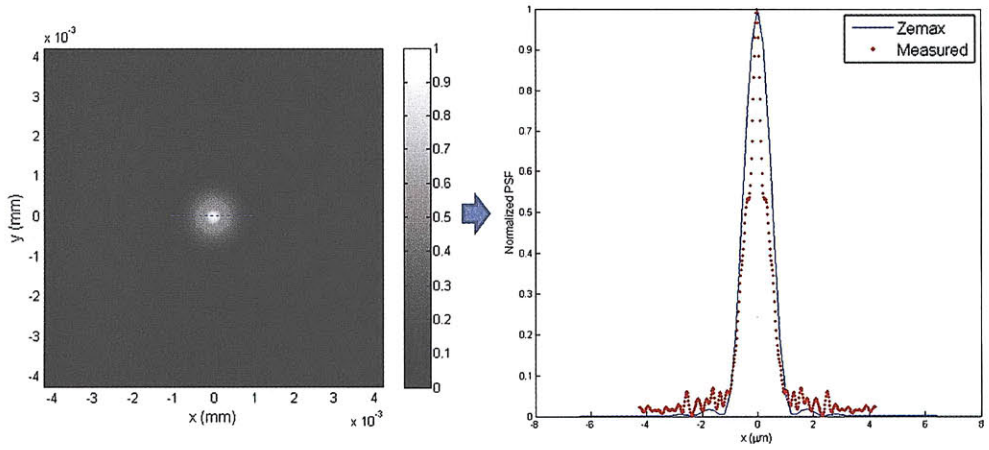


Figure 4-34: Reconstructed corrected PSF.

4.8.3 Sensitivity Analysis

A sensitivity analysis is performed to estimate the effect of potential misalignment errors of the holographic film after chemical processing. Two types of errors are considered: axial and lateral misalignments. Perfect phase conjugation fails if the holographic film is misaligned during the implementation or reconstruction process. The Zemax model described above is used to simulate the resulting PSF for different degrees of axial and lateral misalignment. The error metric used is the $L2$ norm of the difference between the diffraction limited (no misalignment) and aberrated (misaligned) PSFs. To estimate the misalignment tolerance we decouple the axial and lateral misalignments and study the performance of a GRIN lens corrected for on-axis field. Figure 4-35 shows the computed error map with its corresponding isometric view. The estimated tolerance to misalignment is: $\Delta x = \Delta y = \pm 40\mu$m. This level of misalignment can easily be avoided with the designed mount of Figure 4-33-b. The error plot of sensitivity to axial misalignment is shown in Figure 4-36. As the hologram is recorded with a planar-like object wave, it can tolerate large axial deviations without sensibly affecting the lens' performance. This high tolerance to axial misalignment is exploited by removing the beam splitter used during the recording step and placing the hologram in direct contact with the GRIN

lens, reducing the thickness of the imager.

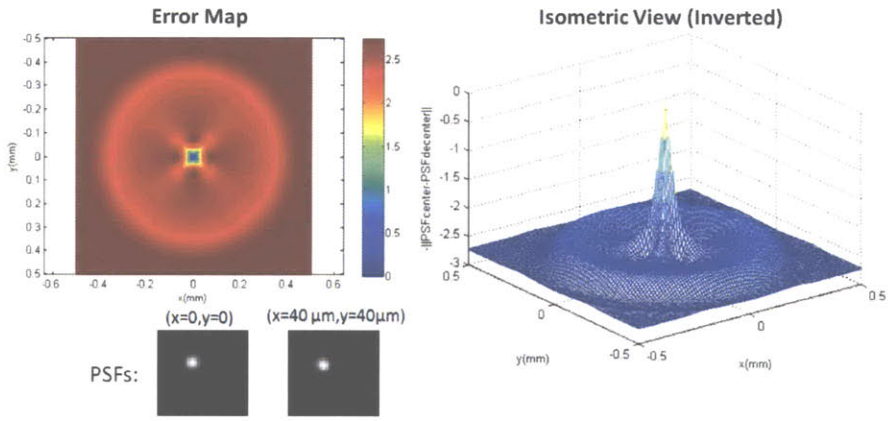


Figure 4-35: Lateral misalignment analysis.

4.8.4 Geometry for Color Imager

Figure 4-37 shows the proposed geometry for polychromatic implementation of the segmented aperture thin imager. A red, green and blue color filter array, such as a Bayer filter, is used to decompose the different color channels. The phase conjugated holograms are recorded on a panchromatic film sensitive to a wide spectral range, such as Slavich's PFG-03C (spectral sensitivity between 457-700nm). Each elemental image produced by a lens in the array is processed in a similar way as the one described above to reconstruct a high-resolution color image. A similar approach is proposed in [242] for conventional microlens array-based systems.

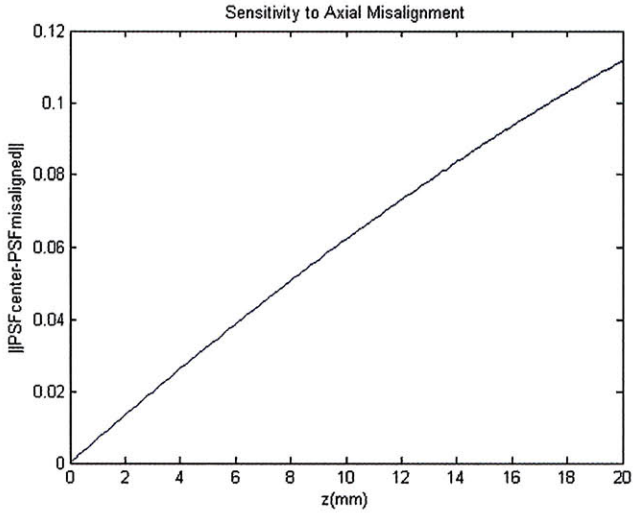


Figure 4-36: Axial misalignment analysis.

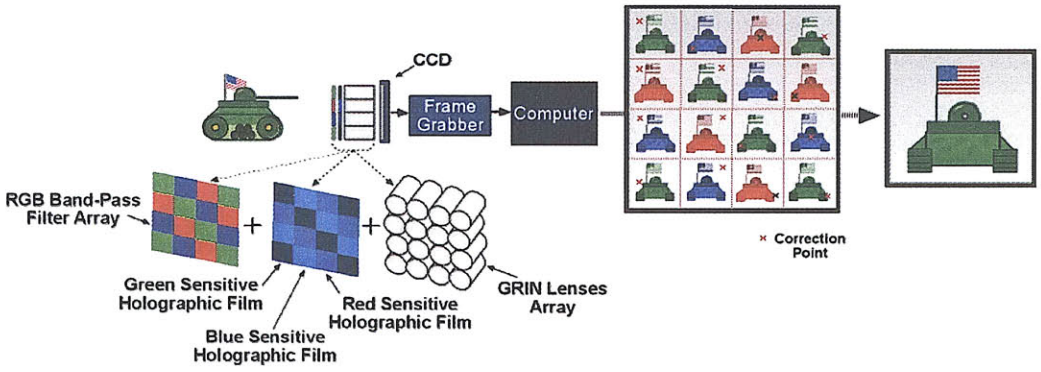


Figure 4-37: Polychromatic holographically corrected segmented aperture thin imager.

Chapter 5

Stability Metric for the Design and Optimization of Digital Holographic Particle Imaging Velocimetry Systems

In the previous chapter, the optimization of a holographically corrected segmented aperture thin imager is presented. The imager is modeled using system's theory by the evaluation of the space-variant point spread function (PSF). The model is written in matrix form defining the structure of the Hopkins transfer matrix. This matrix connects the input and output (object and measured intensities) and includes geometrical, optical and digital detector parameters. The proposed system differs from conventional compound systems in that the GRIN lens array is holographically corrected for optical aberrations using phase conjugation holography. The configuration of the holographic correction positions introduces a new degree of freedom that modifies the structure of the Hopkins matrix resulting in an alternate projection. It is proved that by an adequate system design and control of the correction positions for each lens in the array, the stability for inversion of the Hopkins matrix and the channel capacity of the system are significantly

improved. The resulting alternate projection allows extraction of the scene information more efficiently than conventional isomorphic methods. The system is studied using an information theoretic framework in which the object or scene is treated as a message, the optical system (including digital detector) as the communications channel, and the captured image as the received message. The communication channel is assumed to be Gaussian. A multi-domain optimization (MDO) approach based on genetic algorithms is implemented. The optimization is performed to maximize the system's channel capacity and improve the stability (reduce the condition number) of the Hopkins matrix. An example of a decoding strategy based on the superresolution algorithm is presented. Experimental results and a sensitivity analysis for the design and optimization of the proposed segmented aperture imager are described.

In this chapter, we present a stochastic theoretical model that is used for the design and optimization of digital holographic particle imaging velocimetry systems. Similar to the previous chapter, the encoding strategy is optimized based on information theory to maximize the amount of 3D information from the sample volume encoded by a single captured hologram. The system is modeled as a Gaussian communication channel and is written in matrix form using the Hopkins matrix formulation. The Hopkins matrix comprises an alternate projection of the optical system and includes geometrical, optical and detector related parameters such as number of particles, particle sizes, pixel size and space bandwidth product. In contrast to the holograms presented in previous chapters, the holograms studied here are recorded optically using a digital photodetector and reconstructed numerically. The system model is first presented for the limiting case of point source particles. An extension of this model for particles of various sizes using Mie theory is presented. The developed model is used for optimizing the system based on a stability metric (channel capacity) to find optimum system parameters such as the particle density that leads to maximum information transfer. This is another form of MDO. Simulations as well as experimental results are presented.

5.1 Motivation and Problem Definition

The study of particle distributions using holographic methods for the description of complex 3D non-stationary flow fields has become increasingly popular. In these methods, a fluid volume is seeded with several micron-sized tracing particles, each corresponding to a sample point, which are then measured to retrieve their positions, velocity vectors and trajectories for the characterization of the flow field under study. Recent advancement in experimental and computational fluid dynamics research has increased the demand for instantaneous full-field 3D flow velocity measurements with high spatial and temporal resolutions. Holographic methods offer important advantages over conventional particle image velocimetry (PIV) methods in that information from a complex 3D field can be efficiently encoded by the hologram and can be later decoded for detail post-analysis. In addition, the information capacity provided by the hologram is much higher than that of conventional imaging systems as the phase and amplitude distributions of the optical field are recorded. Traditional PIV methods measure two in-plane components of fluid velocities in a 2D planar domain using a laser light sheet illumination and a conventional imaging system based on photographic films or digital detectors [243]. This technique has been widely used to study complex fluid phenomena such as turbulent flows [244]. Stereoscopic imaging geometries are used to measure three velocity components in a planar domain at the cost of a more complicated optical setup [245]. Despite the advents of traditional PIV methods, only 2D multi-point velocity measurements are possible. Attempts have been made to try to generalize PIV methods for 3D volumetric fields through scanning with severe limitations in spatial and temporal resolution as well as can only be applied to characterize static flows [246]. In particular, measurements of turbulent and complex flows require high accuracy and high spatio-temporal resolutions over large volumetric domains.

Holographic particle imaging velocimetry (HPIV) methods are broadly classified as: conventional and digital recording. In conventional HPIV systems, the holograms are recorded optically on a holographic material, such as the silver halide film studied in

the previous chapter. The hologram is the result of the interference between a reference wave and the wave scattered by the cloud of particles (object wave). Systems based on in-line [247] and off-axis [248] geometries have been studied. As explained in Chapter 2, the off-axis geometry has the advantage that the reconstructed desirable and undesirable diffraction orders are spatially separated and can be filtered out, resulting in reconstructions with a higher signal-to-noise ratio (SNR). In addition, off-axis holographic systems tolerate higher particle densities and have increased resolution and shorter depth of field due to the increased effective numerical aperture from particle side scatterings. However, off-axis systems require more complicated optical setups with powerful laser sources that have high coherence length. Also, the produced holograms are more susceptible to fabrication errors such as shrinkage of the holographic emulsion. As a compromise, hybrid systems have been developed that involve in-line recording and off-axis reconstruction [249]. Holographic systems have also been developed based on high-power lasers for side and back-scattering measurements of particle clouds [250]. The holograms recorded by conventional HPIV systems are chemically processed and reconstructed optically using the conjugate of the reference wave. The real image of the particle field is produced in space and is measured using, for example, a scanning procedure based on a digital photodetector or flat scanner. This process is extremely inefficient due to the large amount of data recorded by the hologram that needs to be analyzed. To obtain velocity vectors, two holograms are recorded at different time instances using, for example, a double exposure or multiplexing techniques. A numerical algorithm is used to correlate the particles from different holograms and estimate their velocity vectors. Several algorithms have been implemented, such as correlation-based [251], genetic algorithms [252] and particle reconstructed by edge detection methods [253].

Digital holographic particle image velocimetry (DHPIV) systems are developed to overcome the efficiency limitations in handling large amounts of data, complicated chemical processing, elaborate optical setups and cost related problems that arise in traditional HPIV systems [254]. The holograms captured by DHPIV systems are recorded

on a digital photodetector such as a CCD or CMOS sensor. Digital holograms have the advantage that the recorded information is already digitized and ready for numerical post-processing. This allows capturing holograms of fast moving flows at high frame rates. However, the performance of digital holography is limited by the detector's space-bandwidth product (SBP), pixel size, dynamic range, and quantum efficiency. In particular, the diffraction limit resolution as well as the ability to accurately recover the particle positions is mainly limited by the detector's SBP and pixel size. The system suffers from relatively large depth of focus compared to the attainable lateral resolution [255]. Also, the information capacity of digital photodetectors is less than that of conventional holographic materials. The large pixel size imposes a constraint on the maximum off-axis angle that can be sampled before aliasing. For a photodetector with a pixel size of 10μm, the maximum off-axis angle is approximately 1 degree (for $\lambda = 632$nm). Such a small angle results in insufficient separation of the desirable and undesirable diffraction orders, limiting the bandwidth of the recordable signal. For this reason, most DHPIV systems operate using the in-line geometry [9], [256]. The hologram reconstruction step is done numerically. Several reconstruction algorithms have been developed such as the convolution method [7] and reconstruction based on wavelet transforms [257], [258]. The convolution method is a back-propagation algorithm based on the Fresnel transfer function of equation 2.48 similar to that used for the simulation of the diffracted field based on scalar diffraction theory explained in Chapter 2. The velocity vectors are computed by numerically processing the reconstructions from two frames separated by a known period of time. Similar algorithms to those used for conventional HPIV systems are used to locate and correlate the tracing particles. Additional applications of DHPIV systems include measurement of water-air mixtures [259], microgravity experiments [260], plankton imaging [7] and airborne systems [261]. In addition, a multiplexing method for the acquisition of large SBP holograms with high temporal resolution is proposed [262].

Despite that DHPIV systems have been used widely in a variety of applications, there is still very limited understanding about how system parameters such as particle density,

particle size and distribution, volume size, recording geometry, number of pixels, pixel size, detector's dynamic range and noise affect the performance of the system. All these parameters are coupled together and affect the system's ability to extract information efficiently from the object space (particle cloud). In particular, it has been experimentally demonstrated that the choice of particle density affects the performance of the reconstruction algorithm [263]. High particle densities are desirable for multi-point sampling of complex flows. However, high densities affect the hologram quality making it difficult to retrieve the particle positions. Early studies of the influence of particle densities in conventional HPIV systems ignored the influence of the virtual image and detector with finite SBP (film recorded holograms) [264], [265]. A later study for DHPIV systems showed that that the volume depth and shadow density affect the percentage of extracted particles [266]; however, they didn't account for number of pixels, pixel size, and dynamic range. The influence of virtual images on the SNR was also studied [267]; however, the particle diameter, distribution and detector's dynamic range were ignored.

