Towards a Unified Treatment of 3D Display using Partially Coherent Light

by

Roarke Horstmeyer

B.S. Physics, Duke University, 2007
Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
in partial fulfillment of the requirements for the degree of
Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author

Program in Media Arts and Sciences, School of Architecture and Planning

August 6, 2011

Certified by

Ramesh Raskar
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by

Prof. Mitchel Resnick
LEGOS Papert Professor of Learning Research, Academic Head,
Program in Media Arts and Sciences
Towards a Unified Treatment of 3D Display using Partially Coherent Light

by

Roarke Horstmeyer

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning on August 6, 2011, in partial fulfillment of the requirements for the degree of Master of Science

Abstract

This thesis develops a novel method of decomposing a 3D phase space description of light into multiple partially coherent modes, and applies this decomposition to the creation of a more flexible 3D display format. Any type of light, whether it is completely coherent, partially coherent or incoherent, can be modeled either as a sum of coherent waves or as rays. A set of functions, known as phase space functions, provide an intuitive model for these waves or rays as they pass through a 3D volume to a display viewer’s eyes. First, this thesis uses phase space functions to mathematically demonstrate the limitations of two popular 3D display setups: parallax barriers and coherent holograms. Second, this thesis develops a 3D image design algorithm based in phase space. The “mode-selection” algorithm can find an optimal holographic display setup to create any desired 3D image. It is based on an iterative algebraic-rank restriction process, and can be extended to model light with an arbitrary degree of partial coherence. Third, insights gained from partially coherent phase space representations lead to the suggestion of a new form of 3D display, implemented with multiple time-sequential diffracting screens. The mode-selection algorithm determines an optimal set of diffracting screens to display within the flicker-fusion rate of a viewer’s eye. It is demonstrated both through simulation and experiment that this time-sequential display offers improved performance over a fixed holographic display, creating 3D images with increased intensity variation along depth. Finally, this thesis investigates the trade-offs involved with multiplexing a holographic display over time with well-known strategies of multiplexing over space, illumination angle and wavelength. The examination of multiplexing trade-offs is extended into the incoherent realm, where comparisons to ray-based 3D displays can hopefully offer a more unified summary of the limitations of controlling light within a volume.

Thesis Supervisor: Ramesh Raskar
Title: Associate Professor of Media Arts and Sciences
Towards a Unified Treatment of 3D Display using Partially Coherent Light

by

Roarke Horstmeyer

The following person served as a reader for this thesis:

Reader: Prof. George Barbastathis
Professor of Mechanical Engineering
MIT Department of Mechanical Engineering
Towards a Unified Treatment of 3D Display using Partially Coherent Light

by

Roarke Horstmeyer

The following person served as a reader for this thesis:

Reader: [Signature]

Prof. Joeseph Paradiso
Associate Professor of Media Art and Sciences
MIT Media Arts and Sciences Program
Towards a Unified Treatment of 3D Display using Partially Coherent Light

by

Roarke Horstmeyer

The following person served as a reader for this thesis:

Reader: ____________________________

V. Michael Bove, Jr.
Principal Research Scientist
MIT Media Lab
Acknowledgements

Many people have contributed greatly to the success of this project. I would like to thank my advisor, Prof. Ramesh Raskar, for his guidance, support, and encouragement over the past two years.

In addition, I would like to thank Se Baek Oh for many hours of talks, lessons, and investigations into all areas of optics. Furthermore, I thank Prof. Paradiso and Prof. Bove for their helpful comments and suggestions. I would also like to thank Douglas Lanman, Tom Cuypers and Otkrist Gupta for their continued efforts towards the project, as well as Ankit Mohan, Yunhee Kim, Andreas Velten, Jaewon Kim, Rohit Pandharkar, Kevin Chiu, Andy Bardagjy, Nikhil Naik and Gordon Wetzstein for their expertise in fields unknown to me.

I would also like to thank Laura Waller, Lei Tian, Hanhong Gao, Shalin Mehta, Jon Petrucelli, Prof. George Barbastathis and others who have been or currently are associated with the 3D Optical Systems group in the MIT Department of Mechanical Engineering for helpful discussions and thought-provoking work. Additionally, Zhengyun Zheng from Stanford University’s Electrical Engineering Department offered numerous helpful suggestions and initially paved the way for many of these proposed concepts and experiments.
Contents

1 Introduction .................................................. 19
   1.1 Motivation .................................................. 19
   1.2 Related Work .............................................. 21

2 3D Display Basics .............................................. 25
   2.1 Background .................................................. 25
   2.2 Parallax Barrier Operation ................................. 28
   2.3 Simplified Hologram Operation ......................... 30
      2.3.1 Hologram Discretization .......................... 31
   2.4 A Numerical Comparison .................................... 33

3 Phase Space Functions ........................................ 35
   3.1 Incoherent Light: The Light Field .................... 36
   3.2 Coherent Light: The Wigner Distribution and Ambiguity Function
      Representations ............................................. 37
      3.2.1 One Wavefront as Many Plane Waves ............... 39
      3.2.2 A Simple Light Propagation Model ................. 41
      3.2.3 The Ambiguity Function ............................ 43
   3.3 Light Fields and WDFs: A Display Example ............ 47
   3.4 Limitations of Incoherent Parallax Barriers and Coherent Holograms 49
      3.4.1 Rank-1 Geometric Light Field .................... 49
      3.4.2 Rank-1 Holographic Light Field ................. 51
4 Multiplexing for Multimode 3D Display Design

4.1 Modes, Partial Coherence and Multiplexing

4.2 Single Mode Selection Algorithm
   4.2.1 Fixed Holograms: Coherent, One Mode

4.3 Single-Mode Selection Performance
   4.3.1 Algorithm Performance, Ground Truth Intensities
   4.3.2 Algorithm Performance, Desired Intensities

4.4 Multi-Mode Selection and Partial Coherence
   4.4.1 Multiple Modes: A Partially Coherent Constraint

4.5 Designing Partially Coherent 3D Images
   4.5.1 Ground Truth Multi-Mode 3D Image Reconstruction
   4.5.2 Desired Multi-Mode 3D Image Design

4.6 Alternative Multiplexed Displays and Partial Coherence
   4.6.1 General Categorization of Multiplexing
   4.6.2 Angular Multiplexing: A Hybrid Design Example

5 Designing 3D Diffractive Displays with Constraints

5.1 Constrained Multi-Mode Selection Failure

5.2 A New Amplitude-Only Constraint

5.3 A New Phase-Only Constraint

6 Experimental Investigations

6.1 Fixed Hologram Design Experiments

6.2 Dynamic Hologram Design Experiments
   6.2.1 Application to Speckle Reduction
   6.2.2 Increased Intensity Variation Along Depth
   6.2.3 Discussion and Limitations

7 Conclusion

7.1 Future Work
   7.1.1 Further Connections Between Rays and Waves
7.1.2 An Improved Algorithm .................................. 108
7.1.3 Benefits Beyond 3D Display .............................. 109

Bibliography ......................................................... 111
## List of Figures

<table>
<thead>
<tr>
<th>Figure No.</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Parallax barrier and hologram introduction</td>
<td>26</td>
</tr>
<tr>
<td>2-2</td>
<td>Parallax barrier and hologram similarities in phase space</td>
<td>27</td>
</tr>
<tr>
<td>2-3</td>
<td>Simple parallax barrier and hologram as labeled diagrams</td>
<td>29</td>
</tr>
<tr>
<td>2-4</td>
<td>Discretization of the parallax barrier and hologram</td>
<td>31</td>
</tr>
<tr>
<td>2-5</td>
<td>Numerical example of parallax barrier and hologram performance</td>
<td>32</td>
</tr>
<tr>
<td>3-1</td>
<td>A simple explanation of the geometric light field</td>
<td>36</td>
</tr>
<tr>
<td>3-2</td>
<td>Spatial frequency and ray angle</td>
<td>40</td>
</tr>
<tr>
<td>3-3</td>
<td>Wigner distribution transformations</td>
<td>42</td>
</tr>
<tr>
<td>3-4</td>
<td>Hologram and ambiguity function relationship</td>
<td>43</td>
</tr>
<tr>
<td>3-5</td>
<td>Parallax barrier and hologram model of 2 points at different depths</td>
<td>48</td>
</tr>
<tr>
<td>3-6</td>
<td>Rank-1 light field</td>
<td>50</td>
</tr>
<tr>
<td>3-7</td>
<td>Rank-1 Wigner distribution</td>
<td>52</td>
</tr>
<tr>
<td>4-1</td>
<td>Diagram of the mode-selection algorithm</td>
<td>58</td>
</tr>
<tr>
<td>4-2</td>
<td>The mode-selection algorithm’s SVD constraint</td>
<td>61</td>
</tr>
<tr>
<td>4-3</td>
<td>Re-creating a binary amplitude hologram with mode-selection</td>
<td>63</td>
</tr>
<tr>
<td>4-4</td>
<td>Re-creating a cubic phase screen with mode-selection</td>
<td>64</td>
</tr>
<tr>
<td>4-5</td>
<td>Algorithm convergence plots, known intensities</td>
<td>65</td>
</tr>
<tr>
<td>4-6</td>
<td>Algorithm performance vs. depth and inputs, known intensities</td>
<td>66</td>
</tr>
<tr>
<td>4-7</td>
<td>Mode-selection design of a desired 3D image</td>
<td>68</td>
</tr>
<tr>
<td>4-8</td>
<td>Optimized holograms with different constraints and performance</td>
<td>69</td>
</tr>
<tr>
<td>4-9</td>
<td>Partially coherent mode-selection with an SVD</td>
<td>71</td>
</tr>
<tr>
<td>4-10</td>
<td>Partially coherent mode-selection to re-create a known field</td>
<td>73</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation

Often, autostereoscopic three-dimensional (3D) displays are discussed and analyzed under two different conditions of light. Parallax barriers, lenticular displays and integral imaging devices are characterized by how well they can direct different rays to different viewing locations, under the assumption of an incoherent light source. Holographic displays, on the other hand, are often characterized by how well they can diffract a wavefront into a desired intensity distribution, typically assuming illumination from a coherent source. While both display genres are applied to the same goal of delivering a 3D image to a viewer, their use of different forms of light hinders direct comparison. What's more, they both suffer certain limitations. For example, parallax barrier-type displays are not able to use the effects of wave interference to produce quickly varying intensity distributions along depth. Likewise, coherent holograms have a strict requirement for all intensities to arise from a single propagating wavefront, yielding effects like speckle and intensity coupling [1]. As autostereoscopic 3D display technologies progress towards smaller feature sizes, the mixture of ray and wave-based effects in image formation will inevitably call for a more unified method of analysis and evaluation.

The overall aim of this thesis is to initiate a framework in which the geometrical and
physical optic-based performance limitations of a 3D display can be mathematically characterized. While this aim is not realized in full, three related steps offer an initial starting point for future work towards the merging of 3D display analyses. First, this thesis motivates the problem of current state-of-the-art display techniques, answering the question, why do 3D displays like parallax barriers and holograms need to be improved upon at all? One interpretation of the restrictions of current incoherent-based parallax barrier displays and coherent-based holographic displays is presented, using a linear algebra-based analysis.

Second, this thesis uses insights gained from the derived mathematical limitations of coherent holographic 3D display to develop a new design model, termed mode-selective. A “mode-selection” algorithm determines the optimal 2D diffractive screen pattern that generates a desired 3D intensity distribution within a given viewing volume. Unlike similar 3D intensity design methods, the proposed algorithm extends quite simply from modeling a perfectly coherent light source to allow for an arbitrary degree of partial coherence. The coherence state of the light can be defined as an additional input to the algorithm, or can be a variable that is also optimized over. If included in the optimization, an optimal diffractive screen pattern and source coherence state can be identified to generate any desired 3D intensity distribution. Within the algorithm, the lights coherence state is represented in terms of a modal distribution, where one coherent mode represents a completely coherent system, and more modes represent an increasing degree of partial coherence. At the limit of many modes, the mode-selection model approaches an incoherent, ray-based model. The validity of the mode-selection model is verified through several experiments, with further detail provided in Chapters 4, 5 and 6.

Third, this thesis contributes a discussion of partially coherent light source alongside the concept of multiplexing a 3D display. As it is difficult in practice to design the specific coherence state of a display’s illumination, various multiplexing methods are instead proposed to mimic partial coherence. Multiplexing a 3D display over time
(i.e., quickly showing multiple diffraction patterns on a screen) has been previously proposed for holographic display in [2], and is investigated in this thesis experimentally. Additionally, spatial and angular multiplexing methods are explored as viable alternatives. Previous examples of multiplexing over space, angle, wavelength and other modalities are categorized in a table to encourage the comparison and possible merger of many varied 3D display architectures.

In summary, this thesis is intended as a broad overview of 3D display from a relatively new perspective, building towards a novel 3D image design method and experimentally tested time-multiplexed display setup. The proposed framework's overarching goal is to join together two very different mathematical treatments of ray and wave optics. While at times mathematically cumbersome, the reward is a novel method of jointly optimizing a holographic display's light and screen to show a given image. Furthermore, as mentioned above, development of both the framework and the design model are by no means complete. Hopefully, they will serve as a starting point towards future work in connecting diffractive, refractive and attenuation-based 3D displays. Future studies could build towards developing 3D display designs that take advantage of both diffractive and refractive elements. Or, for a given 3D display, the mode-selection model can be used to generate optimal display content and parameters given a desired 3D image. Hopefully, the simple partial coherence-based device this thesis experimentally tests will be one of many future display formats to use the proposed optimization framework. It seems this may be possible, as the areas of optical design, dynamic displays and computational optimization continue to merge and overlap.

1.2 Related Work

Following is a brief outline of prior work, roughly grouped into computational procedures, holographic designs and ray-based designs for 3D display. Specific background material on the mode-selective method is also offered, including work on partial co-
herence and matrix decompositions for the unfamiliar reader.

The goal of displaying a 3D image is closely related to fully defining a spatial distribution of light. There are four general categories of determining the amplitude and phase of a wavefront from different intensity measurements along the direction of propagation. The first is phase retrieval [4, 5], which is a non-linear, iterative process. The second is phase space tomography, which uses many intensity measurements and tomographic-based reconstruction [6, 7]. The third is transport of intensity, which estimates the wavefront from two close planes of intensity [8, 9], and the fourth uses projective-based algorithms [10]. The proposed algorithm, discussed in detail in Chapter 4, uses a few desired intensity inputs and applies a unique constraint, and unlike previous work can define a specific degree of coherence during optimization. Additionally, there are numerous methods of designing a 3D distribution of light based solely on geometrical optics, namely, the light field [11]. Recent work has drawn connections between these geometric models and wave-based models in phase space [12, 13] like the Wigner distribution [14] and the ambiguity function [15, 16]. These functions will provide the basis for the proposed design method, and additional references for the interested reader will be provided during their introduction in Chapter 3. Finally, the “augmented light field” [13, 17] offers one framework to join geometric and wave-based optics. This thesis builds on prior work developing the augmented light field within a modeling context.

Focusing on the area of 3D display, this thesis will consider how to optimally compute different display patterns. The area of computer-generated holography (CGH) will be of particular interest. Started by Lohmann [18], the field has evolved to model intensities at multiple planes [1] and into a continuous volume [19]. Good summaries can be found in [20, 21]. As I will describe in the thesis, the concept of partial coherence relates closely to segmenting holograms into different zones. A large amount of early film-based work considered this technique [22, 23]. Additionally, spatially multiplexed holograms, like a holographic stereogram [24], display many discrete viewpoints of an
object similar to a parallax barrier display. They too often exhibit only parallax along the horizontal direction, but not always, and provide a user with the appearance of a fully 3D object [25]. Members of this general category of “advanced” holographic display methods that often utilize a form of multiplexing are discussed in [26, 27, 28], and summarized in [29].

The first ray-based parallax barrier and lenticular 3D displays were developed by Ives [30] and Lippmann [31], respectively. In-depth comparisons between lenticular, barrier and similarly related integral imaging systems can be found in [32], and a physical optics perspective of these devices is in [33]. Several works have drawn some simple comparisons between what is possible with geometric-based and holographic displays [34, 35]. Others have integrated display forms from the two genres [36, 37]. I hope to expand on these general comparisons using a phase space analysis. Holograms in specific have been analyzed from geometric [38] and wave-based [39] phase space perspectives, but not directly compared to parallax barriers. Finally, the general issue of multiplexing an image over a screen for incoherent 3D display is considered in [40]. I hope to extend this analysis to consider both coherent and incoherent-based displays categorized into various groups.

With regards to the specifics of the proposed mode-selection algorithm and time-multiplexed display, several works have led up to their generation and connection. The original algorithm considered modeling spatially coherent light [41] and was initially applied to the design of a camera’s point-spread function [3]. It was recently extended to consider partially coherent light [42]. Many insights into a linear algebraic treatment of partial coherence were gained from Ozaktas et al. [43], and insights into partially coherent phase space functions were gained from Bastiaans [44] and Wolf [45], among others. The specific implementation of a time-varying diffractive pattern to approximate a partially coherent wavefront can be traced back to Desantis et al. [2]. While based upon a large amount of prior work, the coupling of the proposed algorithm with the suggested experimental realization in [2] is a novel con-
cept, as is the construction of an encompassing framework of the spatial, temporal and angular multiplexing, used by coherent and incoherent displays, based in phase space.

