Optics in Neural Computation

Thesis by

Michael Levene

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1998
(Submitted April 10, 1998)

Acknowledgements

I would like to thank Professor Demetri Psaltis for his guidance and patience during my stay here at Caltech. He has been particularly understanding and supportive of my social life, which compelled me to make many trips to Italy. He even went as far as to help arrange a position for me in Milan. For this, I will always be thankful.

I would also like to thank Professors Scott Fraser and Kelvin Wagner. Both Scott and Kelvin provided me with advice and guidance during important points in my research career. Their warmth and openness to helping me, even though I was not their student, was always deeply appreciated.

I have heard it said often that, in graduate school, one learns more from one's peers than from the faculty. In particular, older students play the role of mentors to younger ones. Certainly, I benefited from the guidance amd friendship of older students, especially from Sidney Li, Yong Qiao, Kevin Curtis, Geoff Burr, Jean-Jacques Drolet, Robert Denkewalter, and Rahul Sarpeshkar.

I would also like to thank Xin An, for showing me that it is possible to work for several days without actually sleeping, Allen Pu, for organizing the paintball games (along with Robert Denkewalter), and Ernest Chaung. Thanks also go to Greg Steckman for his collaboration on the Fresnel correlator project.

I would like to thank other members of the Psaltis lab for their friendshp and for creating an enjoyable working environment: David Marks, Jiafu Luo, Chaun Xie, Annette Grot, Greg Billock, Christof Moser, Ali Adibi, Shaw Wang, George Panotopoulos, Jose Mumbru, Demetris Sakellariou, and Yungping Yang.

Nothing in the lab would work without the consistent support of Yayun Liu and Lucinda Acosta. In addition to their professional prowess, both women are extremely warm and kind (although I will be sending Lucinda my dental bill thanks to her birthday cakes and candies!).

I cannot say enough about the wonderful family that I have formed with my

housemates at Mountain Street: Arthur Street, Aaron Batista, Christina Hood, and Erik Winfree. They have provided me with friendship, support, a home, great times, and some really great parties.

I have also relied, heavily at times, on the support of my friends who are scattered across the country, including B.C. Debney, Matt Lang, Melissa Bondi, Stacy Galbreath and Karen Huy.

I could never have come close to arriving at this point in my life without the undying love and support from my parents, my step-mother and my sister and brother.

Finally, I would like to thank Marilena Ricci, for enriching my life with her passionate love and for showing me another part of the world.

Abstract

In all attempts to emulate the considerable powers of the brain, one is struck by both its immense size, parallelism, and complexity. While the fields of neural networks, artificial intelligence, and neuromorphic engineering have all attempted oversimplifications on the considerable complexity, all three can benefit from the inherent scalability and parallelism of optics. This thesis looks at specific aspects of three modes in which optics, and particularly volume holography, can play a part in neural computation.

First, holography serves as the basis of highly-parallel correlators, which are the foundation of optical neural networks. The huge input capability of optical neural networks make them most useful for image processing and image recognition and tracking. These tasks benefit from the shift-invariance of optical correlators. In this thesis, I analyze the capacity of correlators, and then present several techniques for controling the amount of shift invariance. Of particular interest is the Fresnel correlator, in which the hologram is displaced from the Fourier plane. In this case, the amount of shift invariance is limited not just by the thickness of the hologram, but by the distance of the hologram from the Fourier plane.

Second, volume holography can provide the huge storage capacity and high speed, parallel read-out necessary to support large artificial intelligence systems. However, previous methods for storing data in volume holograms have relied on awkward beam-steering or on as-yet non-existent cheap, wide-bandwidth, tunable laser sources. This thesis presents a new technique, shift multiplexing, which is capable of very high densities, but which has the advantage of a very simple implementation. In shift multiplexing, the reference wave consists of a focused spot a few millimeters in front of the hologram. Multiplexing is achieved by simply translating the hologram a few tens of microns or less. This thesis describes the theory for how shift multiplexing works based on an unconventional, but very intuitive, analysis of the optical far-field.

A more detailed analysis based on a path-integral interpretation of the Born approximation is also derived. The capacity of shift multiplexing is compared with that of angle and wavelength multiplexing.

The last part of this thesis deals with the role of optics in neuromorphic engineering. Up until now, most neuromorphic engineering has involved one or a few VLSI circuits emulating early sensory systems. However, optical interconnects will be required in order to push towards more ambitious goals, such as the simulation of early visual cortex. I describe a preliminary approach to designing such a system, and show how shift multiplexing can be used to simultaneously store and implement the immense interconnections required by such a project.

Advisor: Professor Demetri Psaltis

Contents

List of Figures

Chapter 1 Introduction

Contents

The rapid growth in computer and information technology has been one of the most significant aspects of the latter half of this century. Concurrent with that growth, and largely although not entirely independent of it, there has been a stunning expansion in our understanding of the mammalian brain. The recognition that biology has solved difficult tasks such as pattern recognition, scene interpretation, and navigation has led to several attempts to emulate the structure of biological systems in both hardware and software. Until recently these efforts met with only limited success, and so the fields of information processing and neurobiology evolved with only coincidental influence upon each other.

Initial efforts in the 40s, 50s and 60s focused on the observation that brains consist of multitudes of apparently simple processors with many connections between them [1–4]. The bulk of the information processing was thought to be done through these connections. This effort was typified by the Perceptron, what we would now call a one-layer neural network [2]. This approach was mostly abandoned after the publication of *Perceptrons* by Minsky and Papert [5], and was followed by attempts at a more cognitive level.

The rise of artificial intelligence [AI] in the 70s was focused on the decision-making process itself, and consisted of large databases of decision trees and concept maps. If any property of biology could be said to have played a role in this effort, it would be the tremendous size of human brains; in other words, human brains are really big, so maybe if we pack enough information into a box, it will become intelligent, too. Like all previous and successive attempts, AI research started with over-zealous optimism only to be followed by more modest realizations.

The 80's saw the return of the earlier efforts based on connections between many simple neurons. This was typified by the Hopfield network [6] and the back-propagation algorithm [7,8]. With the aid of better algorithms and bigger, faster computers, this generation was able to achieve some moderate successes [9,10]. While there were some hardware implementations of these algorithms, both in optics and in VLSI, most work was done in software.

A new focus began in the 90s with the advent of neuromorphic engineering developed by Carver Mead and others [11]. The idea was to emulate the details of neurophysiology, particularly in early sensory receptors such as the retina and cochlea. By taking inspiration from biology at this detailed level, analog VLSI sensors running below threshold achieved large, adaptive dynamic ranges while consuming very little power. Retinas have been developed that are capable of various types of processing, including motion, contrast sensitivity, and automatic gain control [11–16] and cochleas with dynamic ranges in excess of 60dB running on 0.5mW have been built [17]. This approach to emulating the brain has been almost entirely hardware-oriented, with a focus on using the inherent physics of devices to perform computation. It is the approach that has and will continue to benefit most from our expanding knowledge of biology, and is also capable of helping us to further that understanding.

Neural networks, AI, and neuromorphic engineering represent three different approaches by the engineering community to emulate the functionality of brains. This thesis will describe some aspects of how optics can and has contributed to these three approaches. Up to now, optics has played a role in hardware implementations of neural networks, but has been relatively absent from the other two paradigms. It's lack of involvement in AI research stems from the fact that the only demand this approach has on hardware is computational speed and memory. Optics is only now beginning to make big differences in these areas. Optical interconnections will help improve parallelism, and volume holographic memories are on the verge of commercial viability.

In this thesis I will address the use of volume holograms for data storage and for neural networks. In addition to introducing a new technique for storing data in holograms called shift multiplexing, I will expand upon the close relationship between holographic memories and neural networks. That such a close relationship should exist is not surprising; one of the key lessons of neural network research has been that memory and computation are inseparable in neural architectures. I will focus attention on the trade-off between the number of output units in a given layer and the amount of shift invariance that layer has, and I will demonstrate how various architectures allow one to address this issue. In the last part of the thesis I will address how optics can be used to extend the scope of neuromorphic engineering to include large systems simulating not only early sensory mechanisms but some cortical functions as well.

Chapters 2 and 3 introduce the basic concepts needed to understand volume holography in the setting of angle multiplexing. I will take a lot of time with these chapters, even though most of the work is not my own, for two reasons. One, the Computation and Neural Systems Program includes people from many fields, and so it seems appropriate to include introductory material for people not familiar with optics. Two, Chapter 6, which deals with Shift Multiplexing, requires knowledge of the fundamentals of scalar diffraction theory. It would be impossible to understand some of the more subtle arguments, and their significance, without an understanding of the relationship betweeen spherical waves and plane waves in scalar diffraction theory. Subsequent chapters follow a similar analysis of capacity, parallelism, and shift invariance for alternative multiplexing techniques. Chapter 6 addresses shift multiplexing, a new technique which I invented as part of my thesis. Chapter 7 begins the discussion of the role of optics in neuromorphic engineering.

Chapter 2 Introduction to Holography

Contents

In this chapter I will review the basic ideas behind scalar diffraction theory and volume holography, using angle multiplexing as an example. A discussion of imaging systems and correlators will also be included, with more details to be given in Chapter 3.

2.1 Scalar Diffraction Theory

The fundamental idea behind scalar diffraction theory is the Huygens-Fresnel principle: a given wavefront of light can be thought of as the sum of an infinite number of point sources all along the wavefront, each emitting its own spherical wave (see figure 2.1). We can understand how the wave propagates through space by considering how these individual wavefronts propagate and interfere to form a secondary wavefront [18].

This very intuitive picture is reminiscent of a Green's function, in that the spherical wave is convolved with the complex wavefront. From this intuition one can go on to derive both the Fresnel-Kirchhoff and Rayleigh-Sommerfield scalar diffraction formulas. The two formulas differ only in the form of the obliquity factor, which we will soon choose to ignore. I will skip the derivations here, but recommend that the interested reader consult Goodman's *Introduction to Fourier Optics* [18]. Indeed, most of the discussion here of scalar diffraction will follow the treatment presented

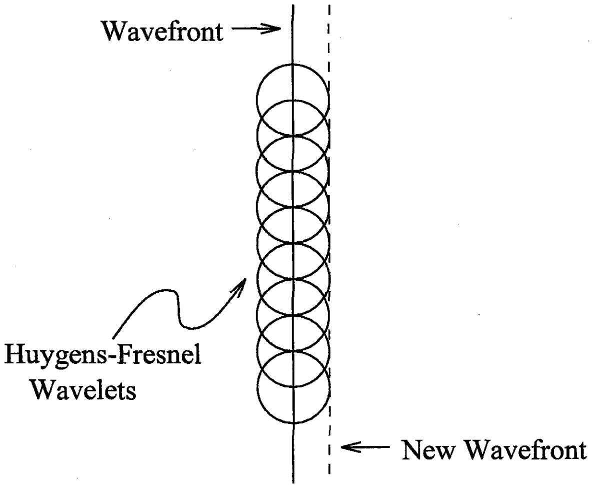

Figure 2.1: Huygens-Fresnel Wavelets.

by Goodman. In keeping with the intuition of the Huygens-Fresnel principle, one can express the Rayleigh-Sommerfield formula as a superposition integral

$$g(x',y',z') = \iint h(x,y,x',y',z')f(x,y)dxdy \tag{2.1}$$

with

$$h(x,y,x',y',z') = \frac{1}{i\lambda}\frac{e^{ik\mathbf{r}}}{\mathbf{r}}\cos(\mathbf{n},\mathbf{r}) \tag{2.2}$$

in which $f(x,y)$ is the complex field distribution across the x,y-plane, $g(x',y',z')$ is the field at the x',y'-plane at some distance z' from $f(x,y)$, $k = 2\pi/\lambda$, $\mathbf{n}$ is the vector normal to the x,y-plane, and $\mathbf{r}$ is the vector joining the points (x,y) and (x',y') (see figure 2.2). I have assumed here and throughout the thesis that the light is monochromatic with wavelength λ. Notice that the convolution kernel $h(x,y,x',y',z')$ is simply the expression for a spherical wave centered at the point (x,y) with an added $\cos(\mathbf{n},\mathbf{r})$ factor. This cosine term is the obliquity factor that comes from the choice of Green's function in the Rayleigh-Sommerfield derivation. It can be thought of as indicating the drop in amplitude for points farther off-axis. If we assume that the

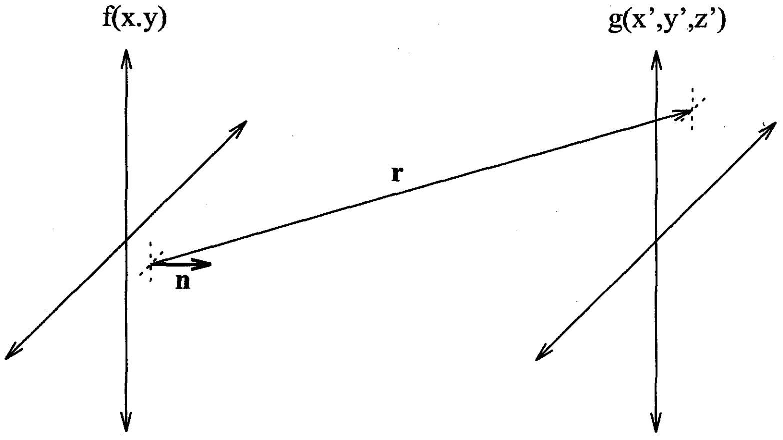

Figure 2.2: Rayleigh-Sommerfield diffraction.

distance z' is large compared to the apertures at the (x, y) and (x', y') planes, then we can drop the obliquity factor altogether and use the paraxial approximation for the vector **r**. These assumptions are known as the Fresnel approximations, and they yield the form:

$$g(x', y', z') = \frac{e^{ikz}}{i\lambda z} \iint f(x, y) e^{i\frac{k}{2z'}\left((x'-x)^2+(y'-y)^2\right)} dx dy \tag{2.3}$$

While the Huygens-Fresnel principle develops an intuition based on spherical waves, an alternative point of view is to think of the wavefront at $f(x, y)$ as a sum of infinite plane waves. The validity of this viewpoint is readily seen by taking the Fourier transform of the field $f(x, y)$

$$\tilde{F}(k_x, k_y) = \iint f(x, y) e^{-i(k_x x + k_y y)} dx dy \tag{2.4}$$

where

$$|\mathbf{k}|^2 = k_x^2 + k_y^2 + k_z^2 \tag{2.5}$$

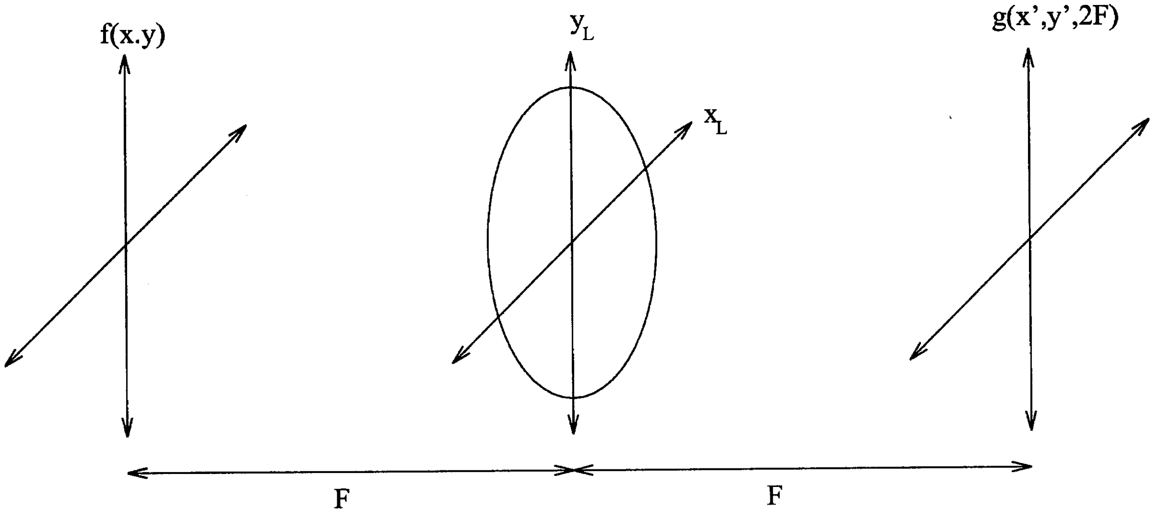

Figure 2.3: Propagation to the Fourier plane.

The inverse Fourier transform is then

$$f(x,y) = \iint \tilde{F}(k_x,k_y)e^{i(k_x x+k_y y)}dk_x dk_y \tag{2.6}$$

But the expression $e^{i(k_x x+k_y y)}$ describes a plane wave travelling in the **k** direction. The quantity $\tilde{F}(k_x,k_y)$ is sometimes referred to as the angular spectrum.

This approach has the advantage that plane waves are the eigenmodes of free-space propagation; they do not change as they propagate through space except for a simple multiplicative phase factor. This phase factor can be derived directly through geometry, keeping track of the individual path lengths each angular component takes as the wavefront propagates along the direction z. Or, since we have already derived the convolution kernel from the Huygens-Fresnel principle, we can simply take the Fourier transform of this kernel to arrive at the eigenvalue for free-space propagation

$$\tilde{H}(k_x,k_y) = e^{ikz}e^{-i\frac{\lambda z}{4\pi}(k_x^2+k_y^2)} \tag{2.7}$$

The intuition that a wavefront can be thought of as either a sum of infinite plane waves or as a sum of spherical waves will be very important throughout this thesis. One other concept that will be critical is the transformation performed on a wavefront by a simple lens with focal length F. Although this derivation is very straightforward, I will go through it here because it will be needed later and because the derivation itself gives some intuition for the effects of the lens [18]. Given a wavefront $f(x, y)$ in the plane one focal length in front of the lens, we will solve for the wavefront $g(x', y', 2F)$ in the plane one focal length behind the lens (see figure 2.3).

If we propagate the wavefront from $f(x, y)$ to a plane just in front of the lens, we have

$$E^-(x_L, y_L) = \frac{e^{ikF}}{i\lambda F} \iint f(x,y) e^{i\frac{\pi}{\lambda F}\left((x_L-x)^2+(y_L-y)^2\right)} dx dy \tag{2.8}$$

The lens will introduce a quadratic phase factor such that the field just after the lens will be

$$\begin{aligned} E^+(x_L, y_L) &= \frac{e^{ikF}}{i\lambda F} \iint f(x,y) e^{i\frac{\pi}{\lambda F}\left((x_L-x)^2+(y_L-y)^2\right)} e^{-i\frac{\pi}{\lambda F}(x_L^2+y_L^2)} dx dy & (2.9) \\ &= \frac{e^{ikF}}{i\lambda F} \iint f(x,y) e^{i\frac{\pi}{\lambda F}(x^2+y^2)} e^{-i\frac{2\pi}{\lambda F}(xx_L+yy_L)} dx dy & (2.10) \end{aligned}$$

If we then propagate another distance F, we get an expression for $g(x', y')$

$$\begin{aligned} g(x', y') &= \frac{e^{ikF}}{i\lambda F} \iint \left(\frac{e^{ikF}}{i\lambda F} \iint f(x,y) e^{i\frac{\pi}{\lambda F}(x^2+y^2)} e^{-i\frac{2\pi}{\lambda F}(xx_L+yy_L)} dx dy \right) & (2.11) \\ &\quad \cdot e^{i\frac{\pi}{\lambda F}\left((x'-x_L)^2+(y'-y_L)^2\right)} dx_L dy_L & (2.12) \end{aligned}$$

Rearranging the order of integration yields

$$g(x',y') = -\frac{e^{i2kF}}{(\lambda F)^2} \iint f(x,y) e^{i\frac{\pi}{\lambda F}(x^2+y^2+(x')^2+(y')^2)} \tag{2.13}$$

$$\cdot \iint e^{-i\frac{\pi}{\lambda F}(x_L^2+y_L^2)} e^{-i\frac{\pi}{\lambda F}((x+x')x_L+(y+y')y_L)} dx_L dy_L dx dy \tag{2.14}$$

The integral over x_L, y_L is in the form of a Fourier transform such that

$$e^{i\frac{\pi}{\lambda F}(x_L^2+y_L^2)} \Rightarrow i\lambda F e^{-i\pi\lambda F(u^2+v^2)} \tag{2.15}$$

where

$$u = \frac{x+x'}{\lambda F} \quad , \quad v = \frac{y+y'}{\lambda F}$$

Substituting this expression for the integral in equation 2.14 gives the final result

$$g(x',y') = \frac{e^{i2kF}}{i\lambda F} \iint f(x,y) e^{-i\frac{2\pi}{\lambda F}(xx'+yy')} dx dy \tag{2.16}$$

But this is just the Fourier transform of $f(x,y)$ (with a complex scaling factor in front of the integral that is not important here)! This is a very important result that can be easily understood in terms of an intuitive picture. We know from high school physics that parallel rays of light will come to a focus one focal length behind a lens and, conversely, light from a point one focal length in front of a lens will produce rays of parallel light after passing through the lens. In other words, spherical wavefronts are converted (approximately) into plane waves, and vice-versa. The angular spectrum of $f(x,y)$, which is just the Fourier transform of $f(x,y)$, produces a set of plane waves that are converted by the lens into converging spherical waves. These spherical waves come to a focus one focal distance behind the lens, in the plane referred to as the

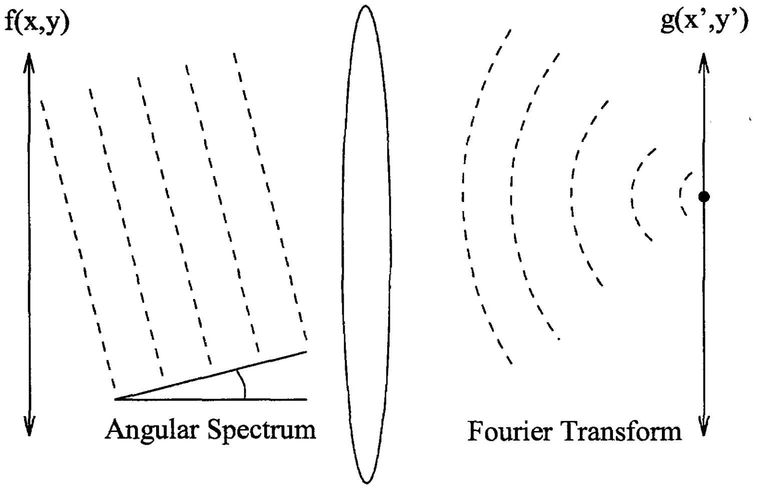

Figure 2.4: Angular spectrum transformed to Fourier plane

Fourier plane. The position of the spherical wave in the Fourier plane is proportional to the angle of the corresponding plane wave before the lens, *i.e.* it is proportional to the spatial frequency of $f(x, y)$ (see figure 2.4).

Although I will not continue with the derivation from here, it should be clear that we can complete an imaging system by placing another lens of focal length F' a distance F' behind the Fourier plane of the first lens. This will, of course, produce the Fourier transform of the Fourier plane a distance F' behind the second lens, which is just the inverted image of $f(x, y)$, since

$$\mathcal{F}\{\mathcal{F}\{f(x)\}\} = f(-x) \tag{2.17}$$

In this case, the image of $f(x, y)$ will be magnified by the factor $-F'/F$, with the negative sign indicating inversion.

This system can be used for more than just imaging, however. We can place transparencies in the Fourier plane of the imaging system to perform image processing operations [18]. If the transparency has a complex transmission function given by $\tilde{T}(u, v)$, this function will multiply the Fourier transform of $f(x, y)$, giving $\tilde{T}(u, v)\tilde{F}(u, v)$. The field at the image plane, $h(x'', y'')$, will then be the Fourier transform of this, or

$$h(x'', y'') = \mathcal{F}\{\tilde{T}(u,v)\tilde{F}(u,v)\} \tag{2.18}$$

$$= \iint f^{*}(-x,-y)t(x''-x, y''-y)dxdy \tag{2.19}$$

where

$$t(x,y) = \mathcal{F}\{\tilde{T}(u,v)\} \tag{2.20}$$

We have just performed a convolution! Thus, this simple optical system is capable of performing two-dimensional linear filtering almost instantaneously.

2.2 Holography

A typical holographic setup is very similar to the imaging system just described. Instead of a transparency, there is a holographic medium in the Fourier plane and there is also a plane wave reference beam that is incident upon the holographic material at an angle of θ (see figure 2.5).

If the reference wave in the Fourier plane is give by

$$R = e^{i\mathbf{k}\cdot\mathbf{r}'} = e^{i(k_x x' + k_y y')} \tag{2.21}$$

then the total field will be

$$E(x', y') = \tilde{F}\left(\frac{x'}{\lambda F}, \frac{y'}{\lambda F}\right) + R(x', y') \tag{2.22}$$

The holographic medium will record the optical power, given by

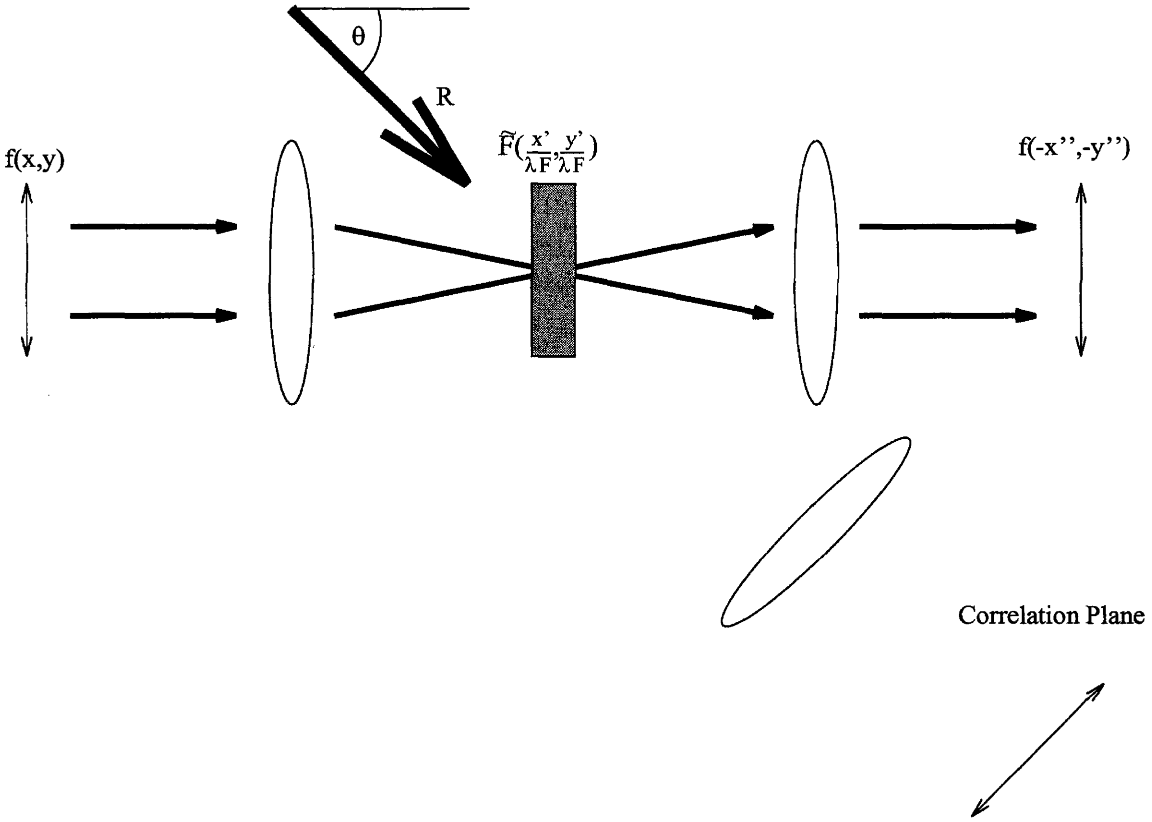

Figure 2.5: Holographic setup.

$$I(x',y') = \left|\tilde{F}\left(\frac{x'}{\lambda F},\frac{y'}{\lambda F}\right) + R(x',y')\right|^2 \tag{2.23}$$

$$= \tilde{F}^*R + R^*\tilde{F} + RR^* + \tilde{F}\tilde{F}^* \tag{2.24}$$

The first two terms represent the grating, while the last two terms are DC modulations of the field, which we will ignore from here on. If we block the signal beam, $\tilde{F}$, and re-illuminate the recorded hologram with the reference beam, we will have

$$R\tilde{F}^*R + RR^*\tilde{F} \tag{2.25}$$

The second term contains both R and R^*, which cancel to leave us with just $\tilde{F}$. Thus, we have reconstructed the signal wavefront $\tilde{F}(\frac{x'}{\lambda F},\frac{y'}{\lambda F})$. The first term reconstructs the complex conjugate signal, which then travels in the opposite direction, so we will not see this reconstruction at the output of our imaging system. In this way, we have stored the (almost) complete information about the signal $f(x,y)$. If we had not used this imaging system, but had placed the holographic medium some distance in front of object O, then we would have recorded the information about the entire wavefront coming from object O. Reconstruction with the reference wave would then reproduce this wavefront, with all the three-dimensional information about the object that was present in the original wavefront. This is the idea behind 3-D holography.

Returning to the system in figure 2.5, what would happen if instead of removing the signal and re-illuminating with the reference, we had removed the reference and re-illuminated with the signal? In this case we would have

$$\tilde{F}\tilde{F}^*R + \tilde{F}\tilde{F}R^* \tag{2.26}$$

The first term is a field given by $\tilde{F}\tilde{F}^*$ superimposed on the original reference carrier wave. If we place a lens along this direction, one focal distance away, we will take the Fourier transform of $\tilde{F}\tilde{F}^*$, which we know from the convolution theorem is the two-dimensional auto-correlation of $f(x, y)$! If we had reconstructed with a different signal, $g(x, y)$, then at the output plane we would have the two-dimensional cross-correlation between the new signal $g(x, y)$ and the stored signal $f(x, y)$! This is the foundation of holographic optical neural networks; the peak value of the cross-correlation is the inner-product between the new signal and the stored signal. A detector in the correlation plane will then respond to the power of the peak. Thus we have a one-neuron neural network in which the weights of the neuron are stored as the first signal, $f(x, y)$, and the inner product between the input, $g(x, y)$, and $f(x, y)$ is followed by a squaring nonlinearity. We can then electronically add additional nonlinearities, such as a threshold.

What about the second term in equation 2.26? This term has $\tilde{F}\tilde{F}$ riding on a carrier that is the complex conjugate of the reference wave. An appropriately placed lens would then produce the auto-convolution of $f(x, y)$. For thin holograms, this output could prove useful; however, we will see that for volume holograms this output will vanish.

A central theme of this thesis is the relationship between holographic memories and holographic correlators. I will discuss several methods for multiplexing holograms in a volume medium, and for each method there will be geometries for holographic data storage and for holographic correlators. It is not surprising that the same system can be used as either a memory or as a recognition system; the processes of recognizing someone and recalling their name are also closely related in the brain. It has been suggested that the two processes are not the same for the brain if for no other reason than the fact that we can recognize far more things than we can remember. Similarly, the capacity for data storage and the capacity for pattern recognition is not the same for holographic systems, as will be shown in Chapter 5.

2.3 Volume Holography

In practice, we would like to be able to store more than one hologram in a given medium. Multiple holograms could be used either as a memory [19–22], with each individual hologram representing an entire page of data, or as multiple correlators for use in a neural network architecture with many neurons per layer [23, 24]. Each hologram contains two dimensions worth of information (the full wavefront description within a given 2-D plane); therefore, in order to store multiple holograms in the same medium, we must add another dimension, *i.e.*, we must use volume holograms.

Using the Born approximation, the volume hologram is represented as a sum of individual thin holograms stacked together to form a volume [25]. We assume that the diffraction efficiency of each individual thin component is weak enough that we need only take into account primary diffraction from the illuminating wave; rediffraction of this light by successive holograms is assumed to be so weak as to be negligible. For the remainder of this discussion we will limit our conversation to plane wave interactions, since we can trivially relate these plane waves in the Fourier plane to point sources in the object and image planes. Figure 2.6a shows the case for Bragg-matched reconstruction of a hologram that was formed with two plane waves. The vertical bars represent thin "layers" of the hologram, each of which produces a diffracted plane wave. When the incident light is Bragg-matched, the diffracted waves from each layer are all in phase, adding constructively to form the net diffracted output from the volume hologram.