From our literature search, we conclude that currently there is no model available which includes all the important system parameters that can assist in the optimization of a DHPIV system. We propose a stochastic model based on information theory that can be used for the design and optimization of DHPIV systems. Similar to the previous chapter, the system model is mainly characterized by a Hopkins matrix, which includes all the important system parameters mentioned above. A stability metric given by the system's channel capacity under the assumption of Gaussian probability distribution is used to maximize the amount of information that can be extracted from the sampled volume. This choice of metric allows improving the stability of the associated inverse problem during the particle recovery or reconstruction processes.

The analyzed problem is equivalent to maximizing the performance of the hologram encoding process. This is illustrated in Figure 5-1 for the recording of an in-line hologram. A plane wave of wavelength, λ, propagating parallel to the optical axis illuminates a volume, V, seeded with tracing particles of size, d_p, at a density, ρ. The wave scattered by

the particles interferes at the hologram plane with the undiffracted wave and is recorded on a CCD sensor with M pixels of size δ_{pix}, dynamic range D, and added electronic Gaussian noise $\sigma_{Gaussian}$. In addition, there is an associated particle density dependent noise, I_N. An abstraction of the encoding problem can then be as follows: maximize the amount of information of the 3D object space (number of particles) that is encoded on a 2D discrete hologram plane subject to fixed system parameters (or constraints). Each particle corresponds to a sampling point within the volume of interest (VOI) and thus, adding more particles increases the amount of information that describes the object space. However, due to the system's limited channel capacity and particle density dependent noise, there is an optimum number of particles that can be encoded and decoded (reconstructed) with arbitrarily low probability of error. This represents a tradeoff between particle density and information capacity and will be analyzed in more detail in the next section. Similar tradeoffs respect to other system parameters can be obtained from the proposed model. The maximum stability metric represents an upper bound for performance of a given system, which maximizes the hologram encoding process. This upper bound is independent of our choice of decoding strategy. The decoding strategy is application dependent and its performance is related to the choice of reconstruction algorithm selected to process the captured holograms. Different applications required specific information to be extracted from the hologram, such as number of particles, particle positions, distribution of positions or size and shape. The good choice of decoding strategy might lead to a performance closer to the upper bound given by the stability metric. By optimizing the system parameters and hence the stability metric, the performance of the decoding strategy is also improved as the associated inverse problem becomes more stable as proved in the previous chapter.

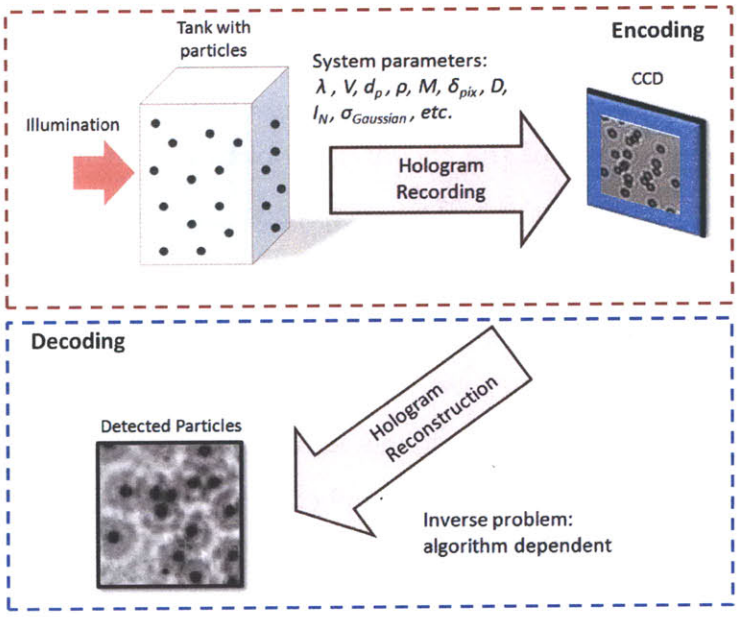


Figure 5-1: Equivalent problem: encoding and decoding.

5.2 Theoretical Model: Point Source Particles

The theoretical model is first derived for the extreme case of point source particles. The model geometry is shown in Figure 5-2. The VOI is illuminated with a uniform quasi-monochromatic spatially coherent plane wave propagating parallel to the optical axis with electric field: $E_i = U_i \exp(ikz)$, where U_i is the wave's amplitude and $k = 2pi/\lambda$. Point particles are randomly distributed inside the VOI and scatter a spherical wave of the form: $E_p = U_p \exp(ikr)$, where U_p is the amplitude of the spherical wave, and $r = \sqrt{x^2 + y^2 + z^2}$ is the radial position from the center of the particle to an arbitrary point in the image plane. The spherical waves scattered by each particle interfere at the image plane with an in-line plane reference wave of the form: $E_r = U_r \exp(ikz)$, where U_r is proportional to U_i. The resulting interferogram is recorded on a digital detector with M pixels of size: $\delta_{pix} \times \delta_{pix}$. The object space (VOI) is divided into V voxels of size: $\delta_x^{voxel} \times \delta_y^{voxel} \times \delta_z^{voxel}$. Each voxel is constrained to only contain a single particle. The minimum voxel size is then: $\delta_x^{voxel} = \delta_y^{voxel} = \delta_z^{voxel} = d_p$, where d_p is the particle's

diameter. By controlling the voxel size, the mean distance between particles is also varied. A binary random variable, $a^{(v)}$, is assigned to each voxel,

$$a^{(v)} = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}, \tag{5.1}$$

where p is voxel's probability of containing a particle. A higher probability value results in higher number of particles inside the VOI. The particle density is then related to the voxel size and its probability of success. The mean expected number of particles is: $N_{mean} = pV$. The corresponding variance is: $N_{var} = Vp(1-p)$.

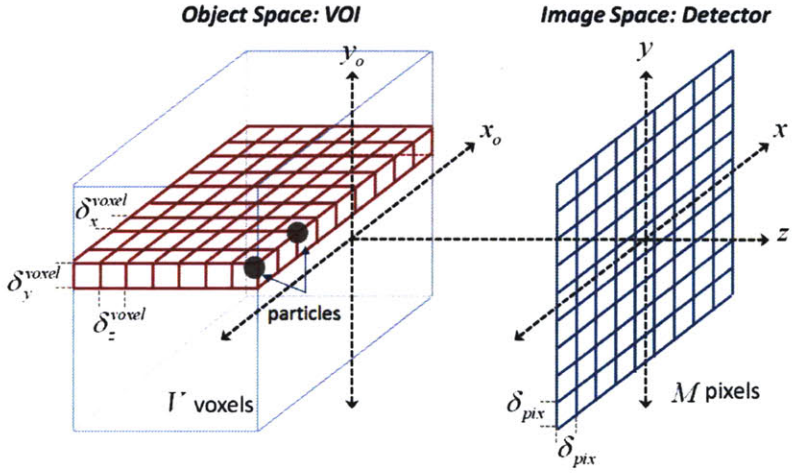


Figure 5-2: Problem geometry.

To analyze the field scattered by a point particle, a Rayleigh scattering formulation is adopted. Rayleigh scattering characterizes the scattering of electromagnetic waves by particles much smaller than a wavelength. This is quantified by: $\alpha << 1$, where α is a characteristic length given by, $\alpha = k(d_p/2)$. The scattered electromagnetic field is computed by modeling the homogeneous dielectric particle as a radiating Hertzian dipole and enforcing the appropriate boundary conditions [153]. In general, the scattered electric

field, $\mathbf{E}^s$, is related to the incident electric field, $\mathbf{E}^i$, by means of the scattering matrix,

$$\begin{bmatrix} E_\theta^s \\ E_\varphi^s \end{bmatrix} = \frac{e^{-ikr+ikz}}{ikr} \begin{bmatrix} S_2 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} E_\theta^i \\ E_\varphi^i \end{bmatrix}, \tag{5.2}$$

where the incident and scattered electric fields are expressed in spherical coordinates; θ is the angle measured from the positive z-axis; φ is the azimuth angle; S_1 and S_2 are the scattering functions. E_θ, and E_φ are the components of the electric vector polarized parallel and perpendicular to the scattering plane. The scattering plane is defined as that containing the $\mathbf{k}_i$ vector of the incident wave and the position vector, $\mathbf{r}$. Under the Rayleigh scattering assumption, S_1 and S_2 take the form,

$$\begin{aligned} S_1 &\approx -i\alpha^3 \left[\frac{n^2-1}{n^2+2}\right], \\ S_2 &\approx -i\alpha^3 \left[\frac{n^2-1}{n^2+2}\right] \cos\theta, \end{aligned} \tag{5.3}$$

where n is the particle's refractive index. The corresponding intensity distributions for the parallel and perpendicularly polarized waves are,

$$\begin{aligned} I_\perp &= \frac{I_o^\perp \alpha^6}{k^2 r^2} \left[\frac{n^2-1}{n^2+2}\right]^2 = \frac{I_o^\perp k^4 r_p^2}{r^2} \left[\frac{n^2-1}{n^2+2}\right]^2 = \frac{I_o^\perp}{r^2}\sigma_s, \\ I_\parallel &= \frac{I_o^\parallel k^4 r_p^2}{r^2} \left[\frac{n^2-1}{n^2+2}\right]^2 \cos^2\theta = \frac{I_o^\parallel}{r^2}\sigma_s \cos^2\theta, \end{aligned} \tag{5.4}$$

where $I_o^\perp$ and $I_o^\parallel$ are the intensity distributions for the vertically and horizontally polarized incident field, r_p is the particle's radius, and σ_s is the scattering cross-section given by,

$$\sigma_s = k^4 r_p^6 \left[\frac{n^2-1}{n^2+2}\right]^2. \tag{5.5}$$

In general, the polarization of the scattered wave is elliptical. In contrast, the polarization of the reference wave is assumed to remain linearly polarized in a direction similar

to that of the incident wave. From equation 5.4 we see that for a perpendicularly polarized incident wave, the scattered power is uniformly distributed around a sphere. In our model, a vertically polarized incident wave is assumed. This state of polarization maximizes the fringe visibility over extended scattering angles. The total power scattered by the wave is found by integrating the intensity of equation 5.4 over a sphere $P_s^{tot} = 4\pi I_o \sigma_s$. The corresponding power over a small spherical cap with angular acceptance, θ, is,

$$\begin{aligned} P_s^{cap} &= P_s^{tot}\left(\frac{A_{cap}}{A_{sphere}}\right) = P_s^{tot}\left(\frac{4\pi r^2 \sin^2(\theta/2)}{4\pi r^2}\right) \\ &= 4\pi I_o \sigma_s \sin^2(\theta/2)\,. \end{aligned} \tag{5.6}$$

For a particle centered in the VOI and located at a distance, z_p, from the detector, the effective numerical aperture is: $NA = \sin\theta = \sin[\arctan(X_{size} z_p/2)]$, where $X_{size} = \sqrt{M}\delta_{pix}$, is the detector size. Under the paraxial approximation ($x, y << z_p$) the spherical cap becomes planar and coincides with the detector. Using equation 5.6, the power form the particle captured by the detector is,

$$P_s^{\det} = \pi I_o \sigma_s \theta^2 = \frac{\pi I_o \sigma_s X_{size}^2}{2 z_p^2}. \tag{5.7}$$

The amplitude of the scattered spherical wave is found by dividing equation 5.7 over the area of the detector,

$$\begin{aligned} U_p^2 &= \frac{\eta (U_i)^2}{z_p^2} \\ &\rightarrow U_p = \frac{\sqrt{\eta} U_i}{z_p}, \end{aligned} \tag{5.8}$$

where $\eta = (\pi\sigma_s)/2$ is a small number and is assumed to be constant over the x-y plane. Also, η is assumed to be the same for all the particles in the VOI. The resulting field at

the mth pixel scattered by the vth particle is,

$$E_p^{(m,v)} = \frac{\sqrt{\eta} U_i a^{(v)}}{z_p^{(v)}} \exp\left[i\phi^{(m,v)}\right], \tag{5.9}$$

where,

$$\phi^{(m,v)} = k\left[z_p^{(v)} + \frac{\left(x^{(m)} - x^{(v)}\right)^2 + \left(y^{(m)} - y^{(v)}\right)^2}{2z_p^{(v)}}\right]. \tag{5.10}$$

The amplitude of the reference wave is found using power conservation,

$$\begin{aligned} P_i &= P_r + \sum_{v=1}^{V} a^{(v)} P_s^{tot}, \\ &\rightarrow P_r = U_r^2(X_{size}^2) = U_i^2(X_{size}^2) - U_i^2 4\pi\sigma_s N, \\ &\rightarrow U_r = U_i\left[1 - \frac{4\pi\sigma_s N}{X_{size}^2}\right]^{1/2}, \end{aligned} \tag{5.11}$$

where N is the total number of particles inside the VOI for a given random experiment: $N = \sum_v a^{(v)}$. In deriving equation 5.11, we assumed that the incident illumination for each particle remains constant. This assumption is valid for small particles and relatively low densities. For larger particles or higher densities, the power of the wave that illuminates a given particle decreases progressively while propagating through the tank.

Using the previous results, we compute the intensity recorded by the detector,

$$\begin{aligned} I^{(m)} &= \left|U_r + \sum_{v=1}^{V} U_p^{(v)} \exp\left[i\phi^{(m,v)}\right]\right|^2 \\ &= |U_r|^2 + \sum_v \left|U_p^{(v)}\right|^2 + 2U_r \sum_v U_p^{(v)} \cos\left[\phi^{(m,v)}\right] \\ &\quad + \sum_v \sum_{v':v\neq v'} U_p^{(v)} U_p^{(v')} \exp\left[i\left(\phi^{(m,v)} - \phi^{(m,v')}\right)\right] \\ &= \kappa\left[1 + \frac{2U_r}{\kappa} \sum_v U_p^{(v)} \cos\left[\phi^{(m,v)}\right] + \frac{1}{\kappa} I_N^{(m)}\right], \end{aligned} \tag{5.12}$$

where $\kappa = |U_r|^2 + \sum |U_p^{(v)}|^2$; and,

$$I_N^{(m)} = \sum_v \sum_{v':v\neq v'} U_p^{(v)} U_p^{(v')} \exp\left[i\left(\phi^{(m,v)} - \phi^{(m,v')}\right)\right]. \tag{5.13}$$

The first two terms of equation 5.12 correspond to the direct component (DC) and halo. The third term contains the information of the signal to be recovered. The fourth term is the cross-talk (interference) between every particle pair. In this analysis, the cross-talk term is treated as noise that does not add any useful information about the measured signal. This term will be referred to as the cross-talk noise and its properties will be discussed later.

The term κ is proportional to the total detected power. Under the weak scattering object wave assumption ($|U_o|^2 << |U_r|^2$, where U_o is the amplitude of the object wave that includes the contributions from all the particles), and for $U_i = 1$: $\kappa \approx 1$. Equation 5.12 then becomes,

$$I^{(m)} = \sum_{v=1}^{V} a^{(v)} \left[\frac{1}{N} + \frac{2U_r\sqrt{\eta}}{z_p^{(v)}} \cos\left(\phi^{(m,v)}\right)\right] + I_N^{(m)}. \tag{5.14}$$

Equation 5.14 can be written in matrix form as: $\mathbf{I} = H\mathbf{a} + \mathbf{I}_N$, where $\mathbf{I}$ is an $M \times 1$ vector of the raster scanned measured intensity, $\mathbf{a}$ is a $V \times 1$ vector with zeros and ones representing the particle locations inside the VOI at an instance in time, $\mathbf{I}_N$ is an $M \times 1$ vector of the raster scanned cross-talk noise, and H is the $M \times V$ Hopkins matrix of the system that connects the input and output signals. The elements of the Hopkins matrix are obtained from the following equation,

$$h^{(m,v)} = \frac{1}{N} + \frac{2\sqrt{\eta}\left[1 - \frac{8\eta N}{X_{size}^2}\right]^{1/2}}{z_p^{(v)}} \cos\left(k\left[z_p^{(v)} + \frac{\left(x^{(m)} - x^{(v)}\right)^2 + \left(y^{(m)} - y^{(v)}\right)^2}{2z_p^{(v)}}\right]\right). \tag{5.15}$$

Similar to the Hopkins matrix derived in the previous chapter, the resulting Hopkins matrix from equation 5.15 contains the impulse response of the systems and includes

optical, geometrical, and detector parameters such as number of pixels and pixel size. Additional parameters, such as detector's dynamic range, can easily be included by quantizing the corresponding intensity values. The general structure of the Hopkins matrix is shown in Figure 5-3. The system's impulse response of equation 5.15 is shift-invariant over the $x-y$ plane and shift-variant over the axial direction. By properly raster scanning the VOI, the Hopkins matrix can be arranged in a block Toeplitz form.