The general study of partial coherence has offered numerous insights into the design and generation of the proposed display. For the interested reader, a good mathematical description of partial coherence is offered in [46]. A more qualitative perspective, including many examples, is in [47]. An initial discussion of applying partial coherence to phase space functions is in [44], with a more in-depth model in [48]. Likewise, methods of determining partially coherent fields from multiple intensities is discussed in [49, 50]. The coherent mode decomposition of a partially coherent source, which this thesis will use often, was first discussed in [45]. Alternative methods of decomposing partially coherent light into different modes is discussed for use with x-rays in [51]. A good summary of different forms of field decompositions is in [10]. Finally, the experimental implementation of the proposed mode-selection algorithm on a spatial light modulator relies heavily on well-researched computational methods like the singular-value decomposition, with properties discussed in [52]. One alternative discussed in this thesis relies on non-negative matrix factorization [53]. The symmetric form of this factorization is explored mathematically in [54].
Chapter 2

3D Display Basics

2.1 Background

Over the past century, a large number of interesting display configurations have attempted to offer 3D viewing of imagery or film. Due to the medium’s complexity, one of the more popular methods of presenting three dimensional content is stereoscopically, or by displaying two horizontally offset images to each eye of a viewer. Glasses-based stereoscopic displays, which use polarizing or spectral filters (i.e., anaglyphs) or alternating shutters to allow only one offset image to pass to each eye, are currently experiencing a resurgence in popularity both at the movie theater and in home theater systems. Another format of stereoscopic display that will increasingly appear in coming years on hand-held devices uses optimized parallax barrier technology, first proposed by Ives [30] over a century ago. These “autostereoscopic” parallax barrier displays interlace two images on the display screen and use a lenticular array or a series of strips placed over each image to ensure it is directed to only one eye of the viewer. In other words, the addition of a second screen of slits effectively replaces the requirement of glasses for viewing the content.

While attractive for their simplicity, two-view autostereoscopic displays currently suffer from a number of setbacks, mostly connected with their inability to present any depth cues beyond image disparity. Specifically, by only approximating the appear-
ance of a 3D scene using two images, viewers often have issues with optical accommodation and convergence. These issues can lead to discomfort and visual fatigue [55]. What’s more, two-view glasses-free displays are optimized for a single viewer in a fixed position. Any movement of the viewer can lead to aliasing and pseudoscopic views. It is a challenge to present correct stereoscopic imagery to more than one viewer or a moving viewer. Efforts have been made to overcome this problem by tracking a viewer’s location, for example [56], which may present its own challenges.

An alternative to two-view autostereoscopic display is multi-view autostereoscopic display, which attempts to physically recreate a 3D volume of light, presenting the appearance of a 3D image at a specific location within a given viewing angle. Multi-view autostereoscopic displays will be the focus of the rest of this thesis, and can be generalized into two categories, although exceptions to a two-category generalization certainly exist. The first category of displays is based upon incoherent light and are referred to as ray-based displays, while the second category is based upon coherent light and are referred to as wave-based displays.

Ray-based displays are modeled assuming light travels as a ray and offer the same

![Common Questions about Holograms](image)

**Common Questions about Holograms**

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can a hologram be represented as a geometric light field?</td>
</tr>
<tr>
<td>Can a hologram create any intensity distribution we want in 3D?</td>
</tr>
<tr>
<td>Why does a hologram create a &quot;wavefront,&quot; while a parallax barrier does not?</td>
</tr>
</tbody>
</table>

Figure 2-1: (a) A parallax barrier (a screen of pixels covered by a series of slits) can easily create a specific ray, but can only create a discrete approximation of a curved wavefront. (b) An amplitude hologram (one screen of wavelength-scale pixels) can create the same ray with a sinusoidal pattern, but will unavoidably generate two additional rays. It can easily create a curved wavefront with a narrow opening.
Figure 2-2: A parallax barrier display (left) emits three rays at different angles from a single slit with three pixels \((p_1, p_2, p_3)\) turned on. A sinusoidal grating (right top) likewise produces three rays \((r_1, r_2, r_3)\) through diffraction. These can be visualized as the angular spectrum of the grating’s wave-based light field, otherwise known as a Wigner distribution \(W(x, u)\). We encourage the merger of the parallax barrier’s light field with the hologram’s Wigner distribution into a general framework for 3D display, as explained in Chapters 3 and 4.

performance regardless of the light’s wavelength. Examples of autostereoscopic 3D displays that use incoherent light include lenticular, integral imaging, rotating-mirror and parallax barrier-based devices, among others [56]. As connections between ray and wave-based optics progress, this thesis will focus on displays that use one or more thin, fixed screens that modulate light, for simplicity. The two thin attenuation screens of a parallax barrier will be a simple starting point, and share many similarities with lenticular and integral imaging display setups. In their most basic form, autostereoscopic parallax barriers are comprised of a plane of pixels and a plane of light-modulating slits (Fig. 2-1(a)). The pixels contain a mix of spatial and angular content in the form of of multiple interlaced images, and the slits direct rays from each image to different viewing locations. Properly calibrated, a volume of space within the intersection cone of all emitting rays can be filled with a discretized 3D image, thus offering glasses-free 3D content to a user at an arbitrary viewing position within the viewing cone.
The second category of 3D autostereoscopic displays we will consider are diffractive, wave-based displays, which are simply summarized as holographic displays. Holograms are based on the principle of diffraction and contain a mixture of spatial and angular information in the form of interference fringes (Fig. 2-1(b)). They typically require illumination by light with at least a small degree of spatial coherence, with some forms requiring light that is highly spatially coherent. A good introduction to the many different forms of holography can be found in [57]. Most recent work in holography has focused on generating holographic fringes computationally, which this paper will also concentrate on. A computed holographic fringe pattern can be displayed as a grayscale 2D image on a high-resolution screen. When illuminated with coherent light source, like a laser, it produces a 3D image from the encoded content. Physically, the grayscale holographic screen is much like the screen of a parallax barrier but at resolution scales closer to the wavelength of visible light (i.e., pixels on the order of a few μm instead of 100’s of μm). Current displays are approaching 10μm pixels [58]. The following two subsections include a simple introduction to the general geometries of a parallax barrier and a hologram, which will be helpful when their limitations are discussed in Chapter 3.

2.2 Parallax Barrier Operation

A generic 1D parallax barrier configuration contains two planes, a screen \( s_1 \) and a mask \( m_2 \), with no additional optical elements (Fig. 2-3(a)). For simplicity, we will assume light from each pixel in \( s_1 \) only travels through one slit in \( m_2 \), and that it does not diffract. For a parallax display with \( N_p \) pixels in \( s_1 \), it is clear that there is a direct tradeoff between the amount of spatial and angular content that can be directed to an optimal viewing plane \( v_p \). Specifically, angular resolution \( \theta_p \) can be given by the number of pixels under each slit, and spatial resolution \( x_p \) can be given as the total number of slits (Fig. 2-4(a)). The total number of pixels in \( s_1 \) (\( N_p \)) is
Parallax Barrier: 2 screens, no diffraction

(a) The geometry of a parallax barrier display. A plane of pixels of height $h_p$ ($s_1$) and a 2nd plane of slits separated by $w$ a distance $d$ away ($m_2$) directs light into specific directions towards a viewer within a field-of-view proportional to $\phi_p$. (b) Coherent illumination of a hologram of height $h_H$ and pixel size $t_H$ creates a virtual image through the process of diffraction and interference. Diffraction spreads rays from a finite area of the hologram much like the geometric operation of a single parallax barrier slit, presenting a virtual image to a viewer a distance $z_H$ away.

thus a combination of this spatial and angular content:

$$x_p \theta_p = N_p.$$  \hspace{1cm} (2.1)

At the optimal viewing position, one ray from each slit will enter one eye, and a discrete number of $\theta_p$ views are visible from different positions along $v_p$.

Discretization is one of the main drawbacks of a parallax barrier, leading to issues like aliasing and pseudoscopic views. High angular resolution is desirable to create a more seamless viewing experience, but parallax barrier displays scale poorly with an increase in resolution for a fixed size $h_p$. As angular resolution increases, $m_2$ decreases in light efficiency for a fixed pixel size, since the optimal slit width in $m_2$ ($r_p$) is equal
to the width of one display pixel in $s_1$ [32]. Lenticular arrays can be used instead of slits in $m_2$ to improve optical efficiency, but they will not completely overcome the second issue of diffraction. A slit of width $r_p$ will diffract a ray across an angle given by,

$$\sin \alpha_p = \frac{\lambda}{2r_p},$$  \hspace{1cm} (2.2)$$

where $\lambda$ is the light’s wavelength. As ray-based systems scale towards smaller pixel and barrier widths, physical optics effects cannot be ignored. For example, the previously mentioned color LCD screens with 11$\mu$m pixels [58] will diffract visible light roughly across a full angle of 4 degrees.

### 2.3 Simplified Hologram Operation

Diffraction is exactly how a hologram achieves image creation. A “conventional” thin, amplitude-only transmission hologram creates a single ray (i.e., a beam of finite width) by replacing the parallax barrier’s single pixel and slit with a small sinusoidal grating and a barrier that blocks two of the three diffraction orders (Fig. 2-1). The finite width of the ray a sinusoidal grating creates through diffraction is given by,

$$\Delta \alpha_H = \frac{\lambda z_H}{t_H},$$ \hspace{1cm} (2.3)$$

where $z_H$ is the image distance and $t_H$ is the grating period. Comparing Eq. 2.3 to Eq. 2.2, we see the hologram’s image sharpness improves with a smaller display pixel ($t_H$), while a parallax barrier’s sharpness decreases. A basic Fourier hologram, which creates one real 2D image from one perspective, is a summation of these sinusoidal gratings that diffract light into different viewing directions. This single 2D image is proportional to the Fourier transform of the screen pattern and is conceptually similar to the 2D image of a parallax barrier display seen from one perspective.

Turning this 2D image into a 3D image requires that we convert the Fourier hologram into a Fresnel hologram. Upon coherent illumination, a Fresnel hologram creates a
Figure 2-4: The discretization of a Fresnel hologram into a series of Fourier holograms presents a spatio-angular tradeoff similar to a parallax barrier's. Taking numerical examples from the successful display scenarios in Fig. 2-5, we can establish the resolution of a strip of a parallax barrier with \(10^3\) pixels (a) will be \(\theta_p = 10\) pixels. For a Fresnel hologram (b) with \(10^5\) pixels, the associated \(\theta_H\) will be 1000 pixels along 1 dimension.

A virtual image in the hologram's "near-zone" discussed in detail in [57]. Unlike a parallax barrier, this image offers continuous angular content and full depth cues, but suffers from speckle noise and the lack of multiple colors common to most forms of display utilizing a single coherent light source.

### 2.3.1 Hologram Discretization

Although not exact, a convenient way to construct a Fresnel hologram is simply by tiling together many Fourier holograms. Hologram discretization is used by more advanced holographic forms like the rainbow hologram [22], holographic stereograms [24], and Lucente's "hogel" based designs [25, 60], which all spatially multiplex the holographic screen in different fashions. For simplicity, the terms "spatial multiplexing" and "discretization" will now be used interchangeably but will receive a proper distinction in Section 4.5. Dividing a Fresnel hologram into discrete "patches" is similar to the division of the parallax barrier into spatial and angular resolution components (Fig. 2-4(b)). Each Fourier hologram patch is the scaled Fourier transform of the 3D image from one unique perspective (i.e., from one viewing angle). Geometrically, rays can be traced to and from each independent Fourier patch at angles dependent upon their composition of sinusoids, much like rays from a parallax
A numerical comparison shows that parallax barriers perform well with larger pixels (100 \( \mu m \)), while a holograms perform well with smaller pixels (1 \( \mu m \)).

The number of required pixels to fully reconstruct the entire parallax content of an image of height \( l_{im} \) using a Fresnel hologram of height \( h_H \) is \( N_H = h_H(h_H + l_{im})/\lambda z_H \) \[57\]. Additionally, amplitude-only Fourier holograms require approximately 4 times the desired angular resolution of the 3D object, due to the creation of more than 1 ray by a sinusoid, as explained in Fig. 2-1. Using these two approximations, the Fresnel hologram’s total resolution can be expressed as the product of the number of Fourier patches \( x_H \) and a desired angular resolution \( \theta_H \) as,

\[
4x_H\theta_H \approx N_H. \tag{2.4}
\]

Comparing Eq. 2.1 with Eq. 2.4, it is clear that a parallax barrier’s spatio-angular tradeoff is almost identical to the tradeoff of a hologram’s virtual image under a discretized approximation, up to a constant multiplier.
2.4 A Numerical Comparison

Fig. 2-5 demonstrates the operation of each display using either $10^3$ or $10^5$ pixels fit onto a 100mm screen in 1D. The parallax barrier in Fig. 2-5(a) is a successful display setup with $10^2$ pixels that are 0.1 mm wide each, consistent with current slit widths [62]. In this example, the screen is split up such that $x_p = 100$ and $\theta_p = 10$. As pixel sizes shrink, diffraction effects lead the parallax barrier to spread rays across an angle $\alpha_p = 30^\circ$, washing out image detail (Fig. 2-5(b)). A hologram utilizes the diffraction from $10^5$, 1$\mu$m-wide pixels to deliver an image across a 30$^\circ$ viewing angle.

From the definition of $N_H$, this setup can fully reconstruct all parallax information of a 5cm object 30cm away. Discretizing the hologram into 100 Fourier patches will match the spatial resolution of the successful parallax display. Each patch will be a 1000-pixel Fourier transform of the desired image from a slightly different angular perspective (Fig. 2-4), offering 250 unique perspectives of the 3D object (Eq. 2.4).

From this brief and simplified comparison between parallax barriers and holograms, two conclusions should be clear:

1. As pixel sizes decrease for a fixed display size, virtual image sharpness and viewing angle conditions improve for a holographic display, while sharpness and light efficiency worsen for a conventional parallax barrier display.

2. Discretizing a hologram into spatial and angular content presents a resolution tradeoff, directly analogous to the space–angle tradeoff in parallax barrier displays.

3. Both forms of display offer a viewing angle and virtual image depth that must obey various geometric constraints, which are an important consideration in design but will not be the focus of the rest of this thesis.

In general, both incoherent parallax barriers and coherent holograms share a number of remarkable similarities, given they achieve 3D display using two completely different physical phenomena (i.e., attenuation vs. diffraction). To further develop their close connection, the notion of optical phase space must first be introduced, which will
lead to the presentation of a shared rank-1 algebraic limitation for each display form. After this shared limitation is demonstrated in phase space, this thesis will turn to focus solely on designs for holographic display, as shrinking pixel trends indicate that diffractive-based 3D display may be the optimal choice in the not too distant future.
Chapter 3

Phase Space Functions

Now that the basic concept of the two most prominent methods of 3D display have been introduced, this thesis now turns to develop a simple way of analyzing their ability to create depth-varying images. In general, a class of functions known as “phase space functions” provide a convenient method of analyzing the propagation of light through an optical system, whether the light is coherent, incoherent, or somewhere in between. In this chapter, we will use these functions to demonstrate how the space of light distributions that both holograms and parallax barriers can create are algebraically limited to rank-1 functions in a certain space. This demonstration is intended to motivate the need for new methods of displaying future 3D autostereoscopic content. In Chapter 4, one new method of holographic display will be suggested. Furthermore, phase space functions will be used to determine the content of this display. The following introduction to light fields and the Wigner distribution will thus help form a mathematical basis for understanding the remainder of the thesis.

The concept of a “phase space,” or a space in which all possible states of a system can be represented, has found application in a wide area of engineering and physics research disciplines. System states can be viewed in a phase space diagram, which is a 2D plot of two related variables describing a 1D event. For example, mechanical motion of a particle is often represented with a plot of all possible position and momentum values, quantum mechanical interactions use energy and time, and elec-
As discussed below, phase space functions applied to optical systems often take two related forms: incoherent light is represented as a function of all geometric ray positions and angles, while coherent light as a function of wave position and local spatial frequency component. Following is an introduction to each of these optical phase space representations, a simple example to demonstrate their similarity, and a discussion of their mutual limitations. To keep things basic, we will mostly limit our discussion to 1D distributions of light propagating in flat space, thus leading to 2D phase space functions that are easy to visualize. Extensions to 2D distributions of light and their 4D phase space representations are straightforward. For the interested reader, a good introduction to different phase space functions for temporal signals is in [63], while a comprehensive discussion of their application to light is in [64].

### 3.1 Incoherent Light: The Light Field

The geometric light field is one parameterization of all possible rays propagating through a volume. For simplicity, we will first restrict our attention to rays leaving a 1D surface along $x$, traveling in the $z$ direction, as shown in Fig. 3-1. Similar to
the well known ray-transfer matrix methods [65], a ray is represented as a position \( x \) and an angle \( \theta \) with a function \( L(x, \theta) \). If a set of rays at position \( z_0 \) originate from a point, then the light field at \( z_0 \), \( L_0(x, \theta) \), can be represented by a vertical line - all possible angles of rays exist at (i.e., are emitted from) a single point in space. At a certain distance away along the propagation axis, \( z_1 \), the light field is represented by a diagonal line. Here all rays have propagated to form a “triangular” distribution or a cone of light.

The transformation of a vertical line representing the light field of a point to a diagonal line after propagation is given by a geometric shearing operation along the \( x \)-axis, which is a well known result of using first-order ray transformation matrices. Other similar transformations include a shearing operation along the \( \theta \)-axis for the passage of light through a thin lens, or a rotation of 90 degrees for propagation across a large distance. More information on these geometric transformations can be found in [57]. These convenient transformations are also obeyed by the Wigner distribution, as explained next.

### 3.2 Coherent Light: The Wigner Distribution and Ambiguity Function

To help us compare the performance of holographic and parallax-based 3D displays, this section develops a phase space model for the transport of coherent light that is as similar as possible to the above geometric light field. A function called the Wigner distribution will serve as a method of connecting the ray and wave-based interpretations of light. The Wigner distribution as considered in this thesis relates the space \( (x) \) and spatial frequency \( (u) \) content of a given function that defines an optical wavefront. For simplicity, we will begin by considering a quasi-monochromatic, completely coherent optical wavefront at a 1D plane. Quasi-monochromatic light follows from the definition in [66] as \( \Delta \lambda / \lambda \geq N_p \), where \( \Delta \lambda \) is the spectrum bandwidth and \( N_p \) is
the number of pixels in the hologram along 1D. As in most display applications, we will assume that the optical signal we are interested does not change quickly with time (i.e., we average out the time variable). The Wigner distribution of a 1D complex optical function, $t(x)$, can be defined as

$$W(x, u) = \int J(x, x') e^{-ix'u} dx', \quad (3.1)$$

where the function in the integrand,

$$J(x, x') = t \left( x + \frac{x'}{2} \right) t^* \left( x - \frac{x'}{2} \right) \quad (3.2)$$

is often called the mutual intensity (MI) function. Here, since we've assumed a completely coherent optical wave, as noted above, our mutual intensity can be represented as a multiplication of two functions $t$ and $t^*$. The case of partial coherence will be discussed in the next chapter. The $^*$ operation represents complex conjugation. Note that after the Fourier transform of the mutual intensity function, the WDF contains only real values, positive as well as negative.