Figure 2.6b shows what happens when the angle of the incident beam is changed slightly, producing Bragg-mismatch. In this case, the diffracted plane waves from each "layer" no longer add up in phase. Instead, they begin to destructively interfere. If the medium is thick enough or the angle large enough, the last half of the medium will produce diffracted waves that are exactly π out of phase with the first half, and so the destructive interference between these beams will lead to no net diffracted power. This is called Bragg-mismatch. Since any wavefront can be decomposed into a sum of plane waves, this same principle can be applied to image-bearing signal beams. The

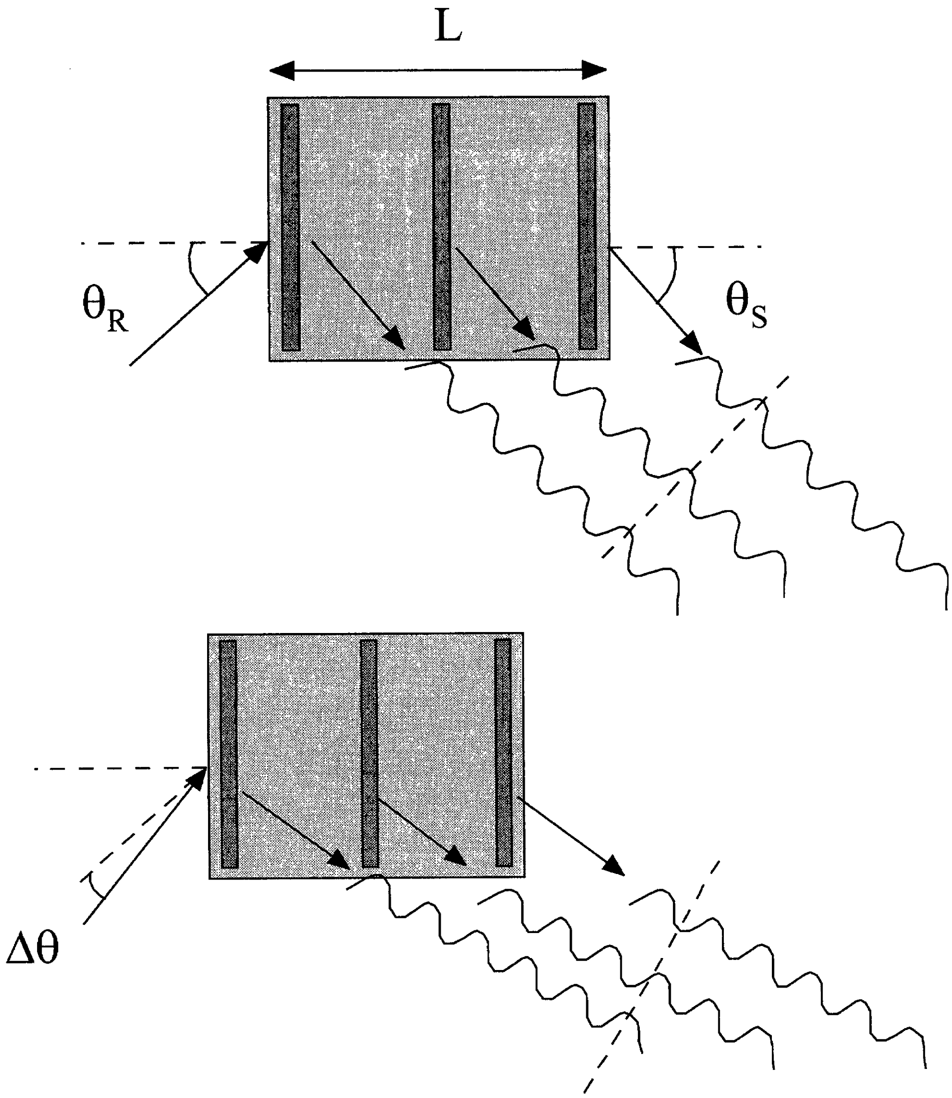

Figure 2.6: Bragg mismatch with change in angle.

only difference is that perfect Bragg-mismatch can only be achieved for one plane wave component of the signal beam at a time. We can store multiple holograms by changing the angle of the reference beam to Bragg mismatch previous holograms, and then storing a new hologram at this angle. This is called angle-multiplexing, and it is the primary way to multiplex holograms both for holographic memories and neural networks. In some cases, the angle change required to reach the first Bragg null, $\Delta\theta$, varies as $\Delta\theta = \lambda/L$. For $\lambda = 500nm$ and $L = 1cm$, $\Delta\theta = 5 \times 10^{-5}$radians, or about 3×10^{-3} degrees. So we can store a lot of holograms this way! As many as 10,000 holograms have been stored in a single crystal [26,27].

A volume hologram maps a set of input plane waves $H_i(\mathbf{k_i})$ to a set of diffracted plane waves $H_d(\mathbf{k_d})$ with a transfer function $A(\mathbf{k_i}, \mathbf{k_d})$ such that [28]

$$H_d(\mathbf{k_d}) = \iint H_i(\mathbf{k_i}) A(\mathbf{k_i}, \mathbf{k_d}) d\mathbf{k_i} \tag{2.27}$$

where

$$A(\mathbf{k_i}, \mathbf{k_d}) = \iiint \frac{\Delta\epsilon(\mathbf{r}') e^{i(\mathbf{k_i}-\mathbf{k_d})\cdot\mathbf{r}'}}{i2k_{dz}} d\mathbf{r}' \tag{2.28}$$

Here, $\Delta\epsilon(\mathbf{r}')$ is the perturbation to the dielectric susceptibility within the holographic material (*i.e.*, the hologram). The derivation of this result is too long to reproduce here. Intuitively, however, we can see that equation 2.28 is just the three-dimensional Fourier transform of the hologram that is stored in the medium. If the medium were infinite in all directions, the single grating would have as its Fourier transform a delta function in three-dimensional space; only one angle of reference beam would produce diffraction, and Bragg mismatch would happen immediately, as soon as the angle of the beam was changed by even the smallest amount. The finite size of the medium causes the Fourier transform of the stored grating to spread a bit in all directions, so that small deviations in the angle of the reference beam will still be able to diffract from the grating.

When calculating the selectivity for real systems, two out of three of the dimensions of the holographic medium are usually taken to be infinite. These two dimensions correspond to the transverse extent of the reference and signal beams on the surface of the medium. Only the thickness of the medium is taken to be finite. This is almost always a very good approximation because the phases of the diffracted signal must build up along the path of the reconstruction, which is usually along the thickness of the medium. With this approximation, the transverse components of the grating vector are delta functions, $\delta(x)\delta(y)$, but the longitudinal component is the Fourier transform of the $\mathrm{rect}(z/L)$ function that describes the thickness of the crystal, L; this is, of course, just a sinc function, given by $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. For a signal beam incident at an angle θ_s and a reference beam at angle θ_r, both with respect to the z-axis, the amplitude of the diffracted signal wave as a function of the deviation of the reconstruction beam, $\Delta\theta$, is [28]

$$A_d(\Delta\theta) = \mathrm{sinc}\left(\frac{nL\sin(\theta_s - \theta_r)}{2\lambda\cos\theta_s}\Delta\theta\right) \tag{2.29}$$

This is best seen with a k-sphere diagram, which is a momentum, or Fourier, space representation of the grating and the incident and diffracted waves. A k-sphere diagram for the case of a Bragg-matched hologram that is infinite in extent is shown in figure 2.7. The vector k_r represents the k-vector of the reference wave. The grating vector is then drawn from the end of the reference vector in the usual fashion of vector diagrams. The grating vector is a double vector, with arrowheads in both directions; each arrowhead represents one term from equation 2.24. The circle represents a cross section of a sphere whose radius is $k = 2\pi/\lambda$; these are the only allowed propogating modes since we have only monochromatic light of wavelength λ and there are no nonlinear processes to create light of another wavelength. For the case of Bragg-matched reconstruction, the tip of the grating vector falls on the surface of this sphere, and so we draw the vector k_s from the center of the sphere to the tip

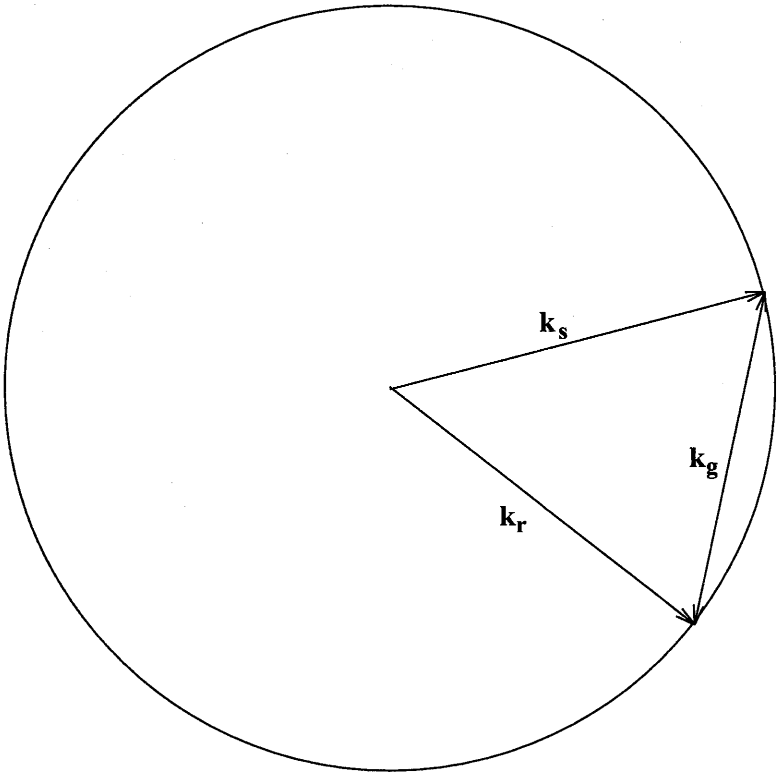

Figure 2.7: K-sphere diagram of Bragg-matched reconstruction.

of the grating, representing the k-vector of the diffracted light. A small change in the angle of incidence of the reference beam would shift the tip of the grating vector off of the surface of the sphere; thus, it would not point to a propagating mode, and there would be no diffraction.

Figure 2.8 shows the k-sphere diagram for a Bragg-mismatched grating that is finite in the z direction. The small sinc^2 curve drawn at the tip of the grating vector represents this uncertainty, or spreading, of the grating vector along the direction of the thickness of the medium. Notice that, for the off-Bragg-matched case, the side lobe of the sinc^2 curve still intersects the k-sphere, and therefore there will be a propagating diffracted wave in this direction, but with a power given by the magnitude of the sinc^2. As the reconstructing reference beam changes its angle of incidence, the intersection between the k-sphere and the tip of the grating vector distribution will move out along the sinc^2 function. Since the sinc^2 distribution is along the z-axis only, a change in the reconstructing reference angle, $\Delta\theta_r$, does not lead to an exactly

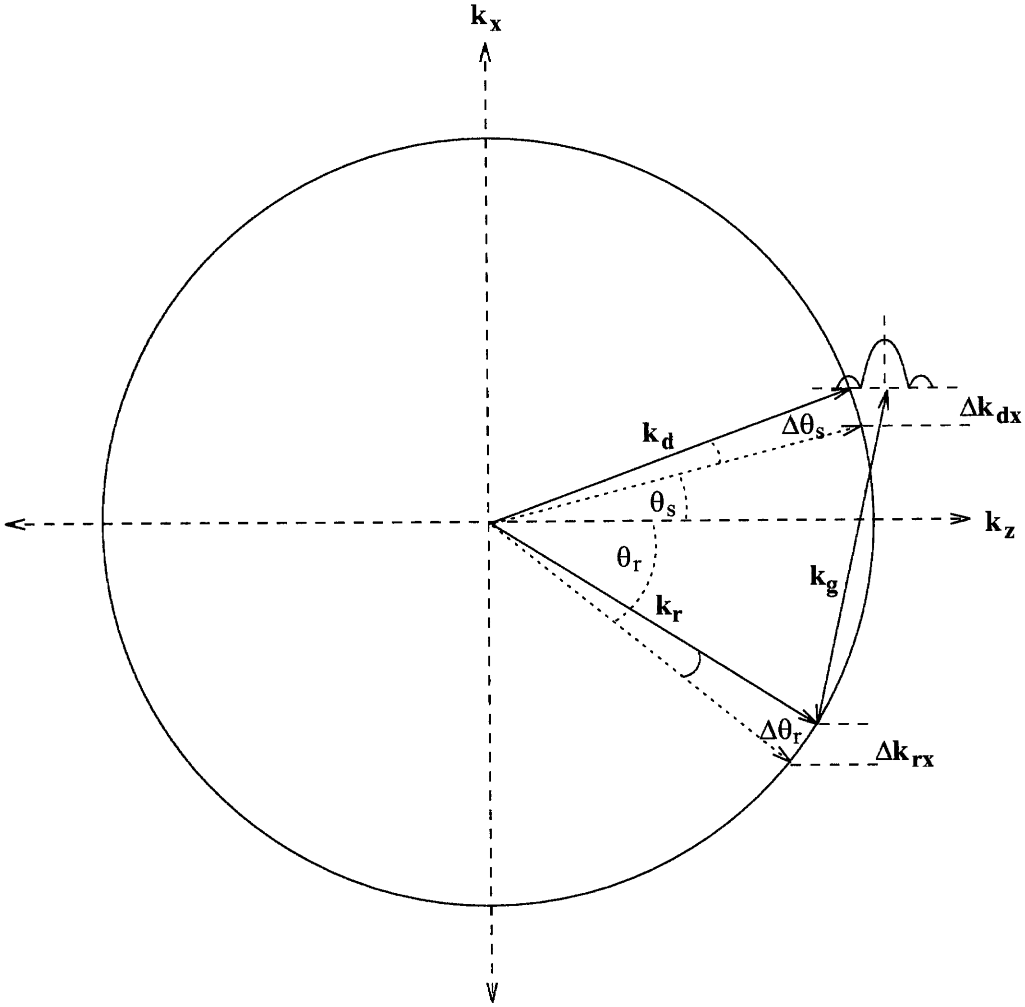

Figure 2.8: K-sphere diagram of Bragg-mismatched reconstruction.

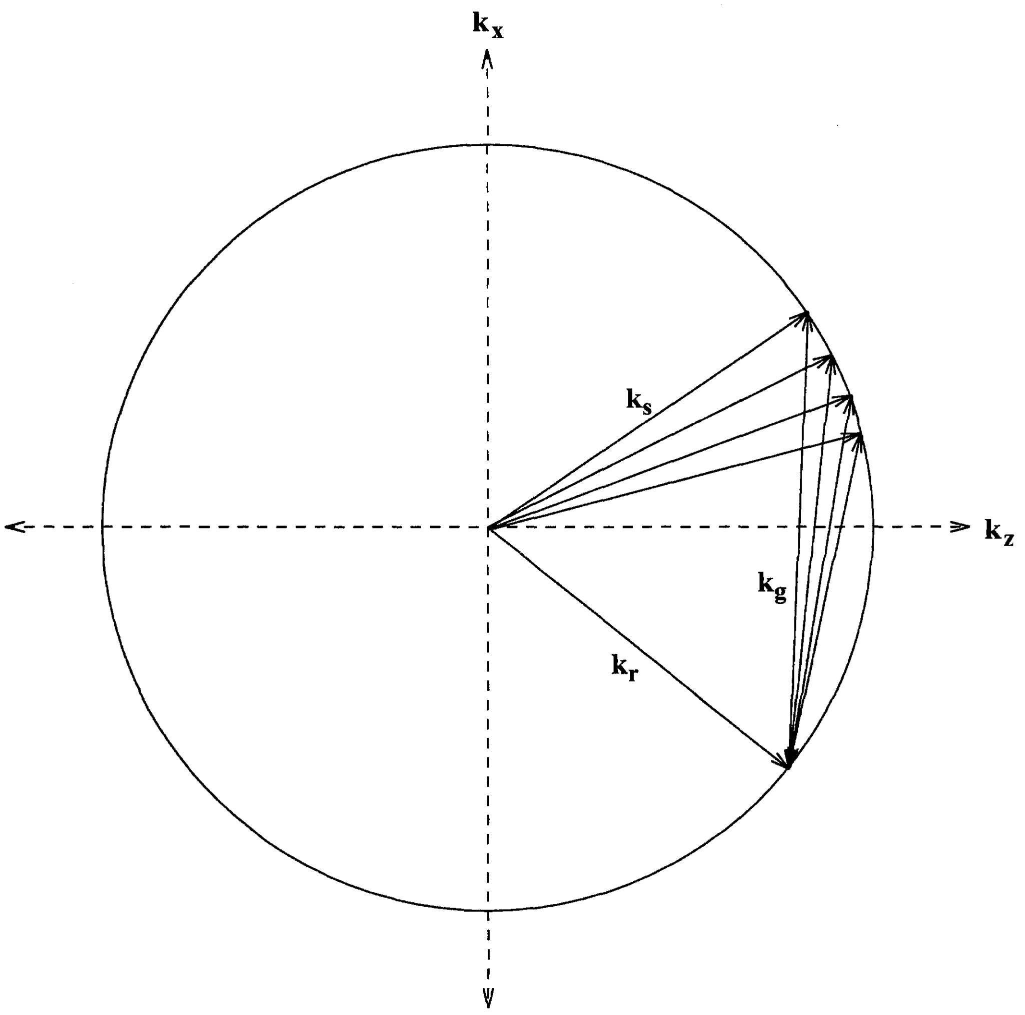

Figure 2.9: K-sphere diagram of Bragg-matched reconstruction with a finite bandwidth signal beam.

equal change in the diffracted signal beam angle, $\Delta\theta_s$. Rather, only the transverse components of the angle changes, $\Delta\theta_{rx}$ and $\Delta\theta_{sx}$, will remain equal. For a given recording geometry, the change in the diffracted angle as a function of the change in the reconstructing angle is given by

$$\Delta\theta_s = \Delta\theta_r \frac{\cos\theta_r}{\cos\theta_s} \tag{2.30}$$

The k-sphere can give us a nice intuitive picture of both holographic memories and correlators. In both cases, the signal beam is no longer a single plane wave; it is an entire fan of plane waves, each from one point in the object plane (as explained above). Consequently, the reference beam writes an entire fan of gratings, one with each component of the signal beam. This is represented in figure 2.9.

If we change the angle of the reference beam upon reconstruction, the entire fan of gratings will move, but only one component of the signal beam can be Bragg-mismatched exactly. The other components will all reside on either side of the null, and so will lead to weakly diffracted reconstruction of these components. This is referred to as cross-talk, because we would like to store another page of data at this location, but when we try to read this page out, we will also read weakly those components of the other hologram that are not completely Bragg-mismatched. For this reason, rather than recording successive holograms at the first nulls of previous holograms, we often multiplex holograms at every third, fourth, or even fifth null. The cross-talk at these more distant nulls is significantly reduced because the power of the sidelobes falls as $1/\Delta\theta^2$ (the selectivity curve for power is the square of the amplitude curve).

Figure 2.9 can also demonstrate how holograms perform the correlation function. In this case, we are using the other grating term of equation 2.24. Figure 2.9 shows how the peak of the correlation is formed; each signal wave reads the grating it wrote with the reference wave, and all of the diffracted beams add together in phase to form one strong plane wave, which will produce a bright spot on the correlation plane behind the lens. Figure 2.10 shows how the sidelobes of the correlation are formed. In this case, individual components of the reconstructing signal beam read out gratings that were written by neighboring components of the recorded signal beam; these grating vectors now come to a point off of the surface of the k-sphere. However, the angle between different components of the signal beams is very small, so Bragg-mismatch is quite small. Consequently, there is still a diffracted beam, but at a slightly different angle than that of the originally recorded reference beam. Each component is reading out a grating whose strength is proportional to the strength of a shifted version of the recorded signal. In other words, the strength of one diffracted

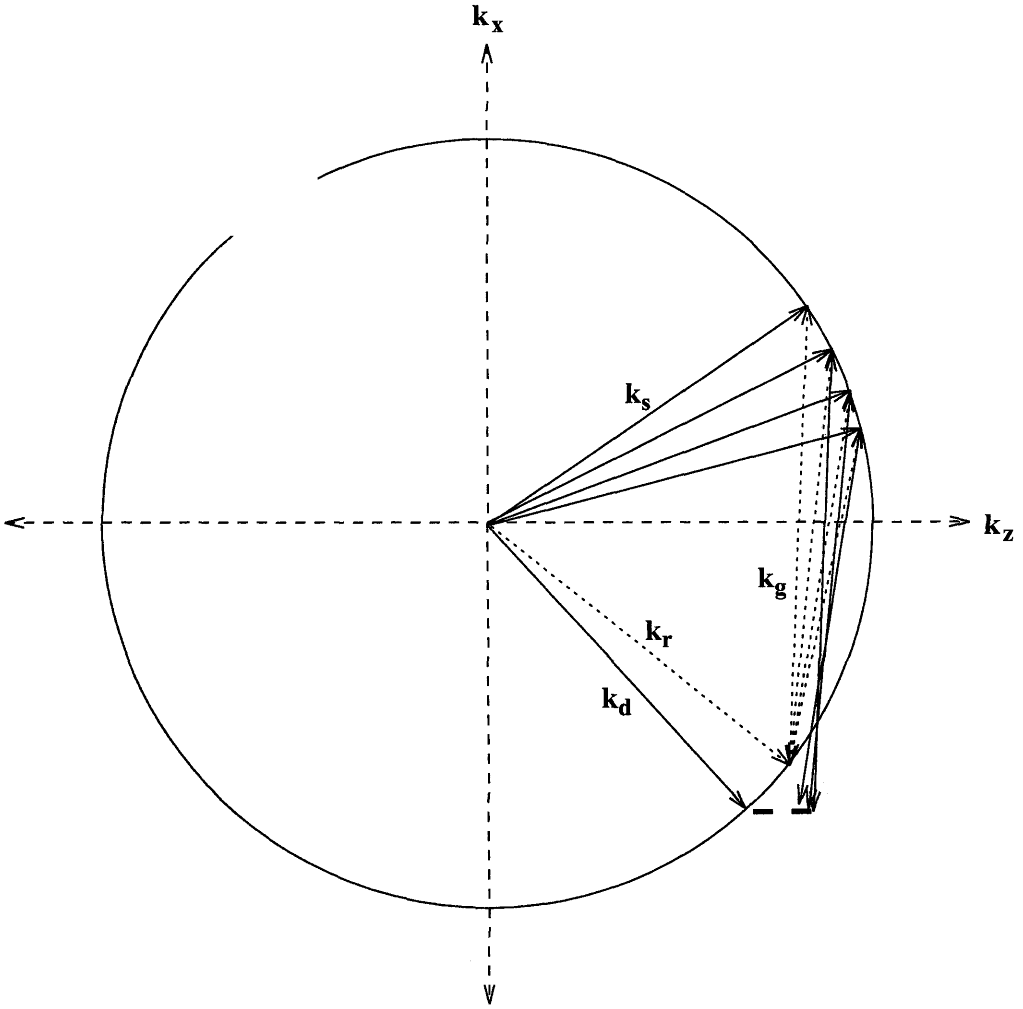

Figure 2.10: K-sphere diagram of side-lobe formation in a correlator

component, D_i, is equal to the product of the reconstructing signal component, S'_i, and a shifted recorded component, $S_{i-\delta}$

$$D_i \propto S'_i S_{i-\delta} \tag{2.31}$$

The total diffracted field would then be the sum

$$D_{total} \propto \sum_i S'_i S_{i-\delta} \tag{2.32}$$

There is actually a continuum of signal components, so writing equation 2.32 as an integral and generalizing to any diffracted k-vector $D(\delta)$ yields

$$D(\delta) \propto \int S'(x) S(x-\delta) dx \tag{2.33}$$

which is exactly the correlation of S with S'. These equations are not precise, because we have neglected Bragg-mismatch; Bragg-mismatch will put an extra scalar multiplier in front of the integral that will be the usual sinc selectivity function. One can also see from figure 2.10 that the tips of the grating vectors do not all come to precisely the same point; this results from the fact that a change in reconstructing angle does not lead to an exactly equivalent change in the diffracted angle. This is greatly exaggerated in the figure, and in practice it is a negligible effect. The main point here is the intuitive understanding of how the correlation comes about in holography.

The k-sphere in figure 2.10 also demonstrates what happens if a shifted version of the original signal is used. The shift in the object plane translates to a change of angle in the Fourier plane, so the same signal fan will be present, but at a different angle. The correlation peak now forms at a different angle, which produces a shifted peak in the correlation plane; thus, the peak shifts to follow shifts in the input image. This is referred to as shift-invariance, and it is very important in many pattern recognition applications. In practice, objects that are to be recognized can appear in any number of locations within the scene and we want to be able to recognize them independently of their location. In addition, shift-invariance can be used to track objects as they move within a scene. The peak will move in proportion to shifts in the input image, but the proportion will not be one-to-one. Rather, it will be given by equation 2.30.

If the reconstructing signal beam shifts too much, however, Bragg-effects will start to effectively reduce the strength of the correlation peak. Eventually, the peak will hit a null in the Bragg selectivity curve. As in the case of holographic memories, we can use this null to record another correlator, or neuron.

Chapter 3 Angle Multiplexing

Contents

In this chapter I will continue to address the fundamentals of holographic data storage using angle multiplexing. I will start with the 90^{circ} geometry. This willbe followed by the disk geometry. I will conclude the chaper with a discussion of correlator capacity.

3.1 The 90° Geometry

The basic ideas behind angle multiplexing were introduced in chapter two. Here I will reiterate that the transfer function $A(\mathbf{k_i}, \mathbf{k_d})$ is given by

$$A(\mathbf{k_i}, \mathbf{k_d}) = \iiint \frac{\Delta\epsilon(\mathbf{r}')e^{i(\mathbf{k_i}-\mathbf{k_d})\cdot\mathbf{r}'}}{2ik_{dz}} d\mathbf{r}' \tag{3.1}$$

For the transmission geometry, in which both the reference and signal beams are incident upon the same face of the medium, we take the integrations in the transverse directions to be infinite and the integration in the longitudinal direction to be over the thickness L, yielding

$$A(\mathbf{k_i}, \mathbf{k_d}) \propto \delta(k_{ix} - k_{dx} + k_{gx})\delta(k_{iy} - k_{dy} + k_{gy})L\mathrm{sinc}\left(L(k_{iz} - k_{dz} + k_{gz})\right) \tag{3.2}$$

which can be reduced to

$$A_d(\Delta\theta) = \mathrm{sinc}\left(\frac{nL\sin(\theta_s - \theta_r)}{2\lambda\cos\theta_s}\Delta\theta\right) \tag{3.3}$$

where $\Delta\theta$ is the change in the reference beam angle upon reconstruction [28].

However, assuming $\theta_s = 0$ (signal beam is normal to the medium), the expression inside the sinc becomes infinite as θ_r approaches $\pi/2$. This is also apparent from the k-sphere. Since we have assumed that the grating is a delta function in the transverse direction, it is clear from figure 3.1 that Bragg-mismatch will occur immediately with any change in the reference beam angle.

Of course, it is impossible to have the signal beam on axis and the reference beam at 90^{circ} if the medium is infinite in the transverse direction. Therefore, this assumption must be changed. Figure 3.2 shows the 90^{circ} geometry. In this case, the signal and reference beams enter adjacent faces of a parellelpiped crystal. We would still like to reduce two of the three dimensions to delta functions to make our integration easier. Looking at figure 3.1, it is clear that the dimension parallel to the signal axis is most important for Bragg selectivity, since the head of the grating vector will move in this direction as the reference beam angle changes. Indeed, for small changes in the reference beam angle, the grating vector will move approximately orthogonally to the surface of the k-sphere; so we will take the other two dimensions to be effectively infinite and reduce the Fourier transforms to delta functions.

Another way to justify this step is to note that the phases that build up either constructively or destructively do so along the path of the signal beam, so it is the

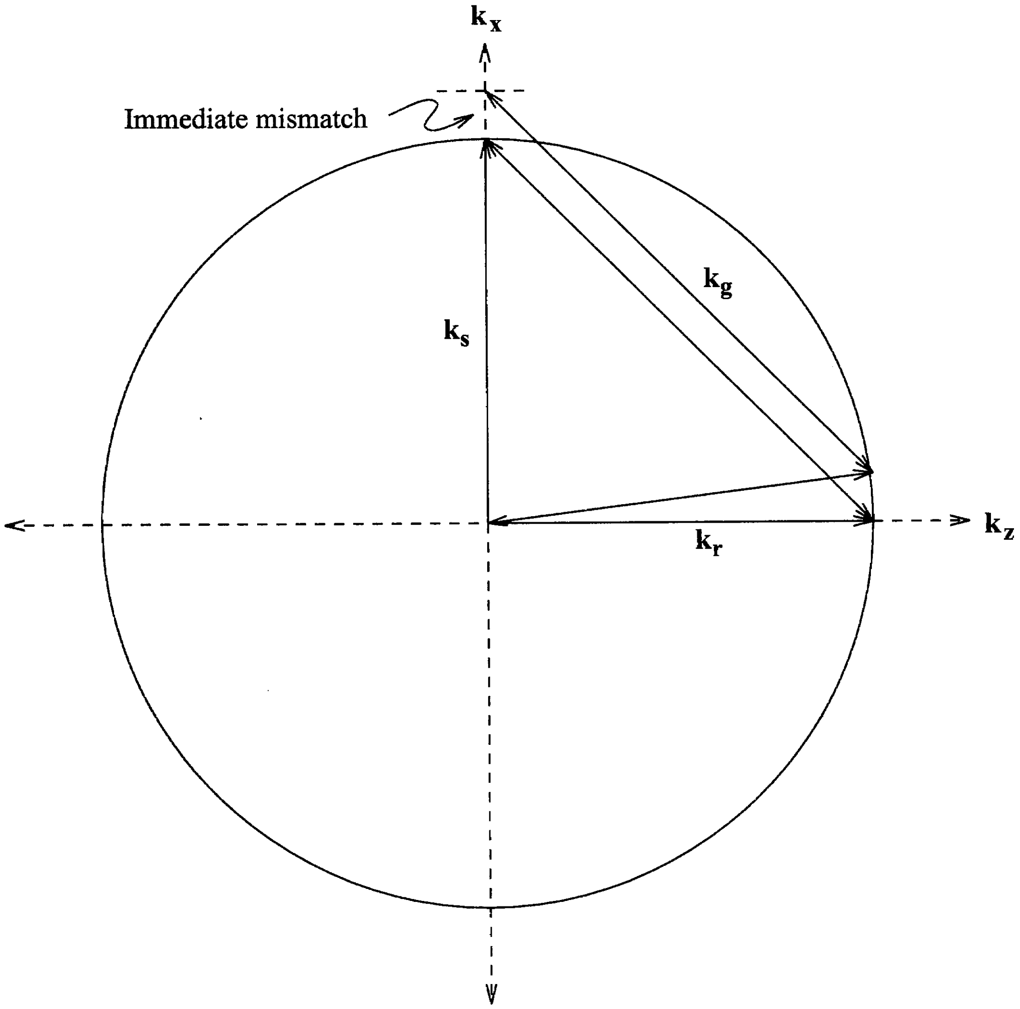

Figure 3.1: K-sphere for the 90^{circ} geometry.