The Hopkins matrix characterizes the system at a given instant in time in which the particle positions are obtained from the random vector **a**. This vector contains ones and zeros and multiplies the Hopkins matrix during the forward problem. The columns of H that are multiplied by the elements of **a** that are zero do not contribute any information to the measurement. This problem is equivalent to modifying the Hopkins matrix: $H \rightarrow \hat{H}$, where $\hat{H}$ has zero columns corresponding to the positions of empty voxels. The rank of $\hat{H}$ is, in general, lower than that of a full H (no zero columns) as it has several linearly dependent columns. The matrix $\hat{H}$ can also be collapsed by removing the zero columns and forming a smaller matrix of the same rank. This is justified as both matrices have the same singular values that are used for the stability metric. The matrix $\hat{H}$ will be referred to as *instantaneous Hopkins matrix*. To fully characterize the performance of the system under a fixed particle density, several random experiments are performed, each resulting in a different instantaneous Hopkins matrix.

To analyze the statistics of the cross-talk noise, equation 5.13 is rewritten as,

$$I_N^{(m)} = \sum_{l=1}^{P} A^{(l)} \cos \Delta\phi^{(m,l)}, \tag{5.16}$$

where the index $l = \{v, v\prime | v > v\prime\}$; $P = (|v|^2 - |v|)/2$; the amplitude and phase difference terms are,

$$\begin{aligned} A^{(l)} &= \frac{2\eta a_l^{(v)} a_l^{(v')}}{z_{p,l}^{(v)} z_{p,l}^{(v')}}, \\ \Delta\phi^{(m,l)} &= \phi_l^{(m,v)} - \phi_l^{(m,v')}. \end{aligned} \tag{5.17}$$

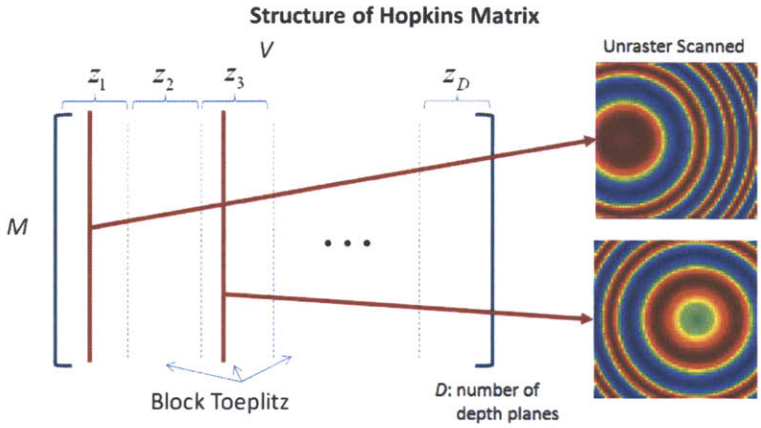


Figure 5-3: Structure of Hopkins matrix

The expectation value of equation5.16 is,

$$\begin{aligned} E\left[I_N^{(m)}\right] &= \sum_l E\left[A^{(l)}\right] \cos\Delta\phi^{(m,l)} \\ &= 2\eta p^2 \sum_l \frac{\cos\Delta\phi^{(m,l)}}{z_{p,l}^{(v)} z_{p,l}^{(v')}}. \end{aligned} \tag{5.18}$$

The mean cross-talk noise of equation 5.18 is composed of the superposition of several small amplitude sinusoids with varying phase shifts. It can be shown that for medium to large particle densities ($P \to \infty$): $E[I_N^{(m)}] \approx 0$. The second central moment of the cross-talk noise is given by,

$$\begin{aligned} E\left[I_N^{(m)^2}\right] &= \sum_l \sum_h E\left[A^{(l)} A^{(h)}\right] \cos\Delta\phi^{(m,l)} \cos\Delta\phi^{(m,h)} \\ &= \begin{cases} 0 & l \neq h \\ 4\eta p^2 \sum_l \left(\frac{\cos\Delta\phi^{(m,l)}}{z_{p,l}^{(v)} z_{p,l}^{(v')}}\right)^2 & l = h \end{cases}. \end{aligned} \tag{5.19}$$

By virtue of the central limit theorem we find that the cross-talk noise follows a Gaussian

probability distribution with zero mean and variance given by (equation 5.19),

$$\sigma^2 \approx 4\eta p^2 \sum_l \left(\frac{\cos \Delta\phi^{(m,l)}}{z_{p,l}^{(v)} z_{p,l}^{(v')}} \right)^2 . \tag{5.20}$$

As shown in Figure 5-4, not every particle contributes to the cross-talk noise at a point O, but only those that lie inside the effective volume, V_{eff}. This is because the cross-talk interference patterns produced by particle pairs outside the effective volume have frequencies higher than the sampling frequency and get integrated by the detector and produce a uniform background. To estimate the volume size, consider the interference pattern produced by two particles at points Q_1 and Q_1' separated by a distance Δx_1 ($y_1 = y_2 = 0$). The resulting interference pattern is,

$$\begin{aligned} I &= \left| U_p e^{i\phi_1} + U_p e^{i\phi_2} \right|^2 \\ &= 2U_p^2 \left[1 + \cos\left(k \frac{x_1^2 - x_2^2 + 2x\Delta x_1}{2z_{ref}} \right) \right] . \end{aligned} \tag{5.21}$$

The local frequency of the interference pattern is: $u_{loc} = \Delta x_1 / \mu z_{ref}$. At the Nyquist limit we get: $\Delta x_1 = \lambda z_{ref} / 2\delta_{pix}$.This corresponds to a half apex angle of: $\theta \approx \lambda / (4\delta_{pix})$. The resulting effective volume is,

$$V_{eff} = \frac{\pi \lambda^2}{48 \delta_{pix}^2} \left[D^3 + 3D \left(z_{ref}^2 + D z_{ref} \right) \right] . \tag{5.22}$$

The effective number of particles that contribute to the cross-talk noise term is: $N_{eff} = \rho V_{eff}$, where ρ is a given particle density.

The system model described above can also be considered as a Gaussian parallel channel similar to the information theoretic analysis of Chapter 3. In addition to the cross-talk noise, additive white Gaussian noise (AWGN) may be added to account for potential noise fluctuations from the detector. A stability metric similar to the channel capacity of equation 4.19 is defined to estimate the stability of the inversion of the system

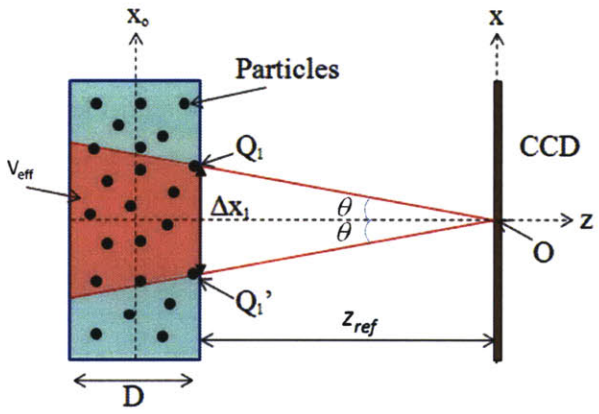


Figure 5-4: Effective volume contributing to the cross-talk noise.

model for a given instant in time. The stability metric is given by,

$$S = \sum\nolimits_i \ln\left(1 + \frac{\mu_i}{\sigma_{tot}^2}\right), \tag{5.23}$$

where μ_i is the ith singular value of the instantaneous Hopkins matrix, $\hat{H}$; σ_{tot}^2 is the total noise variance that includes the particle density dependent cross-talk noise (equation 5.20) and any other sources of noise such as AWGN. The ratio μ_i/σ_{tot}^2 again represents the signal-to-noise ratio (SNR) of the ith eigenmode of the system. The stability metric of equation 5.23 is used to maximize the information transfer from object space to image space subject to various system parameters such as particle density, geometry and detector characteristics.

5.3 Mie Theory: Particles of Various Sizes

The problem of a plane wave scattered by a spherical particle can be solved rigorously using Mie theory [153]. In such a formulation, particles of various sizes and materials can be modeled. We will restrict our analysis to spherical particles composed of homogeneous dielectric material. Consider the scattering from a single particle centered at the optical

axis as shown in Figure 5-5. The illumination wave is assumed to be vertically polarized and propagated in the z-direction: $E_i = \hat{x}U_i \exp(ikz)$. The electric field components perpendicular and parallel to the scattering plane are again found by equation 5.2. However, the scattering functions now take the form,

$$\begin{aligned} S_1 &= \sum_{q=1}^{\infty} \frac{2q+1}{q\,(q+1)} \left[a_q \pi_q\,(\cos\theta) + b_q \tau_q\,(\cos\theta)\right], \\ S_2 &= \sum_{q=1}^{\infty} \frac{2q+1}{q\,(q+1)} \left[b_q \pi_q\,(\cos\theta) + a_q \tau_q\,(\cos\theta)\right], \end{aligned} \tag{5.24}$$

where,

$$\begin{aligned} \pi_q\,(\cos\theta) &= \frac{1}{\sin\theta} P_q^1\,(\cos\theta), \\ \tau_q\,(\cos\theta) &= \frac{d}{d\theta} P_q^1\,(\cos\theta), \end{aligned} \tag{5.25}$$

and $P_q^1\,(\cos\theta)$ is the first degree, qth-order associated Legendre function of the first kind. The scattering functions of equation 5.24 are obtained by solving Maxwell's equations in spherical coordinates and the coefficients, a_q and b_q, are found by applying the boundary conditions. The coefficients are given by,

$$\begin{aligned} a_q &= \frac{\Psi_q'\,(n\alpha)\,\Psi_q\,(\alpha) - \Psi_q\,(n\alpha)\,\Psi_q'\,(\alpha)}{\Psi_q'\,(n\alpha)\,\zeta_q\,(\alpha) - \Psi_q\,(n\alpha)\,\zeta_q'\,(\alpha)}, \\ b_q &= \frac{n\Psi_q'\,(n\alpha)\,\Psi_q\,(\alpha) - \Psi_q\,(n\alpha)\,\Psi_q'\,(\alpha)}{n\Psi_q'\,(n\alpha)\,\zeta_q\,(\alpha) - \Psi_q\,(n\alpha)\,\zeta_q'\,(\alpha)}, \end{aligned} \tag{5.26}$$

where n is the particle's refractive index, α is again the characteristic length: $\alpha = kr_p$, and,

$$\begin{aligned} \Psi_q\,(x) &= \sqrt{\frac{\pi x}{2}} J_{q+1/2}(x), \\ \zeta_q\,(x) &= \sqrt{\frac{\pi x}{2}} H_{q+1/2}^{(2)}(x), \end{aligned} \tag{5.27}$$

are Riccati-Bessel functions. The scattered components of the electric field are (for the far-field solution),

$$\begin{aligned} E_\theta &= \frac{e^{ikr}}{ikr}\cos\varphi S_2(\theta), \\ E_\varphi &= \frac{e^{ikr}}{ikr}\sin\varphi S_1(\theta). \end{aligned} \tag{5.28}$$

We next express the field components in Cartesian coordinates,

$$\begin{aligned} E_x &= \cos\theta\cos\varphi E_\theta - \sin\varphi E_\varphi, \\ E_y &= \cos\theta\sin\varphi E_\theta + \cos\varphi E_\varphi, \\ E_z &= -\sin\varphi E_\theta. \end{aligned} \tag{5.29}$$

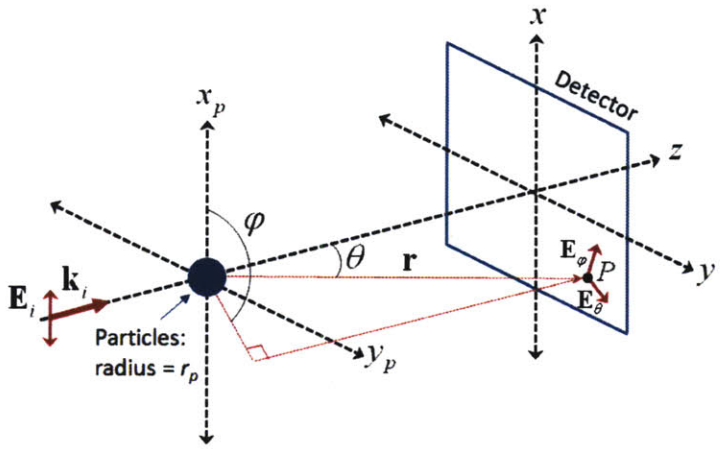


Figure 5-5: Scattering problem geometry.

In general, the wave scattered by the particle is elliptically polarized. In contrast, the reference wave, $\mathbf{E}_r$, is assumed to have the same polarization state as the incident wave. Let $\mathbf{E}_o$ be the total object field scattered by the particle cloud. The object field can be decomposed into components parallel and perpendicular to $\mathbf{E}_r$: $\mathbf{E}_o^{\parallel}$, $\mathbf{E}_o^{\perp}$. The resulting

interference pattern is given by,

$$I = \left(\mathbf{E}_r + \mathbf{E}_o^{\parallel}\right)\left(\mathbf{E}_r + \mathbf{E}_o^{\parallel}\right)^* + \left|\mathbf{E}_o^{\perp}\right|^2. \tag{5.30}$$

From equation 5.30, we see that only the component of the object wave that is parallel to $\mathbf{E}_r$ contributes to the interferogram. Under the weak scattering assumption, the magnitude of the object wave component perpendicular to the reference wave is small and thus can be neglected. If we assume an incident illumination polarized in the x-direction as indicated above, only the x-component of the scatter field is needed to compute the interference pattern. Substituting equation 5.28 in to the x-component of equation 5.29 we get,

$$E_x = \frac{e^{ikr}}{ikr} G, \tag{5.31}$$

where G is a complex amplitude function that depends on the particle size, material, wavelength, and observation position and is given by,

$$G = \cos\theta \cos^2\varphi S_2 - \sin^2\varphi S_1. \tag{5.32}$$

From equation 5.31 we see that the field scattered by the particle is a spherical wave similar to that of the point source particle assumption, but is modulated by the complex amplitude of equation 5.32. This modulating term has amplitude that is proportional to the particle's scattering cross-section and a phase that modifies the otherwise perfect spherical wavefront which can be considered as an intrinsic aberration [268]. Under the paraxial approximation, equation 5.32 takes the same form as equation 5.9, with $\phi^{(m,v)}$ given by equation 5.10, and the amplitude given by,

$$U_p^{(m,v)} = \frac{\hat{G}^{(m,v)} a^{(v)}}{ikz_p^{(v)}}, \tag{5.33}$$

where $\hat{G}^{(m,v)}$ is the paraxial approximated modulating function of equation 5.32 that depends on the positions of the particle and observation pixel as well as the particle's

refractive index and size distribution. Different size distributions may be used, such as log-normal distributions that characterize fine particles [269].