With our assumptions explicitly stated, let's now take a close look at Eq. 3.1 and Eq. 3.2. The WDF of a 1D function is 2D, and as we will see is directly related to the geometrical light field. First, let's consider the spatial dimension $x$, turning to the spatial frequency variable $u$ shortly. Two simple interpretations for $t(x)$ exist: it can be considered a function that describes an optical wave at some plane in space, or it can be considered a function that describes a surface or aperture that a plane wave of light interacts with [13]. The latter interpretation is of more interest from a modeling and design standpoint. Under this assumption, $t(x)$ can describe a surface like the grating-like structure of a CD, the fine mesh of a fabric, or the screen of a holographic display, which is what this thesis will apply it towards. Sharing the same spatial coordinate $x$, it is clear that the Wigner "light field" $W(x, u)$ given by Eq. 3.1 will describe rays with coordinates that start at $x$, immediately after reflection from or transmission through a thin surface. This Wigner light field will be consistent
with physical optics theory up to most approximations of interest. For the interested reader, the Wigner distribution offers an accurate model of optical wave propagation in the paraxial region, away from the near field where evanescent components may exist. For purposes of analyzing and designing holographic displays, these conditions are satisfactory. A more detailed discussion of the function’s validity in different propagation regions is in [68].

For the remainder of this thesis, t(x) will be used to describe our holographic screen of \( N_p \) pixels introduced in Chapter 2. t(x) is a discrete function of pixels at position x. The content of t(x) will describe the surface’s ability to absorb or transmit light. Specifically, the absolute value of t(x) will describe absorption, while the screen’s ability to impart a phase delay to light is given by its complex angle:

\[
A_s = |t(x)| \tag{3.3}
\]

\[
\phi_s = \arctan \frac{Im[t(x)]}{Re[t(x)]}. \tag{3.4}
\]

Here, \((A_s, \phi_s)\) is the amplitude transmittance and phase delay, respectively, of the screen, and \(Re\) and \(Im\) represent the real and complex projection operators. The complex portion of t(x) indicates a phase delay due to either a change in the refraction index or the thickness of the surface’s material at position x. For example, an amplitude grating (i.e., a series of black and transparent stripes) can be represented by a real-valued t(x) that periodically varies between 0 and 1, while a phase grating (i.e., a series of raised glass ridges) can be represented by a t(x) with |t(x)| = 1 for all x and a complex angle \(\phi_s\) that varies between \(-\pi\) and \(\pi\).

### 3.2.1 One Wavefront as Many Plane Waves

Besides containing the same spatial variable as the screen function t(x), the WDF also depends upon the spatial frequency variable \(u\), which can be understood by briefly considering the wave-like nature of light. If a wavefront of light has a single
Figure 3-2: (left) A wavefront can be split up into many plane waves as part of a Fourier decomposition (See Section 3.10 of [57]). This is similar to Huygens’s principle, but uses plane waves instead of spherical waves as a basis. (right) Each plane wave in the decomposition can be related to a ray traveling at a certain angle, shown as an arrow. Spatial frequency is given as the ratio of the sine of this angle and the wavelength of light. Its units are meters$^{-1}$.

wavelength, as we are assuming, then it is considered monochromatic, temporally coherent, or, put simply, of a single color. A great property of a coherent wavefront of monochromatic light is that it can be represented by a sum of plane waves, each traveling at a slightly different angle (Figure 3-2). Decomposing a wave of light into a sum of plane waves at different angles is very similar to decomposing an arbitrary signal into a sum of sine waves at different frequencies with a Fourier transform. However, since we are dealing with a wave over space, each plane wave traveling at a different angle provides a unique spatial frequency to the wavefront. This basic representation of a wavefront is known as its angular spectrum [57] and can be visualized in Figure 3-2. Again, as our Fourier decomposition of a coherent wave happens across space, the definition of the spatial frequency term $u$ is in units of m$^{-1}$.

Besides its elegance in Fourier optics, this plane wave decomposition also offers a simple tie to the ray-picture of geometric optics. From Figure 3-2, it is clear that each plane wave can be described by a single bisecting ray at a certain angle $\theta$. For
example, if the wavefront is propagating directly to the right (with $\theta = 0$), the ray is also at $\theta = 0$, and the spatial frequency $u$ of the wavefront is 0. As the angle with respect to horizontal grows, $u$ grows. The simple formula connecting $\theta$ and $u$ is,

$$ u = \frac{\sin(\theta)}{\lambda} \approx \frac{\theta}{\lambda}, \quad (3.5) $$

where the approximation is valid in the paraxial zone. This relationship allows for the transfer of wave phenomena to a ray treatment. The augmented light field is one framework that builds upon this connection. It creates and renders distributions of rays, augmented with negative values, that exhibit diffractive and refractive effects [13, 17]. In general, the Fourier decomposition of any wavefront into spatial frequency components (i.e., a sum of plane waves) is indirectly a decomposition of any wavefront into a group of rays at different angles. Putting it all together, the WDF function $W(x, u)$ describes a bundle of rays, which start at a surface $t(x)$, and leave at an angle, $\theta_x = \sin^{-1}(\lambda u)$. Each ray is given a weight from the computation of Eq. 3.1. This process effectively provides a Fourier decomposition of $t(x)$ into a bundle of rays at each location along $x$ that arise from diffraction.

### 3.2.2 A Simple Light Propagation Model

As with rays, the WDF also follows many simple linear transformations that can be represented in the well-known ray-transfer matrix formalism [14, 64]. Following is a very simple WDF light propagation model, built using 3 different transformations: propagation through free-space, propagation through a grating, and projection. This model is closely related to rendering schemes used with the augmented light field. It traces light from an initial source, through a diffracting element, and to a screen or image sensor where all rays are projected into an intensity measurement. An example of a camera imaging a point source (while modeling physical-optic effects) using these three transformations is in Fig. 3-3. This thesis will use these transformations to describe the evolution of light from a holographic display to a viewer.
Free Space Propagation: The WDF $W_z(x, u)$ of a complex wavefront will shear due to traveling a distance $z$, similar to the rays of a light field, with,

$$W_z(x, u) = W(x - \lambda z u, u)$$

(3.6)

Propagation Through a Thin Grating: In Section 3.2.1, we saw that the propagation of a plane wave through a grating $t(x)$ is given by its WDF, $W_t(x, u)$, from Eq. 3.1. If something besides a plane wave hits $t(x)$, we can still find the resulting output WDF, $W_o$. It is defined by a convolution along spatial frequency variable $u$ of the incoming WDF, $W_i$, and the grating WDF, $W_t$, with

$$W_o(x, u) = \int W_i(x, a - u)W_t(x, a)da$$

(3.7)

Projection onto a Surface: The intensity of light described by the WDF is found
with its projection along the spatial frequency axis $u$:

$$I(x) = \int W(x, u) \, du$$  \hspace{1cm} (3.8)$$

Even though the Wigner distribution $W(x, u)$ contains negative values, the observed intensity $I(x)$ on a surface is always non-negative [69]. This is demonstrated in Fig. 3-3. The outgoing WDF however does contain negative values, which are marked in blue, positive values are marked in red. The WDF may also be simply extended to describe polychromatic light. A polychromatic source can be separated into a discrete set of weighted wavelength components $\lambda_i$. The WDF of this polychromatic source $W_p(x, u)$ is simply the weighted sum of the WDF of each wavelength component, $W(x, \theta/\lambda)$. Alternatively, different integrals along a single $W(x, u)$ may also yield a polychromatic intensity response [67].

### 3.2.3 The Ambiguity Function

Before the performance of parallax barriers and holograms are specifically examined, one more function that is used extensively in the following two chapters is briefly explained. Simply put, the ambiguity function (AF) is the 2D Fourier transform of the Wigner distribution. However, to only describe the AF as such is to leave out
much insight into its utility. Most notably, the AF has proven extremely useful while modeling a camera’s response to defocus [16, 15]. We will extend these prior camera models to compactly describe the formation of light from a hologram along its axis of propagation (i.e., light in 3D). To begin, we will consider the AF a hologram-lens setup as in Fig. 3-4, where the hologram of interest is placed directly against a lens of focal length \( f \). The AF is a phase space function of both space \((x')\) and spatial frequency \((u)\) and also has close ties to the well known ray space of geometric optics [64]. However, its relationship to space-spatial frequency is quite different than the Wigner distribution’s, as is clear from our change of variables from \( x \) to \( x' \). From Fig. 3-4(a), we represent the 1D plane wave incident upon the holographic mask, \( U(x) \), with the 2D ambiguity function,

\[
A(u, x') = \int U \left( x + \frac{x'}{2} \right) U^* \left( x - \frac{x'}{2} \right) e^{2\pi i x u} dx
\]  

(3.9)

where \( x \) and \( u \) are the same space and spatial frequency coordinates as with the WDF, and \( x' \) is a second spatial parameter proportional to distance along \( z \). Note that since we are considering an incident plane wave, \( U(x) \) at the hologram plane is equivalent to the function that defines the amplitude and phase of the diffracting screen at \( z_0 \). This function is given as \( t(x) \) in the previous section when discussing the WDF but is here represented as \( U(x) \) to keep WDF and AF analyses distinct. The wavefront’s mutual intensity function \( J(x_1, x_2) \) is obtained from the AF through an inverse Fourier transform and coordinate transformation to center-difference coordinates \( x_1 \) and \( x_2 \):

\[
\int A(u, x') e^{-2\pi i x u} du = U \left( x + \frac{x'}{2} \right) U^* \left( x - \frac{x'}{2} \right) = U(x_1)U^*(x_2) = J
\]  

(3.10)

In practice, this transformation is performed on a discrete AF function as an inverse Fourier transform along \( u \), a rotation of \(45^\circ\) and a coordinate re-scaling by one half along \( x_1 \). The mutual intensity function \( J \) will be used as a constraint in the iteration process presented in the next chapter. Setting the \( x_2 \) coordinate in Eq. 3.10 to zero
yields,

$$U(x_1)U^*(0) = \int A(u, x_1)e^{-\pi iux_1}du$$  \hspace{1cm} (3.11)

which shows the wavefront $U(x_1)$, and hence the pattern of a diffracting mask illuminated by a plane wave, can be recovered from the AF up to a constant phase factor. The AF of the hologram mask function $U(x)$ can also represent all OTFs at any distance $z$ of our setup [16]. Specifically, a “defocused” OTF $H$ at any plane $z_n$ along the direction of propagation in Fig. 3-4 is given by,

$$H(x', W_{20}) = \int U\left(x + \frac{x'}{2}\right)e^{ikW_{20}\left(x+\frac{x'}{2}\right)^2}U^*\left(x - \frac{x'}{2}\right)e^{-ikW_{20}\left(x-\frac{x'}{2}\right)^2}dx$$  \hspace{1cm} (3.12)

where $k$ is the wavenumber and $W_{20}$ is a “defocus” coefficient [70] here defined as,

$$W_{20} = \frac{r^2}{2} \frac{\pm \Delta z}{f^2 \pm f \Delta z}.$$  \hspace{1cm} (3.13)

This equation assumes illumination of the holographic mask by a plane wave, where $\Delta z$ denotes defocus distance, $r$ is the radius of the lens, and $f$ is its focal length. The complicated OTF function in Eq. 3.12 can be simply represented as a slice through the middle of the AF function:

$$H(x', W_{20}) = A(x'W_{20}k/\pi, x')$$  \hspace{1cm} (3.14)

In other words, the optical response $H$ at any depth plane after light hits a holographic screen directly before a lens is given as a slice through the center of the 2D AF of the complex hologram’s function $U(x)$. The angle of this slice $\theta$ is proportional to how far away the plane of interest is along $z$ following the equation,

$$\tan(\theta) = W_{20}k/\pi$$  \hspace{1cm} (3.15)

This relationship is shown in Fig. 3-4(b). The utility of this special property was primarily noted while designing apodizers for extended depth-of-field purposes, where one wishes to establish a depth-invariant OTF [71].
In the case of light propagating through a holographic screen that does not have a nearby lens, the equation indicating the light’s OTF as a slice of the AF takes the form,

\[ H(x', z_p) = A(u, x' - \lambda uz_p), \]  

which is also valid under the assumption of paraxial light propagation. Here, note that slices are similarly tilted with increased propagation distance, but the horizontal slice exists towards infinity instead of at the lens’s focal plane, as explained in detail in [72]. Likewise, the slices are at “sheared” angles instead of “rotated” angles, following a different nonlinear relationship with \( z \).

The rest of this thesis will mostly analyze holograms assuming the presence of a thin lens nearby, since diffraction angles from currently available SLMs are still quite small and a more compact and simple experimental setup can be created with a lens. Thus, most of the models will follow from Eqs. 3.9 - 3.15, since they are designed to match the experimental setups used in Chapter 6. However, Eq. 3.14 and Eq. 3.16 show the two cases of lens and no-lens are quite similar, as both demonstrate how an AF represents all planes of propagation into the “depth” dimension \( z \) as different slices. Thus, the only modification needed to transfer between the algorithm discussed in the next sections (that assumes a lens) and the case of a hologram without a lens is the angle at which the OTF slices of interest are filled in.

Additionally, the next chapter of this thesis will not use the AF to simply view the performance of a given apodizer or holographic screen, as much prior work has done. Instead, it will present a method of designing a hologram from a desired set of intensity inputs. In other words, it will address the problem of 3D image design using phase space functions like the AF to solve an inverse problem. Specifically, we will attempt to establish the AF that best matches a desired set of OTFs at different planes along the direction of propagation. Once this optimized AF is known, an
optimal holographic pattern $U(x)$ can then easily be determined, up to a constant phase factor, from Eq. 3.11. Before addressing the design problem, however, this thesis will first demonstrate why new tools for 3D display design are even required. Following, an algebraic analysis presents the inherent limitations of both conventional parallax barrier-based and coherent hologram-based methods of display.

3.3 Light Fields and WDFs: A Display Example

To bring our attention back to displays, we now can use the light field and WDF to model the process of forming a 3D image. A ray-based parallax barrier can be described with a geometric light field since it uses incoherent light, while a hologram illuminated with coherent light is modeled with either a WDF or an AF. While the two display forms are quite different (i.e., a parallax barrier uses two layered screens to modulate light, while a single-screen hologram diffracts light), each phase space function leads to similar models and transformations.

As a simple example, we consider how each display might form the appearance of 2 points at two different depths. Continuing with our assumption of 1D displays, the 2D light field $L(x, \theta)$ and WDF $W(x, u)$ created by each display surface both shear a large distance, represented by a $90^\circ$ rotation, to a viewer’s eyes (Figure 3-5). The screen pattern required for the example parallax barrier and hologram screen are in Fig. 3-5(c)-(d) (here shown as a 2D separable function for ease in interpretation), and both exhibit a similar radial pattern. Both displays in Fig. 3-5 are generated with values from the Section 2.4’s numerical example assuming distances $F_1=10\text{mm}$, $F_2=20\text{mm}$ to each point. From this simple example, it should be observed that the highly discretized parallax barrier offers a lower fidelity 3D reconstruction, while the hologram creates an unavoidable additional order as it attempts to create each point using an amplitude-only screen (note the double “X” pattern in the WDF representation).
Figure 3-5: Incoherent and coherent phase space representations of a (a) parallax barrier and (b) holographic display creating an image of two points at different depths. The parallax barrier uses 4 pixels to generate an initial light field with constant-$\theta$, which shears to the mask $m_2$ and is attenuated. A viewer sees the rotated version of this attenuated light field. Greater angular separation ($\Delta r$ vs. $\Delta b$) maps to greater parallax disparity when viewing, indicating proximity. A hologram will use two Fresnel zone plates to create two points in space. The screen’s WDF will shear and rotate to a viewer the same as a light field to offer the same disparity. The 2D parallax barrier pattern (c) and holographic pattern (d) for $s_1$ both grow radially.
3.4 Limitations of Incoherent Parallax Barriers and Coherent Holograms

So far, a few examples have shown how the discrete spatial and angular content of a parallax barrier and hologram can each be easily visualized in phase space. Furthermore, the above 2-point model demonstrates how a geometric light field can represent the 3D image from a parallax barrier, while the Wigner distribution can represent the 3D image from a hologram (note again our examples are 1D screens displaying in 2D, but extension to 3D is direct). Limitations exist for each display setup such as discretization, and creation of multiple orders. However, it may not be clear exactly how these limitations manifest themselves in the mathematical formulation of the light field and coherent WDF. In the next two sections, the connection of the parallax barrier and hologram’s phase space functions to an algebraic rank-1 limited representation is presented.