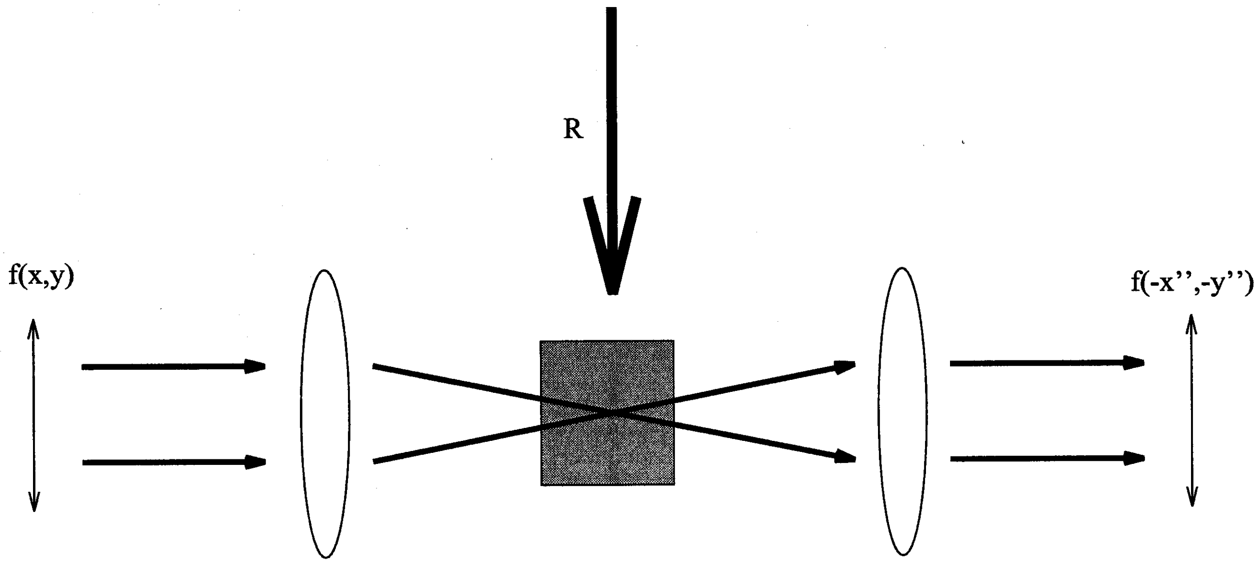

Figure 3.2: The 90° geometry

integration of the phase in this direction that really matters. This argument justifies our considering the hologram to be infinite in the x and y directions even though the hologram is actually the same size in all three dimensions. Note that the size of the hologram in the z direction is not the length of the crystal, but the thickness of the reference beam, since the hologram will only be written where the reference and signal beams overlap.

With these approximations, the selectivity for the 90 degree geometry is

$$\Delta\theta = \frac{\lambda}{L} \tag{3.4}$$

This geometry has the best possible angular selectivity for angle multiplexing. For typical values of λ and L, $\Delta\theta \approx 5 \times 10^{-5}$ radians.

So far we have only looked at what happens when the reference beam angle is rotated in the plane defined by the reference and signal beams. If we change the reference beam angle in the orthogonal direction, out of the interaction plane, equation 3.1 still holds. However, in this case we get

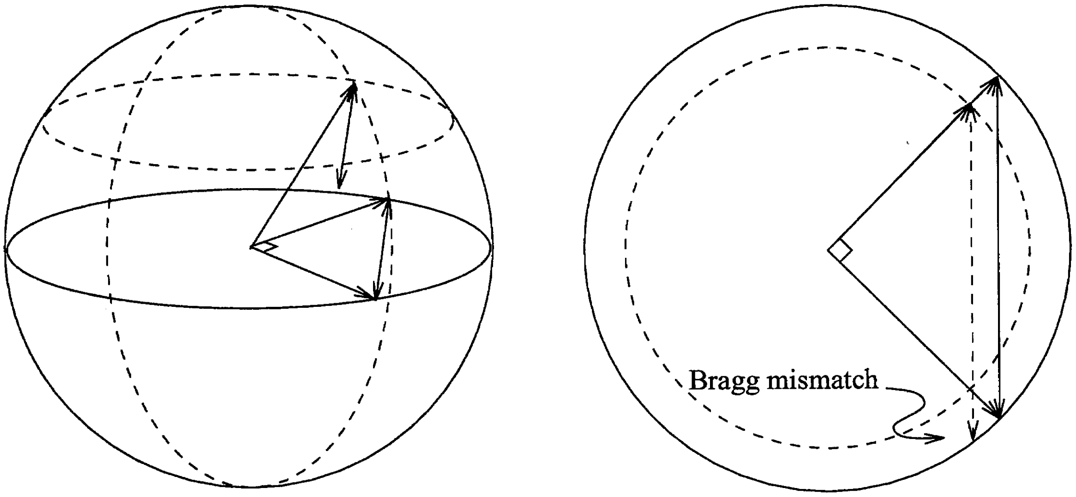

Figure 3.3: K-sphere showing tangential selectivity. a) Side view. b) Top view.

$$\Delta\theta = \sqrt{\frac{2\lambda}{L}} \tag{3.5}$$

The dependence on $\Delta\theta$ is now quadratic, and the selectivity is consequently much broader. The k-sphere diagram for this selectivity is shown in figure 3.3 for the 90 degree geometry; however, the analysis is similar for the transmission geometry. In this case, the motion of the reference beam moves the grating vector tangentially across the surface of the k-sphere. To second order, the surface of the sphere is quadratic, hence the quadratic dependence in equation 3.5. I will refer to selectivity in this direction as "tangential selectivity."

Tangential selectivity, although not as good as the in-plane selectivity, is still useful. Using 5 tangential angles and 2,000 in-plane angles in the 90 degree geometry, 10,000 holograms have been stored in a single crystal, and a large scale system capable of storing 160,000 holograms has been demonstrated [26,27,29].

When used as a memory system, we have several choices about the design of a 90 degree geometry system. Aside from hardware choices, such as what medium to use and what laser, there are choices about where to place the hologram. While our

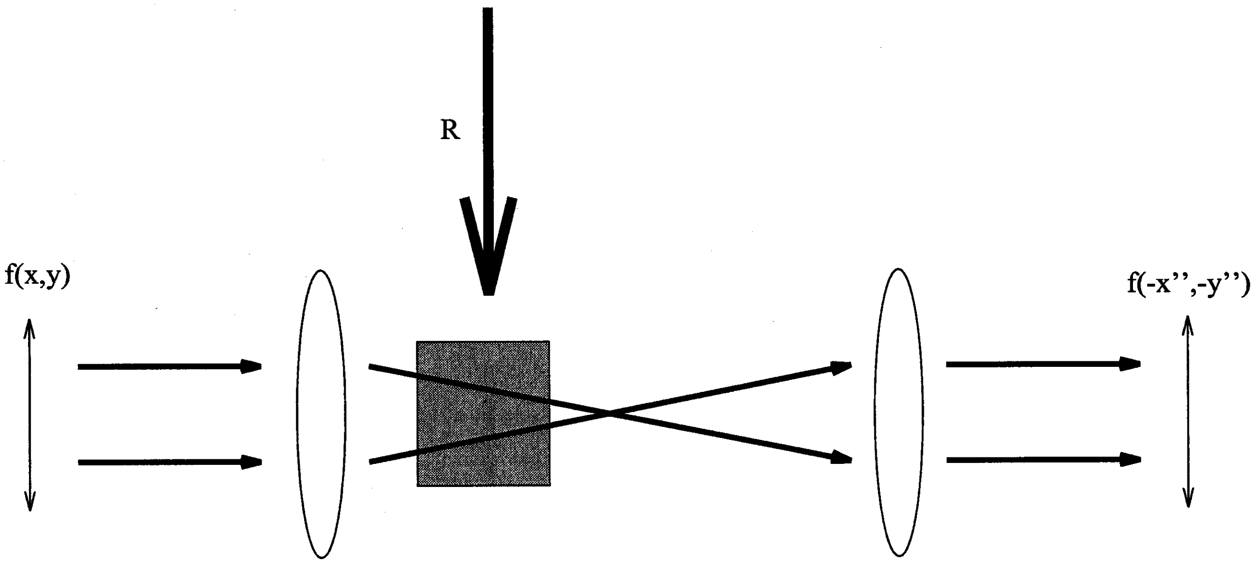

Figure 3.4: Fresnel zone holographic system.

previous analyses were all for systems in which the hologram was placed in the Fourier plane, we could just as well place the hologram at an image plane, or at any place else along the signal path. All that is required is that a complete imaging system is formed from the object to the detector at an image plane, and that the hologram is located somewhere along this path. In practice, one of two positions is used: in the image plane or in the Fresnel zone adjacent to the Fourier plane.

An example of a system with the hologram in the Fresnel zone is shown in figure 3.4. The Fresnel zone is the region between the Fourier plane and either lens. This system is very similar to the one we have been discussing so far; the only reason the hologram has been shifted out of the Fourier plane is to avoid problems in the recording process. The DC level of the data page produces a very bright spot in the center of the Fourier plane. It is nearly impossible to record a good hologram with such an inhomogeneous intensity distribution, because the DC power will burn out all of the dynamic range of the material before the rest of the hologram is written, and in photorefractive materials it can lead to intense fanning noise.

Figure 3.5 shows a typical image plane system. Notice that the system is twice as long as the corresponding Fresnel system. The Fresnel zone geometry also has the advantage that the defects in the material which cause noise affect a spatial frequency

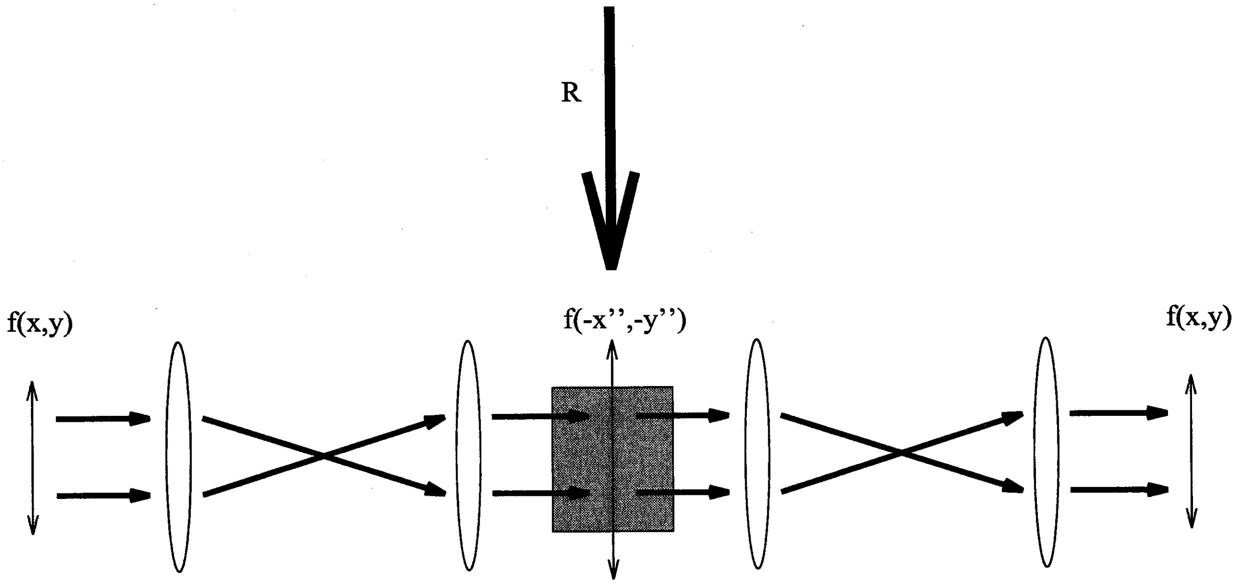

Figure 3.5: Image plane holographic system.

in the image, rather than a spot, and so the noise is spread out over the entire image rather than concentrated at a specific spot. In addition to being more compact and more tolerant to scattering noise in the medium, the Fresnel system usually allows for higher capacity, as explained below.

3.2 Memory Capacity

Computation of the capacity of any holographic memory system involves many assumptions about the geometry of the system, the material response, the available hardware, cross-talk and signal-to-noise tolerances. The calculation is, in principle, very straightforward, if lengthy. For the purpose here, we will perform a rough analysis for the sake of comparing the different forms of multiplexing. The capacity calculations in this and the following chapters follow very closely the work of Sidney Li [30].

The theoretical upper limit on data storage per unit volume of media for holographic data storage is tremendous. The maximum spatial frequency that can be recorded is $2/\lambda$, in the reflection geometry (reference and signal beams at 180 degrees). Taking this to be the bandwidth of our material, the total data stored in a parellelpiped of sides L_x, L_y, L_z is

$$\frac{2L_x}{\lambda} \cdot \frac{2L_y}{\lambda} \cdot \frac{2L_x}{\lambda} = \frac{8V}{\lambda^3} \tag{3.6}$$

which, for $\lambda = 500\text{nm}$, equals $8 \times 10^{12}\text{gratings}/\text{cm}^3$! Intuitively, this result is very appealing in that it is just the volume of the material, V, divided by the minimum resolvable spot size in three dimensions, $\lambda/2$ [31]. Practical limits to achieving this high density come mostly from geometrical factors and the dynamic range of the material. Geometrical factors refers to the fact that although the material may be capable of storing such a huge bandwidth, it is not an easy task to actually address this bandwidth; we need perfect imaging systems, and ways to get the reference beams in at all of the required angles. Dynamic range is the ability of the material to store large numbers of gratings all with sufficient diffraction efficiency to be retrieved with reasonable accuracy and speed. The dynamic range of a material must be divided amongst all M holograms, so if each hologram has an equal share, then the amplitude diffraction efficiency will go like $1/M$. But we can only detect power, which will fall like $1/M^2$. This poses a very serious practical restraint; however, in this thesis I will address only geometrical issues, leaving the material science to chemists.

Actually, one can trivially reach a data density of one bit per wavelength cubed without even using holography. Taking for example a compact disk, each bit is stored on the surface of the disk, one bit at a time. This bit can be stored as a diffraction limited spot, and it is only as deep as a quarter wavelength. If you take the disk to be one wavelength thick, you have will have achieved one bit per wavelength cubed! This ignores, of course, the added area required for tracking. The point is, we can think about data density as the product of an areal density, like on the surface of the compact disk, and a thickness density. As the material is allowed to become thicker, we can begin to multiplex holograms. Ideally, we would be able to multiplex L/λ holograms, while maintaining a perfect imaging system so that our areal density remains at one bit per wavelength squared.

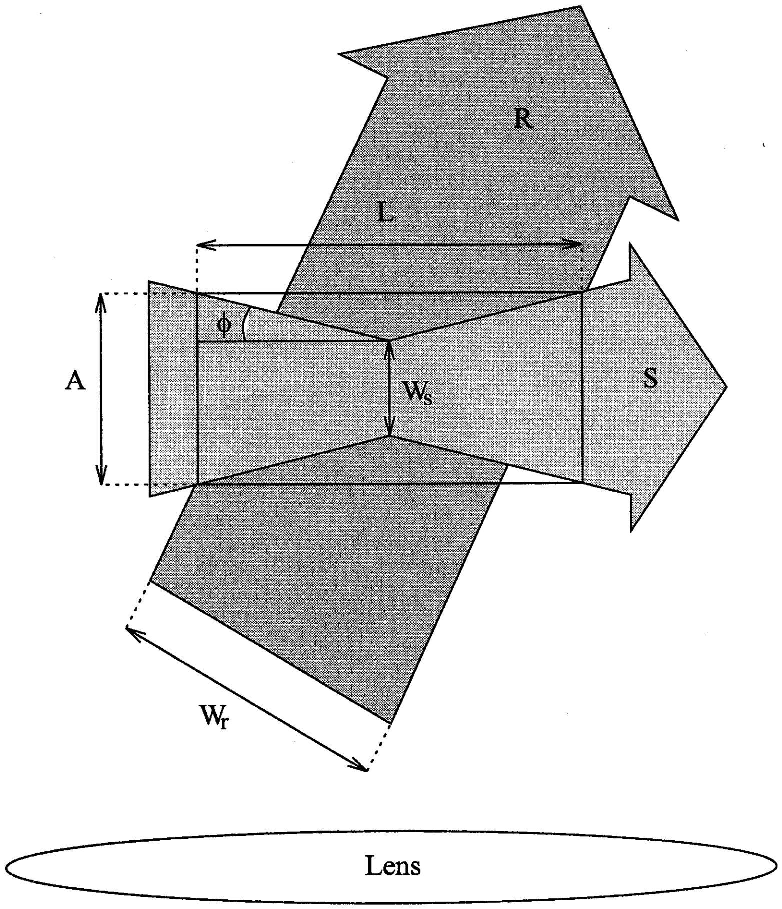

Figure 3.6: Cross section of the beam paths through the hologram in the 90 degree geometry: Image Plane.

3.2.1 Areal Density

In the example of the compact disk, it was easy to maintain an areal density of $1\text{bit}/\lambda^2$ since only one bit was ever imaged at a time. For holography, entire pages of data must be imaged. To properly address this issue would require taking into account lens aberrations and other issues from geometrical optics as one attempts to design a system with diffraction-limited resolution across the entire page of data. I will not spend time on this here.

For volume holography, however, areal density is not simply a matter of the resolving power of the imaging system. Figures 3.6 and 3.7 show the cross section of an imaging beam passing through a holographic medium.

Diffraction causes the beam to spread in both directions from the beam waist,

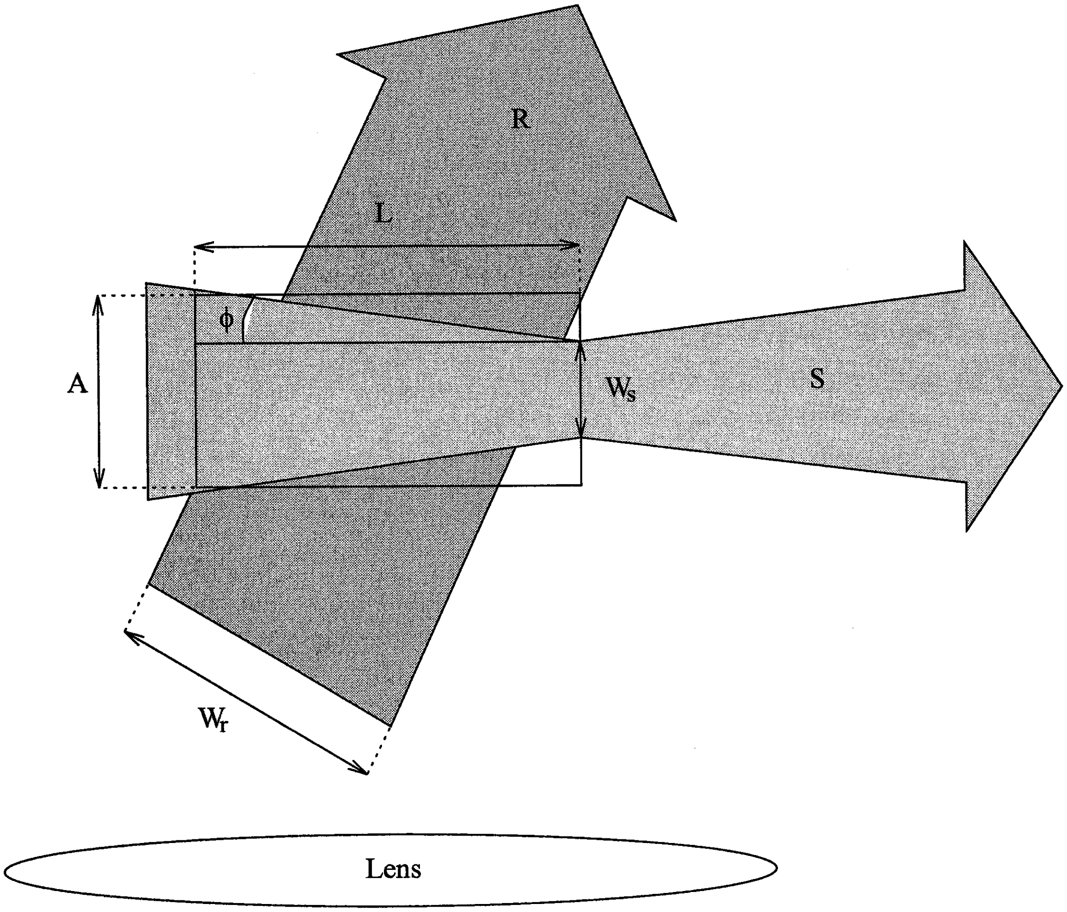

Figure 3.7: Cross section of the beam paths through the hologram in the 90 degree geometry: Fourier zone.

so that the thicker the holographic medium is, the more area is required to fit the same signal beam. Therefore, areal density actually decreases as the thickness of the medium increases. Given a minimum feature size at the beam waist of b, the angle ϕ is

$$\begin{aligned}
\frac{n \sin\phi}{\lambda} &= \frac{1}{b} \\
\phi &= \arcsin(\frac{\lambda}{nb}) \\
&\approx \frac{\lambda}{nb}
\end{aligned}$$

The area A^2 required for the hologram is then

$$\begin{aligned}
A &= 2\left(\frac{L}{2}\tan\phi\right) + W_s \\
A^2 &\approx \left(\frac{L\lambda}{nb} + W_s\right)^2
\end{aligned}$$

For Fourier plane holograms, the beam waist is usually placed at one end of the medium, so that

$$A^2 \approx \left(\frac{2L\lambda}{nb} + W_s\right)^2 \tag{3.7}$$

The number of bits is W_s^2/b^2, giving areal densities of

$$\frac{W_s^2}{b^2\left(\frac{L\lambda}{nb} + W_s\right)^2}, \quad \frac{W_s^2}{b^2\left(\frac{2L\lambda}{nb} + W_s\right)^2} \tag{3.8}$$

for imge plane and Fourier plane holograms, respectively. For an infinitely thin medium, both expressions reduce to $D = 1/b^2$, as expected. The value of W_s and b are obvious for image plane holograms; however, for Fourier plane holograms they are

$$
\begin{aligned}
W_s &= 2F\tan(\arcsin(\lambda/p)) && (3.9)\\
b &= \frac{\lambda}{\sin(\arctan(d/2F))} && (3.10)
\end{aligned}
$$

where F is the focal length of the Fourier transforming lens, p is the pixel size at the object plane, and d is the diameter of the object plane. If we assume that d is the maximum that will still allow proper imaging (ignoring abberations) through a lens of diameter d_L, then

$$d = d_L - 2F\tan(\arcsin(\lambda/p)) \qquad (3.11)$$

Although at first glance it might appear that image plane holograms would require less area, in fact, W_s for Fourier holograms is usually so much smaller that the areal density is actually higher. Fourier holograms are also less sensitive to noise from defects in the medium. For image plane holograms, a defect or piece of dust at the medium will produce a lot of noise for all of the pixels at that location, rendering an error. For Fourier holograms, the defect produces noise at a given spatial frequency, which then affects only slightly the image as a whole rather than concentrating all of the noise power in one location.

3.2.2 Volume Density

Given the areal density of our system, calculation of the density gained by the thickness of the material will give us the total volume density. In some sense, the areal density is just a function of the imaging system, while the thickness leads to Bragg selectivity. Ideally, this Bragg selectivity would allow for L/λ holograms. Then, if we had a perfect imaging system that could maintain the ideal areal density of A^2/λ^2, we could achieve a total volume density of V/λ^3. In practice, however, the constraints of the reference beam lens and other geometrical factors make it impossible to multiplex L/λ holograms.

Figure 3.6 shows the geometry for the reference beam with width W_R. From the figure it is clear that the length of the crystal for a given W_R which is incident at internal angles of between $\pm\theta_r^{max}$ is

$$L = A\tan\theta_r^{max} + \frac{W_r}{\cos\theta_r^{max}} \tag{3.12}$$

Combining eqations 3.8 and 3.12 yields

$$\begin{aligned} L &= \left(\frac{nb}{nb - \lambda\tan\theta_r^{max}}\right)\left(W_s\tan\theta_r^{max} + \frac{W_r}{\cos\theta_r^{max}}\right) && (3.13)\\ A &= \left(\frac{\lambda}{nb - \lambda\tan\theta_r^{max}}\right)\left(W_s\tan\theta_r^{max} + \frac{W_r}{\cos\theta_r^{max}}\right) + W_s && (3.14)\end{aligned}$$

for image plane holograms and

$$\begin{aligned} L &= \left(\frac{nb}{nb - 2\lambda\tan\theta_r^{max}}\right)\left(W_s\tan\theta_r^{max} + \frac{W_r}{\cos\theta_r^{max}}\right) && (3.15)\\ A &= \left(\frac{2\lambda}{nb - 2\lambda\tan\theta_r^{max}}\right)\left(W_s\tan\theta_r^{max} + \frac{W_r}{\cos\theta_r^{max}}\right) + W_s && (3.16)\end{aligned}$$

for Fourier plane holograms.

In order to pivot the reference beam around a point in the middle of the medium, an 4-F system, similar to the one discussed in chapter 2 for imaging, is used. In this case, θ_r^{max} is constrained by

$$\begin{aligned} \tan\theta_r^{max(ext.)} &= \frac{D - W_r}{2F}\\ \theta_r^{max(int.)} &= \arcsin\left\{\frac{1}{n}\sin\left[\arctan\left(\frac{D - W_r}{2F}\right)\right]\right\}\end{aligned}$$

The derivation of the number of holograms that can be stored follows that of Sid Li [30].

Assuming holograms are placed at the mth null of the adjacent hologram, the angle between successive holograms is

$$\Delta\theta_r = \frac{m\lambda\cos\theta_s}{nW_r|\sin(\theta_s - \theta_r)|} \tag{3.17}$$

so that

$$|\sin(\theta_s - \theta_r)|\Delta\theta_r = \frac{m\lambda}{nW_r}\cos\theta_s \tag{3.18}$$

Summing over the total number of multiplexed holograms, N_θ, and taking $\Delta\theta$ to be small enough to approximate as an integral results in

$$\int_{\theta_1}^{\theta_2} \sin(\theta_s - \theta_r)d\theta_r = \frac{m\lambda}{nW_r}\cos\theta_s(N_\theta - 1) \tag{3.19}$$

which yields

$$N_\theta = 1 + \frac{nW_r}{m\lambda}\left|\frac{\cos(\theta_s - \theta_1) - \cos(\theta_s - \theta_2)}{\cos\theta_s}\right| \tag{3.20}$$

where the extra 1 is for the middle hologram. For the 90° geometry, $\theta_s = 0$ and we can take θ_1 and θ_2 to be $\pi/2 \pm \theta_r^{max}$. The total number of holograms recorded in plane is then

$$N_\theta = 1 + \frac{2nW_r}{m\lambda}\sin\theta_r^{max} \tag{3.21}$$

In the out of plane direction, assuming we use the same total angular spread, the number of tangentially multiplexed holograms $N_\theta^{tang.}$ is

$$N_\theta^{tang.} = \mathrm{Top}\left(2\theta_r^{max}\sqrt{\frac{nW_r}{\lambda}}\right) \tag{3.22}$$

where Top rounds down to the closest integer. Assuming a large number of holograms, the thickness density is then

$$\mathcal{D}_\mathcal{L} = \frac{2nW_r \sin\theta_r^{max}}{m\lambda L}\mathrm{Top}\left(2\theta_r^{max}\sqrt{\frac{nW_r}{\lambda}}\right) \tag{3.23}$$

The total volume density is then the product of the areal and thickness densities

$$\mathcal{D}_\mathcal{V} = \frac{W_s^2}{b^2 A^2}\mathcal{D}_\mathcal{L} \tag{3.24}$$

Figures 3.8 and 3.9 show the volume densities for the 90° geometry in both the image and fresnel regions. These plots and all subsequent calculations of capacity do not include tangential selectivity. In practice, tangential selectivity only increases the capacity by a factor of about 5 at best, and is not critical in comparing the effective capacities of different multiplexing schemes. The high areal densities and small beam waists of the fresnel zone lead to much higher volume densities. As W_r increases, the interaction length increases, thus increasing the selectivity and raising the density. As W_r gets too big, however, it begins to take up more of the width of the reference lens, and so the number of accessible angles decreases, until W_r equals the reference lens width, and the density goes almost to zero.

These and all subsequent capacity calulations in this and the following chapters assume an SLM with 12μm pixels. For the image plane, the size of the SLM is assumed to be 2 cm on a side. For the fresnel zone, the SLM is assumed to be as

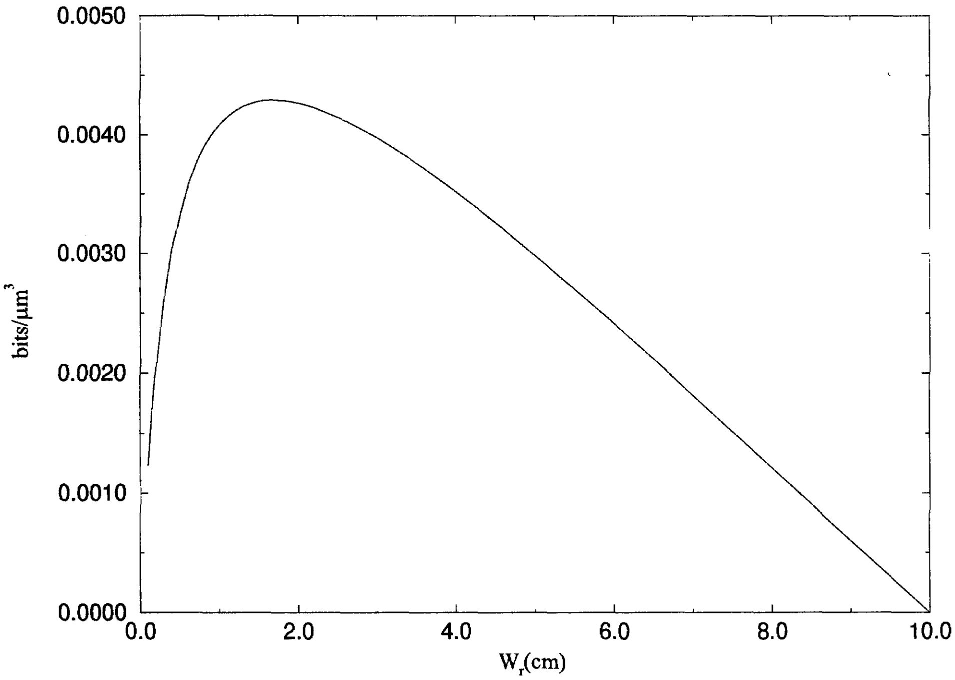

Figure 3.8: Volume density for the 90° geometry in the image plane.

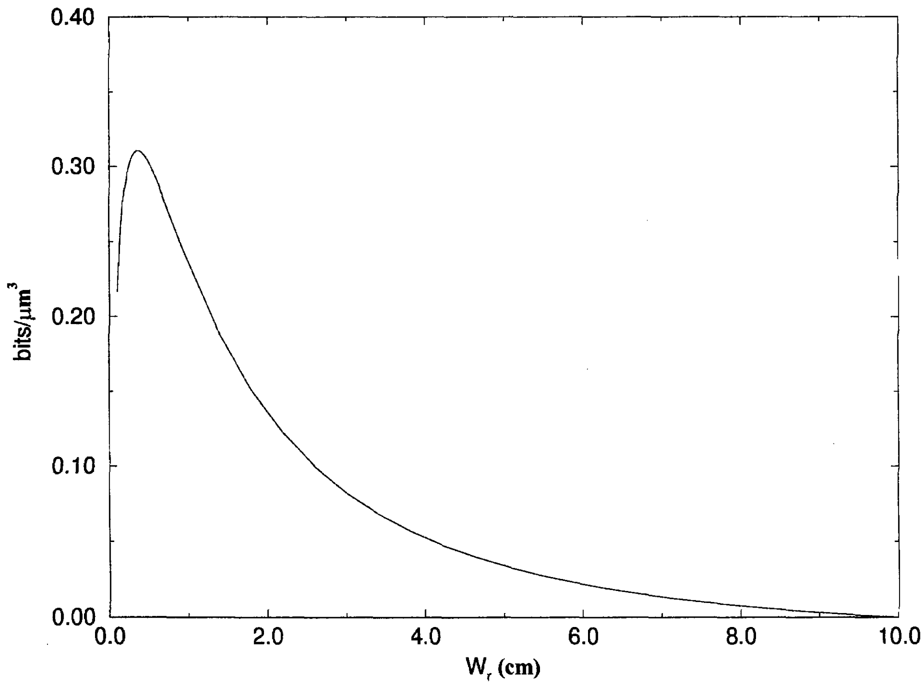

Figure 3.9: Volume density for the 90° geometry in the fresnel zone.

big as possible while still satisfying imaging criteria for the lens. The lenses were all assumed to be 10cm in diameter, with a 10cm focal length. This results in a huge SLM for the fresnel zone, and hence huge data densities. However, it is much easier to make large SLMs or data masks than it is to make small pixels. All subsequent calculations also assume that holograms are stored at the second null ($m = 2$) in order to reduce cross-talk noise.