It only remains to calculate the remaining power of the reference wave after several scattering events of the incident wave. This is calculated in the same way as equation 5.11 but the total power scattered by the particle is replaced with,

$$P_s^{tot} = \frac{8\pi^2 U_i^2}{k^2} \sum_{q=1}^{\infty} (2q+1) \left[|a_q|^2 + |b_q|^2 \right], \tag{5.34}$$

where again we have assumed that each particle is illuminated by an incident wave with the same power. In contrast to the previous case, the power is not uniformly distributed over a sphere and varies with position [270]. For larger particles, most of the scattered energy goes in the forward direction. This introduces an advantage of the in-line geometry versus the off-axis case, as the detector collects most of the signal power. Using these results, the rest of the model can be derived in the same way as that described in the previous section. For particles smaller than the wavelength, the extended model reduces to that of the point source particle case.

5.4 Simulation Results

In this section, simulation results for the numerical computation of the stability metric based on the described theoretical model are presented. The simulation experiments are conducted for the extreme case of point source particles. As discussed previously, the scattered power by these particles is small and hence they can be used to estimate the upper bound performance of the system. To simplify the numerical computations and to reduce the size instantaneous Hopkins matrix, a single x-z cross-section of the VOI is considered. This cross-section is taken at $y = 0$. To control the effective particle density, the probability, p, of having a particle at the vth voxel is varied in the range $[0, 0.2]$. The total number of voxels in the x-z slice is set to be the same as the total

number of detector pixels: $V_{xz} = M = 61 \times 61$. The detector's pixel size is: $\delta_{pix} = 9\mu$m. The mean distance between the detector and the VOI is: $z_M = 50$mm. The operating wavelength and amplitude coefficient are: $\lambda = 632.8$nm, $\sqrt{\eta} = 0.001$. The voxel size along the z-direction is fixed to: $d_z^{voxel} = 50\mu$m. Three simulation experiments are conducted in which the voxel size along the x-direction is set to be smaller, equal and larger than the system's diffraction limit resolution, $\Delta = 0.5(\lambda/NA)$. For each particle density value (probability value), 100 random experiments are performed. For every random experiment, the instantaneous Hopkins matrix, cross-talk noise and stability metric are computed. In addition, the mean and standard deviation of the stability metric corresponding to each random are computed.

Figure 5-6 shows the computed stability metric for the three cases as a function of probability (particle density). As expected, an increase in the number of particles inside the VOI (increase in probability) results in an increased amount of information encoded by the system. Each particle represents a sample point within the volume. Larger number of particles better characterizes the complex flow under study. However, the cross-talk noise also depends on the number of particles and becomes more severe when the density increases. A large amount of cross-talk noise corrupts the encoded hologram making it difficult to recover the original signal. This is the well-known tradeoff between particle density and stability of the inverse problem. The proposed model allows estimating the inflection point in which the stability metric is maximized. This point corresponds to the optimum particle density that guarantees maximum information extraction from the object to the image spaces. Similar tradeoffs may be found respect to other system parameters, such as number of pixels or pixel size. From the three cases shown in Figure 5-6, it can be seen that the stability metric is the smallest when the voxel size along the x-direction is smaller than the diffraction limit resolution. This result is expected as particles in close proximity cannot be correctly resolved by the optical system. If the distance between particles increases, such as the case when $d_x^{voxel} >> \Delta$, the particle reconstruction is less affected by the added noise and hence it has a larger stability metric

value. This influence of voxel size in the stability metric can also be seen in Figure 5-7 for the case of a fixed probability value of: $p = 0.04$. As the voxel size increases, the stability metric also increases and then saturates around a point corresponding to the system's diffraction limit resolution.

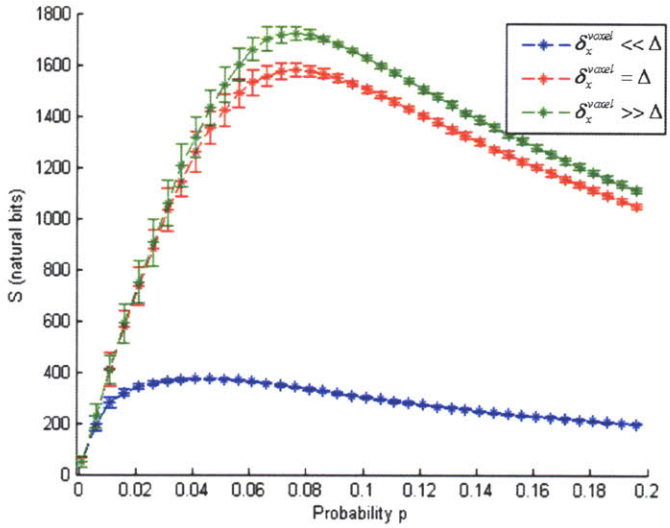


Figure 5-6: Stability metric simulation results.

Figure 5-8-a shows the corresponding increase in cross-talk variance as a function of probability. The estimated number of particles in the x-z plane as a function of probability is shown in Figure 5-8-b.

5.5 Experimental Verification

Experiments are conducted to examine the influence of particle density to the ability of a given decoding strategy to extract information from the recorded holograms. The experimental results are expected to follow a similar trend as that predicted by the stability metric shown in Figure 5-6. The geometry of the implemented optical setup is shown in Figure 5-9. A 658nm, 100mW diode laser is used as the coherent illumination source. The laser beam is first spatially filtered and then collimated using a plano-

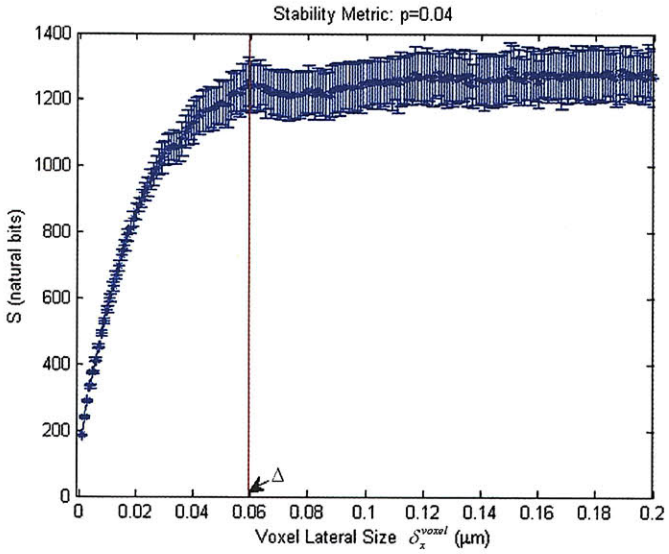


Figure 5-7: Stability metric for different lateral voxel sizes.

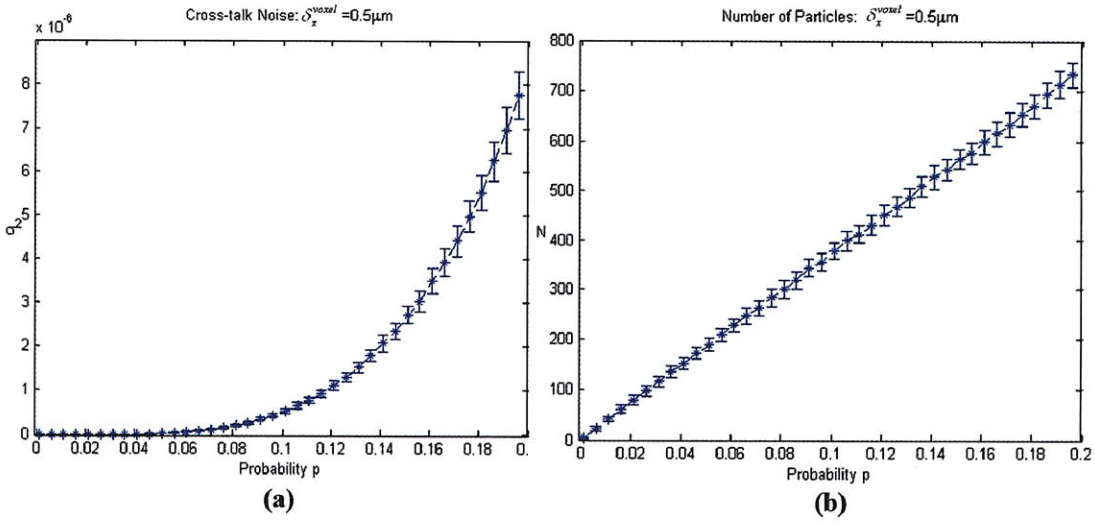


Figure 5-8: (a) Cross-talk noise variance; (b) Number of particles.

convex lens to produce a uniform plane wave. A tank is filled with small tracing particles suspended in water. The tank dimensions are: width = 0.34", height = 0.74", length = 2.13". The plane wave's diameter is truncated such that the sampled volume is: $VOI = 28\text{mm}^3$. The particle's diameter and standard deviation are: $d_p = 78.1\mu\text{m}$, $\sigma = 0.95\mu\text{m}$. The particles are produced by Bangs Laboratories, Inc., and are made out of polystyrene. A small mixer is used to maintain the particles uniformly distributed and prevent them from settling at the bottom of the tank. This guarantees consistency between different realizations of a given experiment. The holograms are captured on Kodak's KAF-16801E CCD sensor with pixel size: $\delta_{pix} = 9\mu\text{m}$. The SBP of the captured holograms is: 200×200 pixels. The holograms are transferred to the computer by means of a frame grabber. To avoid motion blur, the diode laser is modulated using a microcontroller. The microcontroller synchronizes the photodetector's integration time and the execution of a single laser pulse.

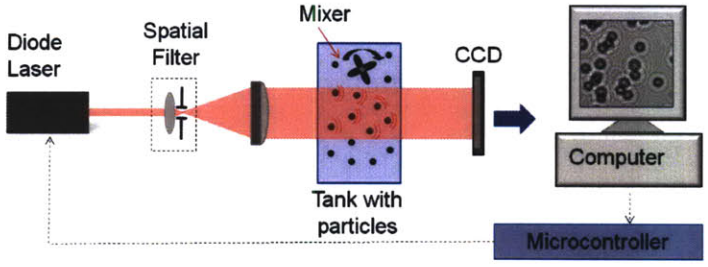


Figure 5-9: Experimental setup.

Holograms are recorded for 17 different particle densities. The particle densities are controlled by carefully extracting a sample volume using a micropipette from a solution source with suspended particles of known density. The extracted samples are further diluted with water to produce constant volumes throughout the experiments. For each dilution level, 100 holograms are captured. This is equivalent to having 100 repetitions per given random experiment. Figure 5-10 shows an example of three captured holograms corresponding to low, medium and high particle densities. For high particle densities,

a speckle-like pattern is observed. This corresponds to the cross-talk noise analyzed in the previous sections. The standard deviation is computed to quantify the statistical variations in intensity over the different captured holograms. The results are shown in Figure 5-11-a. This figure includes the variations of the signal as well as the cross-talk noise, as opposed to the simulation results shown in Figure 5-8-a that only included the variance of the cross-talk noise. The corresponding mean intensity is shown in Figure 5-11-b.

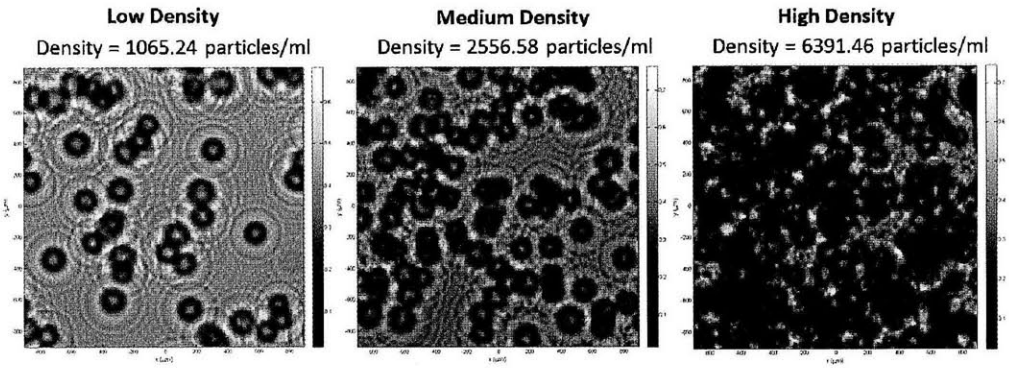


Figure 5-10: Example of captured holograms.

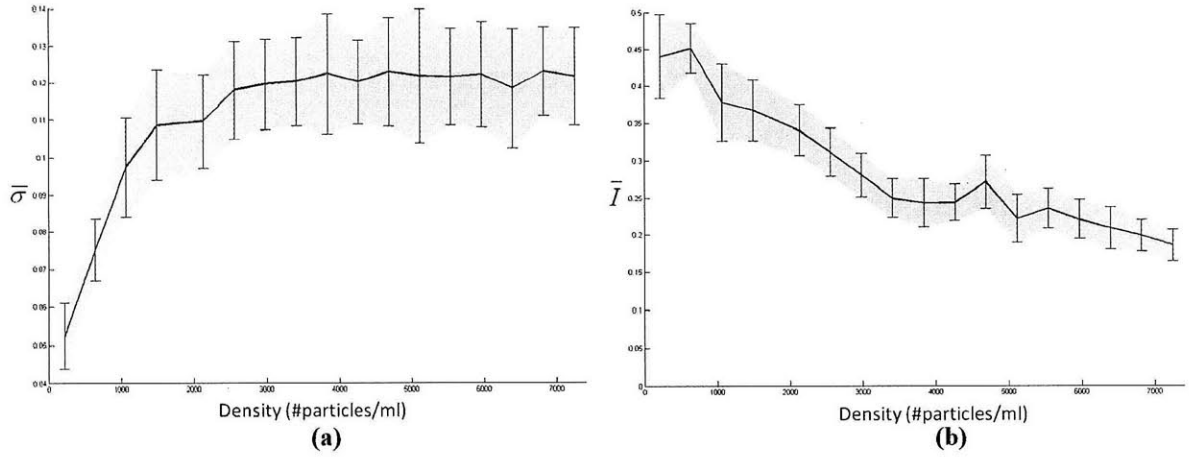


Figure 5-11: Hologram statistics: (a) Variance; (b) Mean intensity.

A decoding strategy is chosen and its ability to extract information from the recorded holograms is studied. This strategy utilizes a template matching scheme based on a cross-correlation algorithm. The algorithm is designed to process all the captured holograms automatically and keep track of the detected number of particles per frame. The block diagram of the template matching algorithm is shown in Figure 5-12. The algorithm begins by loading a given hologram. The 3D reconstruction volume is created by reconstructing the hologram at several planes within the VOI separated by an axial step, Δz. The reconstruction is conducted using the Fresnel back-propagation method of equation 2.48. Next, the template is loaded and used in the cross-correlation algorithm. The template is a small image that contains a single particle in-focus. This template can be extracted from one of the hologram reconstructions or can be digitally created for the known particle diameter. A 2D cross-correlation is computed between the template and a given x-y plane of the reconstructed 3D volume. The resulting correlation maps are stored on the 3D cross-correlation volume. This volume contains peak responses at the locations where the reconstructions match the template. The 3D volume is then binarized using a fixed threshold to isolate the peak responses. This process results in 3D binary objects that elongate along the optical axis. The binary objects are then labeled and their areas and centroids are computed. Three tests are designed to find the "true" responses from particles and discard "false" binary objects. In the first test, binary objects with an area smaller than the expected area are discarded. In the second test, the intensity grey level values corresponding to the centroids of the binary objects are extracted from the 3D reconstruction volume. The objects that have an intensity value larger than a given threshold are discarded. This is because the particles are opaque and the reconstructed intensity distribution is expected to have low grey values inside the particle. The third test compares the correlation peak values from the 3D cross-correlation volume. Binary objects with values lower than a set threshold are discarded. The strength of the correlation peak is also used to estimate double particles that are stuck together.