3.4.1 Rank-1 Geometric Light Field

We will first demonstrate the light field produced by a parallax barrier is rank-1. Note that a joint position \( (x) \) and angle \( (\theta) \) phase space \( L(x, \theta) \) of all rays in 1D is equivalent to a 2D light field parameterization, \( L(s_1, m_2) \), of a ray passing through two parallel planes \( s_1 \) and \( m_2 \). This parameterization is achieved through a simple trigonometric relationship [11]. As mentioned earlier in this chapter, propagation of \( L(x, \theta) \) is represented as light field shear along \( x \), ray values are always positive, and we will assume rays cannot bend (i.e. diffract) upon interference with a mask. Following along with the simplified parallax barrier example in Fig. 3-5(a), an initial screen of pixels creates a light field \( L_1(x, \theta) \) that is constant with broad extent along \( \theta \). \( L_1(x, \theta) \) then shears with free space propagation to become \( L_2(x, \theta) \) directly before the plane of amplitude modulating slits. Here, the slit mask \( m_2(x) \) will either block or allow rays through, defining the light field on the other side of the slits through multiplication: \( L_3(x, \theta) = L_2(x, \theta)m_2(x) \). The light field \( L_3 \) then shears across a large
Figure 3-6: The light field from a parallax barrier is rank-1. (a) A parallax barrier contains a screen of pixels $s_1(x)$ and a screen of slits $m_2(x)$. $s_1(x)$ contains angular $(u_i)$ and spatial $(y_j)$ components under $y_j$ slits. (b) Each ray emitted from the parallax barrier can be plotted as a function of its coordinate in $s_1(x)$ and its coordinate in $m_2(x)$. For example, a ray emitted horizontally from any pixel along $s_1(x)$ will hit the mask $m_2(x)$ at $s_1(x) = m_2(x)$, shown as the diagonal line $\theta(0)$. A parallax barrier only creates a rank-1 approximation of any desired light field, since its light field is expressed as an outer-product of $s_1(x)$ and $m_2(x)$.

distance, represented by a $90^\circ$ rotation of the light field plot, to a viewer’s eye.

The parallax barrier’s light field can be re-parameterized into an outer-product format by noting $L_1$ and $L_2$ are related through the shear relationship,

$$L_2(x, \theta) = L_1(x - d\theta, \theta),$$

where $d$ is the separation between $s_1$ and $m_2$. This yields an expression for the output light field $L_3(x, \theta)$ as a product of two functions,

$$L_3(x, \theta) = L_1(x - d\theta, \theta)m_2(x) = s_1(x - d\theta)m_2(x).$$

Here, we have replaced $L_1(x, \theta)$ with $s_1(x)$, since the initial light field generated by the screen $s_1(x)$ offers no initial control over the $\theta$ dimension. Plotting all rays in $L_3$ in terms of their initial screen coordinate $s_1(x)$ and mask coordinate $m_2(x)$ clarifies
this decomposition (Fig. 3-6(b)). The generated light field $L_3(x, \theta)$ lies at a $45^\circ$ angle with discrete lines of angular content $\theta$ representing rays at different constant angles. With control over only two amplitude-modulating planes in a parallax barrier, we see that the best rotated light field the display can generate will be the product of two real discrete vectors, the screen and the mask of slits:

$$m_1(x)s_2(x)^T = [L_{45^\circ}]$$

(3.19)

In other words, parallax barriers are restricted to display rank-1 light fields. Any light field one wishes to display that is not rank-1 will be under-sampled or presented as an aliased image. A specific consequence of parallax barrier displays is low light efficiency: to map a pixel to a desired direction, all other rays are blocked. This significant light attenuation may not be optimal, and recent attempts have been made to improve it [80] using the above insights.

### 3.4.2 Rank-1 Holographic Light Field

As displays reach resolutions within one order of magnitude of light’s wavelength, the incoherent light field must be transformed into a framework that obeys physical optics. In other words, the assumption that rays cannot bend (i.e., diffract) at a thin screen, like a parallax barrier slit, is not valid at wavelength-scales. As explained above, the Wigner distribution is a direct analogy of the geometric-based light field that includes its diffraction effects for coherent light [64]. In the limit of a very small wavelength, or very large pixels, the WDF approaches a radiance functions, or rays [84]. Looking the other direction towards smaller and smaller pixels (i.e., pixels approaching the order of light’s wavelength), the WDF offers a convenient and direct functional representation of light for those used to working with rays. Rays simply need to be replaced with localized plane waves to describe diffraction, meaning $\theta$ is replaced with the spatial frequency term $u$ following Eq. 3.5. As noted earlier in this chapter, the coherent light field after passing through a holographic screen with transmission function $t(x)$ is the Wigner distribution of $t$: 
As with the parallax barrier example, we can also transform a hologram’s coherent light field into a space where its limited display capabilities becomes clear (Fig. 3-7(a)). This limitation is implicit in the definition of $W(x, u)$ in Eq. 3.20, which only relies on the 1D complex screen function $t(x)$. We can recover the 1D function $t(x)$ up to a constant phase factor [64] with 3 transformation operations on $W(x, u)$. This is similar to the inversion of the AF in Eq. 3.10. First, a 1D inverse-Fourier transform on $W(x, u)$ is performed along the $u$-axis to yield the expression,

$$W(x, u) = \int t \left( x + \frac{x'}{2} \right) t^* \left( x - \frac{x'}{2} \right) e^{-2\pi ix'u} du. \quad (3.20)$$

where $\mathcal{F}_u^{-1}$ represents a discrete inverse Fourier transform operation on the discretized hologram $t$. The next two operations rotate the expression in Eq. 3.21 by 45°, then
re-scale the $x_1$-axis by two. This is equivalent to shifting from the center-difference coordinates $(x, x')$ to the two independent coordinates along the mask $(x_1, x_2)$:

$$R_{45} \left[ D \left[ t \left( x + \frac{x'}{2} \right) t^* \left( x - \frac{x'}{2} \right) \right] \right] \Rightarrow t(x_1) t^*(x_2)$$

(3.22)

The function $t(x_1) t^*(x_2)$ is the mutual intensity function, $J(x_1, x_2)$ in Eq. 3.2. It describes the statistical correlation between any two points on a wave, or here, a holographic screen. Eq. 3.22 is also a rank-1 representation assuming coherent light [43]. Previously cited limitations of using coherent light to design a 3D field include speckle and out-of-focus noise [1]. From the above analysis, we can tie in these effects to a constrained available space of functions (only rank-1 functions) that a hologram can assume in the mutual intensity domain. The limitations of designing a 3D intensity pattern with coherent light can alternatively be understood by realizing that a coherent field with fully defined amplitude and phase at any plane along the direction of propagation will be defined at all subsequent planes. To define the intensity distribution at a single 2D plane, the wave’s amplitude must be defined and fixed at this 2D plane. The only degree of freedom remaining to design the intensity distribution at all subsequent planes along the direction of propagation is the field’s phase at a single plane.

An example of the three-operation process of Eq. 3.21–3.22 applied to a coherent Wigner distribution of two slits is in Figure 3-7(a). The output mutual intensity for this coherent case is clearly rank-1. An example of using partially coherent light is in Figure 3-7(b) (i.e., light passing through one slit is uncorrelated with the light passing through the other slit). The same three mathematical steps in this case lead to a rank-2 mutual intensity function. The relationship between partial coherence, the mutual intensity function’s rank and 3D intensity design flexibility will be considered in further detail in the next three chapters.

Given a mathematical basis for how conventional parallax barrier and holographic
displays are limited, the remainder of this thesis will discuss how to extend holographic 3D displays beyond these limitations. It should be noted that current work is focused on extending parallax barrier displays beyond their rank-1 light field limitation, either by using multiple stacked displays, or by showing many images within the flicker-fusion rate of a viewer’s eye [80, 81]. These display prototypes use similar nonlinear optimization methods as presented in Chapters 4 and 5 to determine optimal screen patterns for a desired 3D image. However, their analysis is based upon an assumption of incoherent light (i.e., attenuation of rays). The following design method and display setup proposed in the next chapters might be considered a holographic counterpart to the work in [80, 81].
Chapter 4

Multiplexing for Multimode 3D Display Design

A variety of methods have previously been suggested to control light through diffraction in 3D [1, 4, 6, 8]. Here, a new computational method is developed based on the phase space functions of the previous chapter. It is referred to as a "mode-selection" algorithm. The first variant of this algorithm is used to model completely coherent light and is termed "single-mode selection." It is simply extended to model partially coherent light (termed "multi-mode selection"), which will lead to an improved 3D display design that relies on time-multiplexing multiple hologram patterns on a screen. Before entering into the specifics of the design algorithm and the display method, a brief clarification of what is meant by multiple modes, partial coherence and multiplexing is first offered.

4.1 Modes, Partial Coherence and Multiplexing

Throughout this chapter, the three terms "modes," "partial coherence" and "multiplexing" are used repeatedly and often in overlapping contexts. While explained in more detail in later sections, it is useful to first define what is meant when each is referred to. "Modes" are connected with representing a wave of light. One mode can be thought of as a single coherent wave field, created for example by one point.
source very far away, or one laser. It was assumed that the coherent hologram in Chapter 2 is illuminated with a single mode (i.e., a completely coherent field from a laser). The Wigner distribution of the previous chapter is also defined for one coherent mode. "Multiple modes" thus refers to multiple, overlapping coherent fields. This could be light generated from multiple, mutually incoherent point sources at a distance, or multiple lasers nearby. Light of this type is represented as a sum of multiple, completely coherent fields, or a sum of multiple Wigner distributions. This type of light is also referred to as partially coherent. Therefore, in some sense, the terms "multiple modes" and "partially coherent" can be used interchangeably. In the limit of adding many modes, light becomes incoherent and can be represented using rays and the geometric light field from Chapter 3. Thus, partial coherence exists on a spectrum light, representing light that is neither perfectly coherent (from one mode) or perfectly incoherent (from an infinite number of modes), but somewhere in between.

Generally defined, "multiplexing" refers to the transformation of a higher dimensional signal into a lower-dimensional representation. In this thesis, the term will explain how multiple modes are generated by a 3D display and sent to a viewer's eyes. For example, the conventional coherent hologram in Chapter 2 is not multiplexed, since it diffracts light from a single coherent mode. If instead two separate lasers illuminate two holograms side-by-side, whose diffracted images mix at a viewer's eye, then the two holograms are "spatially multiplexed." In this setup, two coherent modes are generated at two separate locations in space. They then propagate and mix to form a partially coherent 3D image at a different plane along the direction of propagation. Alternatively, the two coherent modes can be multiplexed by placing the two lasers at different angles and creating an image at their intersection. Or, two coherent holographic images can be shown sequentially over time and added together within a finite sensor integration window. These two setups are examples of angle and time-multiplexing, respectively. Regardless of the specific arrangement, multiplexing mixes more than one coherent field. Thus, this thesis uses "multiplexing" to describe the optical combination of multiple modes to create a partially coherent field at a
The above three terms can now be used to outline the goals of this chapter. First, this chapter explains an algorithm (single mode-selection) to design any fully coherent 3D field using the coherent WDF and AF functions. The input of this algorithm is one desired 3D image to display, and the output is one fixed coherent hologram pattern that will approximately generate it. The mode-selection method is then extended to find an optimal set of $M$ coherent fields, or $M$ modes, referred to as multi-mode selection. This algorithm relies on a partially coherent formulation of the WDF and AF functions. The input of the algorithm is again one desired 3D image, and the output is $M$ different hologram patterns, representing the $M$ optimal modes the algorithm determines. The combination of these modes forms a partially coherent 3D field that optimally approximates the desired 3D image. Any multiplexing scheme can be used to combine these $M$ modes (i.e., $M$ holographic patterns). This thesis theoretically and experimentally investigates the use of time-multiplexing, or quickly showing the $M$ hologram patterns on a dynamic display. Section 4.5 demonstrates how the combination of multiple modes using time-multiplexing generates a desired 3D image of better quality than a single mode does. Finally, Section 4.6 discusses alternate ways to multiplex the $M$ hologram patterns of the multi-mode selection algorithm.

### 4.2 Single Mode Selection Algorithm

Much of the work in this section, as well as the next several sections, was completed with help and many insights from Se Baek Oh and Zhengyun Zhang. This section presents a new method of designing fixed holographic patterns that do not change over time. A fixed, coherently-illuminated computer-generated hologram is traditionally printed or etched as a relief pattern at high resolutions and illuminated with a laser. This type of hologram is connected with the concept of a single coherent mode in Section 4.1. From Chapter 3, it is clear that if one can design a physically valid
Figure 4-1: A schematic diagram of the proposed algorithm operating in 1D. (a) A set of \( n \) desired OTFs (here \( n = 3 \) for a rect-function holographic mask), which are determined from desired intensity (PSF) responses, are used as input. (b) Each OTF populates a slice of the AF from Eq. 3.14. (c) A one-time interpolation from Eq. 4.1 is used to fill in zeros between desired slices. (d) The mutual intensity (MI) can be constrained by taking the first singular value shown in (e). Details of this constraint are in Fig. 4-2. (f) An optimized AF is now obtained, which is re-populated with the desired OTF values in (a) along the specific slices in (b). Iteration is stopped at a specified error value, and Eq. 3.11 is then used to invert the AF into the optimal 1D aperture mask.

coherent WDF or AF from a desired 3D intensity pattern, the amplitude and phase of a holographic mask that re-creates this 3D pattern can be established using Eq. 3.11. Not surprisingly, a large field of work has been dedicated to determining a full phase space function of a given wavefront from multiple intensity measurements at different planes along the direction of propagation. Phase-space tomography [6, 7, 72] borrows tools like the Radon transform and filtered back-projection from tomography to reconstruct a 2D phase space function (WDF or an AF) from a set of its 1D slices. Unfortunately, this tomographic approach typically requires many experimental mea-
urestions, the few desired input intensities this thesis intends to use, which may not even be physically realistic. Similarly, methods based on the transport-of-intensity equation [8, 9] generate a phase space function from two or more closely spaced measurements, but do not facilitate the design of arbitrary intensities at widely spaced planes, which is a more flexible method of designing an image to show in 3D.

To map a few desired input values to a fully valid AF, we propose an iterative solution using constraints available in the mutual intensity domain. Again, this iterative algorithm is termed "mode-selective," with the focus of this chapter being the case of coherent or single-mode selection. The algorithm’s iterative steps are outlined in Fig. 4-1. The mode-selection procedure begins with defining a set of n desired OTFs at different depth planes $z_n$, which we would like to be able to display (Fig. 4-1(a)). Note that these OTFs can be directly determined from desired intensity patterns through a well-known Fourier relationship [57]. There are no fundamental restrictions on $n$, $z_n$, or the shape of the desired intensity patterns, although performance variation with each of these parameters is examined in detail later in this chapter. An approximate AF function is populated with these desired OTFs at slices from Eq. 3.14, each filling two slices in Fig. 4-1(b) given a symmetric aperture, which is then used as an input to an iteration procedure. To obtain a more realistic initial AF approximation (Fig. 4-1(c)), a one-time interpolation is performed between input slices based on a Taylor power series expansion with respect to $u$,

$$A(u, x') = A(u = 0, x') + 2W_{20}x' \frac{\delta A}{\delta u} (u = 0, x') + \frac{(2W_{20}x')^2}{2!} \frac{\delta^2 A}{\delta u^2} (u = 0, x') + \ldots \quad (4.1)$$

which is similar to a previously used expansion along $W_{20}$ [71]. This interpolation simply fills in zeros between populated slices to better pose the function for an iteration process and is typically carried out to the second order.

After this linear interpolation, the mutual intensity of the estimated AF, $J'(x_1, x_2)$, is obtained using Eq. 3.10. After application of a constraint, which will be discussed
next, a more accurate mutual intensity function $J_{opt}(x_1, x_2)$ is created. $J_{opt}$ is transformed back into the AF domain through application of Eq. 3.9, where the desired OTF set again populates the AF at slices at an angle $\theta$ from Eq. 3.15. This procedure iterates until a threshold error value, at which point Eq. 3.11 is applied to determine the optimal amplitude and phase distribution to use as an holographic mask, up to a constant phase factor. The iterative replacement of OTF values is quite similar to the iterative replacement of amplitude values in the well-known phase retrieval methods of Gerchburg and Saxton [4] and Fienup [5]. However, instead of cycling through one depth plane at a time, the proposed mode-selection algorithm replaces all values and constrains the entire system each iteration. Benefits of this include an even weighting of error in the presence of noise and direct control over the systems state of coherence.

4.2.1 Fixed Holograms: Coherent, One Mode

A constraint must be applied to verify that the approximate AF created each iteration step obeys Eq. 3.9 for a given wavefront $U(x)$. The constraint is applied between steps shown in Fig. 4-1(d) and (e). A useful constraint is found from considering the coherence state of the illumination of the theoretical holographic display setup in Fig. 3-4. As we learned in Chapter 3, a fully spatially coherent illumination source (i.e., from a laser or a distance point source) leads to a rank-1 limited mutual intensity function. Thus, $J'(x_1, x_2)$ must be converted to a function $J_c(x_1, x_2)$ that is fully separable, i.e. $J_c(x_1, x_2) = U(x_1)U^*(x_2)$. Taking a linear algebra viewpoint, as with any 2D matrix, the $N \times N$ discrete mutual intensity matrix estimate $J'$ can be represented with a singular value decomposition (SVD):

$$J'(x_1, x_2) = SAV^T = \sum_{i=1}^{N} s_i \lambda_i v_i$$  \hspace{1cm} (4.2)

Here we show $J'$ decomposed into the well known SVD matrices $S$ (with columns $s_i$) and $V$ (with rows $v_i$). $\Lambda$ is a diagonal matrix containing the ordered singular value weights $\lambda_i$ for each rank-1 outer-product $s_i v_i$ that sum to equal $J'$. All off-diagonal elements of $\Lambda$ are 0. Furthermore, from [43], we know that an optimized
Figure 4-2: The decomposition of a mutual intensity function into its coherent modes. An initial mutual intensity guess of a wavefront incident upon an open aperture (left) is decomposed into multiple modes using an SVD (right), with weights given by their singular values. For example, $\lambda_1=1$, $\lambda_2=0.21$, and $\lambda_3=0.13$ after the algorithms first iteration, but quickly approach a single large value for $\lambda_1$. The constraint for single-mode selection simply selects the first mode of this decomposition.