It is clear that, even with ideal parameters, angle multiplexing falls short of the theoretical upper limit for volume density. The areal density is reduced by the growth in the area with increasing thickness that is required to fit the expanding signal beam. The thickness density is limited by the number of accessible angles, and also by the need to the fit all of the reference beams into the medium.

3.3 The Disk Geometry

Although the 90^{circ} geometry typically has the best Bragg selectivity and the highest volume data density, sometimes it is better to use the transmission geometry, in which both the reference and the signal beams enter the same face of the medium. This allows one to use a disk geometry, in which holograms are stored at one location of the disk, and then the disk is rotated so that an entire new set of holograms can be stored at a neighboring spot on the disk [30, 32]. This is sometimes referred to as spatial multiplexing [33–38]. In this case, the total capacity is increased because more medium is used, albeit at a lower density.

3.3.1 Areal Density

Figure 3.10 shows the geometry for an image plane disk. The distance α is important for calculating W_r, and hence the thickness density, as well as the area of the hologram.

$$\alpha = 2\left[\frac{W_s}{2}\cos\theta_s^{int.} + \left(\frac{L}{2} + \frac{W_s}{2}\sin\theta_s^{int.}\right)\tan(\theta_s^{int.} + \phi)\right] \tag{3.25}$$

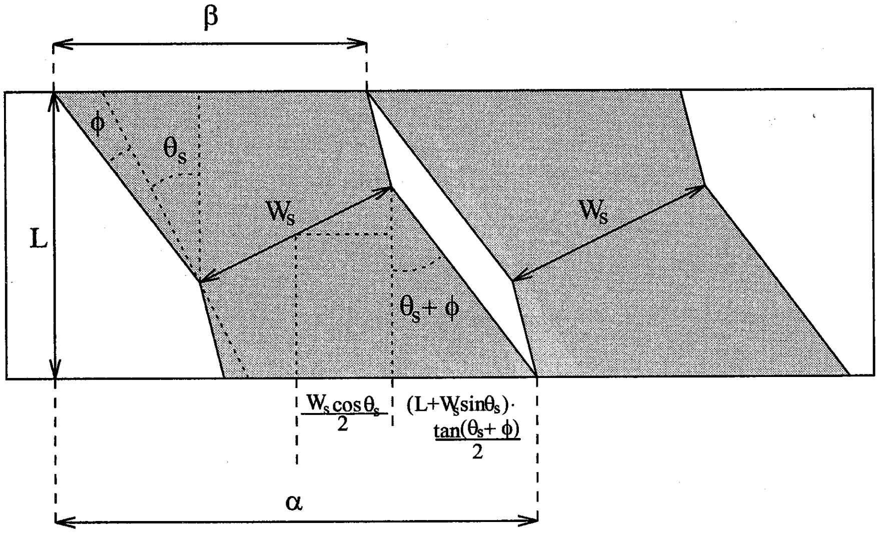

Figure 3.10: Geometry for holographic disk in the image plane.

where $\phi = \arcsin(\lambda/nb)$.

Neighboring holograms should not overlap. Hence, they should be spaced by the distance β, given by

$$\beta = \frac{\alpha}{2} + \frac{W_s}{2}\cos\theta_s^{int.} - \left(\frac{L}{2} - \frac{W_s}{2}\sin\theta_s^{int.}\right)\tan(\theta_s^{int.} - \phi) \tag{3.26}$$

and in the out-of-plane direction by

$$W_s + L\tan\phi \tag{3.27}$$

The total areal density is then

$$\mathcal{D}_{\mathcal{A}} = \frac{W_s^2}{b^2\beta(W_s + L\tan\phi)} \tag{3.28}$$

For Fourier plane holograms (see figure 3.11)

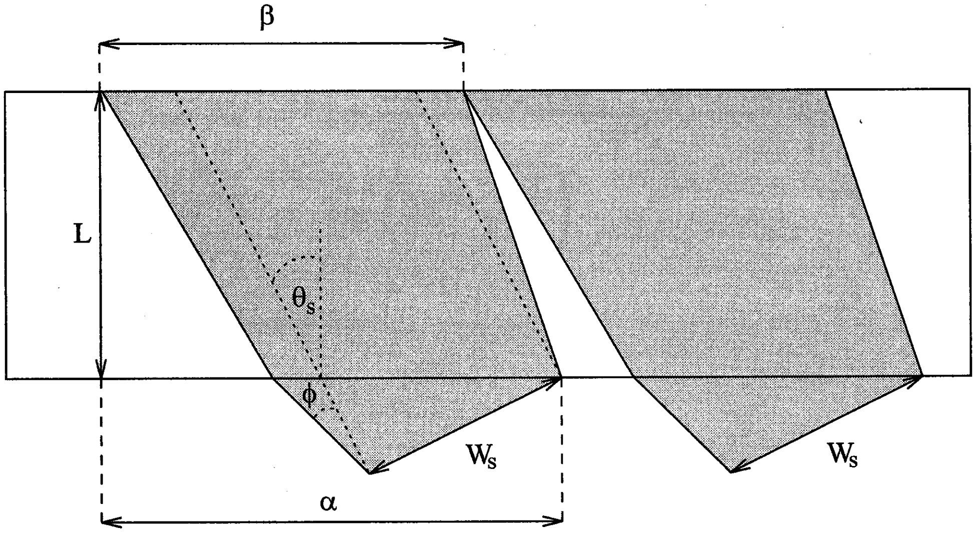

Figure 3.11: Geometry for holographic disk in the Fresnel zone.

$$\begin{aligned} \alpha &= W_s \cos\theta_s^{ext.} + W_s \sin\theta_s^{ext.} \tan\left(\theta_s^{ext.} + \tan(\phi)\right) && (3.29) \\ &+ L\tan\left\{\arcsin\left[\frac{1}{n}\sin(\theta_s^{ext.} + \tan\phi)\right]\right\} && (3.30) \end{aligned}$$

with $\phi = \arcsin(\lambda/b)$.

$$\beta = \alpha - L\sin\left(\arcsin\left(\frac{1}{n}\sin\theta_s^{ext.}\right) - \arcsin\left(\frac{\lambda}{nb}\right)\right) \tag{3.31}$$

and the areal density is still given by

$$\mathcal{D}_{\mathcal{A}} = \frac{W_s^2}{b^2\beta(W_s + L\tan\phi)} \tag{3.32}$$

3.3.2 Volume Density

The number of holograms are now given by

$$N_\theta = 1 + \frac{nL}{m\lambda}\left|\frac{\cos(\theta_s - \theta_1) - \cos(\theta_s - \theta_2)}{\cos\theta_s}\right| \tag{3.33}$$

$$N_\theta^{tang.} = \text{Top}\left(2\theta_r^{max}\sqrt{\frac{nL}{\lambda}}\right) \tag{3.34}$$

And the thickness density by

$$\mathcal{D}_{\mathcal{L}} = \frac{N_\theta N_\theta^{tang.}}{L} \tag{3.35}$$

However, for the above equations we need a new expression for W_r, which will affect θ_1 and θ_2.

$$W_r = \sqrt{\alpha^2 + L^2} \tag{3.36}$$

where α is whichever one is appropriate for the given geometry, as described above.

As in the case of the 90° geometry, increasing the thickness of the material increases W_r and improves the angular selectivity. However, it also decreases the total available angular bandwidth, since we still have to fit the reference beam through a lens of fixed diameter. Figures 3.12 and 3.13 plot the surface capacity versus disk thickness, L. We plot surface capacity rather than volume density for disks in order to better compare performance with standard CD-ROMs. Current CD-ROMs are capable of approximately 1 bit/μm^2, with the new DVDs capable of about 20 bits/μm^2.

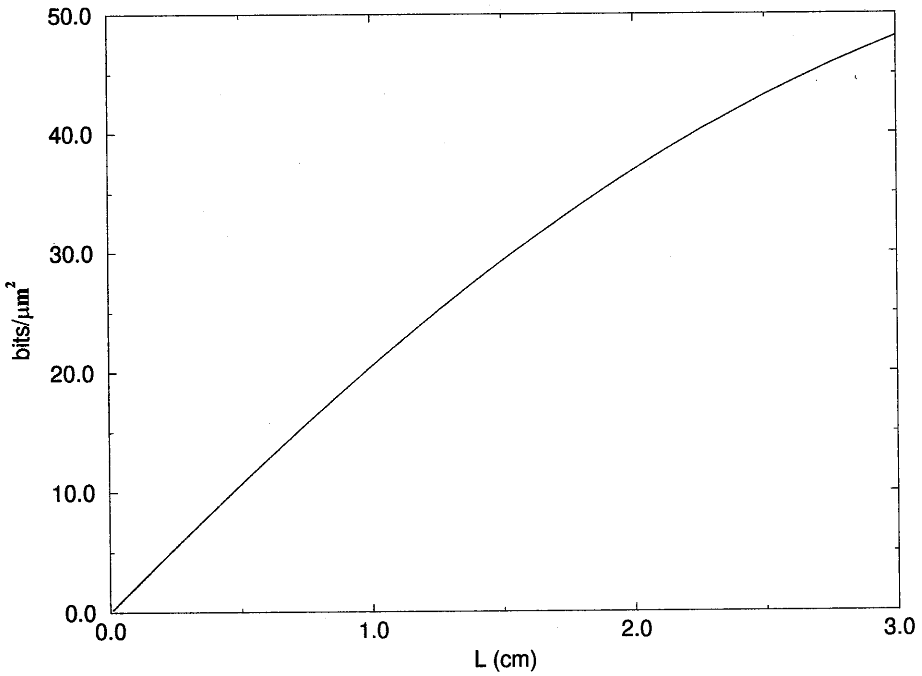

Figure 3.12: Surface density for angle multiplexed holographic disks, image plane.

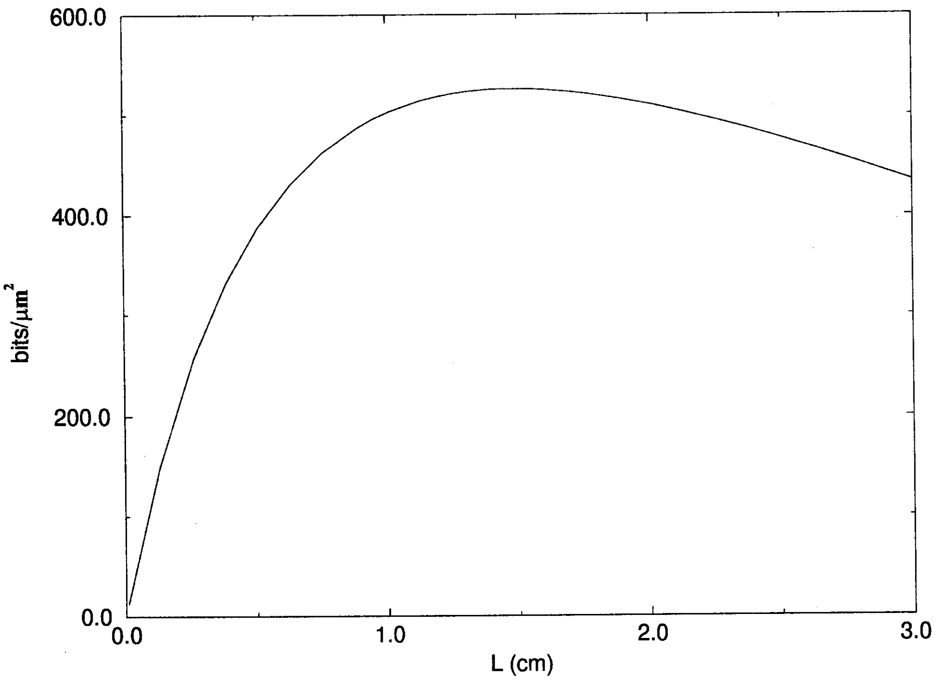

Figure 3.13: Surface density for angle multiplexed holographic disks, fresnel zone.

3.4 Correlator Capacity

As shown in chapter two, a correlator system is almost identical to a memory system, except that images are used to reconstruct all of the holograms at once rather than using a reference wave to reconstruct a single image. In practice, correlator systems differ in that input images usually have their DC component removed in a 4-F image processing system before the correlator. Without this step, the DC signal would swamp everything, so that all images would correlate with all others according only to the amount of DC power in each. For neural networks, the correlation peaks represent hidden units that may then be presented as the input to a second layer of weights in the form of a second correlator system.

The most important difference between a correlator and a memory system is that, in the correlator system, a signal beam of finite, non-zero, bandwidth is used to reconstruct the hologram. The direction of this reconstructing signal beam does not change from hologram to hologram; only the content changes. So all of the holograms are bragg-matched simultaneously. In this way, holographic correlators can implement many correlations in parallel. For most systems, the amount of shift-invariance limits the number of correlation templates that can be used. This is because a given system will have an output plane that must be divided amongst the individual templates in the system. If the system has too much shift-invariance, the output peak from one correlator could shift into an area that has been reserved for a different template; in this case, a shifted version of one object might be mistaken for a well-centered version of a different object.

In the simplest analysis, we can think of the correlator as a memory system and store successive templates with the reference wave at Bragg angle corresponding to those that would be used for a memory [39]. In order to avoid confusion, adjacent correlation templates would be stored such that their Bragg-nulls overlap. This means each correlation is centered at the second null of the neighboring template. The number of holograms that are stored in each direction is therefore given by equation 3.33 and 3.34, with $m = 2$.

However, the correlations have to be detected at an output, or correlation, plane. If we assume the correlation plane is as big as the lenses used in the system, then nothing would really change. However, it is rare to find such a large CCD camera. In practice, a screen is sometimes placed at the correlation plane, and a camera captures the image of the correlation peaks off of the screen using its own imaging lens. This arrangement is less sensitive, however, with dynamic range not being a problem because all of the diffracted power is concentrated into one peak, rather than being spread out over an entire image.

The above analysis, however, ignores the fact that it is the signal beam, rather than the reference beam, that is reconstructing the hologram. So it is the Bragg selectivity of the signal beam, not the reference beam, that determines the shift-invariance of the system. An intuitive picture for how correlations and shift-invariance work was given in chapter two using the k-sphere (see figures 2.9 and 2.10). The angular selectivity is still given, to good approximation, by equation 3.3, except the roles of the reference and signal waves must be swapped, giving the first null at

$$\Delta\theta_s = \frac{\lambda \cos\theta_r}{nL\sin(\theta_r - \theta_s)} \tag{3.37}$$

However, for the transmission geometry, only the x component of this angle stays the same for the diffracted wave giving

$$\sin\theta_s - \sin(\theta_s + \Delta\theta_s) = \sin\theta_r - \sin(\theta_r + \Delta\theta_r) \tag{3.38}$$

which, for small angles $\Delta\theta$, reduces to

$$\Delta\theta_r = \Delta\theta_s \frac{\cos\theta_s}{\cos\theta_r} \tag{3.39}$$

Equations 3.37 and 3.39 determine how closely packed the reference waves can be.

In the 90^{circ} geometry one must to be careful to identify the direction of the z axis. Since the beam that is being reconstructed is the reference beam, the z axis should be defined as normal to the face that the reference beam enters. Equation 3.39 is then zero, or very small. This indicates that while the correlation peaks may fade away if the input shifts too much, the peaks will never move in the in-plane direction. Thus, the peaks can be packed very tightly in the in-plane direction.

The size of the output plane, and consequently the total angle available for output, limits the number of holograms that can be stored. Combining equation 3.34 and equation 3.36 gives

$$\Delta\theta_r = \frac{2\lambda\cos\theta_s}{nL\sin(\theta_s - \theta_r)} \tag{3.40}$$

as for the case of memories. Therefore, our previous derivations (equations 3.30 and 3.31) are also valid for the number of holograms that can be stored for correlator systems.

Defining the capacity of a correlator system, however, is not as straightforward as for the memory system. For memories, the capacity is simply the total number of bits that can be recalled with a good signal-to noise ratio. We could similarly define the capacity of a correlator as the number of holograms, perhaps times the number of pixels per image stored, but this is not necessarily the most relevant metric. Correlators and neural networks are often used for pattern recognition. The shape and size of the shift-invariance domains of a pattern recognition system is an important factor in its usefulness. For the case of the 90 degree geometry, the domains are very long in the tangential direction, and very narrow in the in-plane direction. For typical systems, there is less than one pixel's worth of shift-invariance in the in-plane direction. There are some applications where such domain shapes might be useful, such as when the input images will only ever be shifted in one direction. Some tracking problems in which the object is something that is on the ground, and hence always at the horizontal, are examples of applications where the 90 degree geometry would be very useful.

While the 90^{circ} geometry can store many holograms with very asymmetric domains, neighboring correlation domains could be used to store holograms of shifted versions of the same object, thus "training in" directly the shift-invariance. This would reduce the total number of objects that could be recognized in favor of having greater shift-invariance. In practice, it would be better to have wider domains if that is what the application calls for; fewer holograms would need to be stored, so less dynamic range would be used and stronger signals could be achieved. Additionally, neighboring correlators should have overlapping receptive fields, to avoid having areas in which the object could not be detected. The concept of a receptive field comes from neurobiology directly, and it refers to that area of the visual field that can affect a given neuron. In this case, the size of the shift-invariance domain is projected in image space. If the receptive fields did not overlap, then the object would produce no correlation peak when it is between two domains because the peak would be at exactly the bragg-nulls of each domain. So the capacity in this case would be somewhat smaller than if the system naturally had larger shift invariance domains.

We can define another metric of capacity that takes into account the size of the shift invariance domains. Multiplying the total number of holograms, N_h, by the number of shift invariant positions, N_{shift}, gives the number of different images for which the response of the system is optimum. In order to get a metric with units of bits, we can multiply by the number of bits in each image, M^2 for an M by M image. The correlator capacity, N_c, is then

$$N_c = N_{shift} N_h M^2 \tag{3.41}$$

This definition assumes that the images stored are all of the same size, M by M, but it is trivial to extend it to a summation over images of varying size. This definition also allows for redundancy in the stored representations. Only the area of the input that constitutes the image to be recognized is counted; the rest of the input scene may contain anything. Therefore, some images may be subsets of other stored images. This sort of redundancy can be very useful for fault-tolerant pattern recognition. Indeed, there is thought to be quite a bit of redundancy in the brain. The upper limit of our new metric is thus larger than the total number of possible image scenes, and is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \underbrace{ij}_{M^2} \underbrace{(N-i)(N-j)}_{N_{shift}} \underbrace{2^{ij}}_{N_h} \tag{3.42}$$

for a scene that is N by N binary pixels. The product ij represents the total number of bits in an image that is i by j pixels. The total number of possible such images is given by 2^{ij}. The number of possible shifted positions is given by $(N-i)(N-j)$. In practice, of course, one would never want to store every possible image subset of the scene, but this represents a hard upper bound on the correlator capacity metric.

It is not possible to get a nice analytical form for the correlator capacity; however, we can get a feel for it by making the approximation that $\Delta\theta_r$ and $\Delta\theta_s$ are approximately constant. The number of shift positions, N_{shift}, is then

$$N_{shift} \approx \frac{2nF\Delta\theta_s}{b} = \frac{2F\lambda\cos\theta_r}{bL\sin(\theta_r-\theta_s)} \tag{3.43}$$

where b is the pixel size at the input plane. Multiplying with equation 3.31 gives

$$\begin{aligned} N_c &\propto \frac{2F\lambda\cos\theta_r}{bL\sin(\theta_r-\theta_s)} \cdot \frac{nL|\cos(\theta_s-\theta_1)-\cos(\theta_s-\theta_2)|}{2\lambda\cos\theta_s} && (3.44)\\ &= \frac{nF}{b} \cdot \frac{\cos\theta_r|\cos(\theta_s-\theta_1)-\cos(\theta_s-\theta_2)|}{\sin(\theta_r-\theta_s)\cos\theta_s} && (3.45) \end{aligned}$$

for the transmission geometry, and

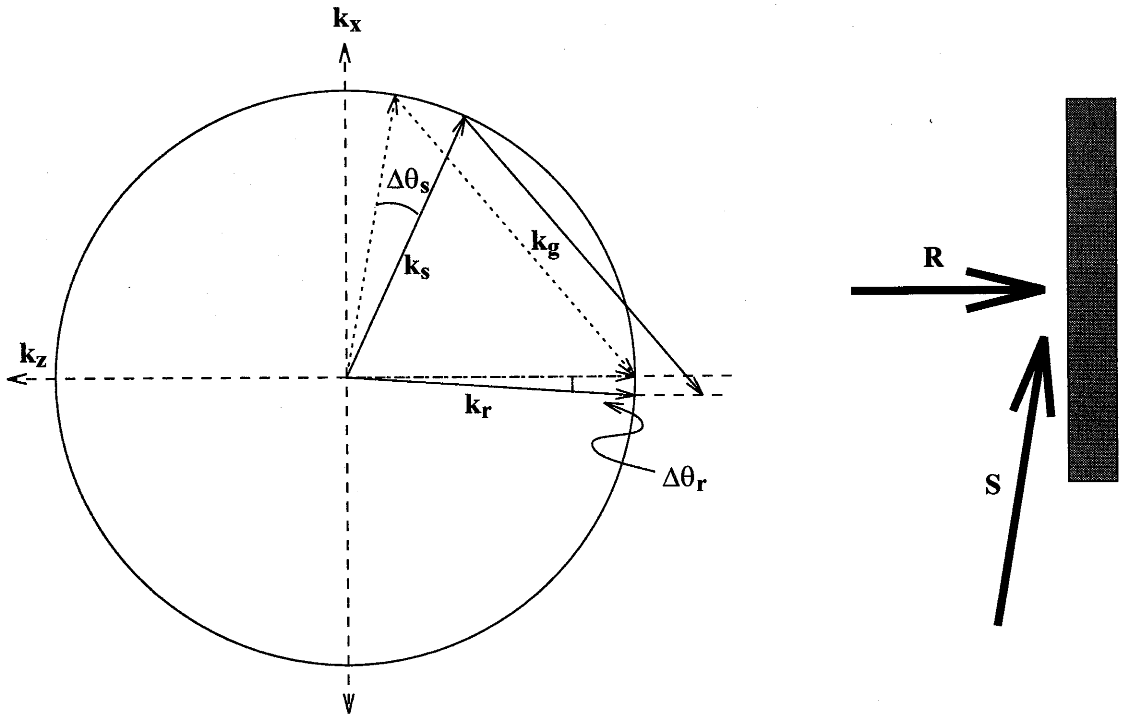

Figure 3.14: Correlator Capacity: a) k-sphere representation. The large angle between $\mathbf{k_s}$ and $\mathbf{k_r}$ makes $\Delta\theta_r$ small even for large input shifts, $\Delta\theta_s$. b) Geometry for greatest correlator capacity. Having the reference beam on-axis makes for the weakest Bragg selectivity.

$$N_c \propto \frac{2F\lambda}{bW_r}\cdot\frac{nW_r}{\lambda}\sin\theta_r^{max} \tag{3.46}$$

$$= \frac{2nF}{b}\sin\theta_r^{max} \tag{3.47}$$

for the 90^{circ} geometry. Both equations assume that the shift-invariance is Bragg-limited, as opposed to limited by the size of the aperture of the input plane. A surprising result is that the capacity is not proportional to L, the thickness of the material. This is because, while N_h is proportional to L, N_{shift} is inversely proportional to L. This breaks down when L becomes small enough that the input field size becomes the limiting factor for shift-invariance.

From equation 3.43 it is apparent that the geometry with the highest correlator capacity in transmission has the reference beam normal to the medium and the signal beam at glancing incidence. This becomes more intuitive from the k-sphere diagram (see figure 3.14). Storing the correlators at close to 90 degrees reduces the shift in the output peak with shifts in the input to a minimum, wile having the reference beam on axis decreases the Bragg selectivity as much as possible.

Although correlator capacity is a useful metric for comparing different systems, it is not the last word. For some applications, large shift invariance is not needed, and it is preferable to trade off some shift-invariance for increased numbers of holograms. Keeping this in mind, the following chapters will analyze different systems for the shape of the shift-invariance domains.

Chapter 4 The Reflection Geometry and Wavelength Multiplexing

Contents

4.1 Wavelength Multiplexing

The previous chapter discussed the transmission and 90 degree geometries for angle multiplexing of volume holograms. This chapter will look at the reflection geometry, in which the reference and signal beams enter opposite sides of the holographic medium. While this geometry is very poor for angle multiplexing, it is ideal for wavelength multiplexing, in which the wavelength of the reconstructing beam is changed [40–48].

4.1.1 Wavelength Selectivity

Figure 4.1 shows the k-sphere diagram for the reflection geometry. When the reference and signal beams are exactly counter-propagating, Bragg mismatch is tangential in all directions. For this reason, the reflection geometry is almost never used for angle multiplexing. However, changing the wavelength of the reconstructing beam will move the grating vector directly off of the surface of the k-sphere, as shown in figure 4.2. A similar derivation to that used for angle multiplexing can be followed to derive

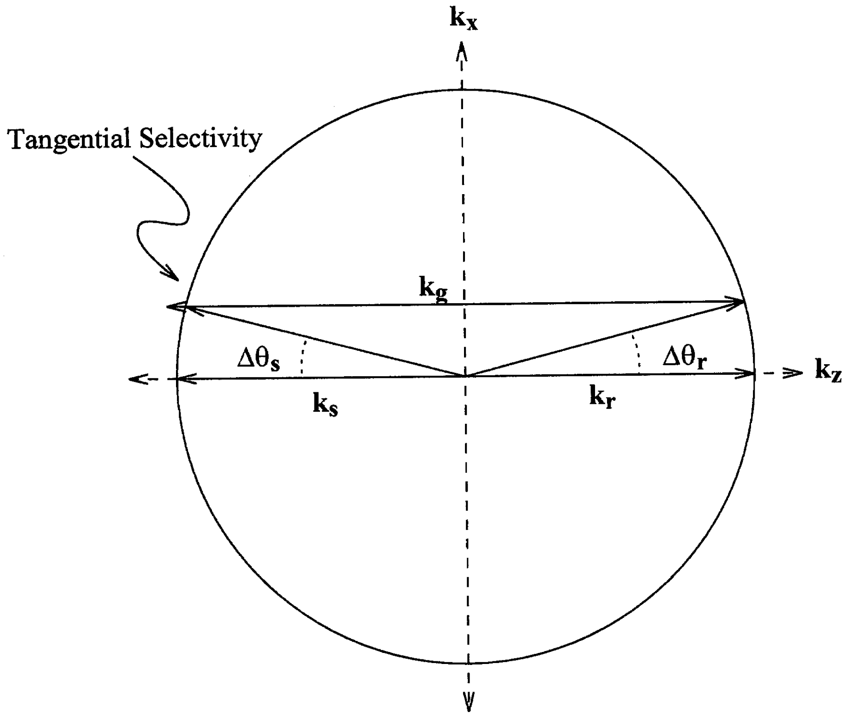

Figure 4.1: K-sphere diagram for the reflection geometry.

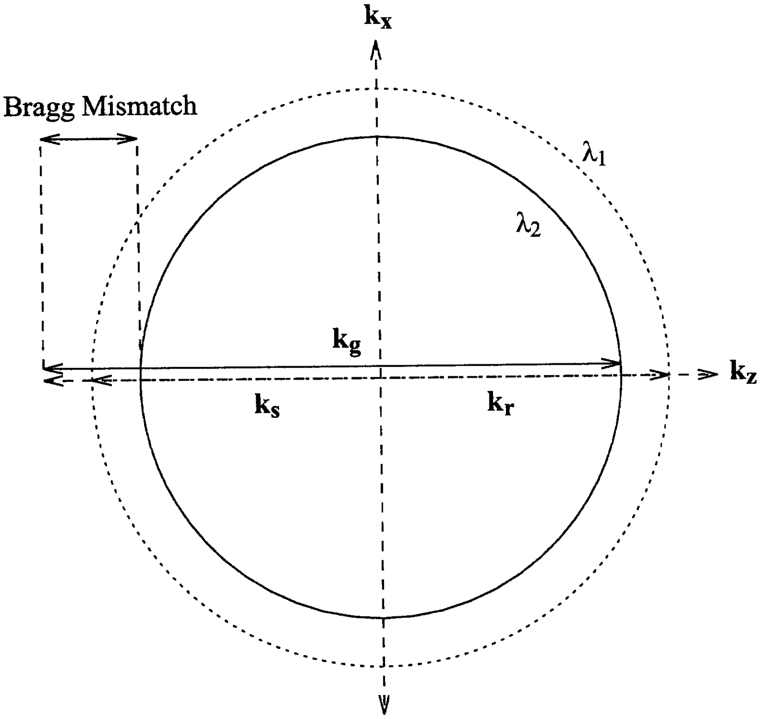

Figure 4.2: K-sphere diagram for wavelength multiplexing.

the wavelength selectivity of a given geometry; for counter-propagating waves, the amplitude of the reconstructed beam, A_λ, is given by

$$A_\lambda \propto \operatorname{sinc}\left(\frac{L}{\lambda^2}\Delta\lambda\right) \tag{4.1}$$

where $\Delta\lambda$ is the change in the reconstructing wavelength. The first zero is at

$$\Delta\lambda = \frac{\lambda^2}{L} \tag{4.2}$$

For memories, wavelength multiplexing has the advantage of requiring no moving parts for awkward beam steering. However, there are no compact, low-cost sources with a wide range of tunable wavelengths currently available. Wavelength multiplexing also suffers from background noise problems due to the reflection of the reference wave back along the direction of the signal beam. Even with a high-quality anti-reflection coating, the back reflection can easily swamp the signal if a large number of holograms are stored. Recording the beams at a slight angle, rather than exactly counter-propagating, can sometimes avoid this problem.

The memory capacity of wavelength multiplexing can be very high. As with angle multiplexing, the density can be broken up into the product of an areal density and a thickness density.

4.1.2 Areal Density

The areal density is very similar to that for angle multiplexing, and if a single spot of the medium is to be used, then the areal density is just equation 3.8 from chapter three. However, if we use wavelength multiplexing in a disk geometry, we can do even better. In this case, we only have to make sure that the individual beam waists do not overlap. For angle multiplexing, each beam had to be in a separate volume so that, upon read-out, holograms recorded with the same reference beam angle at different locations would not overlap at the detector array. Because the image (fourier) planes in wavelength multiplexing are side-by-side, rather than at an angle, it is possible to place an aperture in the system to block light from the unwanted signal. Therefore, the areal density can be $1/b^2$.

For image plane holograms, this means the areal density is just the inverse of the pixel size, *i.e.*, it is the areal density of the original data mask. For the Fourier plane, recall that

$$\begin{aligned} b &= \frac{\lambda}{\sin(\arctan(d/2F))} && (4.3) \\ d &= d_L - 2F\tan(\arcsin(\lambda/p)) && (4.4) \end{aligned}$$

As d approaches F, b approaches 2λ. So, for large images, the areal density can approach $1/4\lambda^2$.

4.1.3 Volume Density

The thickness density is very straightforward for wavelength multiplexing. Since there is no change of angle, it is completely independant of the area of the hologram. Following a similar treatment to that of chapter three gives the number of wavelength multiplexed holograms, N_λ, as [30]

$$N_\lambda = \frac{\lambda_2 - \lambda_1}{\lambda_1\lambda_2}L \qquad (4.5)$$

where λ_1 and λ_2 are the shortest and longest wavelength the source can address.