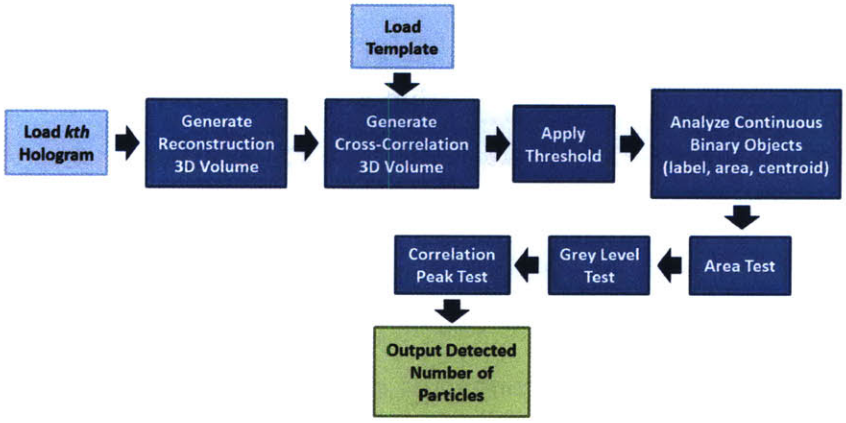


Figure 5-12: Block diagram of template matching algorithm.

Figure 5-13 shows an example of three reconstructed images processed using the template matching algorithm. These images correspond to low, medium and high particle density holograms. The reconstruction shown in Figure 5-13 is not the reconstructed intensity distribution at a given plane, but the project minimum intensity along the axial direction of the 3D reconstruction volume. The red and green crosses indicate detected single and double particles respectively. As can be seen for the high density case, the algorithm fails to detect all the particles contained inside the VOI.

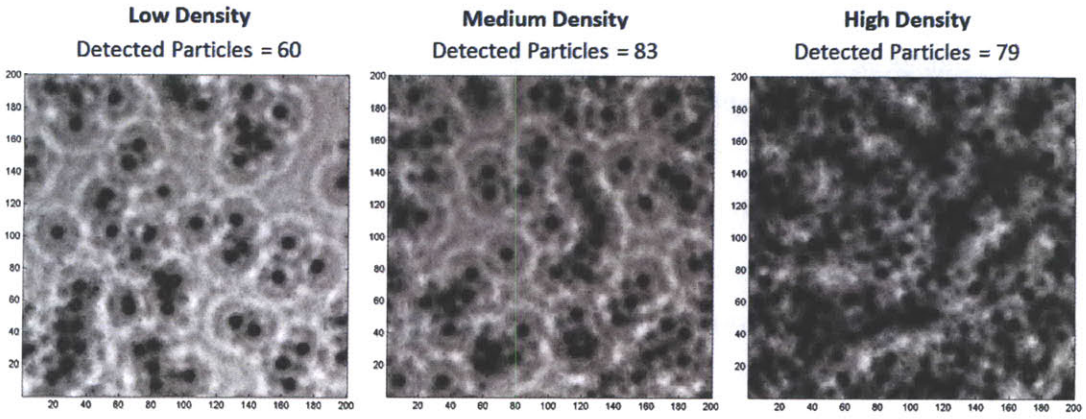


Figure 5-13: Example of reconstructed images.

The comparison between the expected number of particles from the measured densities and the detected number of particles from the template matching algorithm is shown in Figure 5-14. As can be seen, the experimental results follow a similar trend as that predicted by the stability metric shown in Figure 5-6. The expected and detected number of particles increases linearly for low to medium particle densities and then decreases after the inflection point. For high particle densities, the decoding algorithm fails to extract information from the hologram and the reconstruction is dominated by cross-talk noise. The peak from Figure 5-14 corresponds to an optimum particle density of $\sim$ 3409particles/ml (3.4 particles/mm^3). A more efficient decoding strategy may be implemented; however, its performance will be ultimately limited by the upper bound set by the stability metric.

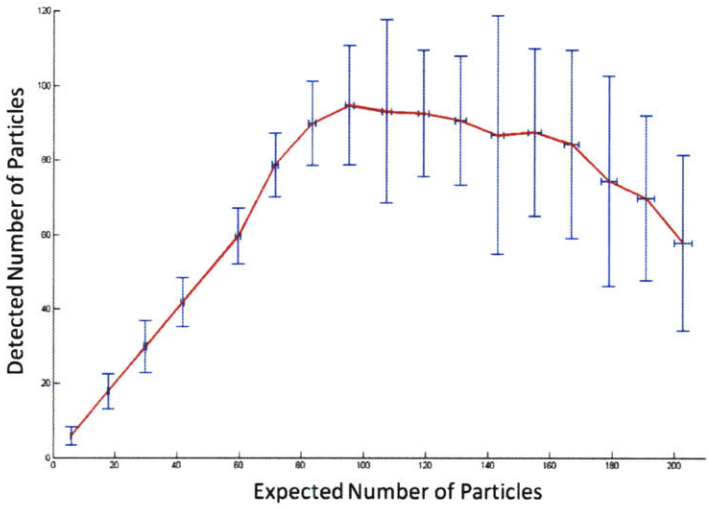


Figure 5-14: Comparison between detected and expected number of particles.

Chapter 6

Conclusions

This thesis presents a variety of multi-domain optimization tools designed to improve the performance of holographic optical systems. Specific examples are described of systems designed for imaging and lithography. Novel optical systems and optimization algorithms are proposed. The developed optimization tools are designed to be robust, computationally efficient and sufficiently general to be extended to other holographic based systems. All the major forms of holographic elements are analyzed: computer generated holograms, thin and thick conventional holograms, numerically simulated holograms, and digital holograms. The design, optimization and experimental implementation of the proposed systems are performed. Each system is accurately modeled using the appropriate scalar or vector diffraction theories. Additional studies are performed using system and information theories. The system's channel capacity and stability of the related inverse problem as a function of various control parameters are evaluated. The performance of the corresponding signal encoding and decoding processes is maximized. A sensitivity analysis is performed to predict and assist in the correction of potential fabrication or calibration errors.

Two lithographic systems are presented, one based on CGHs and the other using optically recorded TIR holograms. The first system is proven to reconstruct high-quality patterns that satisfy the high diffraction efficiency and uniformity demands when opti-

mized using the proposed hybrid optimization algorithm based on GAs and the MER method. The introduction of the local diffusers phase elements (LDPE) and local negative power elliptical phase elements (LNPEPE) masks enabled reducing the complexity of the optimization problem by only controlling a small subset of optimization variables. The algorithm is efficiently implemented on a GPU, resulting in speedups of more than 200× compared to standard CPU implementations. A multiplexing method for the extension of the depth of focus is proposed. A simple CGH fabrication method based on electron-beam lithography is presented, as well as an optimization scheme designed for the local correction of over and under dose errors. Experimental demonstrations of the reconstructions from the fabricated CGHs using coherent and partially coherent illuminations are presented. A sensitivity analysis is conducted. The presented algorithm is extended for the design and optimization of multispectral CGHs applied for high efficiency solar concentration and spectral splitting.

The second lithographic system is studied for the target application of high-resolution, parallel exposure, non-contact, large area flat panel display manufacture. This system is based on TIR holograms recorded on a photopolymer that operate in the near ultraviolet regime. A comparative analysis between scalar and vector diffraction theories for the modeling and simulation of the system is performed. Scalar diffraction theory is proven to be sufficiently accurate for the considered geometry and is chosen for modeling and simulating the system. First order models for simulating the material response and shrinkage of the photopolymer are presented. A novel block-stitching algorithm is introduced for the calculation of large diffraction patterns that allows overcoming current computational limitations of memory and processing time. The numerical model is implemented for optimizing the system's performance as well as redesigning the mask to account for potential fabrication errors. The simulation results are compared to experimentally measured data. In addition, a method for extending the depth of focus of the system is presented.

Two imaging systems are presented: holographically corrected segmented aperture

thin imager and DHPIV system. The first system is proposed to achieve high-resolution imaging on space constrained geometry (maximum thickness of 5mm) better than conventional microlens array compound systems. This imager utilizes phase conjugation holography for correcting the high-order aberrations present in the GRIN lens array. The new degrees of freedom introduced by the holographic elements are utilized for maximizing the information transfer from scene to measurement spaces. The optical performance of the system is evaluated using a combination of Matlab and Zemax. The imager is modeled using system's theory by defining the Hopkins matrix which connects the input and output signals. The system is analyzed using information theory treating it as a Gaussian parallel communication channel. A multi-domain optimization approach is implemented based on GAs for maximizing the system's channel capacity and hence improving the information extraction or encoding process. A decoding or reconstruction strategy is implemented using the superresolution algorithm. Experimental results for the optimization of the hologram's recording process are presented. A tomographic technique based on the Foucault knife-edge test is presented for the measurement of the system's space-variant point spread function. A sensitivity analysis is performed to estimate the effect of potential misalignment errors in the assembly process. A modification of geometry for polychromatic imaging is proposed.

The second studied imaging system utilizes digital holography for the measurement of complex flow fields by tracking micron sized tracing particles. The overall performance of DHPIV systems relies on the proper selection of parameters such as particle density, geometry and detector related parameters. A stochastic theoretical model based on a stability metric similar to the channel capacity for a Gaussian channel defined in information theory is presented. The stability metric is used for optimizing the system, maximizing the amount of 3D information from the VOI that can be encoded by a single hologram. The theoretical model is first derived for the extreme case of point source particles using Rayleigh scattering and scalar diffraction theory formulations. The model is then extended to account for particles of variable sizes using Mie theory for scattering

of homogeneous dielectric spherical particles. The forward problem (recording of digital hologram) is modeled using system's theory by defining the instantaneous Hopkins matrix. The influence and statistics of the particle density dependent cross-talk noise are studied. Simulation results are presented for finding the optimum particle density based on the stability metric. An experimental evaluation is conducted to study the influence of particle density in the information extraction process by a set decoding algorithm. The implemented decoding strategy is based on a template matching scheme designed to automatically process and count the particles present at each frame. The experimental results are compared to the predictions obtained from the stability metric showing similar trends.

Bibliography

[1] D. Gabor, "A New Microscopic Principle," Nature **161**, 777 (1948).

[2] D. Gabor, "Microscopy by reconstructed wavefronts," Proc. Roy. Soc. A **197**, 454 (1949).

[3] D. Gabor, "Microscopy by reconstructed wavefronts: II," Proc. Phys. Soc. B **378**, 449 (1951).

[4] E. N. Leith, and J. Upatnieks, "Reconstructed wavefronts and communication theory," J. Opt. Soc. Am. **52**, 1123 (1962).

[5] E. N. Leith, and J. Upatnieks, "Wavefront reconstruction with diffused illumination and three-dimensional objects," J. Opt. Am. **54**, 1295 (1964).

[6] C. E. Shannon, "A Mathematical Theory of Communication," The Bell Sys. Tech. J. **27**, 379 (1948).

[7] J.A. Dominguez-Caballero, "Digital Holographic Imaging of Aquatic Species," Massachusetts Institute of Technology, Master's Thesis, February (2006).

[8] D. Carl, B. Kemper, G. Wernicke, and G. V. Bally, "Parameter-optimized digital holographic microscope for high-resolution living-cell analysis," App. Opt. **43**, 36 (2004).

[9] G. Pan, and H. Meng, "Digital In-line Holographic PIV for 3D Particulate Flow Diagnostics," 4th Int. Symp. on Particle Image Vel., September (2001).

[10] G. Barbastathis, and D. J. Brady, "Multidimensional tomographic imaging using volume holography," Proc. IEEE **87**, 2098 (1999).

[11] J. F. Heanue, M. C. Bashaw, and L. Hesselink, "Volume holographic storage and retrieval of digital data," Science **265**, 749 (1994).

[12] G. Andersen, "Holographic correction and phasing of large sparse-array telescopes," App. Opt. **44**, 8 (2005).

[13] R. Piestun, et al., "On-axis computer-generated holograms for three-dimensional display," Opt. Lett. **22**, 12 (1997).

[14] K. A. Stetson, "Holography with Total Internally Reflected Light," Appl. Phys. Lett. **11**, 225 (1967).

[15] W. H. Lee, "Binary Computer Generated Holograms," Appl. Opt. **18**, 3661 (1979).

[16] I. Moreno, C. Gorecki, J. Campos, and M. J. Yzuel, "Comparison of computer-generated holograms produced by laser printers and lithography: application to pattern recognition," Opt. Eng. **34**, 12 (1995).

[17] Y. Xie, Z. Lu, and F. Li, "Lithographic fabrication of large curved hologram by laser writer," Opt. Exp. **12**, 9 (2004).

[18] S. C. Baber, "Application of high resolution laser writers to computer generated holograms and binary diffractive optics," Proc. SPIE **1052**, 66 (1989).

[19] W. Cai, T. Reber, and R. Piestun, "Computer-generated volume holograms fabricated by femtosecond laser micromachining," Opt. Lett. **31**, 12 (2006).

[20] Q. Z. Zhao, et al., "Direct writing computer-generated holograms on metal film by an infrared femtosecond laser," Opt. Exp. **13**, 6 (2005).

[21] H. Ichikawa, M. R. Taghizadeh, and J. Turunen, "Contact printing of on-axis computer-generated hologram on silver halide emulsion," IEE Conf. on Holo. Sys., Comp. and App. **16**, 50 (1991).

[22] M. Kajanto, et al., "Photolithographic fabrication method of computer-generated holographic interferograms," Appl. Opt. **28**, 4 (1989).

[23] L. D'Auria, J. P. Huignard, A. M. Roy, and E. Spitz, "Photolithographic fabrication of thin film lenses," Opt. Comm. **5**, 4 (1972).

[24] H. Farhoosh, et al., "Comparison of binary encoding schemes for electron-beam fabrication of computer generated holograms," Appl. Opt. **26**, 20 (1987).

[25] S. M. Arnold, "Electron beam fabrication of computer-generated holograms," Opt. Eng. **24**, 5 (1985).

[26] B. R. Brown, and A. W. Lohmann, "Complex Spatial Filtering with Binary Masks," Appl. Opt. **5**, 967 (1966).

[27] A. W. Lohmann, and D. P. Paris, "Binary Fraunhofer Holograms, Generated by Computer," Appl. Opt. **6**, 1739 (1967).

[28] A. Kozma, and D. L. Kelly, "Spatial Filtering for Detection of Signals Submerged in Noise," Appl. Opt. **4**, 387 (1965).

[29] L. G. Neto, P. S. P. Cardona, G. A. Cirino, R. D. Mansano, and P. Verdonck, "Design, fabrication, and characterization of a full complex-amplitude modulation diffractive optical element," J. Microlith., Microfab., Microsyst. **2**, 2 (2003).

[30] L. G. Neto, P. S. P. Cardona, G. A. Cirino, R. D. Mansano, and P. Verdonck, "Implementation of Fresnel full complex-amplitude digital holograms," Opt. Eng. **43**, 11 (2004).

[31] W. Yu, et al., "Fabrication of multilevel phase computer-generated hologram elements based on effective medium theory," Appl. Opt. **39**, 20 (2000).

[32] J. W. Goodman, and A. M. Silvestri, "Some Effects of Fourier Domain Phase Quantization," IBM J. Res. Dev. **14**, 478 (1970).

[33] R. W. Floyd, and L. Steinberg, "An adaptive algorithm for spatial grayscale," Proc. SID **17**, 78 (1976).