Discrete coherent mutual intensity $J_c$ must fulfill a rank-1 condition. A good rank-1 approximation is given by the first singular value of the SVD in Eq. 4.2,

$$J_c(x_1, x_2) = \sum_{i=1}^{1} s_i \lambda_i v_i = \lambda_1 |s_1\rangle \langle v_1|.$$  \hfill (4.3)

In other words, to fulfill the coherence constraint implicit in a PSF measurement, we can represent $J_c$ as the outer-product between the first column of $S$ ($s_1$) and the first row of $V$ ($v_1$). Since a spatially coherent wave is composed of a single mutual intensity mode, all singular values besides $\lambda_1$ are 0. This constraint reduces our redundant 2D phase space representation to the two 1D vectors $s_1$ and $v_1$, which are equal if $J'$ is positive semi-definite. An example of the application of this constraint is in Fig. 4-2. After this constraint is applied, the coherent mutual intensity must be converted back to an AF so desired slices can again be replaced and iteration can proceed. Continuing with linear algebra notation, a coherent AF (Fig. 4-1(f)) is created from the coherent mutual intensity $J_c$'s first singular value with,

$$A_c(u, x') = F[L_{x_1}[R_{45}[U_c(x_1)U_c^*(x_2)]]],$$  \hfill (4.4)

where $R$ is a $-45^\circ$ matrix rotation, $L$ scales the axis by two, and $F$ is a Fourier
transform along one dimension in the rotated coordinate system. Eq. 4.4 is an implementation of Eq. 3.9 in discrete matrix operation form, much like Eq. 3.21-3.22 applied to the Wigner distribution. After this procedure, the AF provides a physically realistic representation of our coherent illumination setup but may not optimally match the desired inputs. Once again, originally desired OTFs are used to populate the new AF guess at their respective slices, and the outer product constraint of Eq. 4.3 is applied. This procedure iterates until convergence to a final AF, which will match desired responses within a specified error threshold. This AF can be inverted using Eq. 3.11 to solve for the optimal screen function up to a constant phase factor, or can directly determine an OTF at any other depth plane from Eq. 3.14. Since the SVD in Eq. 4.3 provides a rank-1 approximation of the original matrix with minimized Euclidean error (from the Eckart-Young Theorem [59]), quick convergence is expected.

4.3 Single-Mode Selection Performance

The single-mode selection algorithm restricts a set of desired intensity patterns, which may or may not obey the constraints of propagation, to a solution that follows coherent wave propagation. Therefore, two regimes of performance evaluation are necessary. The first regime considered will test performance for a set of OTF inputs that are known to obey the constraints of propagation, which is equivalent to testing the algorithms ability to recreate an entire AF from a few OTF inputs generated from a known hologram. This will demonstrate the mode-selection algorithm’s accurate convergence to known solutions. One could imagine using a known holographic pattern as a design starting point and then altering displayed intensities to determine a new pattern for a different desired intensity pattern. Subsequently, we will test the algorithms ability to converge to arbitrary desired sets of intensity distributions, which may be impossible to recreate exactly.
Figure 4-3: A binary amplitude hologram mask comprised of five slits is used as a known input to test algorithm performance. (b) Three OTFs generated from the aperture mask at different focal depths are used as input. (c) They generate an MI and AF guess, which improve upon iteration. (d) Output OTFs after 25 iterations closely match input OTFs.

4.3.1 Algorithm Performance, Ground Truth Intensities

As a first example, we model the simple binary amplitude mask distribution in Fig. 4-3(a) under the assumption of a lens at the display plane. Three of the OTFs it generates are used as algorithm input: one at the focal plane of the lens, one at an additional distance away corresponding to $W_{20} = 0.25\lambda$, and one at $W_{20} = 0.5\lambda$. These OTFs are in Fig. 4-3(b). For a 10mm-wide hologram and a lens with 50mm focal length, this corresponds roughly to $\Delta z = 0.1\text{mm}$ and $0.2\text{mm}$ for each intensity pattern off the focal plane, respectively. After 25 iterations, the algorithm converges to the OTFs and mask function shown in Fig. 4-3(d). Since the original inputs obey propagation, we expect iterative mode-selection to approach an exact reproduction of the OTFs, which it nearly achieves. The performance metric of mean-squared error (MSE) from desired OTFs is 0.007, which is on the order of error from phase retrieval approaches [73].

As a second example, we use the well-known continuous Cubic Phase Mask (CPM) [70] to generate three 1D OTFs for algorithm input. Unlike the previous example, this
Figure 4-4: Recovery of the CPM from 3 shift-invariant OTFs. (a) Three ground-truth OTFs (G, green) and algorithm reconstructions (R, blue) from a CPM (α=40) at Δz=0mm, 0.2mm and 0.4mm, using the example f/5 setup after 15 iterations. The ground-truth AF (b) and reconstructed AF (c) exhibit a large degree of similarity. (d) The output phase screen is comprised of the expected cubic phase profile.

The mask is phase-only, and is often used to provide a depth-invariant blur for extended depth-of-field (EDOF) imaging systems, or to create the well-known Airy beam. In its basic form, the 1D mask has a phase distribution $\phi(u) = \frac{\alpha^2}{u}$, where $\alpha$ is a scaling parameter. Fig. 4-4(c) displays the reconstructed AF from using three depth-invariant OTFs as input. Since the CPM is a phase-only element, a restriction is utilized during iteration to only select the phase-only contribution (i.e., all values are constrained to lie on the complex circle each iteration). The output mask, given as a separable 2D distribution in Fig. 4-4(d), shows a clear cubic phase profile. The reconstructed OTFs in Fig. 4-4(a) show a total MSE of 0.004 from expected. By providing a method to alter and optimize the AF at individual planes along the direction of propagation, the proposed algorithm has a large potential for assisting the design process of EDOF imaging systems, or to create novel depth-invariant beam profiles for optical trapping and manipulation applications [74].
Figure 4-5: Convergence analysis plots for the above two holographic mask examples (binary mask and CPM mask). (a) With increased iteration, the MSE between ground truth and reconstructed OTFs approaches zero. (b) Singular values, representing modes of partial coherence, approach a single mode with increased number of iterations (shown for the binary mask example). This single mode implies spatially coherent light, which follows from our assumption of modeling a holographic mask under completely coherent illumination.

Since the above examples originate from OTF inputs that obey propagation, the algorithm can converge to an exact solution upon iteration. Fig. 4-5 displays this convergence to near-zero MSE, as well as the mutual intensity function’s ability to approach a single mode (one large singular value). Likewise, both examples used three inputs at three easily definable, uniformly separated depths, for demonstration purposes. In fact, any number of inputs at any plane of depth could be used, and algorithm performance will vary as input parameters change. Clearly, if OTF slices that obey propagation are used, it is desirable to fill in more of the AF with a larger number of slice estimates \( n \). Likewise, a larger maximum plane separation distance \( \Delta z \) (i.e., a larger \( W_{20} \) value) will allow for a wider wedge area of the AF to be filled in, as is known in tomographic reconstruction problems. Both of these trends are demonstrated in Fig. 4-6 but do not remain valid in an arbitrary design situation.

4.3.2 Algorithm Performance, Desired Intensities

The above examples confirm convergence to the correct holographic masks for known intensity distributions. In the case of arbitrary inputs which may not obey propaga-
Figure 4-6: A demonstration of algorithm performance as a function of two free parameters: maximum input defocus parameter and number of input slices. (a) MSE between all input and output OTFs decreases as the maximum input defocus parameter is increased for both example masks. Each MSE value is an average MSE for 3 to 8 equally spaced input planes, each after 15 iterations. (b) MSE of input vs. output OTFs also decreases as the number of pre-determined equally spaced input planes is increased. Here, each MSE value is an average over a maximum $W_{20}$ value of $2\lambda-7/2\lambda$, also after 15 iterations.

Mode-selection can be applied to find a diffracting screen to approximate any 3D intensity distribution. One arbitrary but demonstrative set consisting of one point at one plane, two points at later plane, and three points at a plane further along the
direction of propagation is considered for many examples in this thesis. This type of counting 3D intensity distribution has a potential application as a camera aperture mask for depth detection. Here, it is mostly used as an illustrative example due to its high intensity variation along \( z \). Fig. 4-7 presents the process of the algorithm in simulation, using the same display parameters as in the previous subsection’s simulations. The three desired OTFs from equally separated depth planes of \( \Delta z = 0, 0.1\text{mm} \) and \( 0.2\text{mm} \) from the focal plane populate the AF, which iterates to yield optimized OTFs with an MSE of 0.032 from desired responses. The optimal amplitude and phase distribution to generate the responses in Fig. 4-7(d) is in Fig. 4-8(a). Fig. 4-8 also includes optimal amplitude-only and phase-only masks determined using a different restriction each iteration, which generate a similar but slightly different 3D intensity.

The influence of the number of inputs \( n \) and their distances along the propagation axis \( \Delta z_n \) becomes less predictable for the unknown hologram case. From simulation and through experiment, it appears that the complexity (i.e., rate of change) of the desired OTF set is the most significant influence on MSE performance. MSE versus maximum input plane distance does not follow a general trend and has been examined in part in [10]. Likewise, as opposed to a set that follows the propagation equation, increasing the number of arbitrary desired OTFs can over-constrain the design problem. Since the approximated output wavefield must be compatible with the Fresnel propagation process that ties all depth planes together, more inputs indicates a riskier search. Even specifying intensities at two different planes along \( z \) may not offer an approximate solution, which is especially relevant for closely spaced planes.

While the single-mode selection algorithm can recover known mask patterns and find desired mask patterns with a high degree of accuracy, there are still a number of failure cases and areas for improvement. First, taking the first singular value for a rank-1 mutual intensity function is an approximate restriction. While rarely observed, this first value could lead to mode values with a dynamic range too large to fabri-
Figure 4-7: The design procedure for a desired depth-varying intensity image. (a) Three desired 1D intensities of one, two and three sine functions of 5μm width at three depth planes yield three OTFs in (b) to populate an AF guess. (c) Iterative mode selection is applied to converge to an approximate solution after 50 iterations. (d) Three output intensities at the same depth planes as the intensities in (a) show the expected 1, 2, and 3 peak pattern, but are not exact solutions.

cate. Furthermore, the SVD process may not be optimal if constraints are placed on the holographic mask (e.g., a requirement of amplitude or phase-only content). However, as demonstrated above, an amplitude or phase-only constraint still leads to convergence for the coherent design case. These constraints are common when the mode-selective model is applied to an actual experimental setup, as device limitations commonly prevent displaying both amplitude and phase content. Alternative decomposition methods that address some of these shortcomings are discussed in detail in Chapter 5.

In addition, while an exact solution is approached when the input OTFs are known to obey propagation constraints, an arbitrary desired 3D intensity distribution may iterate to a local instead of global minima. This problem could be overcome with
Figure 4-8: Different hologram masks can be used to approximate the desired intensity set in Fig. 4-7. (a) The 1D optimized amplitude and phase distribution for a hologram that creates the desired 3 intensities in Fig. 4-7(d). (b) MSE drops with number of iterations for an amplitude and phase (A+P), amplitude-only (A-only), and phase-only (P-only) hologram mask, but at different rates. (c) The optimized binary amplitude (d) and phase-only (in radians) hologram masks in 2D.

an exact solution to phase space function inversion, although a direct method for this process is currently unknown. Alternatively, a larger space of solutions may be allowed by relaxing the rank-1 (i.e., single-mode) constraint carried out each iteration by Eq. 4.3. In general, a major advantage of using mode-selection over phase retrieval or phase-space tomography is the control over multiple mutual intensity modes offered by Eq. 4.3. In the following section, we will take a closer look at what relaxing this constraint implies, both in terms of algorithm performance, experimental realization and its relationship to partial coherence.

4.4 Multi-Mode Selection and Partial Coherence

A natural question when examining the mode-selection algorithm is whether the mutual intensity function's rank-1 constraint is a strict requirement. Of course, as explained in the previous section, placing a rank-1 constraint is the same as selecting a single coherent mode of the field, which matches our assumption of completely coherent light illuminating and diffracting through the designed hologram. However, if we assume that our holographic display setup includes a partially coherent illu-
mination source, then the rank-1 constraint is no longer valid. Instead, a rank-$M$
constraint must be placed, where the number of modes $M$ must be selected to match
the number of modes inherent in the partially coherent source [43]. However, it is
quite difficult to design an optical element that emits a desired number of $M$ modes
to illuminate a fixed holographic display. As many partially coherent optical sources
include modal weights that follow a Gaussian distribution, the Gaussian-Schell model
could be incorporated into the mode-selection design process to find an optimized optical source-hologram pair [85]. An alternative method of approximating partially coherent light is through multiplexing several distant coherent sources over space or angle, or using a completely coherent illumination source with a time-multiplexed display pattern [2]. This latter approach offers more experimental control over the partial coherence state display setup and is the focus of this section, with other multiplexing setups considered in Section 4.6.

Following, the mathematical foundation of extending the single-mode selection algo-

rithm to model partial coherence is presented. Then, a couple of examples of designing sets of multiple holographic display patterns using the multi-mode algorithm are pre-
sented. Models assume that these sets of patterns are displayed sequentially over time on a 2D screen to re-create a desired 3D intensity pattern within a finite time window. This time window can be the integration time of a viewer’s eye, or the exposure time of a recording device. The accuracy of integrating over the response from multiple of holograms is much higher than integrating over the response from a single hologram when generating a 3D intensity pattern.

### 4.4.1 Multiple Modes: A Partially Coherent Constraint

As discussed in the previous section, the use of a singular value decomposition (SVD)
of an $N\times N$ discrete $J(x_1, x_2)$ in the mode-selection algorithm is based upon the well-
known coherent mode decomposition [43, 45]. The SVD operation decomposes the mutual intensity function $J$ into a set of $N$ singular value modes, which represent $N$
individual (i.e., orthogonal) coherent optical fields that propagate independently from
one another. Typically, the number of relevant coherent modes $M < N$ is the number of non-zero singular values determined by the SVD of the mutual intensity function. $M$ offers a direct indication of the partial coherence state of the light that is being simulated (or, in other cases, being measured). In summary, the SVD decomposes a function (i.e., a representation of the optical field we wish to display) into $N$ mutually orthogonal components, each with a specific weight $\lambda_i$, for any complex 2D matrix representation. The first $M$ of these components can be selected to approximate a partially coherent field.

In the previous section, we modeled holograms under completely coherent illumination. Thus, the single-mode selection algorithm kept only the first singular value each iteration (i.e., the single most important coherent mode of the modeled field). We can extend the coherence decomposition in Eq. 4.2 to a desired degree of partial coherence by adding up the first $M$ singular values of the SVD of $J(x_1, x_2)$, with $M < N$:

$$J_{pc}(x_1, x_2) = \sum_{i=1}^{M} s_i \lambda_i v_i = \sum_{i=1}^{M} \lambda_i U_i(x_1) U_i^*(x_2)$$

(4.5)

This process is shown in Fig. 4-9. From the Eckart-Young Theorem, we know that $J_{pc}(x_1, x_2)$ is an optimal approximation of $J(x_1, x_2)$, since $J_{pc}$ is the rank-$M$ approximation of $J$ with minimized Euclidean error [59]. In other words, selecting the first

![Figure 4-9: The decomposition of a mutual intensity function into $M$ coherent modes, similar to Fig. 4-2. Here the first $M=3$ modes are selected (with $\lambda_1=1$, $\lambda_2=0.21$, and $\lambda_3=.13$) to obtain a partially coherent estimate of $J$ for the partially coherent mode selection algorithm.](image)
The $M$ largest singular values at each iterative step of the mode selection algorithm will always lead to an optimal constraint, regardless of the partial coherence state we wish to represent! This unique property of the mode-selection algorithm makes it a very attractive method to model any optical field with a known, estimated or desired partial coherence state, extending it into the multi-mode regime.

Furthermore, since each mode is orthogonal, $J_{pc}(x_1, x_2)$ can be thought of as a sum of $M$ unique, coherent mutual intensities. Under this alternative interpretation, $J_{pc}(x_1, x_2)$ creates an AF of a partially coherent source represented by,

$$AF_{pc}(u, x') = \int J_{pc}(x + \frac{x'}{2}, x - \frac{x'}{2})e^{2\pi i xu} dx = \sum_{i=1}^{M} \lambda_i AF_i(u, x')$$ (4.6)

Here we also express $AF_{pc}$ as a summation over coherent, orthogonal $AF_i$'s, which follows from a similar property of the Wigner distribution [44]. Each $AF_i$ obeys Eq. 3.9 for a single orthogonal mode $U_i(x)$. Eq. 4.6 demonstrates that optimizing a partially coherent AF is equivalent to simultaneously optimizing $M$ coherent, mutually orthogonal AFs that must be added to create a desired set of input intensity patterns. These $M$ coherent AFs will provide a set of $M$ holographic masks at the algorithm's output, each weighted by its associated singular value $\lambda_i$.

This summation of AFs to achieve a desired response is directly connected to the longstanding problem of OTF synthesis, studied earlier by Marechal [57]. As with synthesizing OTFs, one way to implement the summation in Eq. 4.6 is to multiplex each coherent mode over time, which has previously been proposed to simulate partially coherent illumination [2]. Specifically, the fixed holograms designed in Section 4.2 (e.g., the hologram in Fig. 3-4(a)) can be replaced with a dynamic screen, like a spatial light modulator (SLM), which can display the mask pattern associated with each coherent mode for a finite amount of time over the duration of one image exposure. The length of time each mode is displayed will be proportional to its singular value $\lambda_i$. The next section demonstrates how several holographic patterns multiplexed
Figure 4-10: Demonstration of the multi-mode selection algorithm applied to a known partially coherent input. (a) 3 mutually incoherent plane waves strike a diffracting screen at different locations, creating 3 orthogonal coherent modes (amplitude shown below). (b) The partially coherent $AF_{pc}$ and mutual intensity $J_{pc}$ (rank 3) for this scenario (maximum is black). (c) 4 OTF slices from $AF_{pc}$, corresponding to 4 different depth planes along $z$, are used as algorithm input. (d) The optimized $AF_{out}$ and $J_{out}$ in such a manner can offer additional flexibility in creating a desired depth-varying 3D intensity pattern as compared to a single holographic mask.