Figure 4.3 shows the thickness density as a function of the full wavelength range. Figures 4.4 and 4.5 show the surface storage density as a function of disk thickness for full width wavelength ranges of 20nm, 40nm, and 80nm. All plots are for wavelength ranges centered at 850nm, with the usual SLM parameters. In order to compete with angle multiplexing, wavelength multiplexing needs a source with at least 80nm

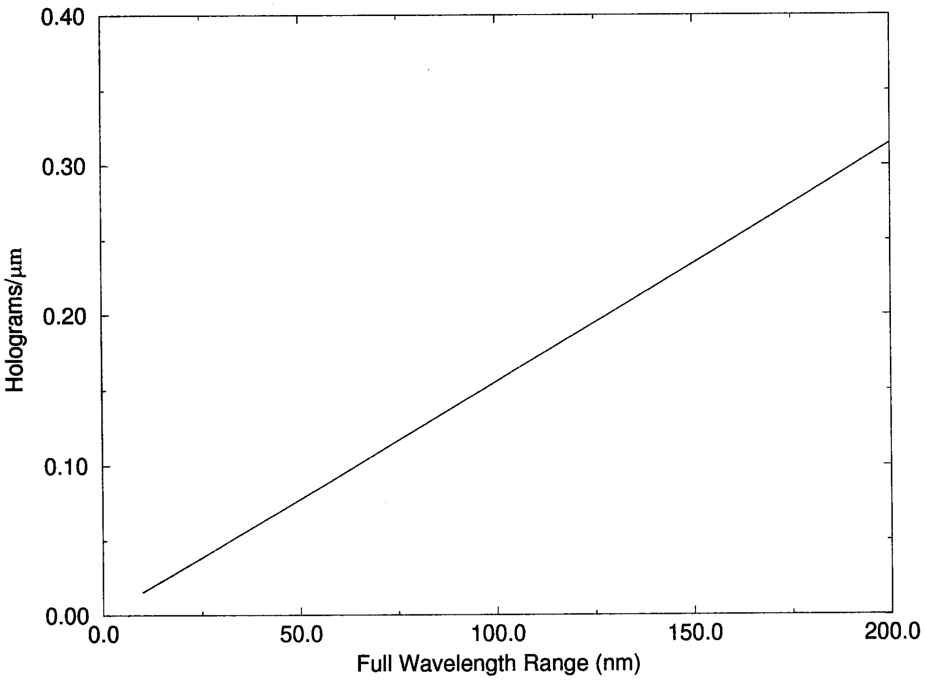

Figure 4.3: Thickness density for wavelength multiplexing. $\overline{\lambda} = 850$nm.

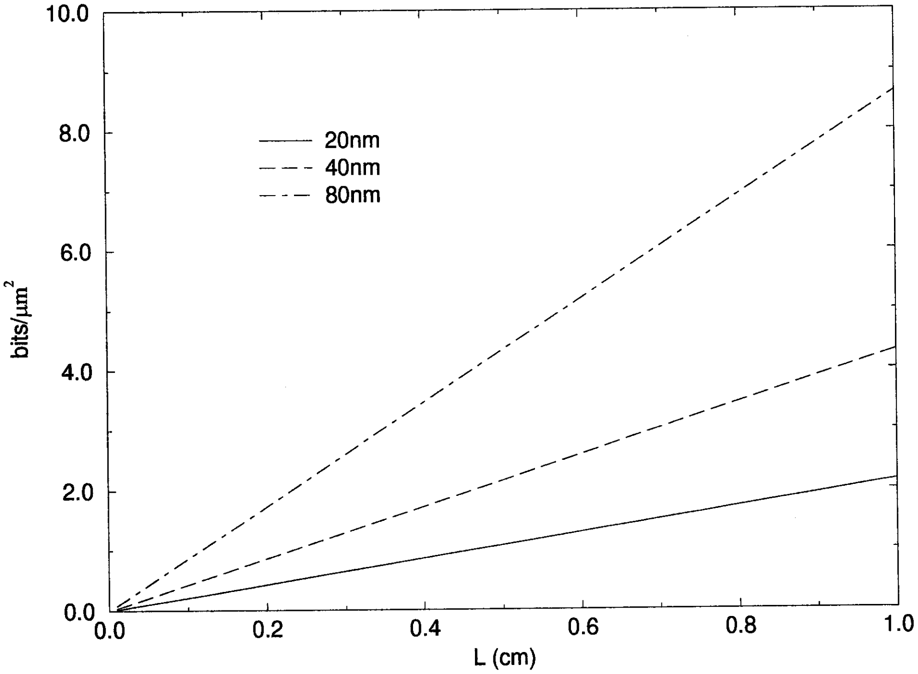

Figure 4.4: Surface storage density for wavelength multiplexing (image plane). $\overline{\lambda}$ = 850nm.

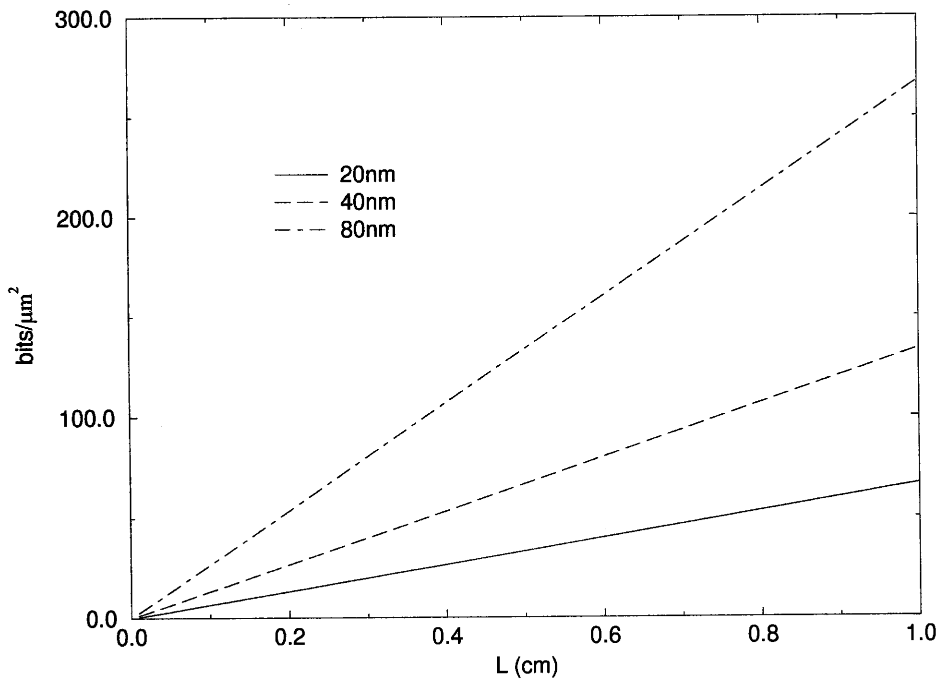

Figure 4.5: Surface storage density for wavelength multiplexing (Fourier plane). $\overline{\lambda} =$ 850nm.

full bandwidth. The center wavelength was chosen to be 850nm because the most practical tunable source would be a GaAs laser diode. Unfortunately, such sources are still expensive and do not have tuning ranges only on the order of 20nm-40nm. There are also currently no good materials that are sensitive in this region of the spectrum.

4.2 Reflection Geometry Correlators

Wavelength multiplexing can be used to store different correlation templates which can be read out sequentially [40], but this makes building a multi-layered neural network difficult. The poor angular selectivity of the reflection geometry made it a poor choice for an angle multiplexed memory; however, it can be useful as a correlator [49]. The center correlator, for which the reference beam is exactly counter-propagating to the signal beam, has a large, symmetric domain. Correlators stored way from the center have successively narrower domains in the radial direction. This is easy to understand, since as the angle between the reference and signal beams becomes less than π, the geometry becomes similar to the transmission geometry, with domains that are narrower in the plane defined by the reference and signal beams.

For the case of counter-propagating waves, we cannot use our previously derived selectivity function

$$\Delta\theta_r = \frac{m\lambda\cos\theta_s}{nL|\sin(\theta_s - \theta_r)|} \tag{4.6}$$

since the denominator goes to zero. This result came from an approximation that was only good to the first order in $\Delta\theta$. In order to get a result which is valid for the reflection geometry, we need to keep higher order terms. These are the same higher order terms that are required to derive the out-of-plane selectivity function, since once again a change in angle moves the tip of the grating vector tangentially across the surface of the k-sphere.

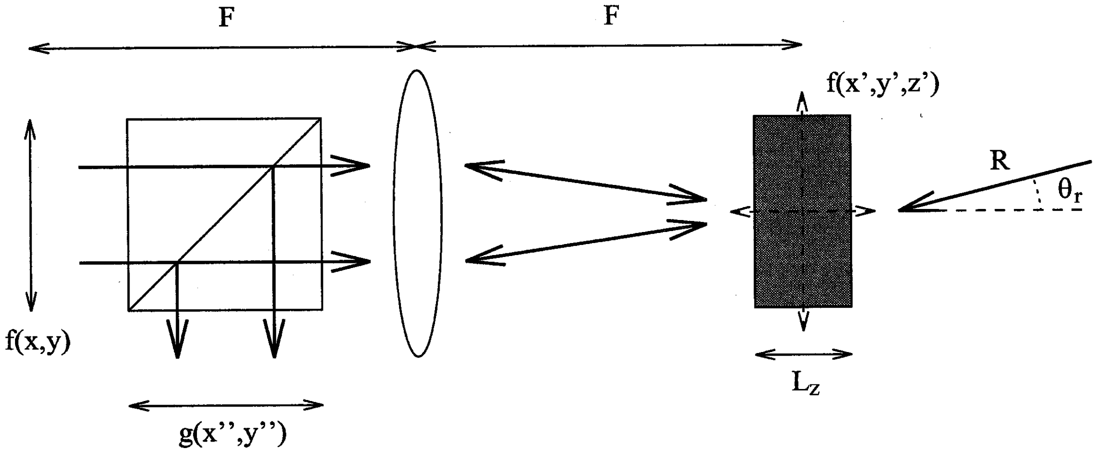

Figure 4.6: Reflection geometry correlator.

Starting with

$$A(\mathbf{k_i}, \mathbf{k_d}) \propto \iiint \Delta\epsilon(\mathbf{r}') e^{i(\mathbf{k_i}-\mathbf{k_d})\cdot\mathbf{r}'} d\mathbf{r}' \tag{4.7}$$

$\Delta\epsilon$ is the grating written by the field from the stored image, $f_1(x_1, y_1)$, and the reference wave $e^{-ik\sin\theta_r x'} e^{-ikcos\theta_r z'}$ given by (see figure 4.6)

$$\Delta\epsilon = \iint f_1^*(x_1, y_1) e^{i\frac{k}{F}(x_1+F\sin\theta_R)x'} e^{-i\frac{k}{F}y_1 y'} e^{-ikcos\theta_r z'} e^{-i\frac{k}{2F^2}(x_1^2+y_1^2)z'} dx_1 dy_1 \tag{4.8}$$

Reading out the hologram with the image $f_2(x_2, y_2)$ results in the diffracted output

$$\begin{aligned} A_d(\mathbf{k_d}) &= \iint dx_1 dy_1 \iint dx_2 dy_2 f_1^*(x_1, y_1) f_2(x_2, y_2) \\ &\cdot \iiint_{volume} e^{-i\frac{k}{2nF^2}(x_1^2-x_2^2+y_1^2-y_2^2)z'} e^{i\frac{k}{F}(x_1-x_2)x'} e^{i\frac{k}{F}(y_1-y_2)y'} \\ &\cdot e^{-\frac{i}{n}(k\sin\theta'-k_{dx})x'} e^{i\frac{k_{dy}}{n}y'} e^{-\frac{i}{n}(k\cos\theta'-k_{dz})z'} dx' dy' dz' \end{aligned}$$

where k_{dx}, k_{dy}, and k_{dz} are the x, y, and z components of the diffracted wave-vector such that

$$k^2 = k_{dx}^2 + k_{dy}^2 + k_{dz}^2 \tag{4.9}$$

Performing the integral over the volume leads to

$$\begin{aligned} A_d(\mathbf{k_d}) = & \iint dx_1 dy_1 \iint dx_2 dy_2 f_1^*(x_1, y_1) f_2(x_2, y_2) \\ & \cdot \operatorname{sinc}\left[\frac{L_z}{2\pi}\left(\frac{k}{2nF^2}(x_1^2 - x_2^2 + y_1^2 - y_2^2) + kn\cos\theta' - nk_{dz}\right)\right] \\ & \cdot \operatorname{sinc}\left[\frac{L_x}{2\pi}\left(\frac{k}{F}(x_1 - x_2) + kn\sin\theta' - nk_{dx}\right)\right] \\ & \cdot \operatorname{sinc}\left[\frac{L_y}{2\pi}\left(\frac{k}{F}(y_1 - y_2) - nk_{dy}\right)\right] \end{aligned}$$

The transverse dimensions L_x and L_y are large enough to approximate the last two sinc functions as delta functions. Integrating over x_2, and y_2 yields

$$\begin{aligned} A_d(\mathbf{k_d}) \propto & \iint f_1^*(x_1, y_1) f_2\left(x_1 + F\sin\theta_r - F\frac{k_{dx}}{k}, y_1 - F\frac{k_{dy}}{k}\right) \\ & \cdot \operatorname{sinc}\left[\frac{L_z}{2\pi}\left\{\frac{k}{2F^2}\left(x_1^2 - \left(x_1 + F\sin\theta_r - F\frac{k_{dx}}{k}\right)^2\right.\right.\right. \\ & \left.\left.\left. + y_1^2 - \left(y_1 - F\frac{k_{dy}}{k}\right)^2\right) - k\cos\theta_r - k_{dz}\right\}\right] dx_1 dy_1 \end{aligned}$$

Imaging the diffracted field to an output plane, x'', y'' with a lens of focal length F allows us to make the following approximations:

$$\begin{aligned} x'' &\approx F\frac{k_{dx}}{k} && (4.10) \\ y'' &\approx F\frac{k_{dy}}{k} && (4.11) \\ k_{dz} &\approx -k + \frac{k}{2F^2}(x''^2 + y''^2) && (4.12) \end{aligned}$$

For $\theta_r = 0$ the output is

$$\begin{aligned} E_d(x'', y'') &= \iint f_1^*(x_1, y_1) f_2(x_1 - x'', y_1 - y'') && (4.13) \\ &\quad \cdot \operatorname{sinc}\left[\frac{L_z}{\lambda F^2}\left[x''(x_1 - x'') + y''(y_1 - y'')\right]\right] dx_1 dy_1 && (4.14) \end{aligned}$$

The argument of the sinc now has multiple roots. Solving for the roots gives

$$\left(x'' - \frac{x_1}{2}\right)^2 + \left(y'' - \frac{y_1}{2}\right)^2 = \frac{x_1^2 + y_1^2}{4} \qquad (4.15)$$

The root we expect to get, which corresponds to the correlation peak, is $x'' = y'' = 0$. The other roots describe a circle, centered at $(x/2, y/2)$ with radius $\rho = \sqrt{x^2 + y^2}/2$, as shown in figure 4.7.

This circle represents a degeneracy in the Bragg selectivity. A given grating represents a chord across the k-sphere. However, this chord can fit inside the surface of the sphere in many positions. Taken all together, these possibilities desribe a cylinder inscribed within the k-sphere as shown in figure 4.8 [50].

Normally this sort of degeneracy is not an issue because the cross section of the cylinder through the angles subtended by the signal ray bundle is small. For the reflection geometry, however, a large portion of this cylinder may pass through the signal beam. In fact, figure 4.9 shows that the center pixel will read out *all* of the gratings degenerately. Normally, individual plane-wave components of the signal beam read out holograms written by neighboring component, thus forming the sidelobes

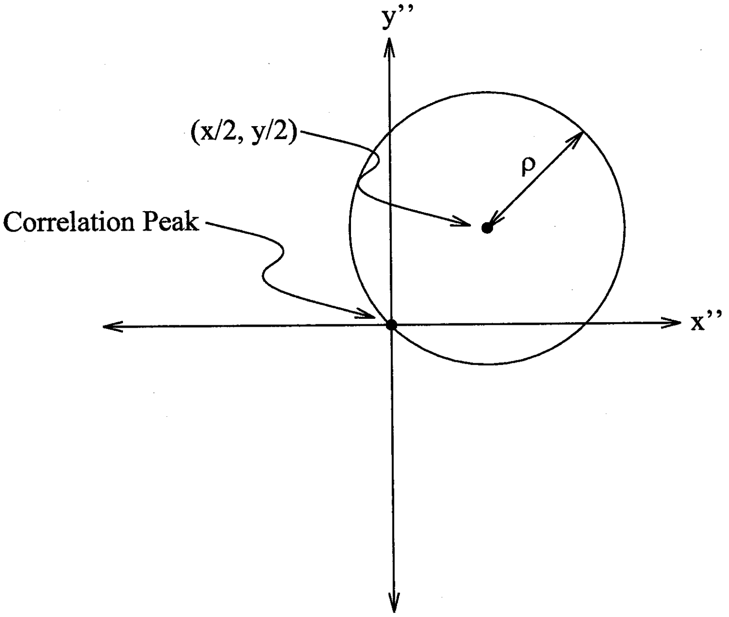

Figure 4.7: Degeneracy circle for reflection correlator.

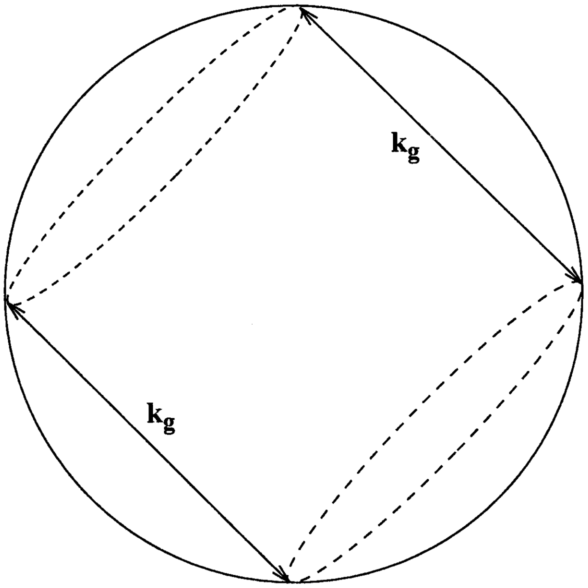

Figure 4.8: K-sphere diagram of a cylinder of degenerate gratings.

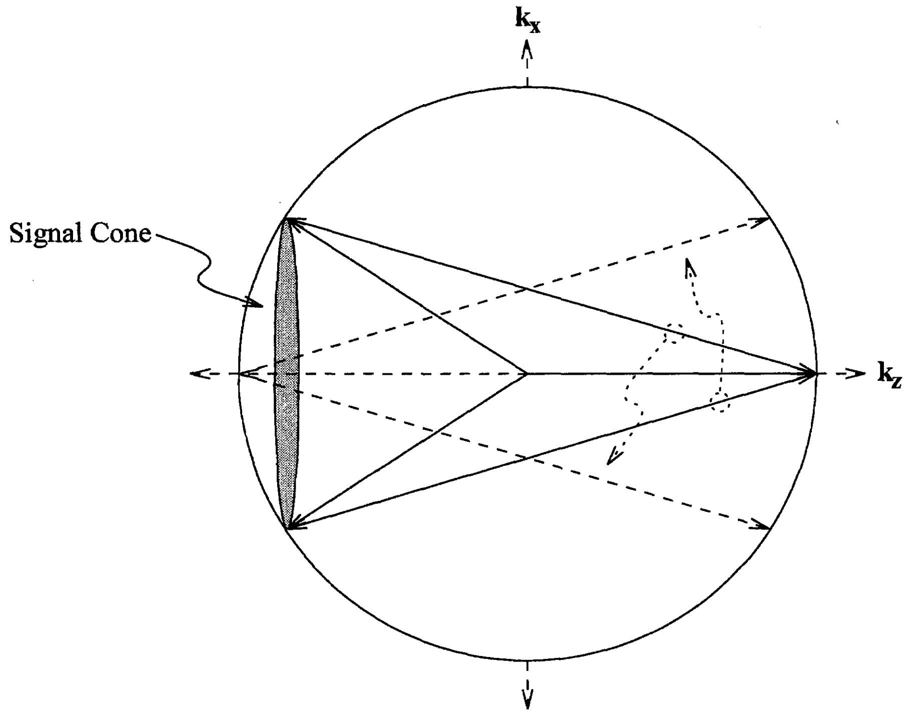

Figure 4.9: K-sphere diagram showing degenerate read-out of all the gratings by the center signal component.

of the correlations as described in chapter two. Bragg selectivity acts to suppress these sidelobes. In the case of the reflection geometry, however, these sidelobes can be Bragg-matched at certain locations, and therefore are not supressed. In practice, however, the sidelobes aren't very strong, even without Bragg-selectivity. Indeed, the affect of this Bragg degeneracy is not noticeable in normal correlator systems.

Given equation 4.14 for the Bragg-selectivity, figure 4.10 shows the correlation domain sizes for the reflection geometry. This figure was made assuming 8mm thick $LiNbO_3$.

The domains are large and symmetric in the center, as one would expect, and narrow progressively the farther the domain is from the center. Such an arrangement might be useful for an active vision system. In this case, the center correlations could be used as generic "blob" finders, utilizing their large shift-invariance to search the scene for objects of interest. This is analagous to the concept of saliency used in psychophyics to describe the way in which some objects in a visual scene grab an

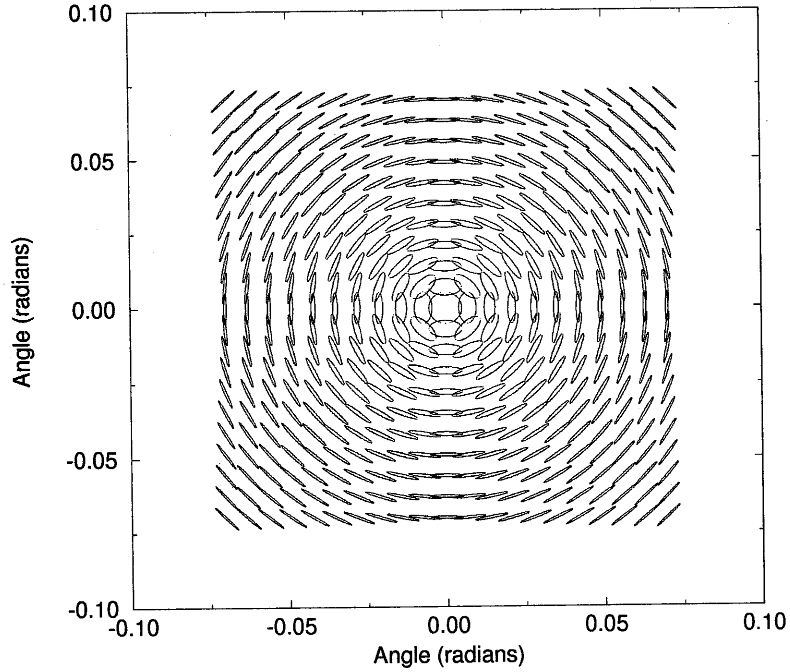

Figure 4.10: Shift invariance domains for a reflection geometry correlator stored in 8mm thick $LiNbO_3$.

observer's attention automatically. Once a salient object is found using the central correlation templates, the active vision system could then turn the camera to look at the object. With the object now well centered, the correlators farther from the center (which have poor shift invariance and therefore require a well-centered object) can be used for more specific object identification.

Chapter 5 Fresnel Correlators

Contents

Chapters two and three demonstrated how Bragg-selectivity can be used to perform the trade-off between shift invariance and the number of templates stored in a holographic correlator. However, the shift domains were either long and narrow, as in the case of the transmission and 90 degree geometries, or asymmetric and highly variable, as in the reflection geometry. This is a problem since most applications require symmetric and consistent correlation domains. The relatively weak control over the size of the domain in the out-of-plane direction also prevents us from storing more templates in this direction. In this chapter we present an alternative method [30,49] for controlling the shift invariance by shifting the hologram away from the Fourier plane, into the Fresnel region.

When the input image shifts in the optical correlator, the plane wave components at the Fourier plane all experience the same phase shift. This property results in shift-invariance for thin holographic correlators stored in the Fourier plane. If the hologram is recorded away from the Fourier plane, however, the phase shift is not uniform across all plane wave components. As a result, the various components of a shifted input image begin to add destructively and the correlation peak eventually disappears. The further the holographic material is from the Fourier plane, the bigger the phase difference between the various component plane waves and the more shift invariance is reduced.

Figure 5.1 shows the basic correlator system with the holographic material shifted a distance z_c from the Fourier plane. A transparency $f_1(x_1, y_1)$, illuminated by monochromatic light of wavelength λ produces the disturbance $g(x, y, z)$ in the Fresnel zone given (within the paraxial approximation and assuming z is small compared with F) by

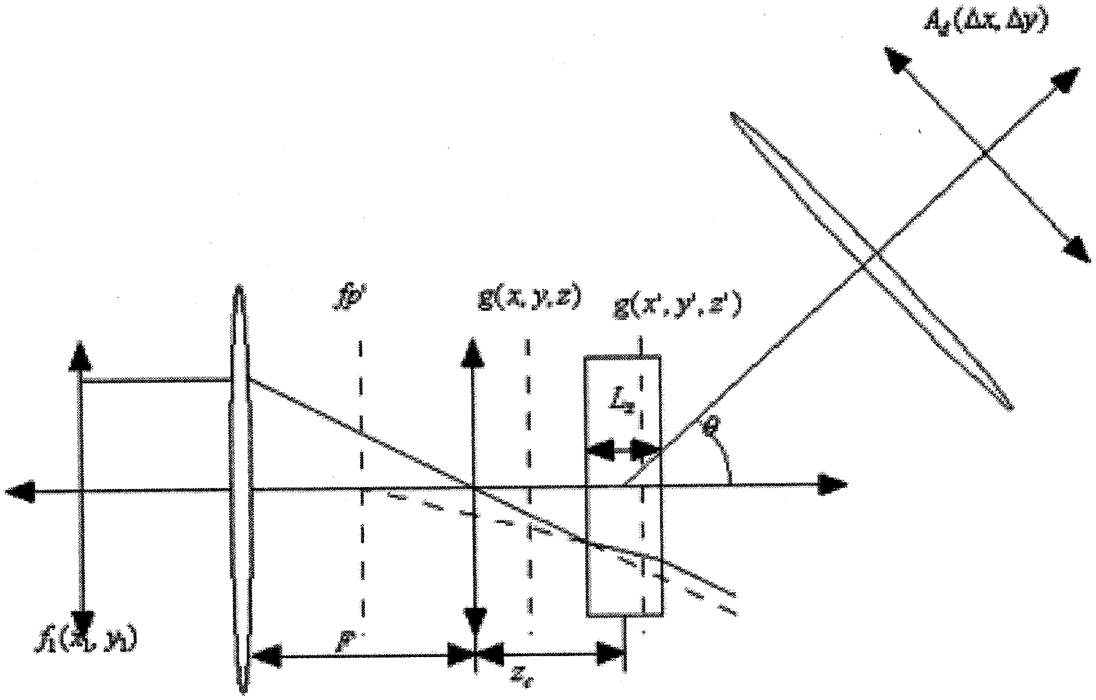

Figure 5.1: Basic Fresnel Correlator Arrangement

$$g(x,y,z) = \iint f_1(x_1,y_1) e^{-i\frac{k}{F}(x_1x+y_1y)} e^{i\frac{k}{2F^2}(x_1^2+y_1^2)z} dx_1 dy_1 \tag{5.1}$$

where $k = 2\pi/\lambda$, F is the focal length of the lens, and z is the distance from the Fourier plane. The refractive index, however, makes the Fourier plane appear to be shifted from its actual position. The field inside the material is then

$$g(x',y',z') = \iint f_1(x_1,y_1) e^{-i\frac{k}{F}(x_1x'+y_1y')} e^{i\frac{k}{2nF^2}(x_1^2+y_1^2)(z'+(z_c-L_z/2)(n-1))} dx_1 dy_1 \tag{5.2}$$

where z_c is the distance from the Fourier plane to the center of the recording material.

A plane wave reference incident at an angle θ records a hologram with the signal beam given by

$$|e^{ik(x'\sin\theta'+z'\cos\theta')} + g(x',y',z')| \tag{5.3}$$

where θ' is the angle of the reference beam inside the medium. Illuminating the hologram with a new input image, $f_2(x_2, y_2)$, produces the diffracted field amplitude A_d along the wave-vector $\mathbf{k_d}$ given (within the Born approximation) by

$$\begin{aligned} A_d(\mathbf{k_d}) \;&=\; \iint dx_1 dy_1 \iint dx_2 dy_2 f_1(x_1, y_1) f_2(x_2, y_2) \\ &\cdot \iiint_{volume} e^{-i\frac{k}{2nF^2}(x_1^2-x_2^2+y_1^2-y_2^2)(z'+(z_c-L_z/2)(n-1))} e^{i\frac{k}{F}(x_1-x_2)x'} e^{i\frac{k}{F}(y_1-y_2)y'} \\ &\cdot e^{in(k\sin\theta'-k_{dx})x'} e^{ink_{dy}y'} e^{in(k\cos\theta'-k_{dz})(z'+(z_c-L_z/2)(n-1))} dx'dy'dz' \end{aligned}$$

where k_{dx}, k_{dy}, and k_{dz} are the x, y, and z components of the diffracted wave-vector such that

$$k^2 = k_{dx}^2 + k_{dy}^2 + k_{dz}^2 \tag{5.4}$$

Performing the integral over the volume of the material with the change of variables

$$\tilde{z} = z' - z_c \tag{5.5}$$

yields

$$\begin{aligned} A_d(\mathbf{k_d}) &= \iint dx_1 dy_1 \iint dx_2 dy_2 f_1(x_1, y_1) f_2(x_2, y_2) e^{-i\frac{k}{2nF^2}(x_1^2-x_2^2+y_1^2-y_2^2)(z'+(z_c-L_z/2)(n-1))} \\ &\cdot e^{in(k\cos\theta'-k_{dz})(z_c+(z_c-L_z/2)(n-1))} \\ &\cdot \text{sinc}\left[\frac{L_z}{2\pi}\left(\frac{k}{2nF^2}((x_1^2-x_2^2+y_1^2-y_2^2)+kn\cos\theta'-nk_{dz}\right)\right] \\ &\cdot \text{sinc}\left[\frac{L_x}{2\pi}\left(\frac{k}{F}(x_1-x_2)+kn\sin\theta'-nk_{dx}\right)\right] \text{sinc}\left[\frac{L_y}{2\pi}\left(\frac{k}{F}(y_1-y_2)-nk_{dy}\right)\right] \end{aligned}$$

The transverse dimensions L_x and L_y are large enough to approximate the last two sinc functions as delta functions. Integrating over x_2, and y_2 yields

$$A_d(\Delta x, \Delta y) = \iint f_1(x_1, y_1) f_2(x_1+\Delta x, y_1+\Delta y) \tag{5.6}$$

$$\cdot\, e^{i(z_c+(z_c-L_z/2)(n-1))\alpha}\text{sinc}\left(\frac{L_z}{2\pi}\alpha\right) dx_1 dy_1 \tag{5.7}$$

where

$$\alpha = \frac{k}{2nF^2}\left(x_1^2-(x_1+\Delta x)^2+y_1^2-(y_1+\Delta y)^2\right)+kn\cos\theta'-nk_{dz} \tag{5.8}$$

$$\Delta x = -nF(\sin\theta'-k_{dx}/k) \tag{5.9}$$

$$\Delta y = -nFk_{dy}/k \tag{5.10}$$

Equation 5.6 is the cross-correlation between f_1 and f_2, with a sinc term from the Bragg-selectivity and an exponential term involving the position of the hologram, z_c. Both the exponential and the sinc term act as a window function on the correlation, attenuating the signal for non-zero values of Δx and Δy, *i.e.*, for deviations from the center of the correlation domain. For volume holograms recorded at the Fourier plane, the exponential term becomes identically equal to one and only the sinc term acts to limit shift invariance. Likewise, for thin holograms recorded away from the Fourier plane, the sinc term becomes negligible and the exponential term becomes the

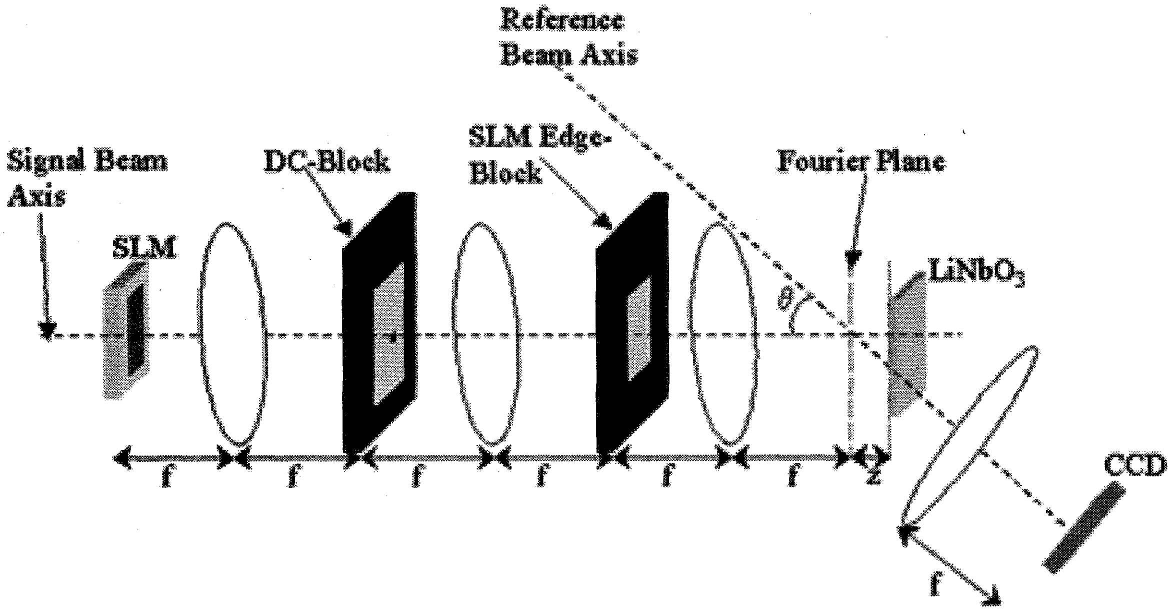

Figure 5.2: Experimental setup.

limiting factor. The presence of this window function within the integral also acts to sharpen the correlation peaks by suppressing the sidelobes, since they occur at non-zero values of Δx and Δy. Issues concerning shift invariance aside, the sidelobes of the correlations place an upper limit on how tightly the correlations can be packed. Attempting to store correlators too closely together results in the sidelobes from one template interfering with the neighboring template, reducing both outputs to noise.