[34] R. Hauck, and O. Bryngdahl, "Computer-generated holograms with pulse-density modulation," JOSA A **1**,1 (1984).

[35] J. R. Fienup, "Iterative method applied to image reconstruction and to computer-generated holograms," Opt. Eng. **19**, 3 (1980).

[36] G. Tricoles, "Computer generated holograms: an historical review," Appl. Opt. **26**, 20 (1987).

[37] W. T. Cathey, Jr., "The Effect of Finite Sampling in Holography," Optik **27**, 317 (1968).

[38] J. Bucklew and N. C. Gallagher, Jr., "Aliasing Error in Digital Holography," Appl. Opt. **15**, 2183 (1976).

[39] B. R. Brown, and A. W. Lohmann, "Computer-generated Binary Holograms," IBM J. Res. Dev. **13**, 160 (1969).

[40] L. B. Lesem, P. M. Hirsch, and J. Jordan, Jr., "Kinoforms: A New Wavefront Reconstruction Device," IBM J. Res. Dev. **14**, 485 (1970).

[41] J. Salo, et al., "Millimeter-wave Bessel beams using computer holograms," Elec. Lett. **37**, 13 (2001).

[42] A. Vasara, J. Turunen, and A. T. Friberg, "Realization of general nondiffracting beams with computer-generated holograms," JOSA A **6**, 1748 (1989).

[43] T. Dresel, M. Beyerlein, and J. Schwider, "Design and fabrication of computer-generated beam-shaping holograms," Appl. Opt. **35**, 23 (1996).

[44] J. Liesener, M. Reicherter, T. Haist, and H. J. Tiziani, "Multi-functional optical tweezers using computer-generated holograms," Opt. Comm. **185**, 77 (2000).

[45] J. P. Kirk, and A. L. Jones, "Phase-only complex-valued spatial filter," JOSA **61**, 8 (1971).

[46] A. W. Lohmann, and D. P. Paris, "Computer Generated Spatial Filters for Coherent Optical Data Processing," Appl. Opt. **7**, 4 (1968).

[47] C. D. Carey, et al., "Computer-generated hologram etched in GaAs for optical interconnection of VLSI circuits," Elec. Lett. **28**, 22 (1992).

[48] A. J. MacGovern, and J. C. Wyant, "Computer-generated holograms for testing optical elements," Appl. Opt. **10**, 3 (1971).

[49] K. M. Leung, J. C. Lindquist, and L. T. Shepherd, "E-Beam Computer-Generated Holograms for Aspheric Testing," Proc. Soc. Photo-Opt. Instrum. Eng. **215**, 70 (1980).

[50] J. R. Freyer, R. J. Perlmutter, and J. W. Goodman, "Digital Holography: Algorithms, E-Beam Lithography, and 3-D Display," Proc. Soc. Photo-Opt. Instrum. Eng. **437**, 38 (1983).

[51] C. Jacobsen, and M. Howells, "Projection x-ray lithography using computer-generated holograms: A study of compatibility with proximity lithography," J. Vac. Sci. Technol. B **10**, 6 (1992).

[52] A. Y. Smuk, and N. M. Lawandy, "Direct laser writing of diffractive optics in glass," Opt. Lett. **22**, 13 (1997).

[53] T. R. Groves, D. Pickard, B. Rafferty, N. Crosland, D. Adam, and G. Schubert, "Maskless electron beam lithography: prospects, progress, and challenges," Micro. Eng. **61**, 285 (2002).

[54] A. A. Patel, "The Development of a Prototype Zone-Plate-Array Lithography (ZPAL) System," Master's thesis, Massachusetts Institute of Technology, Cambridge MA, May 2004.

[55] J. G. Goodberlet, "Patterning 100 nm features using deep-ultraviolet contact photolithography," Appl. Phys. Lett. **76**, 6 (2000).

[56] M. Rothschild, and D. J. Ehrlich, "A review of excimer laser projection lithography," J. Vac. Sci. Technol. B **6**, 1 (1988).

[57] C. Jacobsen, and M. Howells, "Projection x-ray lithography using computer-generated holograms: A study of compatibility with proximity lithography," J. Vac. Sci. Technol. B **10**, 6 (1992).

[58] C. Jacobsen, and M. R. Howells, "A technique for projection x-ray lithography using computer-generated holograms," J. Appl. Phys. **71**, 6 (1992).

[59] M. R. Howells, and C. Jacobsen, "Possibilities for projection x-ray lithography using holographic optical elements," Appl. Opt. **30**, 13 (1991).

[60] F. Wyrowski, E.B. Kley, S. Bühling, A. J. M. Nellissen, L. Wang, and M. Dirkzwager, "Proximity printing by wave-optically designed masks," Proc. Of SPIE **4436**, 130 (2001).

[61] S. Bühling, F. Wyrowski, E.B. Kley, A. J. M. Nellissen, L. Wang, and M. Dirkzwager, "Resolution enhanced proximity printing by phase and amplitude modulating masks," J. Micromech. Microeng. **11**, 603 (2001).

[62] S. Bühling, F. Wyrowski, E.B. Kley, A. J. M. Nellissen, L. Wang, and M. Dirkzwager, "High resolution proximity printing by wave-optically designed complex transmission masks," Proc. Of SPIE **4404**, 221 (2001).

[63] A. Isoyan, Y.-C. Cheng, J. Wallace, and F. Cerrina, "EUV Holographic Lithography: First Results," Micro/Nano Eng. (2006).

[64] G. L. Williams, R. P. McWilliam, J. Toriz-Garcia, R. Curry, A. Maiden, N. L. Seed, A. Purvis, and P. A. Ivey, "A photolithographic process for grossly nonplanar substrates," SPIE Adv. Lith. **6921**, 81 (2008).

[65] A. Maiden, R. McWilliam, A. Purvis, S. Johnson, G. L. Williams, N. L. Seed, and P. A. Ivey, "Nonplanar photolithography with computer-generated holograms," Opt. Lett. **30**, 11 (2005).

[66] A. Purvis, R. McWilliam, S. Johnson, N. L. Seed, G. L. Williams, A. Maiden, and P. A. Ivey, "Photolithographic patterning of bihelical tracks onto conical substrates," J. Micro/Nanolith. MEMS MOEMS **6**, 4 (2007).

[67] J. W. Goodman, "Introduction to Fourier Optics," McGraw-Hill, New York, second edition (1996).

[68] I. B. Baek, et al., "Electron beam lithography patterning of sub-10nm line using hydrogen silsesquioxane for nanoscale device application," J. Vac. Sci. Technol. B **23**, 3140 (2005).

[69] D. E. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning," Addison-Wesley (1989).

[70] M. A. Seldowitz, J. P. Allebach, and D. W. Sweeney, "Synthesis of digital holograms by direct binary search," Appl. Opt. **26**, 14 (1987).

[71] J. R. Fienup, "Iterative method applied to image reconstruction and to computer-generated holograms," Opt. Eng. **19**, 3 (1980).

[72] S. Weissbach, F. Wyrowski, and O. Bryngdahl, "Digital Phase Holograms: Coding and Quantization with an Error Diffusion Concept," Opt. Comm. **72**, 31 (1989).

[73] R. W. Gerchberg, and W. O. Saxton, "A practical Algorithm for the Determination of Phase from Image and Diffraction Plane Pictures," Optik **35**, 237 (1972).

[74] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," J. Opt. Soc. Am. A **19**, 7 (2002).

[75] O. K. Ersoy, J. Y. Zhuang, and J. Brede, "An iterative interlacing approach for synthesis of computer-generated holograms," ECE Technical Report, Purdue University (1992).

[76] M. S. Kim, and C. C. Guest, "Simulated annealing algorithm for binary phase only filters in pattern classification," Appl. Opt. **29**, 1203 (1990).

[77] E. N. Leith, and J. Upatnieks, "Imagery with Pseudo-Randomly Diffused Coherent Illumination," Appl. Opt. **7**, 2085 (1968).

[78] H. J. Gerritsen, W. J. Hannan, and E. G. Ramberg, "Ellimination of Speckle Noise in Holograms with Redundancy," Appl. Opt. **7**, 2301 (1968).

[79] D. Gabor, "Laser Speckle and its Elimination," IBM J. Res. Dev. **14**, 509 (1970).

[80] P. M. Hirsh, J. A. Jordan, Jr., and L. B. Lesem, "Method of Making an Object-Dependent Diffuser," U.S. Patent No. 3,619,022 (Nov. 9, 1971; filed Sept. 17, 1970).

[81] R. H. Katyl, "Moiré Screens Coded with Pseudo-Random Sequences," Appl. Opt. **11**, 2278 (1972).

[82] W. J. Dallas, "Deterministic Diffusers for Holography," Appl. Opt. **12**, 6 (1973).

[83] H. Akahori, "Comparison of Deterministic Phase Coding with Random Phase Coding in Terms of Dynamic Range," Appl. Opt. **12**, 2336 (1973).

[84] F. Wyrowski, and O. Bryngdahl, "Speckle-free reconstruction in digital holography," JOSA A **6**, 1171 (1989).

[85] F. Wyrowski, and O. Bryngdahl, "Iterative Fourier-transform algorithm applied to computer holography," JOSA A **5**, 1058 (1988).

[86] A. V. Oppenheim, R. W. Schafer, and J. A. Buck, "Discrete-time signal processing," Prentice Hall, (1999).

[87] S. S. Young, "Alias-free image subsampling using Fourier-based windowing methods," Opt. Eng. **43**, 843 (2004).

[88] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," Proc. of the IEEE **66**, 51 (1978).

[89] J. H. Holland, "Adaptation in natural and artificial systems," Ann Arbor: The University of Michigan Press (1975).

[90] R. S. Rosenberg, "Simulation of genetic populations with biochemical properties," Doctoral thesis, University of Michigan (1967).

[91] S. Forrest, B. Javornik, R. E. Smith, and A. S. Perelson, "Using genetic algorithms to explore pattern recognition in the immune system," in Evolutionary Computation, MIT Press (1993).

[92] S. K. Pal, and P. P. Wang, "Genetic algorithms for pattern recognition," Chapman & Hall Press, First edition (1996).

[93] L. Chambers, "The Practical Handbook of Genetic Algorithms: Applications," Chapman & Hall Press, Second edition (2001).

[94] F. Allen, and R. Karjalainen, "Using genetic algorithms to find technical trading rules," J. of Fin. Econ. **51**, 245 (1999).

[95] A. Hill and C. J. Taylor, "Model-Based Image Interpretation Using Genetic Algorithms," Im. And Vis. Comp. **10**, 5 (1992).

[96] J. Periaux, H. Q. Chen, B. Mantel, M. Sefrioui, and H. T. Sui, "Combining game theory and genetic algorithms with application to DDM-nozzle optimization problems," Fin. Elem. in Anal. & Des. **37**, 5 (2001).

[97] D. Cojoc, and A. Alexandrescu, "Optimization of the computer generated binary holograms using genetic algorithms," Proc. SPIE **3904**, 256 (1999).

[98] N. Yoshikawa, M. Itoh, and T. Yatagai, "Use of genetic algorithms for computer-generated holograms," Proc. SPIE **2577**, 150 (1995).

[99] N. Yoshikawa, M. Itoh, and T. Yatagai, "Quantized phase optimization of two-dimensional Fourier Kinoforms by a genetic algorithm," Opt. Lett. **20**, 7 (1995).

[100] M. Wen, J. Yao, D. W. K. Wong, and G. C. K. Chen, "Holographic diffuser design using a modified genetic algorithm," Opt. Eng. **44**, 8 (2005).

[101] J. N. Gillet, and Y. Sheng, "Multiplexed computer-generated holograms with polygonal-aperture layouts optimized by genetic algorithm," Appl. Opt. **42**, 20 (2003).

[102] E. G. Johnson, A. D. Kathman, D. H. Hochmuth, A. Cook, D. R. Brown, and W. Delaney, "Advances of genetic algorithm optimization methods in diffractive optics design," in Diff. and Mini. Opt., S.H. Lee ed. Proc. SPIE **49**, 54 (1993).

[103] G. Zhou, Y. Chen, Z. Wang, and H. Song, "Genetic local search algorithm for optimization design of diffractive optical elements," Appl. Opt. **38**, 20 (1999).

[105] G. Zhou, X. Yuan, P. Dowd, Y. L. Lam, and Y. C. Chan, "Design of diffractive phase elements for beam shaping: hybrid approach," J. Opt. Soc. Am. A **18**, 4 (2001).

[106] J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 15 (1982).

[107] R. W. Gerchberg, "Super-Resolution Through Error Energy Reduction," Opt. Acta **21**, 709 (1974).

[108] W. O. Saxton, "Computer Techniques for Image Processing in Electron Miscroscopy," Academic Press, (1978).

[109] N. C. Gallagher, and B. Liu, "Method for Computing Kinoforms that Reduces Image Reconstruction Error," Appl. Opt. **12**, 2328 (1973).

[110] B. Liu, and N. C. Gallagher, "Convergence of a Spectrum Shaping Algorithm," Appl. Opt. **13**, 2470 (1974).

[111] J. R. Fienup, "Reduction of Quantization Noise in Kinoforms and Computer-Generated Holograms," J. Opt. Soc. Am. **64**, 1395 (1974).

[112] J. R. Fienup, "Improved Synthesis and Computational Methods for Computer-Generated Holograms," Ph.D. thesis, Stanford University, May 1975.

[113] J. A. Dominguez-Caballero, S. Takahashi, S. Lee, and G. Barbastathis, "Design and Analysis of Fresnel Domain Computer Generated Holograms," To be published at the International Symposium of Nanomanufacturing (2009).

[114] I. Avcibas, et al., "Statistical evaluation of image quality measures," J. of Elec. Imag. **11**, 2 (2002).

[115] D. N. Bhat, and S. K. Nayar, "Ordinal Measures for Image Correspondence," IEEE Trans. of Pat. Anal. and Mach. Int. **20**, 4 (1998).

[116] F. Wyrowski, "Upper bound of diffraction efficiency of diffractive phase elements," Opt. Lett. **16**, 24 (1991).

[117] U. Krackhardt, J. N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31**, 1 (1992).

[118] G. Zhou, X. Yuan, P. Dowd, Y. L. Lam, and Y. C. Chan, "Efficient method for evaluation of the diffraction efficiency upper bound of diffractive phase elements," Opt. Lett. **25**, 17 (2000).

[119] NVIDIA CUDATM, "Programming Guide Version 2.3", July (2009).

[120] The MathWorksTM, "MEX-files Guide", URL: http://www.mathworks.com/support/tech-notes/1600/1605.html#intro, Accessed on September (2009).

[121] Liang R., Pan Z., Krokos M., Chen M., Bao J., and Li C., "Fast hardware-accelerated volume rendering of CT scans," J. Display Technol. **4**, 431 (2008).

[122] Kasson P. M., Ensign D. L., and Pande V. S., "Combining Molecular Dynamics with Bayesian Analysis To Predict and Evaluate Ligand-Binding Mutations in Influenza Hemagglutinin," J. Am. Chem. Soc. **131**, 32 (2009).

[123] Molemaker J., Cohen J. M., Patel S., and Noh J., "Low Viscosity Flow Simulation for Animation," Euro./ ACM SIGGRAPH, (2008).

[124] Corrigan A., Wallin J., and Vesenjak M., "Visualization of Meshless Simulations Using Fourier Volume Rendering," Prog. on Mesh. Meth, Comp. Meth. in Appl. Scie. Ser. **11**, 291 (2009).

[125] Bennemann C., Beinker M. W., Egloff D., and Gauckler M., "Teraflops for Games and Derivatives Pricing," WILMOTT magazine, page 50 (2008).