4.5 Designing Partially Coherent 3D Images

To incorporate partial coherence effects into the single-mode selection algorithm, only the rank constraint connecting Fig. 4-1(d) and Fig. 4-1(e) must be modified using Eq. 4.5. For a given set of input intensity patterns and a desired number of coherent modes $M$, this simple change will allow mode-selection to find $M$ optimal weighted holographic mask functions. In most cases, these $M$ holographic patterns will lead to a 3D intensity pattern that is a closer match to the desired input as compared with the 3D intensity pattern created by a single hologram. If $M > 1$, then the mode-selection algorithm is operating within the multi-mode domain.
4.5.1 Ground Truth Multi-Mode 3D Image Reconstruction

First, as a demonstration of multi-mode selection's ability to accurately converge, performance is tested on a set of ground-truth OTFs that are known to obey the constraints of partially coherent propagation. This is equivalent to testing multi-mode selection's ability to recreate an entire partially coherent AF from a few OTF inputs. Here, each OTF will be a sum of the OTFs produced by a set of known holographic masks at a given depth plane. Fig. 4-10(b) displays an example partially coherent $J_{pc}$ and $AF_{pc}$ used to generate 4 input OTFs: one from an initial plane 50mm away, and three additional intensity patterns at $W_{20} = .5\lambda$, $W_{20} = \lambda$ and $W_{20} = 1.5\lambda$ from this plane, respectively. For a 10mm holographic mask and a lens with 50mm focal length, this corresponds roughly to three additional depth planes at 50.2mm, 50.4mm and 50.6mm distance from the hologram. A faithful reproduction is achieved after 50 algorithm iterations. The performance metric of mean-squared error (MSE) from desired OTFs is $8x10^{-4}$, which is slightly better than the error achieved using a completely coherent constraint as considered in the previous chapter (i.e., a rank-1 constraint between Fig. 4-1(d) and Fig. 4-1(e)). MSE is defined as the normalized squared difference between model input (i.e., ground truth OTFs) and output. As with the coherent design case, errors can be attributed to the limited maximum angle of the OTF slices, as well as a large amount of rapid changes from three overlapping central AF cross-terms in this particular example.

4.5.2 Desired Multi-Mode 3D Image Design

Second, the multi-mode selection algorithm is applied to the problem of designing a desired 3D image that may not obey the constraints of propagation (Fig. 4-11). Specifically, a set of desired intensity patterns at different depths and a desired partial coherence state $M$ are used as input. In the example in Fig. 4-11, the same "test" 3D image consisting of one point turning into two and then three points at three depth planes is used. The algorithm iterates to find an optimal rank-$M$ mutual intensity function, which corresponds to a set of $M$ holograms that can be displayed over time.
Figure 4-11: The multi-mode selection algorithm applied to determining an optimal set of $M$ holographic masks from a desired 3D intensity pattern (left), with $M = 3$. The algorithm uses a rank-3 constraint between the steps shown in Fig. 4-1(d) and Fig. 4-1(e). The algorithm’s output is 3 1D amplitude and phase functions (right), each representing one holographic mask pattern. Alternatively, one can interpret each optimized hologram pattern as the optimal representation for each individual coherent mode of the $M$ modes of $J$.

to recreate this desired 3D image. While $M=3$ modes worked well in the example ($\text{MSE} = 5.4 \times 10^{-3}$), $M$ can be varied depending upon system specifics to typically (but not always) provide a better estimate for more modes. In this example, three modes perform much better than one mode, whose results are shown in Fig. 4-7 and offered a $\text{MSE}=0.032$. Normed weights for the three modes are $\lambda_1=1$, $\lambda_2=0.61$ and $\lambda_3=0.11$.

For a large set of tested intensity patterns, using $M > 1$ modes almost always yields a decreased MSE between the desired and realized 3D pattern as compared to using a single mode ($M=1$). This finding offers the theoretical basis for the success of the proposed display device and can be understood from several viewpoints. From a mathematical standpoint, it is not surprising that an algorithm with a rank-$M$ constraint performs better than an algorithm with a rank-1 constraint. Each iteration, the rank-$M$ constraint works within a much larger space of possible functions than the rank-1 constraint. Specifically, it operates in a space that is $M-1$ times larger.
than the single mode-selection algorithm’s space (i.e., coherent field design space), offering increased flexibility in design parameters. From a physical standpoint, it is well understood that a partially coherent optical source can create a larger subset of desired intensity patterns as compared to a single inflexibly wavefront (see [75, 1], for example). Finally, from an engineering standpoint, displaying multiple holographic patterns to generate a desired intensity pattern offers an increased number of degrees of freedom to work with than just having a single pattern to display. Specifically, if a coherent holographic display has $K^2$ pixels, then its partially coherent counterpart will be able to utilize $MK^2$ pixels, predictably offering enhanced performance.

3D image improvement with an increasing $M$ does not continue unbounded. While the above simulations demonstrate improved MSE for small values of $M$, larger $M$’s lead to a narrow depth range across which a 3D intensity can be designed [20]. This can be understood by considering the MI function $J$, with a rank that increases proportional to $M$ but loses off-diagonal non-zero elements as $J$ approaches a single diagonal matrix at the incoherent limit. This effect is also observed with an increase in temporal incoherence in areas like optical coherence tomography [86], which obtains a narrower spot along $z$ with less coherent light. Note that the use of polychromatic illumination in a holographic display setup will wash out image detail due to a range of wavelength-dependent diffraction angles creating a single 3D image. Generally, a monochromatic source should be used and an optimal number of modes for given 3D image should be found by cycling through many values of $M$ and finding the value that minimizes MSE between the desired and realized 3D image.

4.6 Alternative Multiplexed Displays and Partial Coherence

So far, this chapter attempted to develop a direct relationship between the number of mutually incoherent modes of a field, its partial coherence state and its performance
at displaying a 3D pattern. Limited amounts of partial coherence improve 3D intensity design when implemented in a time-multiplexed setup. However, a number of alternative experimental methods can be used to create or simulate a specific partial coherence state by delivering multiple modes to a viewer. For example, instead of displaying modes sequentially over time, modes could simultaneously be shown from multiple positions or angles. In this section, we briefly explore some of these alternative multiplexed display methods. While often not directly acknowledged, many previously proposed holographic and 3D display setups either directly or indirectly rely on multiplexing multiple images over space, time, angle or wavelength. Following, an outline of 3D display multiplexing is developed under several broad categories, and an example of simple conceptual display is offered as a thought experiment.

4.6.1 General Categorization of Multiplexing

Multiplexing is a borrowed term from the field of communications, defined as the combination of multiple signals over a shared medium. As noted in Section 4.1, it generally refers to the reduction of a higher-dimensional signal into fewer dimensions in a recoverable process. In the context of 3D display, it implies the mixing of 3D imagery data onto a 2D screen, which can re-create the appearance of a 3D image when illuminated. It is used quite often in holography, as a typically thin 2D recording surface is used to preserve the 3D spatial and sometimes spectral information of an object. It is also used in other areas of computational imaging and display, where some explore the general problem of projecting the 4D light field of an object and its 1D spectral and 1D polarization content onto a 2D sensor, then demultiplexing each dimension to gain more information about an object or scene of interest [87, 88].

Fig. 4-12 presents a table of general multiplexing categories and various previous realizations for both ray-based and holographic 3D display formats. This table is by no means complete but is instead meant to place the proposed display method in a context of many other 3D display formats that attempt to improve performance. Furthermore, many of the entries could potentially belong to multiple categories, or
## 3D Displays Categorized by Multiplexing Method

<table>
<thead>
<tr>
<th>Space</th>
<th>Incoherent, 2D Screen</th>
<th>Incoherent, Other</th>
<th>Holographic, 2D Screen</th>
<th>Holographic, Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Integral Imaging</td>
<td>Swan’s cube* (A)</td>
<td>Hogel (Zebra) [Lucente '94]</td>
<td>Multi-SLM [Yaras'08]</td>
</tr>
<tr>
<td></td>
<td>[Fraul '04, Javidi '05]</td>
<td>[Swan 1862]</td>
<td>DOE [Chen '98]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Stereogram [Debietto '88, Halle '91]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rainbow Hologram [Benton '69]</td>
<td></td>
</tr>
<tr>
<td>Angle</td>
<td>Parallax Barrier</td>
<td>Multi-projector</td>
<td>Horizontal Scan*</td>
<td>Active Tiling</td>
</tr>
<tr>
<td></td>
<td>[Ives 1903]</td>
<td>[Yoshida '11]</td>
<td>[Takaki '09]</td>
<td>[Stanley '03]</td>
</tr>
<tr>
<td></td>
<td>Lenticular</td>
<td>Projector Array</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Lippmann 1908]</td>
<td>[Said '09]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Content Adaptive</td>
<td>Rotating mirror* (D)</td>
<td>Proposed Display</td>
<td>MIT HoloVideo*</td>
</tr>
<tr>
<td></td>
<td>[Lehman '10]</td>
<td>[Simon '77, Jones '07]</td>
<td>[DeSantis '86, Horstmeyer '11]</td>
<td>[St. Hilaire '91, Lucente '93]</td>
</tr>
<tr>
<td></td>
<td>Dynallax</td>
<td>Varifocal* (D)</td>
<td>Polymer-based* (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Peterka '07]</td>
<td>[Traub '67]</td>
<td>[Blanche '10]</td>
<td></td>
</tr>
<tr>
<td>Wavelength</td>
<td>Color-Sequential</td>
<td>Reflection Hologram* (D)</td>
<td>Volume Storage*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Lee '06]</td>
<td>[Denisyuk '62, see note in text]</td>
<td>[Bashaw '95, Gerke '10]</td>
<td></td>
</tr>
<tr>
<td>Depth</td>
<td>Layered 3D</td>
<td>Fiber Voxel</td>
<td>Tandem Holograms</td>
<td>Aperiodic</td>
</tr>
<tr>
<td></td>
<td>[Wetzstein '11]</td>
<td>[MacFarlane, '94]</td>
<td>[Bartlett '94, '95]</td>
<td>Elements* (W,A)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fluorescence</td>
<td></td>
<td>[Gerke '10]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Lewis '71]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Solid (FELIX)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Langhans '03]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

An asterisk (*) denotes the listed display may also belong to another category in the same column.

A letter in parenthesis denotes the multiplexing scheme that is considered when generating display content. (S)=Space, (A)=Angle, (T)=Time, (W)=Wavelength and (D)=Depth.

Figure 4-12: A table of various 3D display techniques categorized into different multiplexing schemes. Several technologies do not fix explicitly into a single category. These display methods are included with an asterisk. Furthermore, when content generation may be achieved through an alternate multiplexing scheme, a letter in parenthesis indicates the related category (using the first letter of the categories in the first column).
might be classified into a different category under an alternate interpretation of operation. Such entries are appended with an asterisk. Sometimes, content is generated with one multiplexing method in mind while the display’s optical or mechanical implementation uses another multiplexing method. If this is the case, the entry is listed under its optical/mechanical category, with its content generation category listed in parenthesis. For the purpose of simplicity, spatial multiplexing refers to the division of a hologram across its surface with a single illumination source, while angular multiplexing relies upon multiple sources or moving or rotating sources. The general categories of angle, space, time, wavelength and depth are used. All of the categories except depth are discussed in previous literature. Depth is added to include several 3D display forms that rely on many optical interactions between the light source and the viewer’s eye to improve performance. All displays mentioned in the depth category rely on designing multiple stacked screens or emitters, and are thus categorized here as a subclass of the general “volumetric” display category that also includes single screens or elements that may move or rotate.

First considering non-holographic 3D displays, the previously discussed parallax barrier and lenticular forms of display fall into the category of angular multiplexing (i.e., different views are delivered along different ray angles to a viewer). Integral imaging, on the other hand, mixes elemental images of an object on the screen [40]. Temporal multiplexing is often used to improve the visibility of parallax images but can also increase light throughput and help with user tracking [80, 89]. Similarly, color can also be delivered sequentially over time [90]. Also, multiple dynamic screens were recently stacked together to use the depth dimension (z) to improve 3D image appearance.

As for alternative incoherent displays, one of the first methods proposed to present 3D scenery (the “Swan cube”) shows two different offset images combined with a beamsplitter [91]. While this display places two images at different spatial locations, it determines the image content considering angular disparity. Alternatively, multiple projectors arranged in a ring behind a screen performs the task of multiplexing images
over angle [92, 93], allowing viewers at largely varying viewing angles to observer horizontal parallax. Rotating mirrors effectively achieve the same effect by guiding many images to different angles quickly over time [94, 95], as do deformable lenses [96]. These two displays also extend into the depth dimension and are often regarded as volumetric display methods. Finally, incoherent displays that include multiple depth planes containing individual fibers [97] or fluorescent material [98, 99] are not quite multiplexing devices but are included for completeness.

Turning to holographic multiplexing methods, we will first consider (mostly) thin 2D holographic elements. There are many holographic techniques that in some form or another multiplex multiple holograms onto different spatial locations of a material’s 2D surface [25, 100, 23, 34]. Squeezing multiple holographic elements into a thin area below a lenticular array to be spread across a wide angle was demonstrated by [101]. Multiplexing a 2D hologram over time is where the technique proposed in this thesis fits in with all others. An unrelated method of displaying holograms over time (although not multiple modes for a single holographic image) was recently demonstrated by [102] using a polymer-based recording medium. The pattern design method for this display relates to the MIT HoloTV, discussed later. Multiplexing over wavelength is indirectly achieved in Denisyuk’s reflection holograms, which offer a high selectivity of wavelength over angle and the possibility of recording multiple wavelength images onto a single surface. However, note that reflection holograms are “thick” holograms, but are here considered within the category of a single hologram screen. The near-vertical fringes of this screen facilitate wavelength selectivity (much like a dielectric stack). Thinner holograms with more horizontal fringes include (in descending order of thickness) edge-lit [103, 106], off-axis [104], and in-line holograms [105]. Finally, multiple in-line holograms have been stacked in tandem along the optical axis to offer increased performance [107, 108], which this thesis categorizes as multiplexing holographic content over depth.

Many alternative holographic displays have also been proposed. Multiplexing the
holographic image over multiple SLMs (spread out spatially) improves resolution and can offer color [109]. An LCD and lenslet combination has been placed behind an SLM, effectively multiplexing the SLM over time and angle to provide a high resolution output [111]. The MIT HoloTV is a good example of a holographic display method that multiplexes its images over time [26, 28], although its operation is connected closely to spreading its content over angle as well. Finally, thick volume elements (much thicker than Denisyuk’s reflection hologram) have been used to multiplex multiple images over the incident light’s wavelength [110], effectively extending holographic design into a volume [112].

In summary, loosely grouping various 3D display methodologies within the context of multiplexing establishes connections between otherwise unrelated display setups. Regardless of its coherence, light emitting from a 2D screen into a 3D volume greatly benefits from some additional dimension of manipulation. From the above table, it should be clear that space, angle, time, wavelength and depth are several common dimensions exploited by clever optical setups. Other potential directions for investigation may include either polarization or code-division multiplexing (CDMA), which is a common multiplexing method specific to communications. A key distinction between previous multiplexing setups and the proposed display is the content of each elemental multiplexed component. The mode-selection process finds an optimal set of components (i.e., the principle components), while most other methods do not. Any multiplexing scheme could potentially be applied to transmit the optimized modes of the partially coherent mode-selection algorithm. In the next two chapters, temporal multiplexing will remain the focus of this thesis. Before turning to this setup, an alternative multiplexing scheme is briefly presented as a conceptual example.

4.6.2 Angular Multiplexing: A Hybrid Design Example

An alternative multiplexing setup that generates holographic images with high variation along viewing angle demonstrates the value of multiple modes. This 3D display is a hybrid parallax barrier-hologram setup, delivering unique 4D light fields into dif-
Figure 4-13: (a) One model of a hybrid parallax barrier-hologram display design. For the modeled setup, $d=0.5\text{mm}$, $z_1=100\text{mm}$, $z_2=100\text{mm}$, $h_H=40\text{mm}$, $t_H=6.6\mu\text{m}$, $w=0.06\text{mm}$, $r_p=20\mu\text{m}$, and the resolution of each desired image is $100^2$ pixels. Higher orders are not shown. (b) The associated amplitude and phase distributions to generate each of the coherent light fields in (a), which are interlaced together every 3 pixels to generate the total screen pattern $s_1$. Color bars extend from $[0,1]$ for $A$ and $[0,2\pi]$ for $\phi$. (c) A simplified experimental demonstration of the proposed hybrid parallax barrier-hologram design, performed with binary printed holographic film, with parameters marked.
ferent viewing directions, effectively generating a 5D output (Fig. 4-13). The display replaces the interlaced images of a parallax barrier with high-resolution segments of separate Fresnel holograms to create depth and angle variant imagery. It takes advantage of the shared spatio-angular resolution relationship of both parallax barriers (Eq. 2.1) as well as Fresnel holograms (Eq. 2.4). Instead of serving as a final display solution, however, it is simply intended as a novel example of 3D display to encourage others to examine 3D display methods within a multiplexing context.

The proposed hybrid display utilizes angular multiplexing to generate $M$ coherent light fields that change dramatically with viewing angle. Displaying unique modes that change with angle is similar to the holographic methods in the "angular" category of the above table (see [111, 112]). The display's initial hologram plane contains $M$ Fresnel holograms of resolution $N_h$ split vertically into $K$ segments and interlaced much like the elemental images that comprise screen $s_1$ of a parallax barrier. Each elemental strip of the $M$ holograms contains $N_h/K$ pixels instead of the single-pixel strips found with parallax barriers. Selection in display angle is set by a plane of vertical slits a distance $d$ away that directs diffracted light into specific directions much like a parallax barrier. The slits in this plane are much wider (specifically $N_h/K$ times wider) than the hologram screen's pixels, and can therefore be considered under a ray-based model. Multiple illumination sources can replace a moving source to present all modes at a single moment. Unlike previous display setups, each coherent mode (i.e., each interlaced hologram) can be optimized within the mode-selection framework to approximate a desired 3D image. Thus, the $M$ interlaced holograms direct $M$ unique 3D images into different viewing angles.