The experimental setup for the correlator is shown in Figure 5.2. An image of random white and black rectangles, shown in Figure 5.3, was displayed on a portion

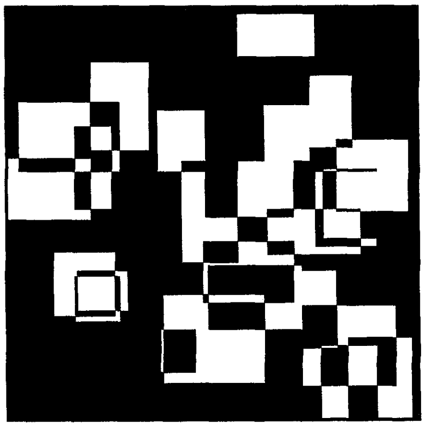

Figure 5.3: Object mask.

of the liquid crystal spatial light modulator, which has a resolution of 640 by 480 pixels and a 24μm pixel pitch. A DC block in the Fourier plane of the first lens edge enhances the image before correlation. The filter behind the second lens blocks the edges of the SLM, created by the edge-enhancement process of the DC-block. If not blocked, the SLM creates an undesirable constant DC offset to the strength of the correlation regardless of what image is presented on the SLM. The holographic material (a 250μm thick $LiNbO_3$ crystal) is mounted on a motorized translation stage so as to enable computerized control of the location relative to the Fourier plane. The signal beam is coincident with, and the reference beam at a 25° angle to, the recording material surface normal. A lens is placed along the path of the reference beam and in its back focal plane a CCD camera is used to capture the intensity and position of the correlation peak. The video signal from the CCD camera is digitized and analyzed by a computer.

For each hologram displacement distance z_c, a hologram of the input pattern centered on the SLM is recorded. After recording, the reference beam is turned off and the image on the SLM is correlated with the stored hologram. The input image is shifted, electronically, on the SLM. The image is first shifted horizontally (the in-plane direction) while centered vertically. For each horizontal location, the peak of the correlation and its location on the CCD is measured. The image is then shifted vertically (the out-of-plane direction) while centered horizontally, and again the peak intensity and position are measured. The correlation measurements are taken under very weak illumination to both prevent saturation of the CCD and erasure of the hologram.

Figures 5.4 and 5.5 show typical curves of peak intensity versus image location for both horizontal and vertical displacements. The shift-selectivity is measured as the width of the curve when it attains half of its maximum value. Plots of the shift-selectivity for both the in-plane and out-of-plane directions together with the theoretical predictions are shown in figures 5.6 and 5.7 as functions of the recording material location relative to the Fourier plane.The correlation integral derived in the previous section was computed with a Monte Carlo technique with experimental

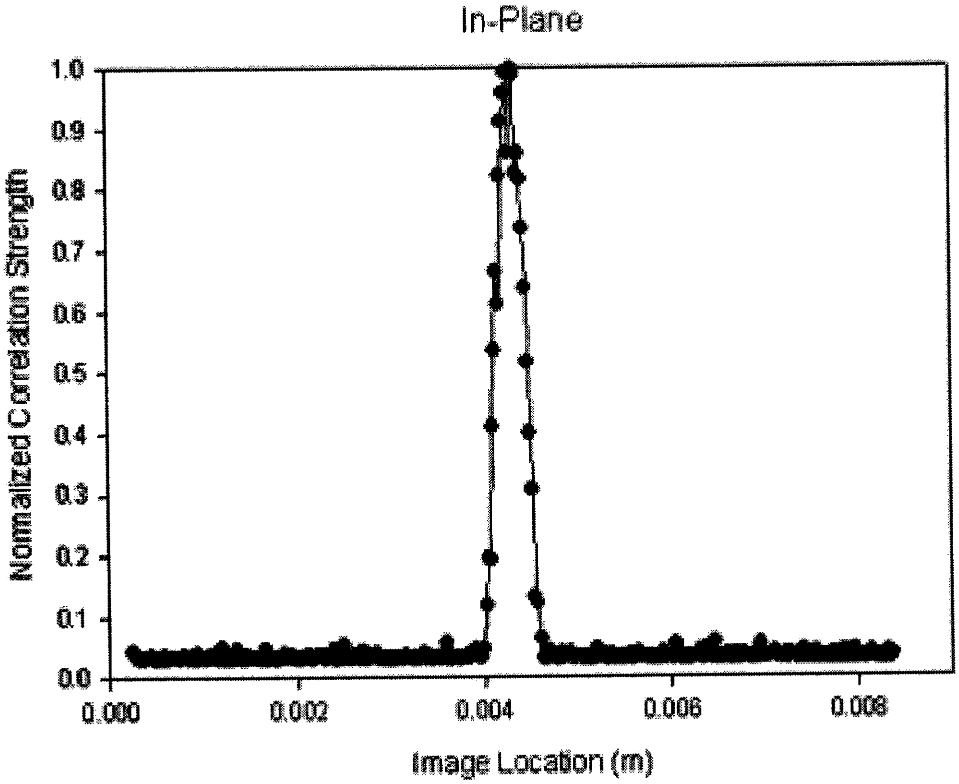

Figure 5.4: Peak Intensity vs. Object Shift (In-Plane).

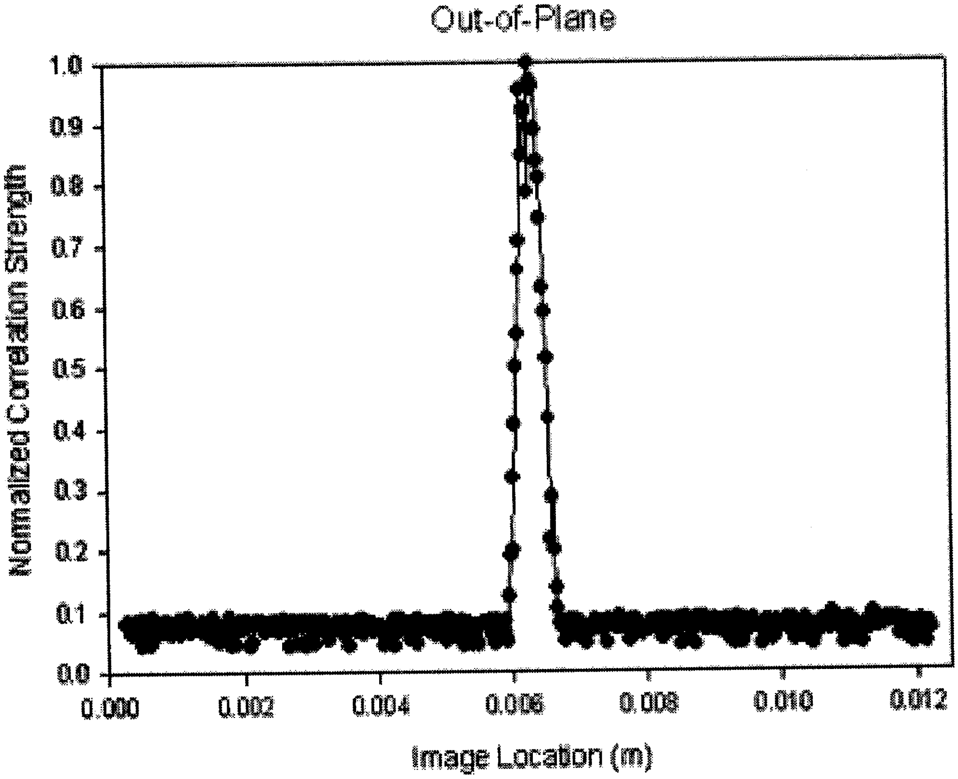

Figure 5.5: Peak Intensity vs. Object Shift (Out-of-Plane).

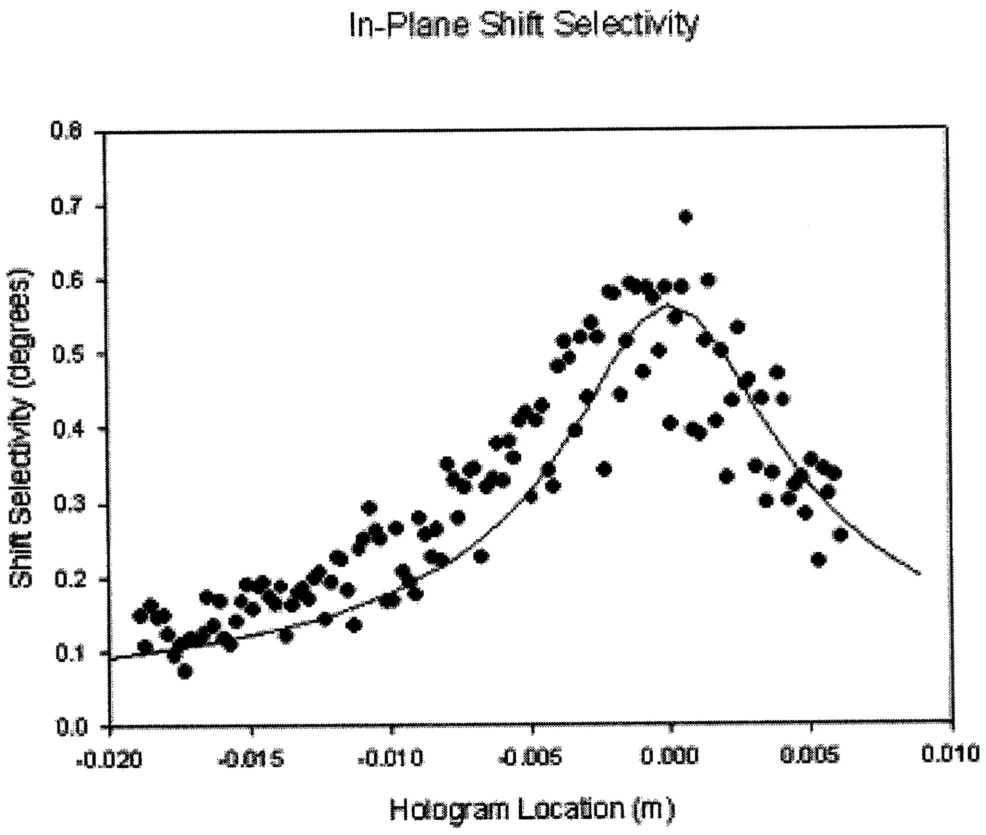

Figure 5.6: Shift Invariance, In-Plane.

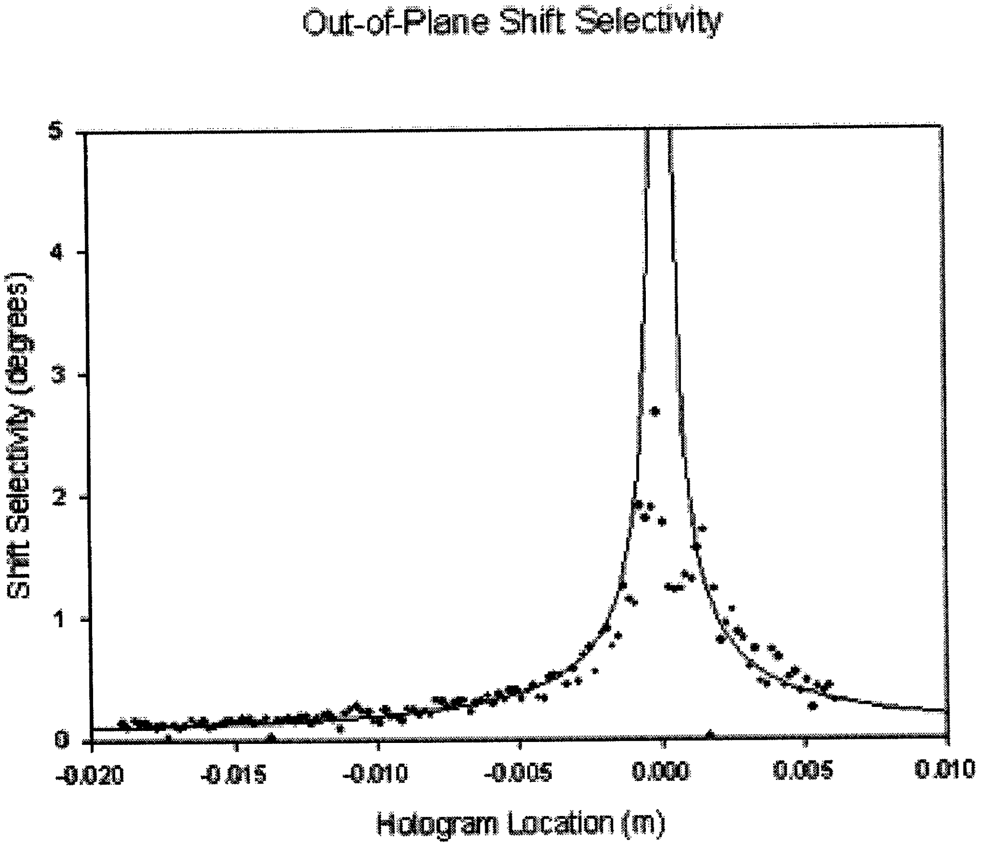

Figure 5.7: Shift Invariance, Out-of-Plane.

values for beam angle (25°), material thickness (250μm), and index of refraction (2.24). The experiment agrees well with the theoretical calculations over a large range of material displacements. Theory and experiment deviate most for out-of-plane shifts close to and at the Fourier plane, where the predicted value of the shift-invariance shoots up to 25°. The figure shown does not contain the full vertical range of the theoretical curve so that the details of the wings would be evident.

Figures 5.8a shows the output from an array of 81 correlators stored in 250μm $LiNbO_3$ displaced 1 cm in front of the Fourier plane. Only two different faces were used as templates, in an alternating fashion, so that the overall array size could be easily viewed. In this experiment the central reference beam angle was 50° and each reference beam was separated by 0.08°. Figure 5.8b shows the output when the input images are shifted just enough so that their correlation peaks would fall in the area reserved for the neighboring template; the peaks have disappeared, as intended, due to the positioning of the hologram in the Fresnel zone. Figure 5.8c shows the output when the holograms are stored and used in the Fourier plane. In this case, the Bragg-selectivity is not enough to prevent the sidelobes from interfering with neighboring templates and the output of the system is noisy even for well-centered input images. This shows that, using the Fresnel corrrelator system, more correlators can be stored than would be possible in the Fourier plane. Figure 5.9 shows cross sections of autocorrelations for both the Fresnel and Fourier plane holograms. The sidelobes of the Fresnel hologram are clearly suppressed relative to those for the Fourier hologram.

While it is possible to rely on Bragg-selectivity alone to control the shift invariance of holographic correlator systems, simply recording the holograms in the Fresnel zone allows for convenient control without the need to order material of the precise thickness necessary for a given application. Additionally, Fresnel correlators use thin holographic media. Currently, thin materials are available (*e.g.*, DuPont Photopolymer HRF-150) that are very sensitive with high diffraction efficiencies, and are capable of storing permanent holograms. It is interesting to note that, for large z_c, the constraint on the size of the shift domains is dominated by the exponential term. In other words, we can store more correlators than if we relied on Bragg selectivity

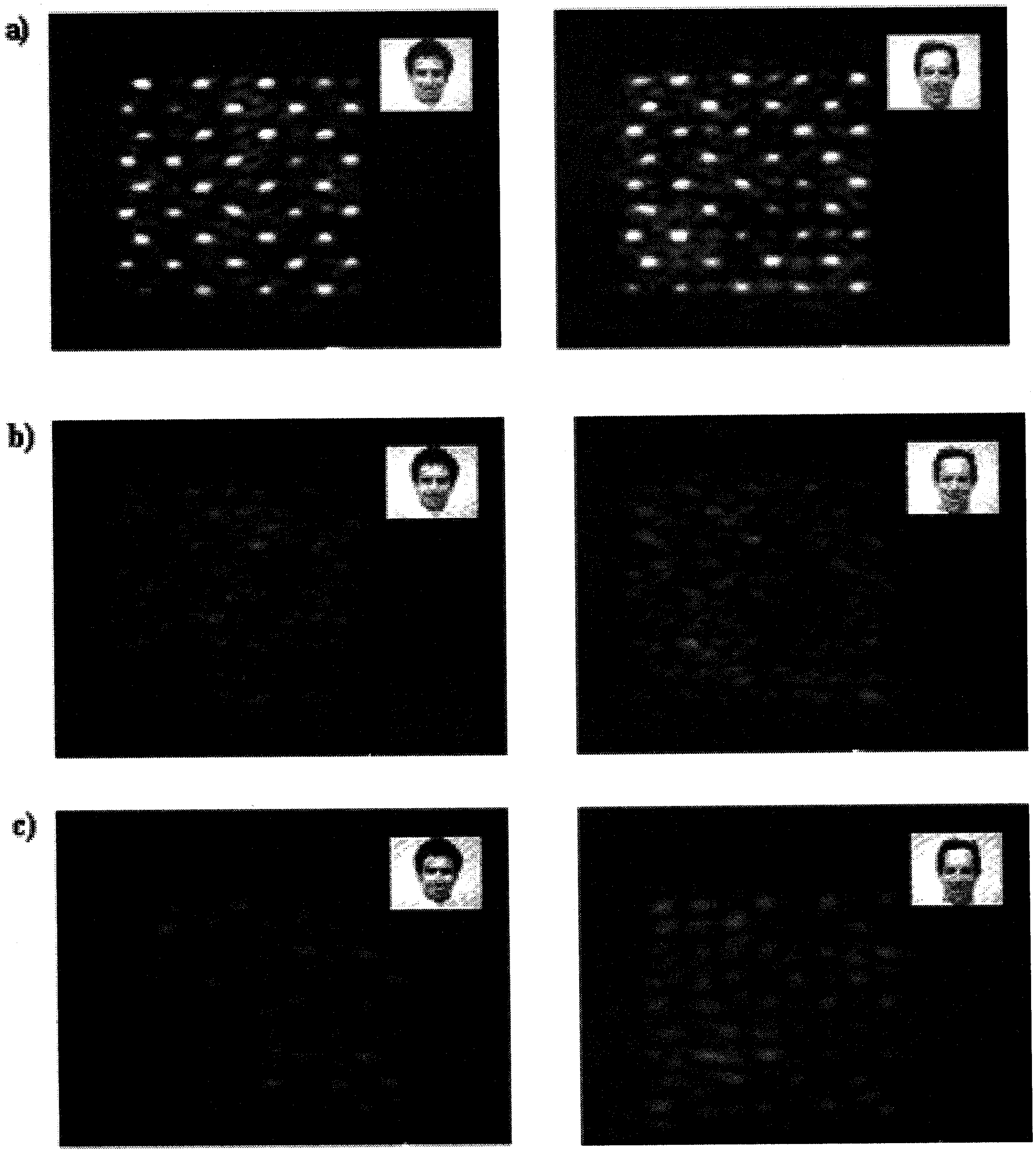

Figure 5.8: Correlation output for 81 stored templates. a) For original input images stored 1cm into the Fresnel zone. b) For shifted input images, stored 1cm into the Fresnel zone. c) Attempt to store the same templates in the Fourier plane.

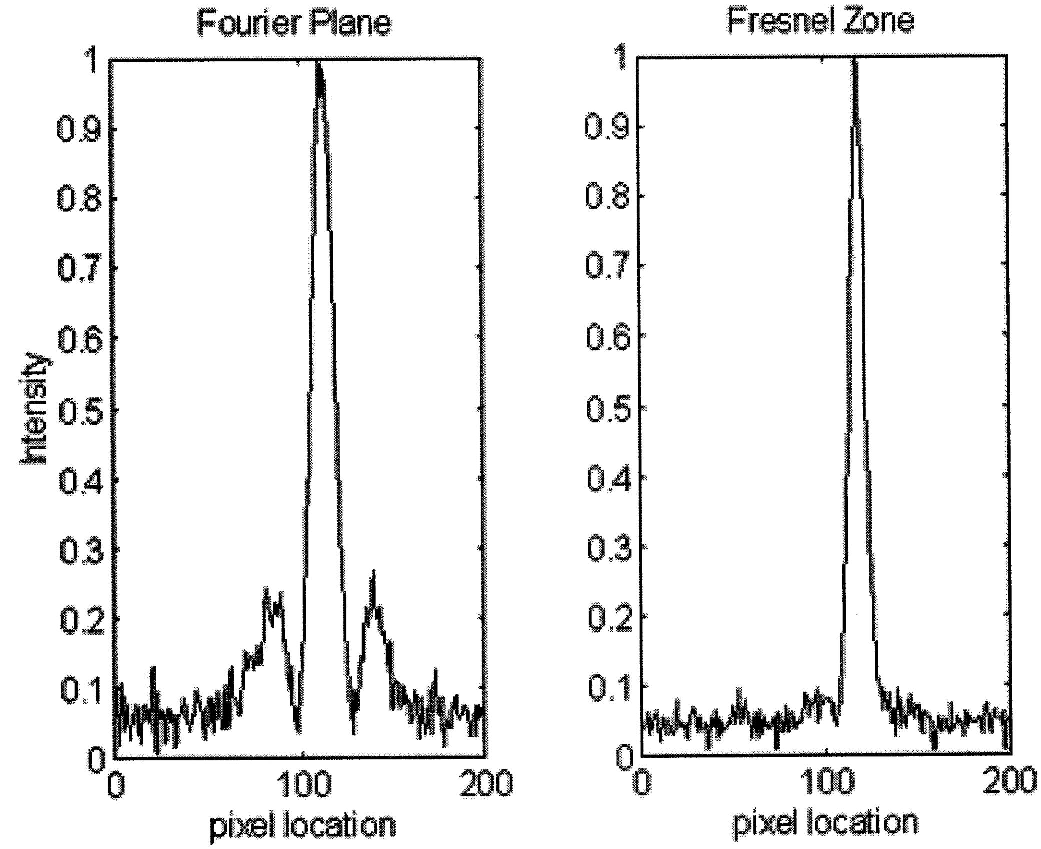

Figure 5.9: Autocorrelation peak cross sections of a face, for holograms resorded in the Fourier Plane and in the Fresnel region

alone. However, we could not read out these correlators as independant pages of a memory. In this case, the capacity of the correlator system is larger than a memory system that uses the same material! That this should be possible is partially supported by the observation that we can recognize more objects (such as faces) than we can freely recall.

Chapter 6 Shift Multiplexing

Contents

Angle and wavelength multiplexing were discussed in chapters three and four. For memories, both angle and wavelength multiplexing are capable pf very high capacities. However, they both suffer from some practical disadvantages.

Wavelength multiplexing is usually done in the reflection geometry for the highest densities. However, even small back-reflections of the reference wave off of the surface of the material can result in noise that can swamp the signal if many holograms are stored. For diffraction efficiencies in the range of 10^{-4} to 10^{-6}, even a good antireflection coating may not be enough to prevent the back-reflection from swamping the signal. This problem can be compensated in part by tilting the medium relative to the signal and reference waves, but this is not always easy to do for very high-bandwidth signals. Such signals require large tilts in order to move the back-reflection completely away from the reconstructed signal. Perhaps the most serious problem for wavelength multiplexing is the lack of good laser sources. A good source needs to have a very wide spectral tuning range, and it must be able to change lasing wavelength very quickly. Additionally, it should be cheap and compact for use in compact, disk architectures. Also, it needs to lase near 500nm, where current materials are sensitive. Currently, the best option for a wavelength tunable source is laser diodes with tunable Bragg reflectors. Several companies now offer such devices [51]. They can change their lasing wavelength within 20ns, and have 10 - 20nm total bandwidth. However, lack of demand makes them very expensive, with current prices near $20,000 a piece. Also, they currently lase near 1.5μm. Although they could probably be designed for shorter wavelengths, it will likely be several years before they will be available in the visible range, much less in the blue-green spectral range where current materials are sensitive.

Angle multiplexing suffers from the requirement for awkward beam steering devices [52]. Large lenses with good $f/\#$ are required to provide many angles with a reference beam of suitable width. These lenses are very expensive and heavy. This makes them particularly unattractive for use in disk systems. It is not possible to build a reading head for a disk system that contains two large lenses and a mirror for beam steering. Therefore, the tracking on the disk must be achieved by moving the disk radially, as opposed to moving a small reading head as is done for CD-ROMs. This arrangement causes mechanical difficulties that increase both the size and cost of a disk system.

I have invented a technique called shift multiplexing that avoids these difficulties. In shift multiplexing, the reference beam consists of a spherical wave with the focus just in front of, or behind, the hologram. Multiplexing is achieved by either shifting the hologram relative to the focused spot, or vice-versa. Typical shift distances can be on the order of several tens of microns. This technique is particularly appropriate for disks, in which the rotation of the disk serves as the shifting required for multiplexing. Such a system is shown in figure 6.1. Previous authors have treated holography with spherical reference beams [53–56]. The standard approach to volume holography is

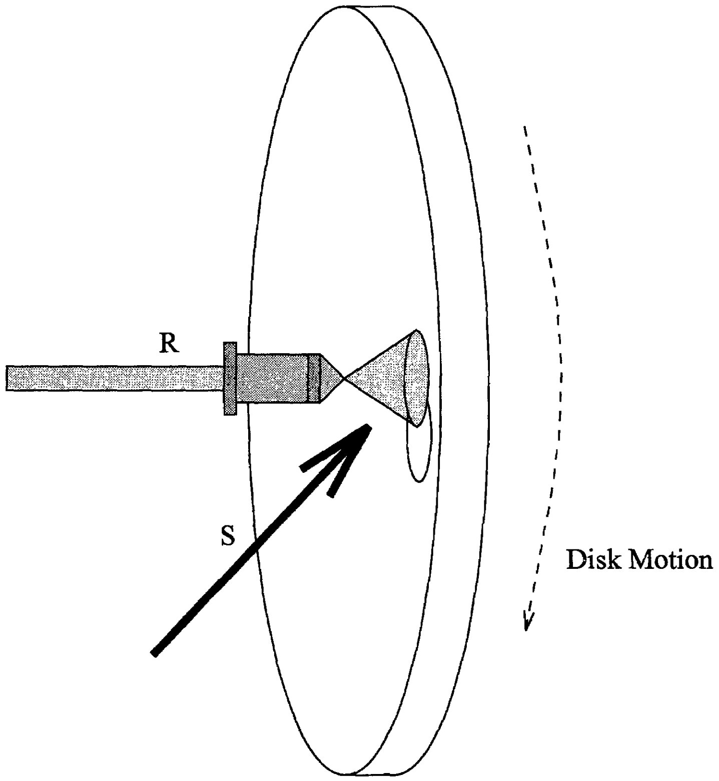

Figure 6.1: Shift multiplexed disk. The motion of the disk as it rotates serves to multiplex overlapping holograms.

to represent the spherical wave as an integral of a continuum of plane waves. In this case, a shift in the source of the spherical wave would be represented as a change in the relative phases of the component plane waves. In order to calculate the bragg sensitivity to shifts in the source, one would have to do an integral over the entire volume. This is non-intuitive, and requires many pages of algebra and assumptions to arrive at an analytical result. Kulich [56] described the wavefront of a spherical reference wave as being locally planar. However, he did not justify this approximation, nor did he consider the effect of changing the position of the reconstructing spherical wave.

In this chapter, I will first go back to the far-field approximations of Fraunhofer to justify the concept of a "local plane wave." I will then use this to trivially derive the same analytical form for shift selectivity that results from an otherwise lengthy and non-intuitive calculation. In the third section, I will go on to use the far-field concept to further derive a refined shift selectivity that is more accurate. Although the refined theory requires numerical integration, it only requires integration in one dimension. Using the standard approach would require a full volume integral that would require considerably more computational resources. In the last section of this chapter, I will analyze the capacity of memories using shift-multiplexing, and compare it with other techniques.

6.1 The Far-Field

To understand how shift multiplexing works, we need to go back to the Fresnel diffraction equation (see equation in chapter 2).

$$g(x',y',z') = \frac{e^{ikz}e^{i\frac{\pi}{\lambda z}(x'^2+y'^2)}}{i\lambda z}\iint f(x,y)e^{i\frac{\pi}{\lambda z}(x^2+y^2)}e^{-i\frac{2\pi}{\lambda z}(xx'+yy')}dxdy \tag{6.1}$$

If we look far away from the diffracting aperture, so that $z \gg x^2/\lambda$, then the

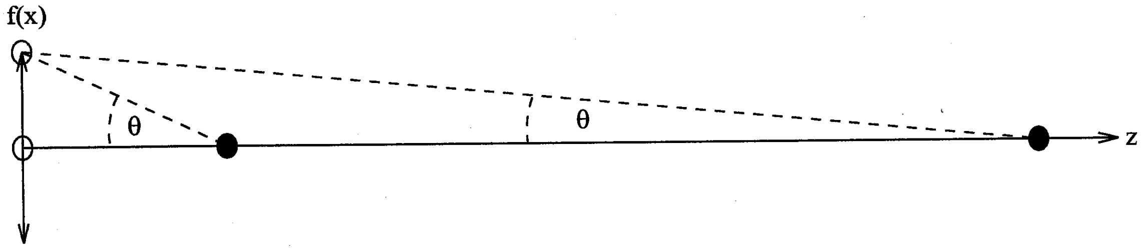

Figure 6.2: Optical path difference decreases with increasing distance from $f(x)$. The far-field is where the optical path difference becomes $\ll \lambda$.

quadratic exponential is approximately equal to one, and we are left with the Fraunhofer far-field expression

$$g(x', y', z') \approx \iint f(x, y) e^{-i\frac{2\pi}{\lambda z}(xx' + yy')} dx dy \tag{6.2}$$

But this is just a Fourier transform! In other words, if we look at the optical field far enough from the diffracting aperture, the field has the form of the Fourier transform of $f(x, y)$, scaled by $\lambda z / x$. This pattern will then continue to propagate relatively unchanged, with an ever-increasing scaling factor.