[126] Masuda N., Ito T., Tanaka T., Shiraki A., and Sugie T., "Computer generated holography using a graphics processing unit," Opt. Exp. **14**, 603 (2006).

[127] Yaras F., Kang H., and Onural L., "Real-time multiple SLM color holographic display using multiple GPU acceleration," DH and 3D Imag. Conf., Vancouver (2009).

[128] Shiraki A., Takada N., Niwa M., Ichihashi Y., Shimobaba T., Masuda N., and Ito T., "Simplified electroholographic color reconstruction system using graphics processing unit and liquid crystal display projector," Opt. Express **17**, 16038 (2009).

[129] Kang H., Yaras F., Onural L., and Yoshikawa H., "Real-time Fringe Pattern Generation with High Quality," DH and 3D Imag. Conf., Vancouver (2009).

[130] Bianchi S., and Leonardo R. D., "Real-time optical manipulation of micron sized structures using GPU generated holograms," Jul 2009. [Online]. Available: http://arxiv.org/abs/0907.4027

[131] Shimobaba T., Sato Y., Miura J., Takenouchi M., and Ito T., "Real-time digital holographic microscopy using the graphic processing unit," Opt. Exp. **16**, 11776 (2008).

[132] M. Shigeki, et al., "Proximity-effect correction software for EPL using the pattern classify method," Proc. of the SPIE **5446**, 897 (2004).

[133] D. Gil, "Maskless nanolithography and imaging with diffractive optical arrays," Massachusetts Institute of Technology Ph.D. thesis, Dept. of EECS, May 2003.

[134] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," IEEE Trans. on Sys., Man, and Cyb. **9**, 62 (1979).

[135] Zweibel K., Mason J., and Fthenakis V., "A Solar Grand Plan," Sci. Amer., pp. 64-74, January (2008).

[136] Luque A., and Hegedus S., eds., Handbook of Photovoltaic Science and Engineering, CH 11, "Photovoltaic Concentrators," Swanson R. M., Wiley J., and Sons, (2003).

[137] Nuwayhid R. Y., Mrad F., and Abu-Said R., "The realization of a simple solar tracking concentrator for university research applications," Rene. Ener. 24, 207 (2001).

[138] Kostuk K. R., and Rosenberg G., "Analysis and Design of Holographic Solar Concentrators," Proc. of SPIE **7043**, 704301 (2008).

[139] Bloss W. H., Griesinger M., and Reinhardt, "Dispersive concentrating systems based on transmission phase holograms for solar applications," Appl. Opt. **21**, 3739 (1982).

[140] Ludman J. E., et al., "Photovoltaic systems based on spectrally selective holographic concentrators," SPIE Prac. Holo. VI **1667**, 182 (1992).

[141] K. A. Stetson, "Improved Resolution and Signal-To-Noise Ratios in Total Internal Reflection Holograms," Appl. Phys. Lett. **12**, 11 (1968).

[142] S. Syinov, R. Stoicheva, and P. I. Markovski, "Total-internal-reflection holograms recorded in thin As2S3 films," Sov. J. Quan. Elec. **10**, 366 (1980).

[143] H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," The Bell Sys. Tech. J. **48**, 2909 (1969).

[144] S. Sainov, and R. Stoycheva-Topalova, "Total internal reflection holographic recording in very thin films," J. Opt. A: Pure Appl. Opt. **2**, 117 (2000).

[145] R. T. Chen, "Limitations of Submicron Holographic Lithography," SPIE **2337**, 138 (1994).

[146] R. T. Chen, et al., "Microcircuit Lithography Using Holographic Imaging," SPIE Opt./Laser Micro. III **1264**, 342 (1990).

[147] F. Clube, S. Gray, B. Le Gratiet, N. Magnon, S. Malfoy, D. Struchen, and A. R. Nobari, "Fine-Pattern Lithography for Large Substrates Using a Holographic Mask-Aligner," Micro. Eng. 41-42, 149 (1998).

[148] J. H. Choi, et al., "Fabrication technologies of field emitter arrays," Mat. Res. Soc. Symp. Proc. **471**, 211 (1997).

[149] M. Barge, S. Bruynooghe, F. Clube, A. Nobari, J.-L. Saussol, E. Grass, H. Mayer, B. Schnabel, and E.-B. Kley, "120-nm lithography using off-axis TIR holography and 364 nm exposure wavelength," Micro. Eng. 57-58, 59 (2001).

[150] E. Hecht, "Optics," Ed. Pearson Education, Fourth Edition (2004).

[151] A. Frosh, et al. "Method of Making Totally Internal Reflected Holograms," US Patent: 3,796,476 (1974).

[152] W.-S. Han, et al., "Method and apparatus for recording a hologram from a mask pattern by the use of total internal reflection holography and hologram manufactured by the method," US Patent: 7,092,134 B1 (2006).

[153] J. A. Kong, "Electromagnetic Wave Theory," EMW Publishing, Cambridge MA (2005).

[154] F. S. M. Clube, "Method and Apparatus for Forming a Surface-Relief Hologram Mask," US Patent: US 2006/0232838 A1 (2006).

[155] M. G. Moharam, and T. K. Gaylord, "Rigorous coupled-wave analysis of planar-grating diffraction," J. Opt. Soc. Am. **71**, 811 (1981).

[156] M. G. Moharam, and T. K. Gaylord, "Rigorous coupled-wave analysis of planar grating diffraction – E-mode polarization and losses." J. Opt. Soc. Am. **73**, 451 (1983).

[157] M. G. Moharam, E. B. Grann, D. A. Pommet, and T. K. Gaylord, "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," J. Opt. Soc. Am. A **12**, 1068 (1995).

[158] M. G. Moharam, and T. K. Gaylord, "Rigorous coupled-wave analysis of metallic surface-relief gratings," J. Opt. Soc. Am. A **3**, 1780 (1986).

[159] M. G. Moharam, D. A. Pommet, E. B. Grann, and T. K. Gaylord, "Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach," J. Opt. soc. Am. A **12**, 1077 (1995).

[160] E. J. Restall, and A. C. Walker, "Rigorous coupled-wave method applied to fan-out gratings," IEE Proc.-Optoelectron. **145**, 165 (1998).

[161] E. N. Glytsis, and T. K. Gaylord, "Rigorous three-dimensional coupled-wave diffraction analysis of single and cascaded anisotropic gratings," J. Opt. Soc. Am. A. **4**, 2061 (1987).

[162] P. B. Johnson, and R. W. Christy, "Optical constants of transition metals: Ti, V, Cr, Mn, Fe, Co, Ni, and Pd," Phys. Rev. B **9**, 5056 (1974).

[163] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," IEEE Trans. on Ant. and Prop. **14**, 302 (1966).

[164] A. Taflove, and S. C. Hagness, "Computational Electrodynamics: The Finite-Difference Time-Domain Method," Artech House Publishers, 3rd ed. (2005).

[165] MIT. MEEP FDTD Software. http://ab-initio.mit.edu/wiki/index.php/meep (accessed on September 2009).

[166] B. Baker, and E. Copson, "The Mathematical Theory of Huygens' Principle," Oxford Univ. Press, 2nd ed., London (1950).

[167] M. Born, and E. Wolf, "Principles of Optics," McGraw-Hill, 6th ed., Cambridge (1997).

[168] P. Clemmow, "The Plane Wave Spectrum Representation of Electromagnetic Fields," Pergammon Press, London (1966).

[169] E. P. Wigner, "On the quantum correlation for thermodynamic equilibrium," Phys. Rev. **40**, 749 (1932).

[170] A. Belendez, C. Neipp, and I. Pascual, "Silver halide sensitized gelatin holograms in Slavich PFG-01 red-sensitive emulsion," J. of Mod. Opt. **46**, 1913 (1999).

[171] Q. Cao, and J. W. Goodman, "Wave-front inversion using a thin phase hologram: a computer simulation," Appl. Opt. **23**, 4575 (1984).

[172] Wu S., and E. N. Glytsis, "Holographic grating formation in photopolymers: analysis and experimental results based on a nonlocal diffusion model and rigorous coupled-wave analysis," JOSA B **20**, 12 (1974).

[173] U. S. Rhee, H. J. Caulfield, J. Shamir, C. S. Vikram, and M. M. Mirsalehi, " Characteristics of the DuPont photopolymer for angularly multiplexed page-oriented holographic memories," Opt. Eng. **32**, 1839 (1993).

[174] S. D. Wu, and E. N. Glytsis, "Characteristics of DuPont photopolymers for slanted holographic grating formations," JOSA B **21**, 1722 (2004).

[175] J. R. Lawrence, F. T. O'Neil, and J. Sheridan, "Photopolymer holographic recording material parameter estimation using a nonlocal diffusion based model," J. of Appl. Phys. **90**, 3142 (2001).

[176] G. Zhao, and P. Mouroulis, "Diffusion model of hologram formation in dry photopolymer materials," J. of Mod. Opt. **41**, 1929 (1994).

[177] U. S. Rhee, H. J. Caulfield, C. S. Vikram, and J. Shamir, "Dynamics of hologram recording in DuPont photopolymer," Appl. Opt. **34**, 846 (1995).

[178] C. Zhao, et al. "Shrinkage correction of volume phase holograms for optical interconnects," SPIE **3005**, 224 (1997).

[179] J. M. Watson, "Evaluation of Spatial-Spectral Filtering in Non-Paraxial Volume Holographic Imaging Systems," Master's Thesis, Massachusetts Institute of Technology, June (2008).

[180] L. W. Liebmann, and J. A. Carballo, "Layout Methodology Impact of Resolution Enhancement Techniques," Proc. of the 2003 Int. Symp. Of Phys. Des., 110 (2003).

[181] A. K.-K. Wong, "Resolution Enhancement Techniques in Optical Lithography," SPIE Press, Bellingham, WI (2001).

[182] S. Lee, J. A. Dominguez-Caballero, and G. Barbastathis, "Surface Relief Hologram Mask recording Simulation and Optimization base don SDTA in the Fresnel Diffraction Zone," Trans. of the KSME A **33**, 8 (2009).

[183] M. D. Stenner, A. Ashok, and M. A. Neifeld, "Multi-Domain Optimization for Ultra-Thin Cameras," in Frontiers in Optics, FWH5 (2006).

[184] Zemax Development Corporation, http://zemax.com, accessed on October (2009).

[185] J. N. Mait, et al., "Evolutionary paths in imaging and recent trends," Opt. Exp. **11**, 2093 (2003).

[186] DARPA, "Multiple Optical Non-redundant Aperture Generalized Sensors (MONTAGE) Program," http://www.darpa.mil/MTO/Programs/montage/index.html, accessed on September (2009).

[187] E. J. Tremblay, et al., "Ultrathin cameras using annular folded optics," Appl. Opt. **46**, 463 (2007).

[188] E. J. Tremblay, et al., "Relaxing the alignment and fabrication of thin annular folded imaging systems using wavefront coding," Appl. Opt. **46**, 6751 (2007).

[189] D. J. Brady, et al., "Compressive optical MONTAGE photography," Proc. SPIE **5907**, 590708 (2005).

[190] M. R. Shankar, et al., "Thin Infrared Imaging Systems through Multi-Channel Sampling," Appl. Opt. **47**, 10 (2008).

[191] M. A. Neifeld, et al., "Task-specific information for imaging system analysis," J. Opt. Soc. Am. A **24**, 12 (2007).

[192] J. Tanida, et al., "Compact image capturing system based on compound imaging and digital reconstruction," Proc. of SPIE **4455**, 34 (2001).

[193] J. Duparré, D. Radtke, A. Brückner, and A. Bräuer, "Latest Developments in Microoptical Artificial Compound Eyes: A Promising Approach for Next Generation Ultra-Compact Machine Vision," Proc. of SPIE-IS&T Elec. Imag. **6503**, 650301 (2007).

[194] J. Duparré, P. Dannberg, P. Schreiber, A. Bräuer, and A. Tünnermann, "Artificial apposition compound eye fabricated by micro-optics technology," Appl. Opt. **43**, 4303 (2004).

[195] E. R. Dowski Jr., and W. T. Cathey, "Extended depth of field through wavefront coding," Appl. Opt. **34**, 1859 (1995).

[196] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocusing," ACM SIGGRAPH (2007).

[197] G. Lippmann, "Epreuves reversibles donnant la sensation du relief," J. Phys. **7**, 821 (1908).

[198] J. Tanida, et al., "Thin observation module by bound optics (TOMBO): an optoelectronic image capturing system," SPIE **4089**, 1030 (2000).

[199] Y. Kitamura, et al., "Reconstruction of a high-resolution image on a compound-eye image-capturing system," Appl. Opt. **43**, 1719 (2004).

[200] J. Tanida, et al., "Thin Observation Module by Bound Optics (TOMBO): Concept and Experimental Verification," Appl. Opt. **40**, 1806 (2001).

[201] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-held Plenoptic Camera," Stanford Tech. Report CTSR (2005).

[202] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light Field Microscopy," Proc. SIGGRAPH **25**, 3 (2006).

[203] B. C. Platt, and R. Shack, "History and Principles of Shack-Hartmann Wavefron Sensing," J. of Ref. Sur. **17**, (2001).

[204] R. Shogenji et al., "Bimodal fingerprint capturing system based on compound-eye imaging module," Appl. Opt. **43**, 1355 (2004).

[205] J. Upatnieks, A. V. Lugt, and E. Leith, "Correction of Lens Aberrations by Means of Holograms," Appl. Opt. **5**, 589 (1966).

[206] J. Munch, and R. Wuerker, "Holographic technique for correcting aberrations in a telescope," Appl. Opt. **28**, 1312 (1989).

[207] J. Munch, R. Wuerker, and L. Heflinger, "Wideband holographic correction of an aberrated telescope," Appl. Opt. **29**, 2440 (1990).

[208] G. Andersen, and R. J. Knize, "Holographically corrected telescope for high-bandwidth optical communications," Appl. Opt. **38**, 6833 (1999).

[209] G. Andersen, and R. J. Knize, "A high resolution, holographically corrected microscope with a Fresnel lens objective at large working distances," Opt. Exp. **2**, 546 (1998).

[210] G. Andersen, and R. J. Knize, "Holographically corrected microscope with a large working distance," Appl. Opt. **37**, 1849 (1998).

[211] H. M-Zolbanine, and C. Froehly, "Holographic correction of both chromatic and spherical aberrations of single glass lenses," Appl. Opt. **18**, 2385 (1979).

[212] E. N. Leith, and J. Upatnieks, "Holographic Imagery Through Diffusing Media," J. of the Opt. Soc. of Am. **56**, 523 (1966).

[213] P. L. Ransom, "Proposal for holographic imaging through phase-distorting media without alignment," Opt. Lett. **5**, 327 (1980).

[214] NSG America, Inc., Website: http://www.nsgamerica.com/, Accessed on October (2009).

[215] V. N. Mahajan, "Aberrations Theory Made Simple," SPIE Opt. Eng. Press (1991).

[216] M. A. Neifeld, "Information, resolution, and space-bandwidth product," Opt. Lett. **23**, 1477 (1998).

[217] T. M. Cover, and J. A. Thomas, "Elements of Information Theory," Wiley Series in Telecommunications, 2nd ed. (1991).

[218] S. Hranilovic, and F. R. Kschischang, "Capacity Bounds for Power- and Band-Limited Optical Intensity Channels Corrupted by Gaussian Noise," IEEE Trans. on Inf. Theory **50**, 5 (2004).

[219] E. H. Linfoot, "Information Theory and Optical Images," J. of the Opt. Soc. of Am. **45**, 10 (1955).

[220] C. Eckart, and G. Young, "The approximation of one matrix by another of low rank," Psychometrika **1**, 3 (1936).