Figure 4-13(a) offers a simulation of a hybrid display, with parameters listed in the caption. Each $2048^2$ pixel hologram is generated using the desired input images at the bottom of Fig. 4-13(a) using the single-mode selection procedure. $M$ holographic patterns are generated independently for $M$ desired 3D images, here created using binary desired intensity patterns from 2 planes along $z$, for simplicity. In this exam-
ple \( M=3 \), with the hologram patterns shown in Fig. 4-13(b). The MSE between the desired and modeled 3D images is .0194. The modeled display containing \( 6144^2 \) 7\( \mu \)m pixels is beyond current dynamic display fabrication techniques. A lower resolution film-based setup (Fig. 4-13(c)) offers a proof-of-concept example.

The experimental proof-of-concept uses two binary transparencies printed at 25\( \mu \)m resolution with dimensions 24mm by 36mm. The hologram plane \( s_1 \) contains \( M=3 \) interlaced holograms consisting of \( N_h=480 \) pixels in 1D, segmented into 96 segments (5 pixels per segment). A plane of 96, 125\( \mu \)m slits spaced apart by 250\( \mu \)m is placed 1mm behind the hologram plane. Due to resolution constraints, each elemental hologram is a 2D Fourier hologram yielding one 2D image and is not specified through the mode-selection process. As the illumination varies, images at a viewing plane (10cm from the hologram plane) dramatically change appearance (i.e., become completely different letters). This simple demonstration of holographic multiplexing along angle using only two planes of amplitude modulation suggests the larger format modeled in Fig. 4-13 may be experimentally possible in the future.

In summary, the angle-multiplexed hybrid display is included here to encourage creative new techniques to display the multiple output holograms of the multi-mode selection algorithm. At the same time, it highlights many similarities between parallax barriers and holograms. Drawbacks include the unnecessary creation of multiple image orders common to all amplitude-only based hologram generation. Additional orders are typically blocked. Light efficiency also decreases with the addition of slits, and multiple illumination sources are needed for simultaneous viewing. Furthermore, cross-talk between neighboring interlaced holograms must be limited through padding or baffling in an experimental implementation. However, this simple form of hybrid display is one step towards the merger of two methods of presenting 3D images that, as pixels continue to scale down, are bound to intersect.
Chapter 5

Designing 3D Diffractive Displays with Constraints

While the multi-mode selection algorithm optimizes a set of $M$ holographic masks accurately in simulation, slight modifications are required to successfully apply it to a practical display situation. As noted earlier, almost all methods of dynamically addressing a display, whether with an LCD, a transparent or reflective SLM, a MEMs mirror array or otherwise, have difficulty displaying amplitude and phase content simultaneously. Several interesting solutions have been proposed to overcome this well-known problem in holography and beam-shaping [76, 77]. However, as the proposed algorithm is a method of design, the specific limitations of any display setup can be taken into account simply by applying additional constraints to the iterative optimization process. For example, in the coherent design case of the previous chapter, Fig. 4-8 shows successful output holographic screens with amplitude-only (c) or phase-only (d) content. These experimentally successful screen designs were achieved simply by constraining the screen to its absolute value (for amplitude-only) or the closest value on the complex unit circle with magnitude 1 (for phase-only) at each iterative step. The trends of convergence in Fig. 4-8(b) make clear that direct constraints work well with an unmodified single-mode-selection algorithm. However, the same constraints do not work well for multi-mode selection when the desired number of modes $M > 1$. Following is a closer examination of the problem of constraining
multiple modes (i.e., a partially coherent field) leads to several unique coherent mode decomposition techniques.

5.1 Constrained Multi-Mode Selection Failure

In the following analysis, we will mostly focus on the situation where our available dynamic modulation screen can only display pixels that vary in amplitude, as with many commercially available LCD displays. As noted above, this type of screen requires the several holographic patterns that are the output of the mode-selection algorithm to contain no phase information. The case where the diffractive screen used in experiment can only show phase is briefly considered in the next section.

Unlike the single-mode selection algorithm, the multi-mode selection algorithm produces a partially coherent mutual intensity (MI) function that is not directly generated by our holographic display. Instead, the partially coherent MI leads to a set of screens, whose response after integration over time generates the desired 3D image output. Thus, it is slightly more difficult to connect the amplitude-only optical constraint of our setup with a specific constraint to use during algorithm iteration. If a constraint process similar to the single-mode selection algorithm is used (e.g. creating Fig. 4-8(c) for \( M > 1 \) modes), two choices are available:

1. Constrain the entire rank-\( M \) MI to its absolute value each iterative step, then find its modal decomposition.

2. First find its modal decomposition, then individually constrain each rank-1 mode of the MI (i.e., each orthogonal component of the singular value decomposition).

Intuitively, it seems that option 1 should be selected, since constraining the entire MI worked well with the single-mode algorithm. However, selecting option 1 leads to an immediate problem. While the rank-\( M \) MI function will contain only real values and thus obey the constraints of our display setup, its orthogonal modes will not.
Putting it another way, even though a rank-$M$ matrix contains only positive and real values, each of its rank-1 modes created by an SVD do not necessarily have to contain positive, real values. Often, each mode will contain many negative values, which are impossible to optically create on our amplitude-only display according to Eq. 3.3.

Selecting the alternative option 2 does not help much. While taking the absolute value of the $M$ separate coherent modes leads to screen patterns that can be shown on an experimental display, it prevents the algorithm from converging to a solution (i.e., the decreasing trends in Fig. 4-8(b) are not observed). Although the cause of this phenomena is mathematically complex, it can be intuitively understood somewhat directly. Placing an amplitude-only or phase-only constraint at each iterative step of the algorithm is quite similar to filling in one additional desired intensity at an additional slice of the AF. Specifically, the intensity at the screen plane is given as the vertical slice through the AF, which becomes clear if $\Delta z$ is set to $-f$ in Eq. 3.13, or if $z_p$ is set to 0 in Eq. 3.16. A phase-only constraint is identical to assigning the intensity at this slice, since it is required that the intensity $I(x)=1$ everywhere within the screen area and that $I(x)=0$ outside the screen area. An amplitude-only constraint leaves a similar number of degrees of freedom available (less a factor of $1/2\pi$), but is not as directly explicit within the AF framework. In either case, when designing for $n$ desired intensity patterns, a direct parallel can be drawn between placing a global constraint on the hologram screen and the performance of the algorithm when designing for $n+1$ desired intensity patterns. As the single-mode selection algorithm successfully converges for any number of desired intensity patterns, convergence with screen constraints is thus also expected, as demonstrated by the successful convergence to the screens in Fig. 4-8(c)-(d).

For the multi-mode case, placing constraints on each coherent mode of the MI is not the same as filling in one additional slice of the partially coherent AF. Instead, it is equivalent to filling in one slice of each coherent AF that comprise the sum in Eq. 4.6. Since the desired intensity slices are assigned to the partially coherent AF
each iteration, this means two different functions are constrained each iteration. In other words, the algorithmic step in Fig. 4-1(a) constrains $AF_{pc}$, while an amplitude-only requirement at step Fig. 4-1(e) constrains each individual $AF_i$ in the multi-mode case. This causes the iteration process to oscillate between two different paths as it proceeds to converge, and thus fails to lead to a final solution.

5.2 A New Amplitude-Only Constraint

To address the shortcomings of both of these options, the decomposition step of the mode-selection algorithm must be modified. Although the Eckert-Young Theorem informs us that an SVD yields a set $M$ of modes that sum to an MI estimate with minimum mean-squared error, this does not remain true if constraints or prior knowledge are applied. Fortunately, the solution to this problem exists for our amplitude-only requirement, where the decomposed modes must be comprised of positive, real values. Non-negative matrix factorization (NNMF), recently developed by Lee and Seung [53], decomposes a matrix $J$ into a set of $M$ positive, real rank-1 matrices that sum to form $J$'s optimal estimate with minimized mean-squared error. Mathematically, the solution can be presented as a convex optimization procedure solving for,

$$\min ||J - WH^T||, W \geq 0$$

(5.1)

Here, $J$ is our $N \times N$ MI function, and $W$ and $H$ are $N \times M$ matrices, where $M$ is an input desired number rank-1 modes (i.e., the desired amount of partial coherence). Note that this convex optimization procedure is approximate in nature and is not unique. However, an accurate result with minimized MSE is often found with any number of algorithms, including multiplicative updates, least squares with projected gradients, and alternating least squares [82].

The problem of decomposing a physically accurate MI function $J$ requires a slight modification of Eq. 5.1. From Eq. 3.2, it is clear that not only is $J$ rank-1, it is also symmetric (assuming no phase content). Thus, a symmetric NMF solution must be
Figure 5-1: Symmetric NMF decomposition in the multi-mode selection algorithm. (top) The same desired intensity set of 1-3 points at 3 depths as input, here as 2D separable patterns. The same parameters are used in simulation as in Chapter 4 (f=50mm, hologram size=10mm, pixel size = 100µ, sensor pix.=5µ). (middle) Symmetric NMF with M=3 produces the 3 holograms shown to the left with associated weights. Simulating their display over a duration of time proportional to each weight creates the three intensities shown, with an MSE of 8x10^{-3} from above inputs. (bottom) In comparison, an SVD creates these holograms (phase not shown) and intensities, with an MSE of 8x10^{-4}.

found, which has previously been examined in [54]:

$$\min ||J - WW^T||, W \geq 0$$  \hspace{1cm} (5.2)

This thesis implements the symmetric NMF search for W through a modification of Lee and Seung’s multiplicative updates method (which is openly available [83]). The modification relies on the iterative nature of convex optimization, adding a constraint to the multiplicative update rules that W and H should be equivalent. The updates
on $W$ and $H$ thus become,

$$H = W^T$$

$$W = W \odot (H^TJ \odot ((HH^T)H + \delta))$$

$$H = H \odot (W^TJ \odot (H(WW^T) + \delta))$$

(5.3)

where $\odot$ and $\odot$ represent a Hadamard product and division, respectively, and $\delta$ is a small value added to avoid possible division by 0. With this update, the NMF process yields an optimal $W$ and $H$ that are approximately equal, with residual error on the order of a few percent due to a finite iteration length. Thus, to obtain a set of $M$ optimized amplitude-only holograms, the partially coherent MI function $J$ is first set to its absolute value to remove any phase content. Then, the symmetric NMF process replaces the SVD decomposition that links step (d) and (e) in Fig. 4-1 (i.e., the decomposition laid out in Fig. 4-2) for a real, positive $J$. This new process overcomes the non-convergence problem discussed in the previous section. It offers an experimentally realizable set of amplitude-only holograms that lead to an optimal estimate of a desired 3D image. An example set of optimized output patterns is in Fig. 5-1, shown in comparison to the amplitude values of a set of amplitude and phase patterns. The desired input in this example is the same 1, 2, 3 - point pattern used in Fig. 4-11, here shown as a 2D separable function. Generally, performance of amplitude-only patterns is slightly worse than joint amplitude and phase patterns, which is to be expected since less degrees of freedom are available for designing the 3D image. However, multiple amplitude screen patterns perform better than a single amplitude screen pattern. For the example in Fig. 5-1, three amplitude and phase screen patterns displayed sequentially over time decrease the MSE between the desired and realized 3D image by a factor of six as compared to a single screen pattern (MSE drops from .032 to .0054). Three amplitude-only screen patterns decrease MSE by roughly a factor of three as compared to a single pattern (MSE drops to .0096). Some of the difference in performance can be attributed to a build-up of background bias in the case of amplitude-only screen patterns.
5.3 A New Phase-Only Constraint

Placing a phase-only constraint on a partially coherent MI is quite similar to the restrictions surrounding the constrained amplitude-only case. Again, the decomposition linking steps (d) and (e) in Fig. 4-1 must be modified. Instead of an SVD, an alternative iterative procedure was developed to decompose a matrix $J$ into a set of $M$ phase-only rank-1 symmetric matrices, here called “phase factorization.” Each entry of the complex phase-factorized matrices has an absolute value of 1. Their weighted sum approaches an approximation of $J$ with minimized MSE. The solution is found through an iterative convex optimization procedure, which finds one rank-1 matrix and corresponding weight at a time and uses residual error to drive the search for the next weighted mode. Otkrist Gupta played an invaluable role in designing this algorithmic implementation of the phase-factorization concept.

Further details of this decomposition process are not necessary for the full development of this thesis. It should be noted that while phase factorization is successful in simulation, producing similar MSE values, it is more error-prone than the symmetric NMF process. For example, by definition all diagonal elements of each symmetric mode will have a value of 1. Furthermore, a larger number of phase-only modes (i.e., phase-only holograms) are often required for convergence. Typically, roughly ten phase-only modes are required to reconstruct a mutual intensity function as accurately as approximately five amplitude-only modes. This requirement is mostly a manifestation of the phase-factorization algorithm, and should not hold as generally true for all constrained solutions. For these reasons, phase factorization is not used to produce the experimental results in the next chapter. Future efforts will focus on improving the phase-factorization process, as a phase-only holographic device is preferable over an amplitude-only device to increase both the available degrees of freedom and optical efficiency, among other reasons.
Chapter 6

Experimental Investigations

In this chapter, experimental verification of both the coherent (fixed holographic display) and partially coherent (time-varying holographic display) design examples are presented. Many challenges currently face implementation of dynamic holographic 3D displays. For example, the unavoidable generation of speckle noise, multiple diffractive orders and the lack of color are three commonly cited shortcomings that are also noted in the following demonstrations. However, as will be demonstrated, the greatest challenge with creating quality holographic 3D images is finding a dynamic optical element with many small pixels (i.e., a high space-bandwidth product). Due to these challenges, the experiments in this chapter are not meant as complete solutions for future 3D display architectures. Unsurprisingly, the experimental results demonstrating performance of the single-mode selection algorithm offer the same performance as computer-generated holograms designed by many others over the past several decades. The multi-mode results are new demonstrations but are also grounded in a well-known phenomenon [2]. Instead, these experiments provide initial verifications of this thesis’s proposed algorithm, which offers a more accurate method of 3D intensity design. They also provide a basis for further investigation into the use of temporally multiplexed holograms to increase the accuracy of 3D image creation. As such, several points should be noted about the chosen experimental setups, which constitute platforms for performance evaluation instead of direct display.
First, while the mode-selection algorithm can be applied to the design of both real and virtual image intensities with a minor modification, experiments were performed assuming the formation of a real image for measurement simplicity. In practical display scenarios, a virtual image is typically generated for viewing. Note, however, that since a real and virtual 3D image are related by a conjugation of phase, the ability of a set of holograms to produce either form of image is nearly identical. Second, while it is anticipated that the pixel size of commercially available displays will shrink towards the wavelength of light in the future (currently with pixels approximately 15 times the wavelength of visible light), these experiments use much larger printing and display resolutions for reasons of cost and availability. To increase the diffraction efficiency of the available displays, a lens is placed close to the hologram plane. The operation of a lens in these experiments is to effectively bring the Fourier plane (i.e., the far field, where the entire optical signal can mix) much closer to the holographic display. Without it, the anticipated design method still works but on scales much larger than an optical bench, producing a much larger 3D image and proportionally reduced intensity. As noted in Chapter 3, the use of a lens changes the mode-selection algorithm only slightly (i.e., Eq. 3.14 must be replaced with Eq. 3.15, simply filling in desired OTFs at slightly different slice locations in the AF). Future display setups will most likely not use a large lens, and thus should be optimized using Eq. 3.15 in the mode-selection process.

6.1 **Fixed Hologram Design Experiments**

With these two simplifications, the single-mode selection algorithm is first tested against the performance of a brute-force search algorithm, which attempts to find an optimal AF and holographic mask pattern by considering all possible patterns within a limited search space. To reduce the search area, a set of discrete 1D functions (2D separable functions) with 40 discrete binary amplitude-only segments are considered as possible hologram patterns. Fig. 6-1(a) displays the three desired intensity patterns used in this optimization. The 3D intensity test function is comprised of one
Figure 6-1: A desired 3D intensity pattern designed through brute-force search. (a) A 1, 2, 4 point sinc pattern intensity set at $z=50\text{mm}$, $50.1\text{mm}$ and $50.3\text{mm}$ are used to define the desired 3D intensity. Optimization is in 1D, while experiments use 2D separable holograms. (b) AF and holographic mask after a search over all 40-element binary amplitude patterns. (c) The modeled intensities produced by the hologram at $z=50$ (blue), $50.1$ (green) and $50.3\text{mm}$ (red). (d) Experimentally measured intensities at the same distances as in (c). (e) Raw experimental images.

Point (sinc pattern) at the focal plane turning into two separated points, and then four points at two later planes along the axis of propagation. The desired sinc functions are chosen to be near diffraction-limited width for a setup with a 10mm-wide holographic mask in the aperture plane of a $f=50\text{mm}$ focal length lens. The three depth planes correspond to $50\text{mm}$, $50.1\text{mm}$ and $50.3\text{mm}$ distance from the hologram. For the experiment setup (diagrammed previously in Fig. 3-4(a)), the 10mm-wide holographic pattern is printed as a binary amplitude-only pattern on a transparency at $50\mu\text{m}$ resolution and illuminated with a coherent quasi-monochromatic source ($\lambda=532\text{nm}$). Measurements are taken with a $5\mu\text{m}$-pixel monochromatic CMOS sensor placed at varying distances to the hologram-lens pair.

Fig. 6-1(b) displays the AF and binary mask pattern that yield a minimum MSE be-
Figure 6-2: The same 3 intensity distributions in Fig. 6-1(a) are input into the single-mode selection algorithm following the procedure diagrammed in Fig. 4-7. (a) The optimal AF and binary amplitude holographic mask are shown after 50 iterations of the algorithm. Note the similarity between the optimized AF and binary hologram pattern after this procedure and the brute force search in Fig. 6-1. (b) Experimental intensity patterns at the corresponding depth planes under the exact same parameters used in Fig. 6-1 show improved results.

tween desired and produced intensities (MSE=.093). The modeled and experimental intensities are in Fig 6-1(c)-(e), and display a certain degree of agreement in general shape. The experimental results, however, do not match desired inputs as well as simulated results. In general, the performance of a brute-force search is quite limited both in simulation and experiment.