Figure 6.2 demonstrates how the Huygens-Fresnel wavelet interpretation can be used to understand the far-field approximation. For the point close to $f(x)$, the optical path length difference between spherical waves from the edges of $f(x)$ and those from the center varies more strongly as a function of distance than for points far from $f(x)$. This means that the relative phases of the spherical waves from each part of $f(x)$ change rapidly as one moves away from the aperture. However, for large enough distances, these phase differences change more slowly. The path difference can be approximated as the difference between z and $z \cos \theta$, where θ is the angle between a wave from the outer edge of $f(x)$ and a wave from the center of $f(x)$. As z grows, we can use the paraxial approximation for θ

$$\theta \approx \frac{x}{2z} \tag{6.3}$$

The optical path difference, OPD, is then

$$\text{OPD} = z(1-\cos\theta) \approx z(1-\cos(\frac{x}{2z})) \approx \frac{x^2}{4z} \qquad (6.4)$$

Once z gets large enough that the phase difference is much less than a wavelength, the pattern will cease to change since the amount of constructive and destructive interference between each of the Huygens-Fresnel spherical waves will not change. This happens when

$$z \ll \frac{x^2}{\lambda} \qquad (6.5)$$

which is exactly the approximation used to derive the Fraunhofer expression.

This has a particularly interesting interpretation in the Fourier domain. Each spatial frequency in $f(x,y)$ leads to a plane wave in the angular spectrum. These plane waves are, by definition, infinite in extent. However, according to the Fraunhofer equation, the amplitude of the wavefront in the far-field at any given location is proportional to only *one* plane wave component. The *infinite* plane wave contributes to the amplitude of the wavefront in only *one* location. This seems to be a paradox. How can it be infinite, yet appear in only one location?

Take, for example, two plane waves interfering on a screen. In this case, one observes a sinusoidal interference pattern. The plane waves are both infinite, but they have an effective amplitude at only certain locations. If one were to remove the screen, and place a small obstacle in one of the nulls of the intereference pattern, the two plane waves would propagate unaltered.

In the case of a far-field pattern, each plane wave component has an effective amplitude in just one location on the wavefront. All of the other component plane waves, taken together, interfere to produce a null at that location on the wavefront, just as the two plane waves produced many nulls in the above example. So the amplitude of the field at one point on the wavefront in the far field is due to only one plane wave component. Everywhere else on the wavefront, this component experiences destructive interference with the rest of the plane waves in the angular spectrum of the field.

We know that, if we place a small obstacle in front of the far-field wavefront, it will cease to qualify as a far-field pattern, and Fresnel diffraction from the obstacle will be observed. This would seem to contradict the above arguments. However, if we think about removing one plane wave component from the original, diffracting transparency, the size of the pattern on the transparency would become infinite because the plane wave is, by definition, infinite. This means that the location where our small obstacle is placed would no longer qualify as the far-field, because the original image size, x, has now expanded to infinity.

Thinking about the inverse problem, we can ask, what happens if we remove part of the angular spectrum at the source? We want to remove as little as possible, while still maintaining the aperture size so that the location of the far-field pattern does not change. If just this one band of spatial frequencies could be removed from $f(x)$, the only difference in the far field would be that absence of amplitude at one location on the wavefront. Another way to ask the same question is to ask at what location do the various bands of the angular spectrum occupy separate portions of the wavefront.

If we imagine $f(x)$ is an infinite pattern, followed immediately by an aperture the size of the true $f(x)$, then each plane wave component of the angular spectrum passes through this aperture (see figure 6.3). The diffraction from this aperture causes each component of the angular spectrum to be broadened as each plane wave expands due to diffraction. The angular broadening of each component is just given by

$$\begin{aligned}\frac{\sin\theta}{\lambda} &= \frac{1}{x}\\ \theta &\approx \frac{\lambda}{x}\end{aligned}$$

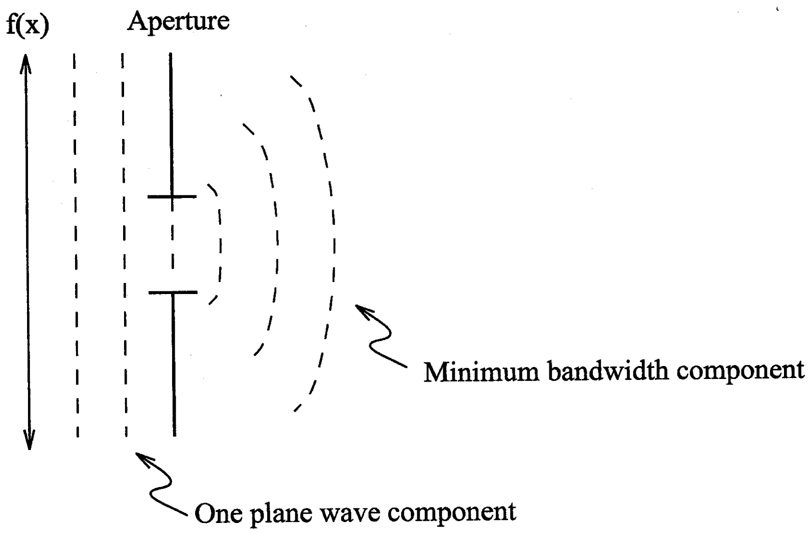

Figure 6.3: Minimum angular bandwidth component of $f(x)$ that can be removed without changing the far-field distance.

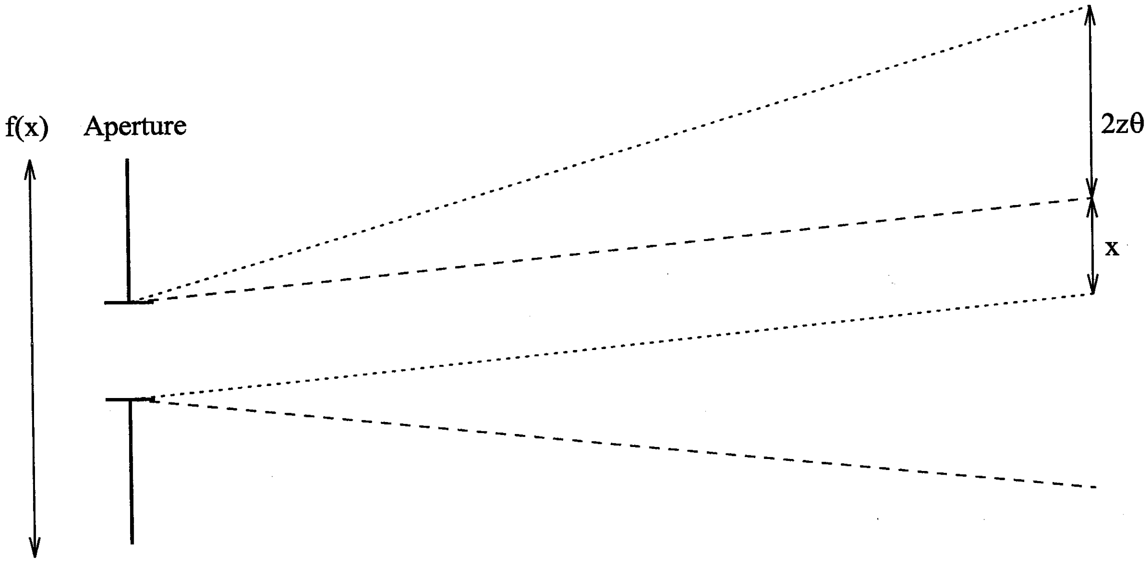

Figure 6.4: Divergence of components of the angular spectrum with propagation to the far-field.

This defines the minimum angular bandwidth that can be removed from the spectrum while still maintaining the aperture size. The lowest spatial frequency in $f(x)$ corresponds to the DC component before the aperture. The next smallest component of the angular bandwidth corresponds to a period of $x/2$ at the aperture and propagates at an angle of $2\theta = 2\lambda/x$ (see figure 6.4). Since both components shown in figure 6.4 pass through the same aperture, they are both broadened by $\theta = \lambda/x$. Therefore, the minimum component of the higher spatial frequency band is parallel to the highest component of the DC band. The distance between these components will remain equal to x for all z, as shown in the figure. However, each band will broaden by a factor of $2z\theta$. The different bands can be considered separated when their overlap is much smaller than their width. This occurs when

$$\begin{aligned} 2z\theta &\gg x \\ \frac{2z\lambda}{x} &\gg x \\ z &\gg \frac{x^2}{\lambda} \end{aligned}$$

Once again, we arrive at the same condition on z. This time by requiring that the different bands of the angular spectrum propagate until they are spatially separated from each other. In other words, by requiring that each spatial frequency occupies only one location on the wavefront.

For a tightly focused spot, the far-field approximation holds even at very short distances from the focus. If the spot size is on the order of ten wavelengths, then the far-field condition holds for distances longer than only 100 wavelengths. For 488nm light, this means the far-field is just 50μm away! Certainly, if we look at a small part of the surface of a sphere, it appears to be flat (after all, it took a long time before humans realized the Earth is round!). So, locally, a spherical wavefront looks like a plane wave. For holography, this means that diffraction from a far-field wavefront behaves as if it were diffraction from a plane wave propagating along the normal of the wavefront.

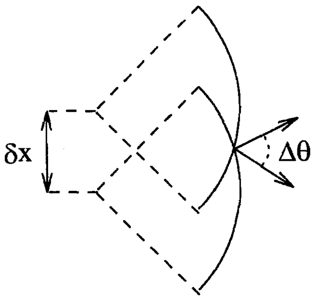

Figure 6.5: Shift as a change in angle.

6.2 Shift Selectivity

6.2.1 In-Plane Selectivity

The above arguments are the basis for shift multiplexing. Each part of the hologram behaves as if it were made with a plane wave reference beam whose direction is normal to the spherical wavefront. As the source of the wavefront shifts relative to the hologram, the angle of the apparent wavefront changes. This is shown in figure 6.5.

The change in angle, $\Delta\theta$, is just

$$\Delta\theta = \frac{\delta x}{z} \tag{6.6}$$

for $z \gg \delta x$. Substituting for $\Delta\theta$ in the normal Bragg selectivity curve yields directly an expression for shift selectivity

$$\delta x = \frac{z\lambda}{L \tan\theta_s} \tag{6.7}$$

assuming that the reference wave is normal to the hologram so that $\theta_r = 0$. If the reference beam is not normal, then we need to divide z by $\cos\theta_r$ to account for the actual distance from the focused spot to the hologram. We also need to account for the fact that the hologram is no longer shifting normal to the spherical wavefront, but at the angle θ_r from normal (see figure 6.6). Therefore, our final expression for

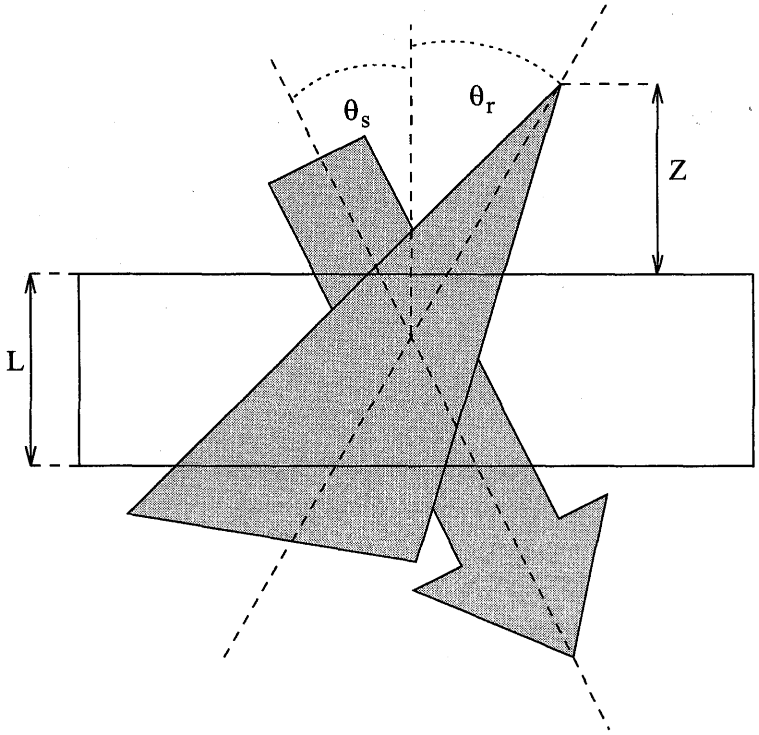

Figure 6.6: Geometry for shift multiplexing.

the shift selectivity is

$$\delta x = \frac{z}{\cos^2\theta_r} \cdot \frac{\lambda \cos\theta_s}{L\sin(\theta_s - \theta_r)} \tag{6.8}$$

The selectivity for the 90 degree geometry takes a particularly appealing form. As shown in figure 6.7, the path length that the signal takes depends on the distance, z, from the focused spot.

From the normal Bragg expression we have

$$\delta x = z\frac{\lambda}{L} \tag{6.9}$$

but now

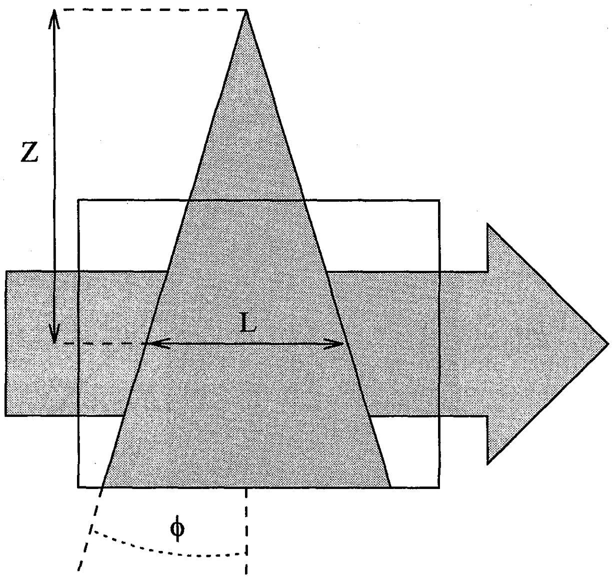

Figure 6.7: Geometry for shift multiplexing capacity in the 90° geometry.

$$L = 2z \tan \phi \tag{6.10}$$

where ϕ is the largest angle of the spherical wave (*i.e.*, numerical aperture = $n \sin \phi$). Thus,

$$\delta x = \frac{z\lambda}{2z \tan \phi} = \lambda f^{/\#} \tag{6.11}$$

In other words, the shift selectivity for the 90^{circ} geometry is just the spot size of spherical reference wave's focus. It is also interesting to note that the selectivity is now independant of the distance z.

6.2.2 Out-of-Plane Selectivity

Just as angle multiplexing has tangential selectivity for reference beam rotations out of the interaction plane, shift multiplexing has selectivity for shifts out of the interaction plane. Continuing with the shift-as-angle-change analogy, the out-of-plane

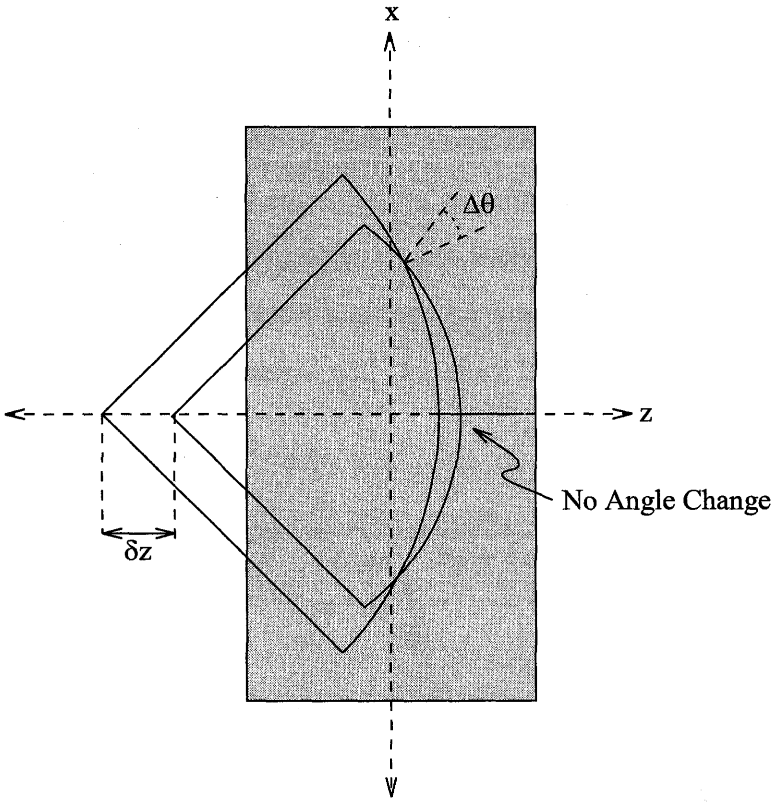

Figure 6.8: Z selectivity

shift selectivity will just be

$$\delta x \propto z_0\sqrt{\frac{2\lambda}{L}} \tag{6.12}$$

In practice, this means we can pack our tracks on a disk system so that the holograms between tracks actually overlap. As for angle multplexing, however, in practice this adds only a little bit to the overall density. Since this selctivity is similar for both systems, it will not be used in estimates of capacity, since it is not important for the comparison with angle multiplexing.

6.2.3 Z Selectivity

Having examined the selectivity for shifts in the x and y directions, we can also ask what happens if the focused spot is moved along the z axis, towards or away from the hologram. Once again, the k-sphere gives an immediate answer (see figure 6.8).

As the distance z_0 changes, locations on the wavefront see an effective angle $\theta_1 = x/z_0$. If the spot moves a distance δz along the z axis, the new angle will be $\theta_2 = x/(z_0 + \delta z)$. The net change in angle will therefore be

$$\begin{aligned} \Delta\theta &= \frac{x}{z_0} - \frac{x}{z_0 + \delta z} && (6.13) \\ &\approx \delta z \frac{x}{z_0^2} && (6.14) \end{aligned}$$

It is clear from equation 6.19 that the change in angle depends on the coordinate within the hologram, x. For locations at the edge of the hologram, the change in angle will be larger, and will lead to Bragg-mismatch faster, than for locations near the center. Indeed, the center portion of the hologram will *never* be Bragg-mismatched. For these reasons, shifts in z are not useful for multiplexing. Even at the edge of the hologram, bragg mismatch will be slow, as $x \ll z_0^2$. However, for disk systems this can actually be a good thing; accommodating for wobble as the disk spins would be made more difficult if the system were too sensitive to change in the distance between the disk and the read head.

6.3 Refined Shift Selectivity

Deriving the exact shift selectivity for a spherical reference wave with a plane wave signal requires relaxing our approximation that the change in angle, $\Delta\theta$, is uniform throughout the hologram. It is neither uniform through the depth of the hologram nor across the width of the reference beam. In order to take this non-uniformity into account, it is necessary to re-interpret our expression for the diffracted field, $A_d(\mathbf{k_d})$, as a path integral along the direction of the diffracted wave vector. We will then parameterize $\Delta\theta$ along this path and perform the integral numerically.

6.3.1 The Path Integral Interpretation

Our full expression for $A_d(\mathbf{k_d})$ is

$$A(\mathbf{k_i}, \mathbf{k_d}) = \iiint \frac{\Delta\epsilon(\mathbf{r}')e^{i(\mathbf{k_i}-\mathbf{k_d})\cdot\mathbf{r}'}}{2ik_{dz}} d\mathbf{r}' \tag{6.15}$$

In chapter two I discussed how $A(\mathbf{k_i}, \mathbf{k_d})$ can be interpreted as the three-dimensional Fourier transform of the hologram, $\Delta\epsilon$. However, writing the order of integration as

$$A(\mathbf{k_i}, \mathbf{k_d}) = \iiint \Delta\epsilon(x', y')e^{i(k_{ix}-k_{dx})x'}e^{i(k_{iy}-k_{dy})y'}dx'dy'\frac{\Delta\epsilon(z')e^{i(k_{iz}-k_{dz})z'}}{2ik\cos\theta_d}dz' \tag{6.16}$$

is more reminiscent of our intuitive picture of the Born approximation, in which successive thin slices of hologram scatter light along $\mathbf{k_d}$, and these scattered light components are integrated along the length of the material in the z direction. Of course, for most applications the transverse integrations are assumed infinite and are performed first, leaving the integration along the z-axis for last. The first two integrations, being infinite, lead to δ functions that uniquely determine θ_d in terms of the original reference and signal angles, θ_r and θ_s, and the change in the reference angle, $\Delta\theta_r$. This leads to

$$A(\mathbf{k_i}, \mathbf{k_d}) \propto \int \frac{e^{ik\frac{\Delta\theta_r\sin(\theta_s-\theta_r)}{\cos\theta_s}z'}}{k\cos\theta_d}dz' \tag{6.17}$$

It is clear that this is the origin of our Bragg selectivity expression for $\Delta\theta_r$. However, there is the factor of $k\cos\theta_d$ in the denominator. We can get rid of this factor by making a change of variables

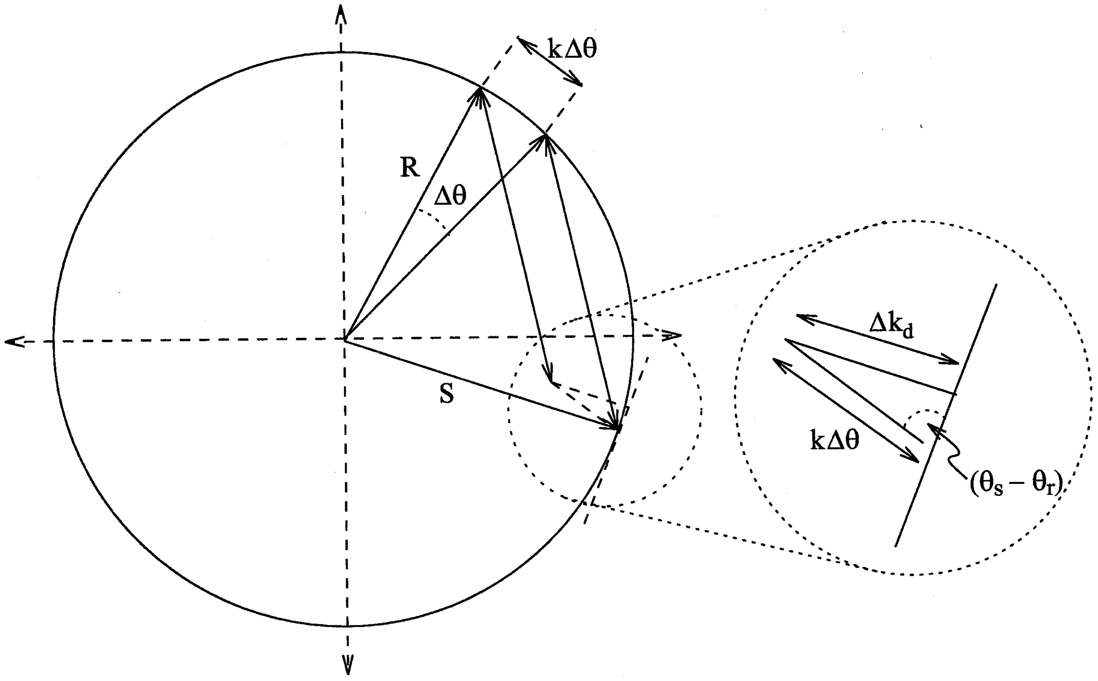

Figure 6.9: K-sphere diagram of the path integral interpretation.

$$z_d = \frac{z'}{\cos\theta_d} \tag{6.18}$$

$$dz_d = \frac{dz'}{\cos\theta_d} \tag{6.19}$$

making equation 6.22

$$A(\mathbf{k_i}, \mathbf{k_d}) \propto \int e^{ik\Delta\theta_r \sin(\theta_s - \theta_r) z_d} dz_d \tag{6.20}$$

Here we have illiminated the $\cos\theta_s$ term in the exponent because it is very close to $\cos\theta_d$. What is left is just an integral along the path of the diffracted wave. In essence, Bragg selectivity derives from an integration of the accumulated phase along the path of the diffracted beam. In almost all cases of practical interest, this path is a straight line within the hologram. The $\cos\theta_d$ term merely served to change the coordinate frame to the z-axis. With the aid of the k-sphere (figure 6.9) we can understand the remaining terms in this new light. In the exponent, the $\Delta\theta$ term represents the paraxial approximation to the distance travelled by the tip of the reference vector $\mathbf{k_i}$ as it rotates. The $\sin(\theta_s - \theta_r)$ term represents the projection of that motion onto the axis of the diffracted wave, which is approximately the same as the axis of the original signal wave. The $\cos\theta_s$ term in the exponent comes from the change of axis, but represents the fact that the path of integration is actually longer than L by a factor of approximately $\cos\theta_r$. We will now apply the path integral idea to a more accurate derivation of shift-selectivity. We will start with the 90° geometry, because it is the simplest. For the following derivations, the signal beam is assumed to be sufficiently narrow that the effective change in angle with shift is a constant across the diameter of the beam. For Fourier plane holograms, this is a good assumption since the signal beam can be quite narrow. In the following examples, the experimental configuration was chosen to push well beyond the limits of the approximate theory in order to adequately test the refined version.

6.3.2 The 90° geometry

Figure 6.10 shows the path of two rays of light as they enter the holographic material from the spherical reference wave from two different shift locations. We would like to integrate the phase along the path of the signal beam, the x-axis. However, because of Snell's law, there is no simple way to express the change of angle at a location in the crystal as a function of x. So we will need to perform a change of variables to allow us to integrate the variable θ_0, which is the external reference beam angle within the spherical wave. We want to compute the amplitude of the diffracted wave, $A_d(\Delta\theta)$ as

$$A_d(\Delta\theta) = \int_{-\theta_0^{max}}^{\theta_0^{max}} e^{-ik\Delta\theta_1(\delta x,\theta_0)x(\theta_0)\frac{\sin(\theta_1-\theta_s)}{\cos\theta_s}} \frac{\partial x}{\partial\theta_0} d\theta_0 \qquad (6.21)$$

where θ_0^{max} is the maximum angle in the reference wave and θ_1 is the internal angle given by Snell's law, $\theta_1 = 1/n \arcsin(\theta_0)$. The partial derivative of x by θ_0 is

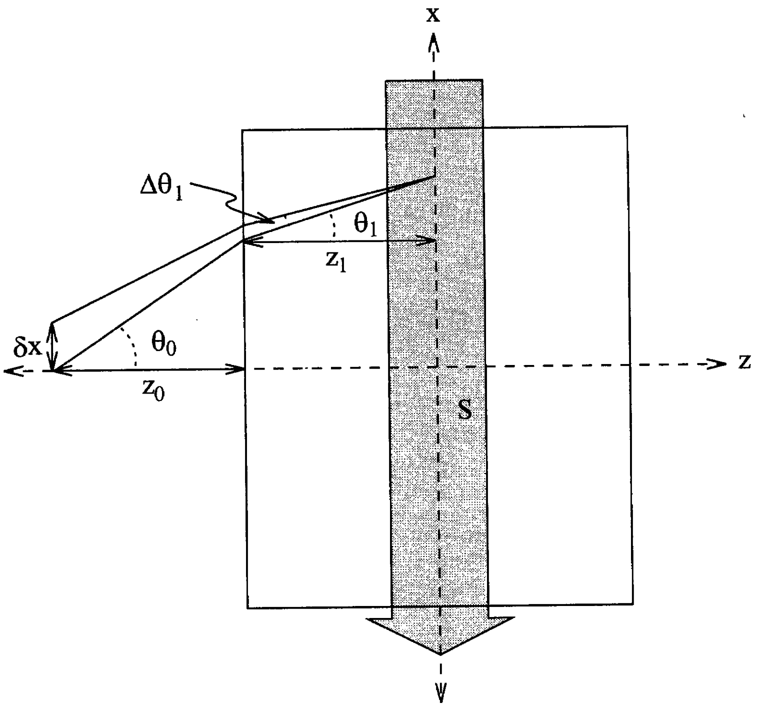

Figure 6.10: Geometry for refined shift multiplexing calculation (90° geometry).

just the Jacobian required by the change of variables.

For the 90° geometry, the trigonometry in the exponent simplifies to

$$A_d(\Delta\theta) = \int_{-\theta_0^{max}}^{\theta_0^{max}} e^{-ik\Delta\theta_1(\delta x,\theta_0)x(\theta_0)\cos\theta_1} \frac{\partial x}{\partial\theta_0} d\theta_0 \tag{6.22}$$

If z_0 is the distance from the focus to the medium, and z_1 is the distance from the surface of the medium to the signal beam, then

$$x(\theta_0) = z_0 \tan\theta_0 + z_1 \tan\theta_1. \tag{6.23}$$

From here, we can solve for $\Delta\theta_1(\delta x, \theta_0)$ as:

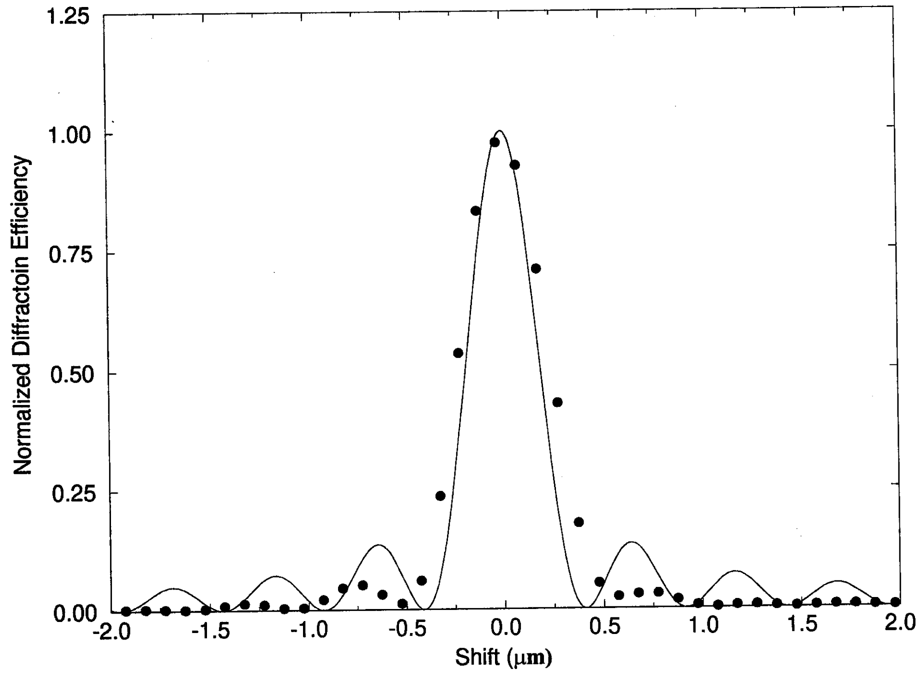

Figure 6.11: Selectivity curve for the 90° geometry. Recorded in 8mm thick $LiNbO_3$ with $z_0 = 2mm$. The numerical aperture was 0.65.

$$\begin{aligned}
x - \delta x &= z_0 \tan(\theta_0 + \Delta\theta_0) + z_1 \tan(\theta_1 + \Delta\theta_1) \\
\delta x &= z_0 (\tan\theta_0 - \tan(\theta_0 + \Delta\theta_0)) + z_0 (\tan\theta_1 - \tan(\theta_1 + \Delta\theta_1)) \\
&\vdots \\
\Delta\theta_1 &= \frac{\delta x \cos^3\theta_0 \cos^2\theta_1}{\frac{\delta x}{2}(\cos^3\theta_0 \sin 2\theta_1 + n\cos^3\theta_1 \sin 2\theta_0) - (z_0 n \cos^3\theta_1 + z_1 \cos^3\theta_0)}
\end{aligned}$$

Here we have dropped terms that are second order in $\Delta\theta_0$.