[221] M. Bertero, and P. Boccacci, "Introduction to Inverse Problems in Imaging," Inst. of Phys. Pub. (2002).

[222] A. Ashok, and M. A. Neifeld, "Recent progress on multi-domain optimization for ultra-thin cameras," Proc. of SPIE **232**, 62320N (2006).

[223] S. S. Young, and R. G. Driggers, "Superresolution image reconstruction from a sequence of aliased imagery," Appl. Opt. **45**, 5073 (2006).

[224] A. V. Kanaev, D. A. Scribner, J. R. Ackerman, and E. F. Fleet, "Analysis and application of multiframe superresolution processing for conventional imaging systems and lenslet arrays," Appl. Opt. **46**, 4320 (2007).

[225] R. C. Hardie, K. J. Banard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," IEEE Trans. Image Process. **6**, 1621 (1997).

[226] M. S. Alam, J. G. Bognar, S. Cain, and B. J. Yasuda, "Fast registration and reconstruction of aliased, low-resolution frames by use of a modified maximum-likelihood approach," Appl. Opt. **37**, 1319 (1998).

[227] M. S. Alam, J. G. Bognar, R. C. Hardie, and B. J. Yasuda, "High resolution image reconstruction using multiple, randomly shifted, low resolution, aliased frames," Proc. SPIE **3063**, 102 (1997).

[228] S. Letrattanapanich, and N. K. Bose, "High resolution image formation from low resolution frames using Delaunay Triangulation," IEEE Trans. Image Proc. **11**, 1427 (2002).

[229] R. Y. Tsai, and T. S. Huang, "Multiple frame image restoration and registration," in Adv. in Comp. Vis. and Image Proc. **314**, (1984).

[230] R. W. Gerchberg, "Superresolution through error energy reduction," Opt. Acta **21**, 709 (1974).

[231] K. Bolla, J. Schmidt, J. T. Sheridan, N. Streibl, and R. Volkel, "Holographic optical beam splitters in dichromated gelatin," J. of Mod. Opt. **39**, 881 (1992).

[232] R. K. Kostuk, "Dynamic Hologram Recording Characteristics in DuPont Photopolymers," Appl. Opt. **38**, 1357 (1999).

[233] C. Neipp, A. Marquez, I. Pascual, and A. Beléndez, "Thick phase holographic gratings recorded on BB-640 and PFG-01 silver halide materials," J. Opt. A: Pure Appl. Opt. **5**, S183 (2003).

[234] P. Hariharan, and C. M. Chidley, "Rehalogenating bleaches for photographic phase holograms: the influence of halide type and concentration on diffraction efficiency and scattering," Appl. Opt. **26**, 3895 (1987).

[235] C. Neipp, I. Pascual, and A. Beléndez, "Bleached silver halide volume holograms recorded on Slavich PFG-01 emulsion: the influence of the developer," J. of Mod. Opt. **48**, 1479 (2001).

[236] A. Beléndez, C. Neipp, and I. Pascual, "Silver halide sensitized gelatin holograms in Slavich PFG-01 red-sensitive emulsion," J. of Mod. Opt. **46**, 1913 (1999).

[237] J. M. Kim, et al., "Holographic optical elements recorded in silver halide sensitized gelatin emulsions. Part I. Transmission holographic optical elements," Appl. Opt. **40**, 622 (2001).

[238] A. Beléndez, T. Beléndez, C. Neipp, and I. Pascual, "Determination of the refractive index and thickness of holographic silver halide materials by use of polarized reflectances," Appl. Opt. **41**, 6802 (2002).

[239] L. M. Foucault, "Mémoire sur la construction des telescopes en verre argenté," Ann. Obs. Imp. Paris **5**, 197 (1859).

[240] F. Zernike, "Diffraction theory of the knife-edge test and its improved form, the phase-contrast method," JM^3 **1**, 87 (2002).

[241] W. D. Furlan, et al., "Optical aberrations measurement with a low cost optometric instrument," Am. J. Phys. **70**, 857 (2002).

[242] J. Tanida, et al., "Color imaging with an integrated compound imaging system," Opt. Exp. **11**, 2109 (2003).

[243] R. J. Adrian, "Particle-imaging techniques for experimental fluid mechanics," Annu. Rev. Fluid Mech. **23**, 261 (1991).

[244] R. J. Adrian, S. M. Soloff, Z. C. Liu, C. D. Meinhart, and W. Lai, "Stereoscopic PIV Applications to the Study of Turbulence," Workshop on PIV-Fukui, July 8-11, Japan (1997).

[245] M. P. Arroyo, and C. A. Greated, "Stereoscopic particle velocimetry," Meas. Sci. Technol. **2**, 1181 (1991).

[246] Y. G. Guezennec, Y. Zhao, and T. J. Gieseke, "High-speed 3-D scanning particle image velocimetry (3-D SPIV) technique," Proc. Laser Symp. **26**, 1 (1994).

[247] K. D. Hinsh, "Holographic particle image velocimetry," Meas. Sci. Technol. **13**, R61 (2002).

[248] Y. Pu, and H. Meng, "An advanced off-axis holographic particle image velocimetry (HPIV) system," Exp. in Fluids **29**, 184 (2000).

[249] H. Meng, and F. Hussian, "In-line Recording and Off-axis Viewing (IROV) technique for holographic particle velocimetry," Appl. Opt. **34**, 1827 (1995).

[250] Y. Pu, and H. Meng, "Four-dimensional dynamic flow measurement by holographic image velocimetry," Appl. Opt. **44**, 7696 (2005).

[251] K. Huang, J. Slepicka, and S. S. Cha, "Cross-correlation of three-dimensional images for three-dimensional three-component fluid velocity measurements," SPIE **2005**, 655 (1993).

[252] J. Sheng, and H. Meng, "A Genetic Algorithm approach for 3D velocity field extraction in holographic particle image velocimetry," Exp. in Fluids **25**, 461 (1998).

[253] Y. Pu, X. Song, and H. Meng, "Off-axis holographic particle image velocimetry for diagnosing particulate flows," Exp. in Fluids (Suppl.) S117 (2000).

[254] H. Meng, et al., "Holographic particle image velocimetry: from film to digital recording," Meas. Sci. Technol. **15**, 673 (2004).

[255] W. Yang, A. B. Kostinski, and R. A. Shaw, "Depth-of-focus reduction for digital in-line holography of particle fields," Opt. Lett. **30**, 1203 (2005).

[256] C. Fournier, C. Ducottet, and T. Fournel, "Digital in-line holography: influence of the reconstruction function on the axial profile of a reconstructed particle image," Meas. Sci. Technol. **15**, 686 (2004).

[257] S. Coëtmellec, et al., "Application of in-line digital holography to multiple plane velocimetry," Meas. Sci. Technol. **12**, 1392 (2001).

[258] M. Malek, et al., "Digital in-line holography for three-dimensional-two-components particle tracking velocimetry," Meas. Sci. Technol. **15**, 699 (2004).

[259] N. Salah, et al., "Application of multiple exposure digital in-line holography to particle tracking in a Bénard-von Kármán vortex flow," Meas. Sci. Technol. **19**, 074001 (2008).

[260] V. Kebbel, et al., "Digital holography as a versatile optical diagnostic method for microgravity experiments," Meas. Sci. Technol. **10**, 893 (1999).

[261] J. P. Fugal, et al., "Airborne digital holographic system for cloud particle measurements," Appl. Opt. **43**, 5987 (2004).

[262] J. A. Domínguez-Caballero, et al., "Techniques Based on Digital Multiplexing Holography for Three-Dimensional Object Tracking," Conf. on Lasers and Elec.-Opt., Baltimore, (2007).

[263] S. Kim, and S. J. Lee, "Effect of particle number density in in-line digital holographic particle velocimetry," Exp. Fluids **44**, 623 (2008).

[264] H. Meng, W. L. Anderson, F. Hussain, and D. D. Liu, "Intrinsic speckle noise in in-line particle holography," J. Opt. Soc. Am. A **10**, 2046 (1993).

[265] Y. Pu, and H. Meng, "Intrinsic speckle noise in off-axis particle holography," J. Opt. Soc. Am. A **21**, 1221 (2004).

[266] M. Malek, D. Allano, S. Coëtmellec, and D. Lebrun, "Digital in-line holography: influence of the shadow density on particle field extraction," Opt. Exp. **12**, 2270 (2004).

[267] W. D. Koek, N. Bhattacharya, J. M. Braat, T. A. Ooms, and J. Westerweel, "Influence of virtual images on the signal-to-noise ration in digital in-line particle holography," Opt. Exp. **13**, 2578 (2005).

[268] Y. Pu, and H. Meng, "Intrinsic aberrations due to Mie scattering in particle holography," J. Opt. soc. Am. A **20**, 1920 (2003).

[269] L. B. Kiss, et al., "New approach to the origin of lognormal size distributions of nanoparticles," Nanotech. **10**, 25 (1999).

[270] J. Sheng, et al., "Single beam two-views holographic particle image velocimetry," Appl. Opt. **42**, 235 (2003).

Appendix A

Additional CGH Optimization Results

In this appendix, additional optimization examples of the design of CGHs are presented. The first example consists of optimizing an in-line binary phase CGH designed to reconstruct a gate pattern that is used in the fabrication of liquid crystal displays (LCDs). This CGH is optimized using the encoding strategy based on the LDPE mask. Four different values of the diffuser factor are considered showing how the LDPE mask helps to spread the encoded information over the entire hologram window and improve the quality of the reconstructed signal. The diffuser factor values range from 0 to 1. In the case of $D_{\text{factor}} = 0$, the problem reduces to the diffracted field encoding strategy. The simulation parameters are indicated in Table A.1. Figure A-1 shows the resulting optimized CGHs for progressively increasing diffuser factors. For the CGH optimized with $D_{\text{factor}} = 0$, the encoded information is not uniformly spread over the hologram window. This results in large empty patches showing that the hologram is not using the available 16 million degrees of freedom (16 million bits of information capacity). The information spread increases with larger diffuser factors progressively covering the entire hologram window. The corresponding reconstructed amplitude distributions close ups are shown in Figure A-2. This reconstruction example demonstrates how the LDPE mask helps in

the reduction and elimination of background noise, due to undesirable diffraction effects and improving the hologram's diffraction efficiency and uniformity of its reconstruction. These results are not fully optimized and a better solution can be found by implementing the HOA to find the optimum diffuser and frequency factors.

Table A.1: Optimization Parameters: In-line CGH - Gate Pattern.

Wavelength (λ)	364nm	**Object Window** (O_{size})	350μm
Working Distance (d)	200μm	**Hologram Size** (H_{size})	400μm
Pixel Size (δ_{pix})	100nm	**Number of Iterations**	100

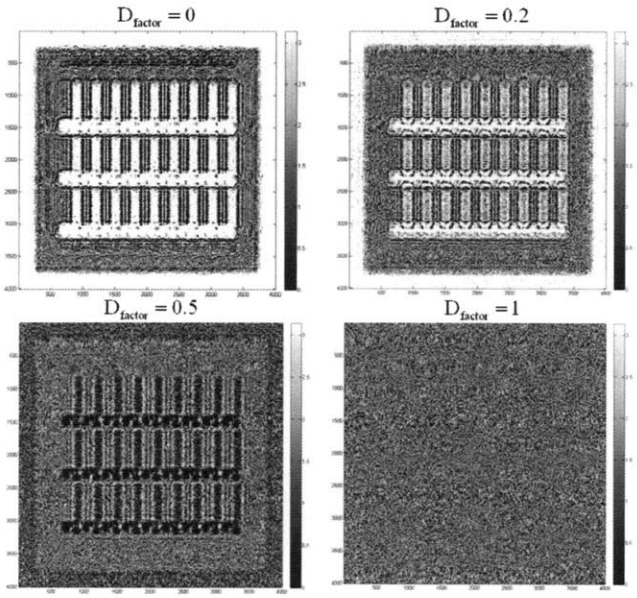


Figure A-1: Optimized CGHs for different diffuser factors.

The second example is the optimization of a binary phase CGH using the LDPE mask, designed to reconstruct a 2×2 array of the resolution target of Figure 2-84-a. The simulation parameters are indicated in Table A.2. The final optimized CGH is shown in Figure A-3-a. The corresponding reconstructed amplitude distribution is shown in Figure A-3-b.

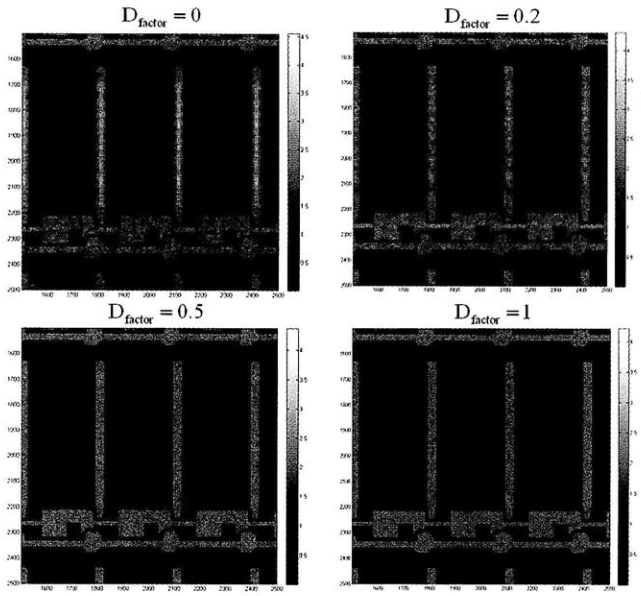


Figure A-2: Reconstructed amplitudes from CGHs designed with different diffuser factors.

Table A.2: Optimization Parameters: In-line CGH - Resolution Target Array.

Wavelength (λ)	364nm	**SBP (After Padding)**	2000×2000
Working Distance (d)	200μm	**Hologram Size (H_{size})**	400μm
Pixel Size (δ_{pix})	100nm	**Number of Iterations**	150

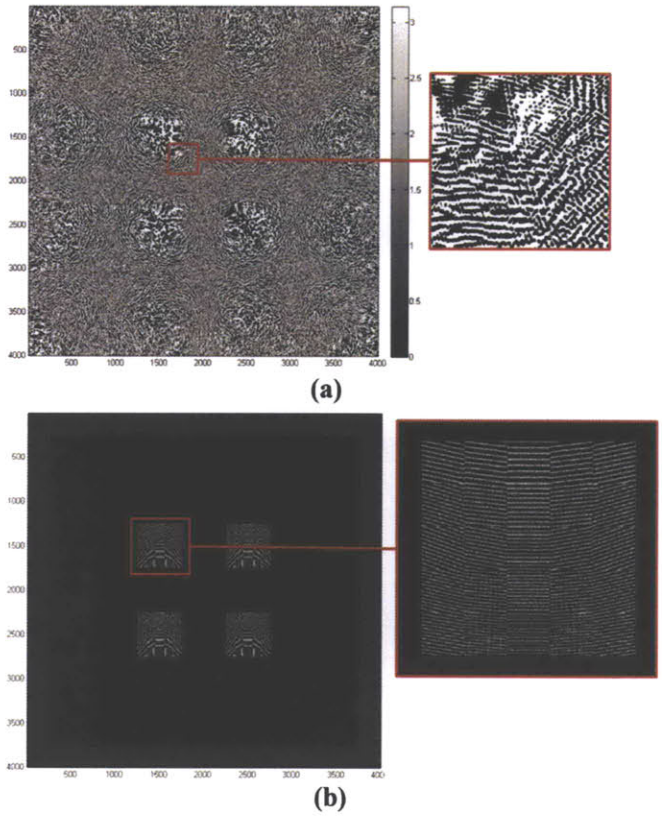


Figure A-3: (a) Optimized CGH; (b) Reconstructed amplitude.