To contrast against a brute-force approach, the single-mode-selection algorithm is applied to the same desired 1, 2, 4-point 3D intensity pattern and experimental conditions as in Fig. 6-1. The results of this exercise are in Fig. 6-2. Immediate benefits of using a mode-selection approach over a brute-force search include a decrease in computation time (seconds instead of hours on a typical laptop) and a continuous-valued mask, which could also contain phase content. The brute force search computation time is directly proportional to the search space (i.e., hologram resolution), which is limited. Optimizing over 20 discrete pixels requires eight hours of computation time on a typical desktop computer. In contrast, the mode-selection algorithm search time is related only to a lower MSE between desired and produced 3D intensities. Thus, after 50 iterations (requiring several seconds of computation), the second benefit of a lowered MSE is quickly achieved. In this example, mode-selection yields a modeled
Figure 6-3: Three simulated intensities of the hologram in Fig. 4-8(c) placed in front of a 50mm lens at (a) $z = 50\text{mm}$, (b) $z = 50.1\text{mm}$, and (c) $z = 50.2\text{mm}$. (d)-(f) Intensity measurements obtained with the experimental setup described in this section at the same depths as (a)-(c). The scale bar in the lower right represents 50$\mu$m.

MSE of .039, approximately three times lower than the brute force search. Experimental results also offer much better agreement in Fig. 6-2(b), but are also somewhat limited due to printing restrictions requiring an amplitude-only pattern.

As a second example of the single-mode selection algorithm, a slightly different desired 3D intensity distribution of one point at one depth, that turns into four points, and then nine points at two further depth planes is experimentally tested. This is the same desired 3D image of one point turning into two and then three points, here shown as a 2D separable function. It is used in the optimization scheme discussed in Fig. 4-7 with outputs and performance shown in Fig. 4-8. It again appears in the constrained simulations in Fig. 5-1, which provides a nice comparison between model and experiment. This example’s optimized amplitude-only mask, shown in Fig. 4-8(c), is printed as a 10mm-wide binary transparency at 50$\mu$m resolution and placed directly next to a 50mm lens. The same sensor as above captured the real image of the hologram’s diffraction pattern at three different depth planes (50mm, 50.1mm, 50.2mm), with images in Fig. 6-3(d)-(f). Additionally, the performance of the 2D
Figure 6-4: Summary of algorithm performance. A brute force search approach performs poorly in all categories. Multi-mode selection offers enhanced reconstruction, as demonstrated experimentally. Placing amplitude-only or phase-only constraints on hologram design both increases computation time and decreases performance, as expected.

mask is checked with an independent Fresnel propagation model, which produces the intensities shown in Fig. 6-3(a)-(c) at the same corresponding depths. Agreement between models and experiment are not exact, but both follow the expected trend of one point turning to four, and then nine points. Like the brute-force search experiment, differences can be attributed to a source of finite size and the non-ideal transmission of the printed binary transparency. In summary, the single-mode selection algorithm is simply tested with binary printed holograms. These optimized holograms follow desired input trends in 3D, qualitatively matching modeled results. They suffer from the same drawbacks often noted regarding coherent holographic display, connected in this thesis to a rank-1 limitation on the light’s mutual intensity function. Fig. 6-4 summarizes the various properties of brute force, single-mode and multi-mode selection algorithm performance.

6.2 Dynamic Hologram Design Experiments

To experimentally verify the benefit of the multi-mode selection algorithm, a display that can change quickly over time to show an optimized set of $M$ optimized
holograms is required. Furthermore, the resolution of the display will ideally be on the order of 1μm to maximize diffraction efficiency. Dynamic displays of both high speed and resolution are not common. In the experiments below, a transparent liquid crystal-based spatial light modulator (SLM) is used. Both an amplitude-only and phase-only SLM were tested, with performance being roughly equivalent for testing ground-truth holographic patterns. However, as noted in Chapter 5, each of these constrained display devices required a modification to the multi-mode selection algorithm to design desired 3D images. As the phase-only modification required a larger number of modes for enhanced performance, and since the available phase SLM extended only to modulate from 0 to π radians, amplitude-only modulation was chosen for the demonstrations below.

For future setups intended for viewing, a time-multiplexed holographic display must be able to display $M$ unique patterns within the integration time of a viewer's eye. This short window, typically referred to as a viewer's flicker-fusion rate, is on the order of 15-30ms [79]. For example, to meet this requirement assuming $M=4$ patterns, a display with a refresh rate of roughly 240Hz is necessary. Current commercially available LCD displays offer this rate. However, a faster rate is desirable to enhance the benefit of time-multiplexed display (i.e., to increase $M$).

The setup used to perform time-multiplexed experiments is in Fig. 6-5. It includes a Holoeye LC2002 SLM with 800x600 resolution and 32μm pixels, illuminated by a 532nm 10mW laser. The narrow laser is first sent through a microscope objective (10x) that expands the beam to cover the entire SLM screen. As noted above, an $f=150$mm lens is placed shortly after the SLM to increase its diffraction efficiency. At the equivalent Fourier plane (roughly 1.5m away for a short SLM-lens distance), a Lambertian screen is placed perpendicular to the optical axis to show the hologram's real image. A camera (Canon RebelEye 50mm f/1.8 SLR) is used to capture the real image over a finite exposure time, long enough to capture all hologram modes displayed on the SLM.
Figure 6-5: (a) Diagram of the experimental display setup, consisting of a quasi-monochromatic coherent source (532nm 10mW laser), a 10x microscope objective, the Holoeye amplitude SLM and an f=150mm lens. (b) Image of the setup from the side and (c) from behind, where the viewing screen and camera (Canon SLR 50mm f/1.8) that captures real image data are seen at the other end of the optic bench.

The recording screen is moved along the optical axis to bring different planes of the 3D image into view. Alternatively, since a viewer could look into the SLM-laser pair to "see" a virtual image, another possible experimental setup is to place the camera close the the SLM and focus through the screen to a more distant virtual image plane. However, the geometry of the small diffracting screen prevents this setup from being practical. Furthermore, it is difficult to test intensity variation along depth in this manner. A variation in viewing angle could instead be used to estimate performance. The real image geometry is thus used in experimental tests to overcome these virtual image shortcomings.
Figure 6-6: A time-multiplexed holographic display reduces speckle. (a) A single holographic image of the letter "A", generated from a binary desired input intensity. Below is a plot of the pixel-to-pixel intensity fluctuation of the letter across the orange line, showing a normed standard deviation of 0.23. (b) The same "A" produced by a time-multiplexed holographic display quickly showing 8 different holograms, with a much lower standard deviation (0.13) (i.e., reduced speckle). (c) The 1 holographic pattern to create image (a) boxed in red and the 8 patterns for image (b) boxed in blue.

6.2.1 Application to Speckle Reduction

A directly observable effect produced by time-multiplexing a holographic image is its ability to limit the effects of speckle. Speckle is an inherent property of coherent light propagation and has been studied in detail for a variety of applications [78]. For holographic display, it is typically considered a source of unwanted "noise" in resulting 3D imagery [20]. Increasing the partial coherence of display illumination helps reduce speckle noise. It should then be no surprise that simulating partial coherence over time will also reduce speckle noise.
The data in Fig. 6-6 demonstrates how a single holographic image produced by the SLM will include a significant amount of fluctuation in areas of the image that are designed to exhibit a constant intensity response (i.e., the line crossing the “A”). Time-multiplexing with $M=8$ modes reduces the effects of speckle from a single image by a factor of approximately 50 percent. Little buildup across the background is observed as a result of this process. Future time-multiplexing methods could also incorporate color into this mode generation process, delivering the red, green, and blue channels of the image over time along with reducing speckle noise. With a fast enough modulator, these two large shortcomings of current holographic display techniques (i.e., speckle noise and lack of color) could potentially be addressed.

### 6.2.2 Increased Intensity Variation Along Depth

Creation of a higher fidelity 3D image with enhanced depth-variability is offered as a second experiment. Again, as the limitations of many currently available dynamic diffractive screens are pretty severe, demonstration of partially coherent concepts is difficult and error prone. To accommodate the low resolution and diffraction efficiency of the Holoeye SLM, the mode-selection algorithm's amplitude-only mode creation process was slightly modified. Instead of the NMF method introduced in Chapter 5, which experimentally resulted in low-frequency modal patterns centered close to the zeroth order, an altered modal decomposition method was tested. Specifically, an amplitude-only Fourier hologram was first computed for desired intensity inputs (appended with per-pixel random phase), resulting in a hologram with increased spatial frequency. These patterns were then used as input to the mode-selection process, with input angles proportional to each desired image’s distance from the Fourier plane. Multiple modes were calculated for additional Fourier holograms computed with an initial phase set to the residual phase of the previous iteration. Without this modification, the NMF process yielded patterns too close to the 0th order to be detected within the dynamic range of the sensor. Its inclusion, however, possibly pre-
Figure 6-7: Experimental demonstration of the multi-mode selection algorithm creating a desired 3D image. (a) Two inputs of the letter “A” and “B” are used to define the desired intensity at $z_1=-40\text{mm}$ in front and $z_2=80\text{mm}$ behind the Fourier plane. (b) An example set of 4 amplitude-only hologram patterns (4 partially coherent modes) shown on the SLM in setup (c) to create the A-B 3D image. (d) Independent simulation of the performance of one holographic mode creates an approximation to desired results. Experimental images (scale is 2cm wide) show a gradual progression towards a more faithful reproduction of the A-B image with an increased number of modes $M$. Multiple modes were captured by setting the integration time of the camera to equal the number of modes being sequentially displayed on the SLM times their display duration.
vents the output set of patterns from being a globally optimal solution. Future work could benefit from modifying the NMF process to increase each mode's variations, or shifting to a setup that does not require the NMF decomposition.

With this caveat placed on experimental verification of partially coherent mode-selection, two desired intensity patterns of the letter “A” and “B” are used as input to the iterative optimization process. The desired intensities were chosen (somewhat arbitrarily) to be \( z_1 = -40 \text{mm} \) closer to (A) and \( z_2 = 80 \text{mm} \) further from (B) the hologram from the focal plane of the 150mm lens. This corresponds to a 1.28m distance from the hologram to “A” and 1.40m from the hologram to “B”. Displaying one amplitude-only holographic pattern on the roughly 26mm-wide screen yields the two measured intensity patterns for \( M=1 \) mode in Fig. 6-7 at \( z_1 \) and \( z_2 \). Each image shown is cropped to be 20mm wide, although image size is simply set by varying the hologram-lens distance. For comparison, independently simulated results produce the intensity patterns in the first column of Fig. 6-7. Summing up \( M=3 \) modes improves the experimental results, which will present the three equally weighted holographic images within the flicker-fusion rate of a viewer’s eye on current 240Hz display setups. Finally, extending the display to \( M=8 \) equally weighted modes, which would require a display rate around 480Hz, yields relatively clear results. The SLM in this experiment has a rate of 60Hz, so the integration time of the camera was set accordingly to capture the correct number of modes for each image. Furthermore, all images are normalized to a maximum intensity value, which varies with the integration time. In summary, this experiment offers a simple demonstration of how multiple holographic screens displayed over time can improve creation of a desired 3D image, but has the potential to be greatly improved upon with a higher performance SLM.

### 6.2.3 Discussion and Limitations

In the experiments in Fig. 6-6 and Fig. 6-7, generating a depth-varying image with a coherent source leads to a significant amount of speckle and background noise. This background noise is a product of the poor optical transmission characteristics of the
liquid crystal-based transmissive SLM and the production of multiple orders from the amplitude-only screen. Note that additional orders are not shown in Fig. 6-6 or Fig. 6-7's measurements. Unwanted orders could potentially be blocked or removed by switching to a phase-only or combined amplitude and phase SLM setup. As with the results in Fig. 6-6, increasing the number of modes $M$ (i.e., increasing the display's simulated partial coherence) improves the speckle conditions (i.e., reduces the MSE between desired and achieved intensity patterns). Furthermore, it improves erroneous intensities (e.g., note the reduction of the spot within the lower circle of the "B" with added modes). However, additional improvements are possible with an increase in the number of SLM pixels, a decrease in their size, an improvement in their fill-factor and maximum/minimum transmission, and again by gaining some control over the phase of the light. All of these improvements are limited by the current state of available liquid crystal-based diffracting screens. Some of these display traits will improve in the future. For example, pixel count and size are certainly progressing towards values that may make liquid crystal SLM holograms a future diffractive display possibility [58]. If current trends continue, the largest challenge of creating displays with high space-bandwidth product (i.e., many, small pixels) may eventually be realized. However, improved transmission curves and fill factors are two important aspects that must also be improved. Alternative light modulation methods, such as MEMs or AOMs, exhibit very high diffraction efficiency and may offer much better performance than liquid crystal-based modulation for certain setups. As noted in Section 4.5, color holographic images can be directly generated with any of these devices by multiplexing each color channel over time, as in [90], for example. Thus, using the proposed time-sequential holographic display to reduce speckle noise, increase image fidelity and produce color may help holographic video slowly enter mainstream use. As discussed next, gaining control over light in 3D has many additional applications outside of the area of display, providing additional future directions of investigation for the mode-selection process.
Chapter 7

Conclusion

This thesis first explained and mathematically characterized the limitations of the two most popular current 3D display methods, holograms and parallax barriers. It then offered a new model for understanding these two display designs, as well as many other methods of display based in phase space. This phase space model was extended to describe a new “mode-selection” algorithm that designs a holographic display from desired 3D imagery. Finally, this algorithm led to a method of improving coherent holographic display by quickly alternating through an optimal set of diffractive patterns over time, which was demonstrated experimentally. This general overview and extension of state-of-the-art 3D displays has many possible directions for continuation into the future.

7.1 Future Work

7.1.1 Further Connections Between Rays and Waves

To begin, while this thesis shows that coherent and incoherent methods of 3D display are limited to rank-1 functions in a certain space, a full connection of these limitations has yet to be developed. Being able to derive one limitation from the other would lead to a more robust mathematical model of what ray-based light fields and 3D wave fields each can display. This may lead to an optimized method of hybrid
display, which relies both on ray-attenuation and wave-diffraction effects. Or, it may lead to a better understanding of how to utilize the partial coherence state of illumination to provide a better looking image to viewers for both ray and wave-based displays.

The most direct first step towards building upon this thesis is to improve the experimental demonstration of the time-varying holographic display. As noted at the end of the last chapter, an SLM with improved pixel count, pitch and size would yield a more robust setup that could test performance of designing intensities at \( N > 2 \) planes along \( z \). Furthermore, testing performance of a phase-only SLM versus its amplitude-only counterpart could experimentally determine which may create better 3D imagery. Optimal display speeds, image depths, number of modes and other parameters that each display type operates best with should also be tested. More advanced experimental tests could combine amplitude and phase modulation, attempt to measure performance in virtual image generation, or even add in color or extend the display over multiple viewing angles.

7.1.2 An Improved Algorithm

The mode-selection algorithm could also benefit from future study. To begin, the alternative decomposition procedures offered in Chapter 5 work well in practice. However, it is still left to determine what exactly their relationship is with partial coherence. Specifically, the mathematical computation of an SVD, whether discrete or in continuous form, is well connected to the physical definition of a partially coherent field as a sum of orthogonal coherent modes. The proposed NMF and phase-only decomposition steps are iterative and approximate in nature and thus do not have a continuous counterpart. What’s more, they produce modes that are not orthogonal, and thus may not propagate independently. However, they offer better MI approximations when constraints are present. Future work could examine these relationships. Furthermore, the mode-selection process is iterative. An improved algorithm could use a convex approach to cut down computational time. Finally, the partially coherent mode-selection algorithm could be implemented within a camera to design its
PSF. In general, solving for a set of aperture patterns to use on a dynamic screen that changes during an image’s exposure offers an interesting and exciting future direction.

7.1.3 Benefits Beyond 3D Display

Finally, apart from the area of display, many other areas of optics benefit from an improved control of light in 3D. For example, beam-shaping setups are finding increased use in laser design, astronomy and microscopy setups. However, the goal is often to design only a single 2D profile of the beam, or to find a profile that is depth-invariant. Having specific control over a beam’s 3D shape could open up many new investigation areas. Few have connected beam shaping with the beam’s partial coherence state, and even fewer have offered any tests of the concept. For example, with specific 2D beam shapes finding applications in optical trapping and particle manipulation, extending control into the 3rd dimension would provide quite a powerful tool for moving around small objects or constraining them within a volume. Additionally, interesting beam profiles that follow a rotating or vortex-like trajectory could be further optimized over a specific depth range, or improved by adding a specific degree of partial coherence. Furthermore, the area of lithography often considers the use of partially coherent light in improving system resolution. It would be interesting to connect mode-selection and time-multiplexed display with lithography setups, including their joint source-mask optimization methods and simulated annealing algorithms. Doing so might allow lithography to benefit from considering the full 3D profile of its projected light, or even design patterns along the depth dimension in the material it is processing within a single exposure. Finally, the area of biomedical optics has yet to adapt many 3D design methods into measurement or illumination schemes. If the effects of scattering can be accounted for within the phase-space framework, then interesting challenges of delivering light into or measuring light from a 3D area within human tissue might also be addressed.

In general, computational and mathematical models have progressed to the point that
it is direct to design high resolution 3D distributions of light. Furthermore, dynamic and adaptive optical elements are continually improving their ability to produce high-quality diffraction. The convergence of 3D design with dynamic elements appears an inevitable juncture, and will offer improved control of volumetric light distributions over time for many optical applications. Hopefully, this thesis provides a convenient and intuitive framework for these applications to base future work upon, whether in the area of 3D display or otherwise.
Bibliography