Figure 6.11 shows the curve for a shift multiplexed hologram in the 90^{circ} geometry. The signal consists of a single plane wave, the hologram was recorded in 8mm thick $LiNbO_3$, and a microscope objective with numerical aperture of 0.65 was used to focus the reference wave 2mm in front of the crystal. The location of the zeros

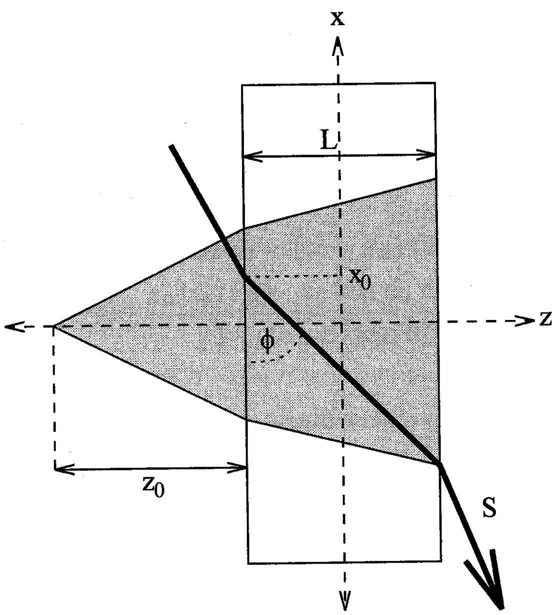

Figure 6.12: Geometry for refined shift multiplexing calculation (transmission geometry).

is slightly broader than the theoretical prediction, and the sidelobes are somewhat smaller. The smaller sidelobes are probably due to a scattering noise floor, while the underestimate of the shift distance may come from absorption by the crystal or inaccuracies in the experimental setup. The apporoximate theory from the previous section would predict a shift selectivity of just 285nm.

6.3.3 The Transmission Geometry

Figure 6.12 shows the geometry for a transmission hologram. In this case, the path is diagonal, through the material. Once again, we will change variables to integrate over the reference beam angle. We need expressions for $z_1(\theta_0)$ to plug into equation 6.28, and we need $x(\theta_0)$ for the integral equation 6.26.

Solving for both:

$$\begin{aligned} x &= z_0 \tan\theta_0 + z_1 \tan\theta_1 & (6.24) \\ z_1(\theta_0) &= (x - x_0)\tan\phi_s & (6.25) \\ x &= z_0 \tan\theta_0 + (x - x_0)\tan\phi_s \tan\theta_1 & (6.26) \\ x(\theta_0) &= \frac{z_0 \tan\theta_0 - x_0 \tan\phi_s \tan\theta_1}{1 - \tan\phi_s \tan\theta_1} & (6.27) \end{aligned}$$

where ϕ_s is the internal signal beam angle relative to the surface of the medium and x_0 is the entrance point of the signal beam. Although, ideally, the signal beam would enter at one edge of the reference beam fan, and exit at the opposite edge, in practice this will not always be the case. For the following experiment, the signal beam was aligned so that it coincided with the edge of the reference beam upon *exiting* the crystal, so that

$$x_0 = z_0 \tan\theta_0^{max} + L \tan\theta_1^{max} - \frac{L}{\tan\phi_s} \qquad (6.28)$$

The limits of integration for equation 6.23 are now from $-\theta_0^{max}$ to $\arctan(x_0/z_0)$.

Figure 6.13 shows the results for a transmission geometry, shift multiplexed hologram in 8mm $LiNbO_3$ with $z_0 = 5mm$ and a numerical aperture of 0.65. The signal beam was a single plane wave, incident at an external angle of 66°, with a value of $x_0 = 7.3mm$. The reference wave was normal to the crystal. The approximate theory predicts the shift selectivity should be 134nm.

6.4 Memory Capacity

6.4.1 Shift Multiplexed Disk

In calculating the capacity for a shift-multiplexed disk, we must once again take into careful consideration the physical size and shape of the hologram in the material. In previous chapters, the capacity was divided into an areal and thickness capacity.

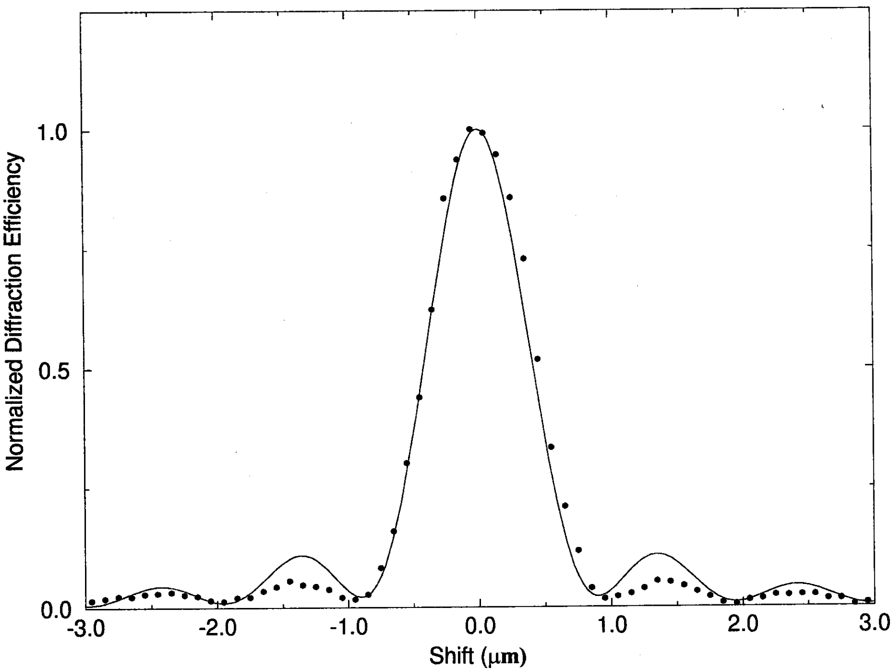

Figure 6.13: Selectivity curve for the transmission geometry. Recorded in 8mm thick $LiNbO_3$ with $z_0 = 5mm$ and $\theta_s^{ext} = 66°$ the numerical aperture was 0.65, and $x_0 = 7.3mm$.

Although it is possible with shift multiplexed disks, such a division is not as relevant because the entire disk, and hence the area occupied by each hologram, moves with each successive hologram. It is better to calculate the bits per hologram, and the shift distance per hologram. Deriving the bits per unit area is then just a matter of knowing how closely packed the tracks are. Because disks can often be quite thin, we will use the approximate shift selectivity for deriving our results, keeping in mind that this approximation becomes poor as the shift selectivity approaches λ.

The number of bits per hologram is just W_s^2/b^2 for image plane, and d^2/p^2 for Fourier plane holograms. The shift distance is determined by the distance, z, from the focused spot to the hologram. The focus must be far enough from the hologram to allow the reference wave to overlap entirely with the signal beam. The smaller the area of the hologram and the larger the numerical aperture of the spherical wave, the smaller z can be and the better the shift selectivity, δx.

Figure 6.14 shows the configuration for a shift multiplexed disk with the focus

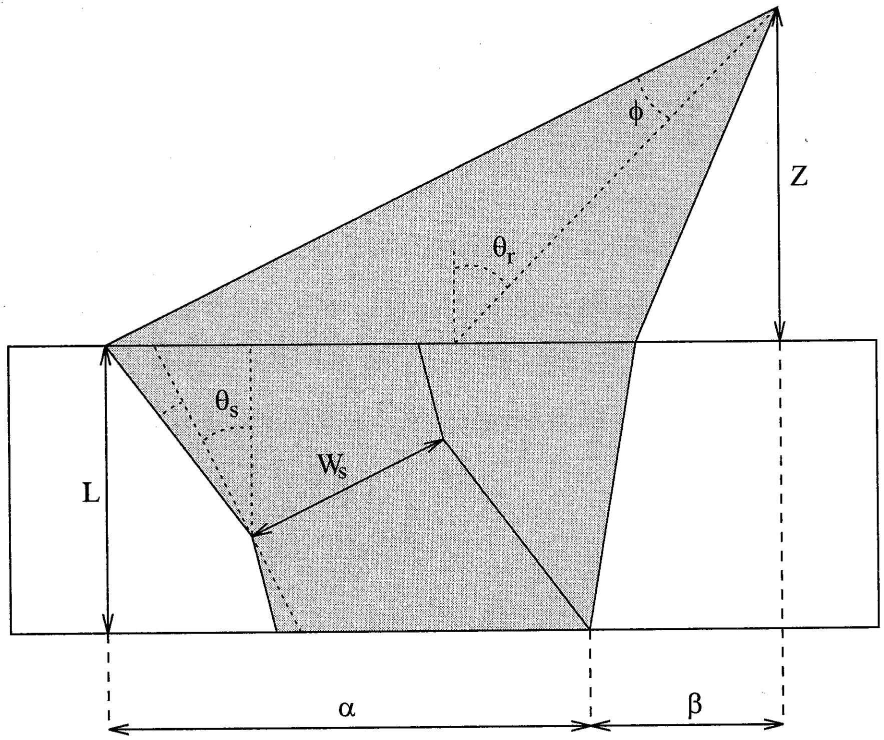

Figure 6.14: Geometry for shift multiplexing capacity for the disk geometry

in front of the disk with a numerical aperture = $\sin\phi$. Although the configuration shown is for the image plane, the distance α is the same as was computed in chapter three (see equations 3.25 and 3.29) and depends on whether it is a Fourier plane or image plane hologram. The figure shows

$$\alpha+\beta = z\tan(\theta_r+\phi) \tag{6.29}$$

$$\beta = z\tan(\theta_r-\phi)+L\tan\left\{\arcsin\left[\frac{1}{n}\sin(\theta_r-\phi)\right]\right\} \tag{6.30}$$

Subtracting gives

$$\alpha = z[\tan(\theta_r+\phi)-\tan(\theta_r-\phi)]-L\tan\left\{\arcsin\left[\frac{1}{n}\sin(\theta_r-\phi)\right]\right\} \tag{6.31}$$

Solving for z results in

$$z = \frac{\alpha + L \tan\left\{\arcsin\left[\frac{1}{n}\sin(\theta_r - \phi)\right]\right\}}{\tan(\theta_r + \phi) - \tan(\theta_r - \phi)} \tag{6.32}$$

The refractive index of the material, however, will make the focus actually appear to be farther away than it really is. Although the apparent distance will vary across the aperture due to Snell's law, we will approximate it by analyzing the central ray of the reference wave. In this case, the apparent distance from the center of the hologram to the focus, z', is

$$z' = z\frac{\tan\left(\frac{\pi}{4} - \theta_r^{int}\right)}{\tan\left(\frac{\pi}{4} - \theta_r^{ext}\right)} + \frac{L}{2} \tag{6.33}$$

where θ_r^{ext} and θ_r^{int} are the external and internal central reference beam angles such that $n\sin\theta_r^{int} = \sin\theta_r^{ext}$. The shift selectivity is then

$$\delta x = \frac{z'\lambda\cos\theta_s}{nL\cos^2\theta_r\sin(\theta_s - \theta_r)} \tag{6.34}$$

Figure 6.15 shows the shift selectivity for image and fresnel zone disks as a function of thickness, L. The shift selectivity is better for fresnel zone holograms because they have a smaller area, so z_0 can be smaller.

Figures 6.16 and 6.17 show the surface densities for the image and fresnel zone disks. Plots of the surface densities for angle and wavelength multiplexing are included for comparison. Shift multiplexing achieves higher densities than angle multiplexing in two ways. First, the holograms all overlap so there is no "dead volume" in between holograms as in the angle multiplexing case. This effect is more pronounced for the fresnel zone disk because the "dead volume" is bigger. Second, shift multiplexing uses the entire bandwidth of the reference lens, whereas angle multiplexing loses angular bandwidth because it needs to fit the width of the reference beam into the lens.

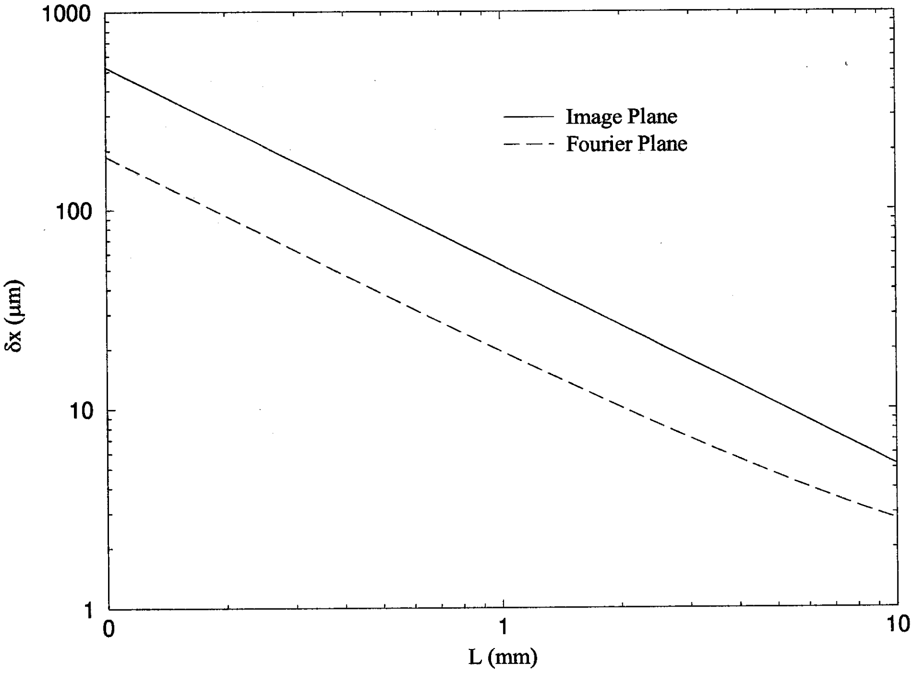

Figure 6.15: Shift selectivity for image and Fresnel holographic disks

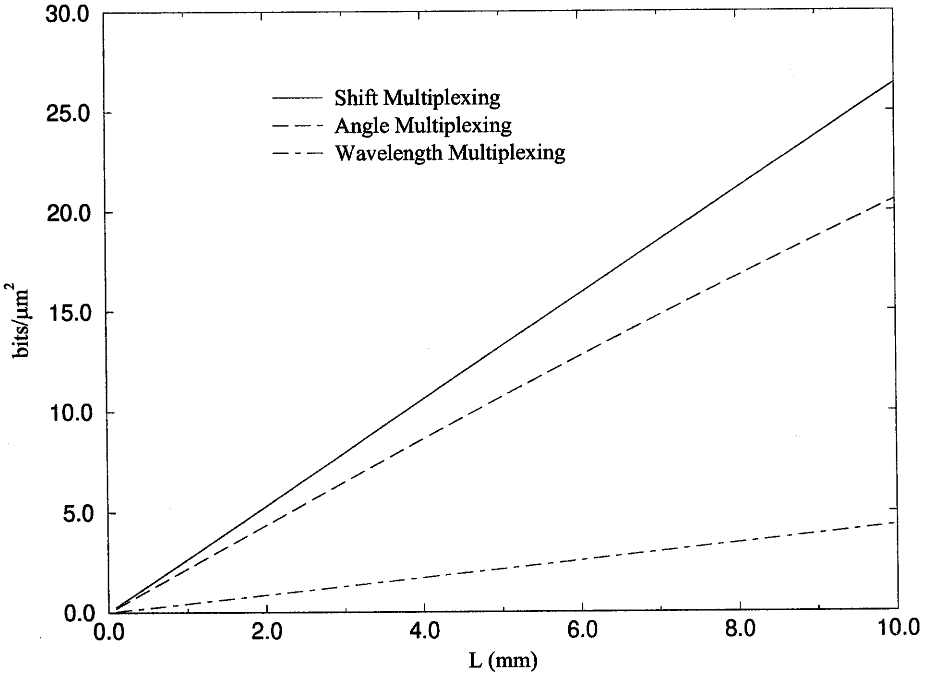

Figure 6.16: Surface density for image plane holographic disks.

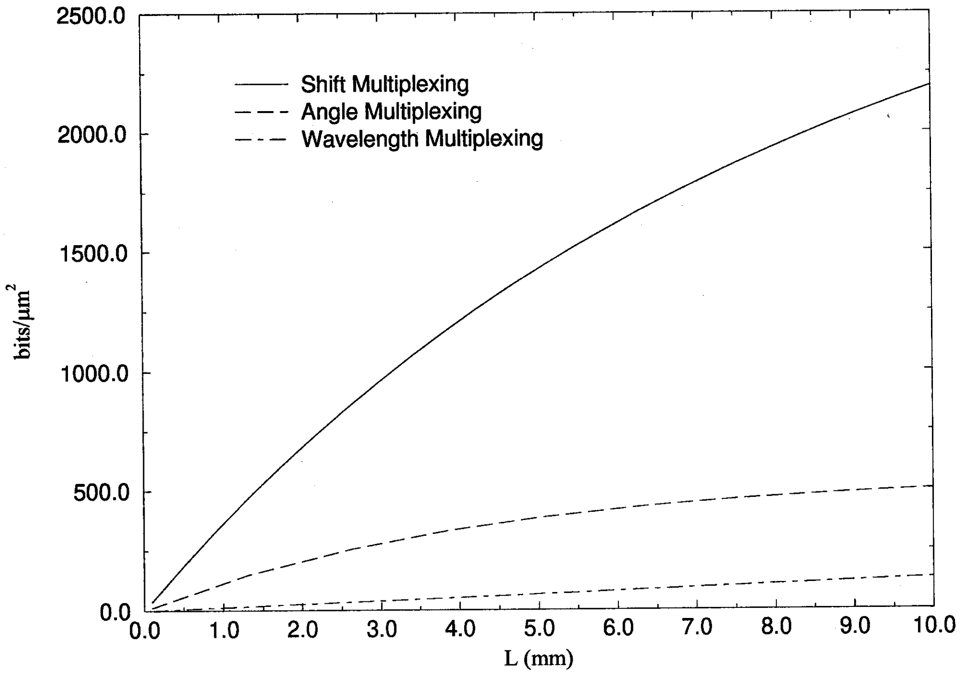

Figure 6.17: Surface density for Fresnel zone holographic disks.

For these calculations, the same reference lens was used for both the angle and shift multiplexed disks. Again, the external reference and signal beam angles were assumed to be ±40°.

6.4.2 The 90° geometry

For the 90° geometry, we cannot use the approximate solution. Therefore, we will use the results from the refined theory, and assume a numerical aperture of 0.65 giving a shift selectivity of 500 nm. We will assume, again, that holograms are stored at the second null. The derivation of the area required for the signal beam is the same as for angle multiplexing, and is given by equation 3.7 in chapter three. Here, L is no longer a function of other paramters and is taken to be the length of the crystal along the signal axis. In the case of shift multiplexing, however, we need to fit extra area on top of and below the signal beam in order to assure that the reference beam overlaps the entire signal. This is shown in figure 6.18.

In this case, the height, H, of the hologram is given by

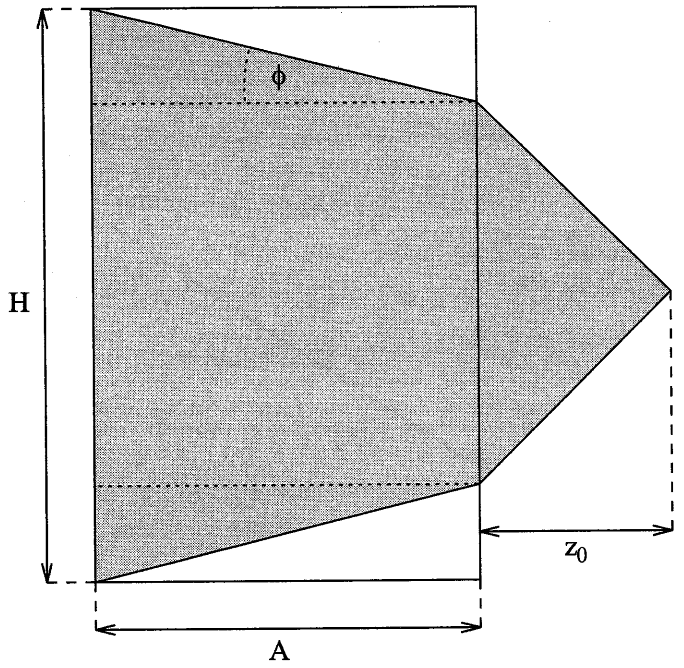

Figure 6.18: Area required for signal beam and reference wave to overlap for shift multiplexing in the 90° geometry. The signal beam is propagating into the page, with the reference beam as shown.

$$H = A + 2\tan(\phi) \tag{6.35}$$

where A is the width of the signal beam entering the crystal as given in chapter three, and ϕ is the internal maximum angle of the reference wave. The areal densities for the fresnel and image plane holograms are shown in figures 6.19 and 6.20. As for angle multiplexing, the areal density is much higher for the Fresnel zone than for the image plane. Again, this is because we can use large SLMs but there are no SLMs available with very small pixels.

The thickness density for shift mutliplexing would be just $1/\delta x$, except that we need to take into account the added length required at the ends to accomodate the reference beam. As in the case of figure 6.18, the added length is just $2\tan(\phi)$. The thickness density, $\mathcal{D}_{\mathcal{L}}$, is then

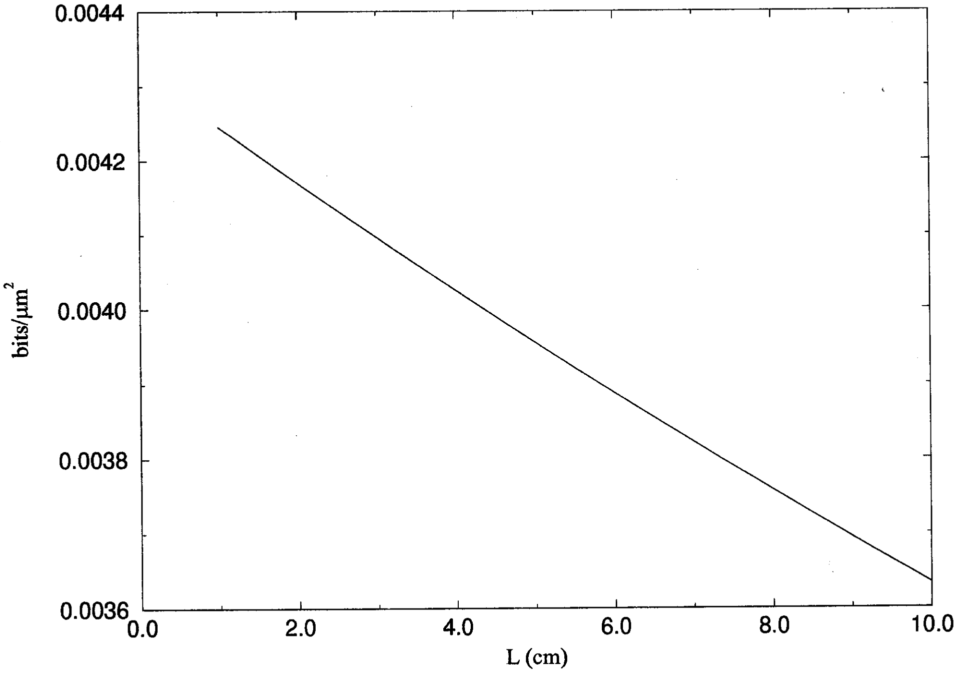

Figure 6.19: Areal density for shift multiplexing in the 90° geometry at the image plane.

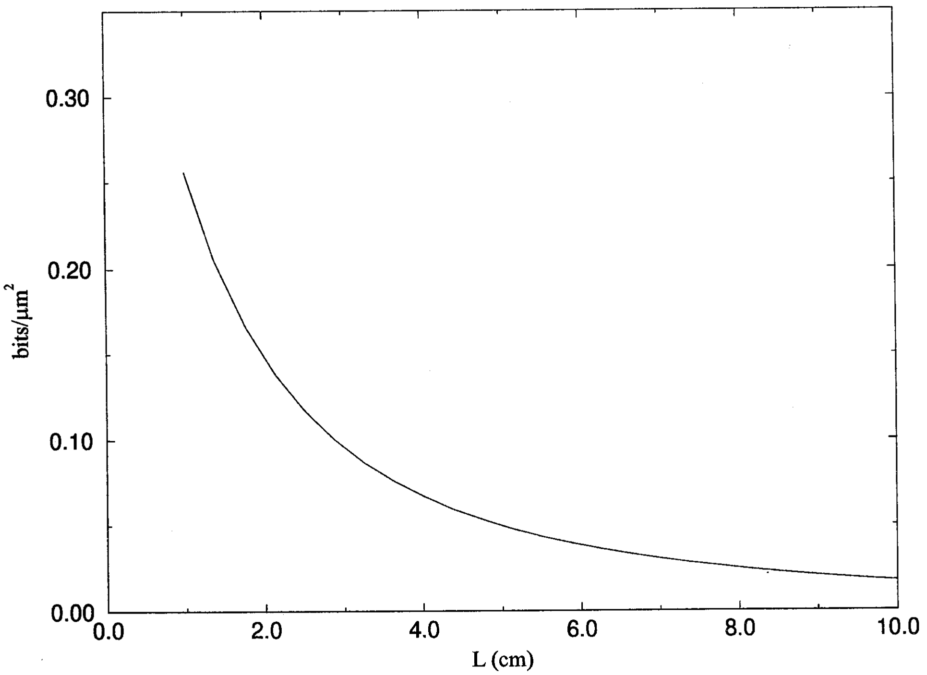

Figure 6.20: Areal density for shift multiplexing in the 90° geometry in the Fresnel zone.

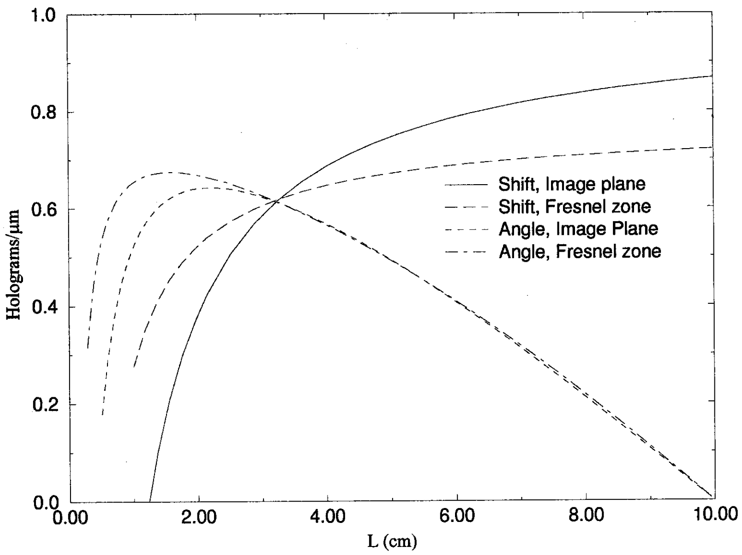

Figure 6.21: Thickness density for shift multiplexing in the Fresnel and image planes.

$$\mathcal{D}_L = \frac{L - 2A\tan(\phi)}{m\delta x L} \approx \frac{1}{m\delta x} - \frac{2}{m\delta x}\left(\frac{\lambda}{nb} + \frac{W_s}{L}\right)\tan\phi \qquad (6.36)$$

where we will take $m = 2$. Figure 6.21 shows the curves for both the Fresnel and image plane geometries. As L increases, the edge effects become more negligible, and the density saturates. The saturation value is

$$\mathcal{D}_L^{sat.} = \frac{1}{m\delta x}\left(1 + \frac{2\lambda}{nb}\right) \qquad (6.37)$$

The saturation value decreases with decreasing pixel size. For thinner materials, the cross section of the Fresnel holograms is smaller than for image plane holograms; however, as the thickness increases the rapidly diverging Fresnel beam becomes wider than the image beam, and so the thickness density becomes slightly less.

Multiplying the areal and thickness densities give us the volume densities as shown in figures 6.22 and 6.23.

Although shift multiplexing achieves slightly lower densities than angle multiplexing in the 90° geometry, it is capable of higher surface densities for disks. The real

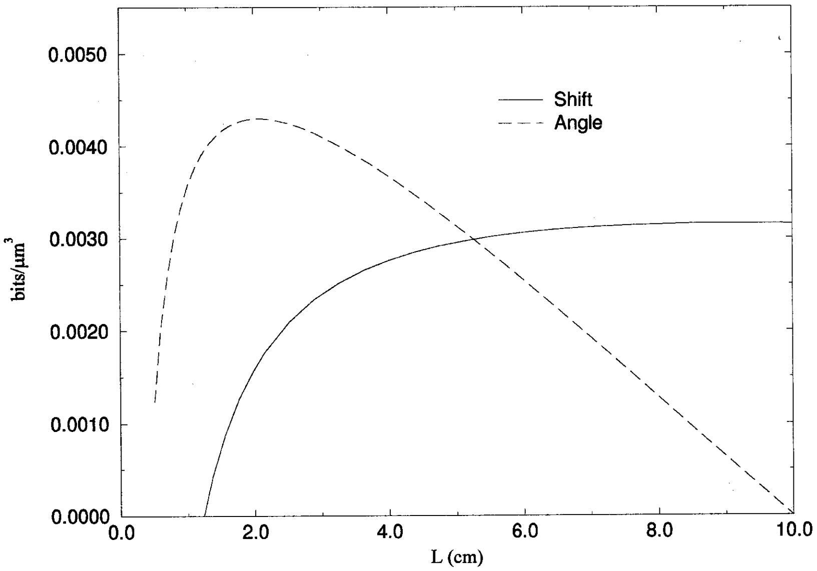

Figure 6.22: Volume density for shift multiplexing in the image plane.

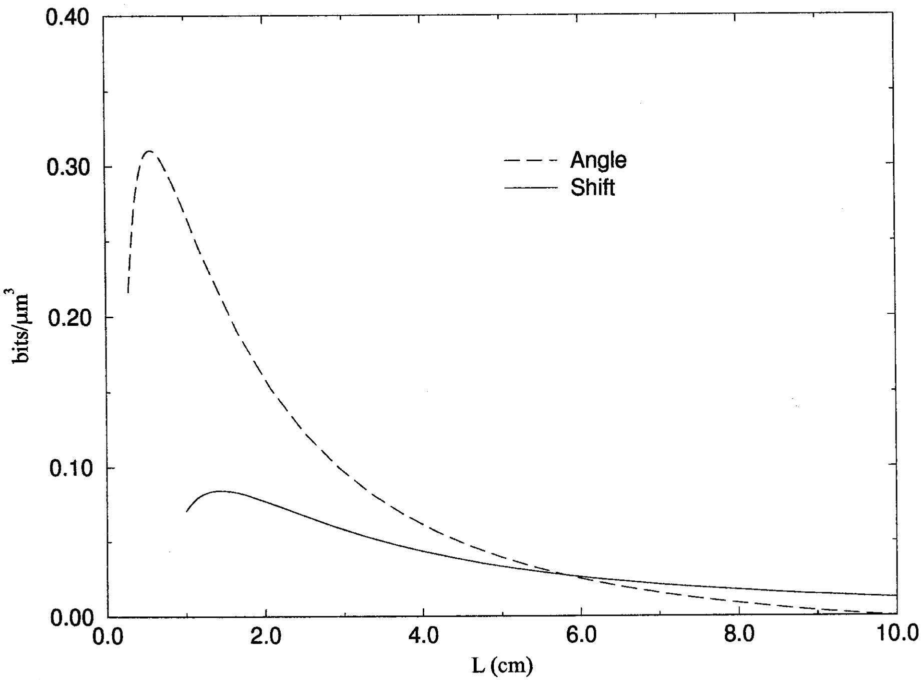

Figure 6.23: Volume density for shift multiplexing in the Fresnel zone.

advantage of shift multiplexing is in its simplicity. It allows for small, cheap, very high angular bandwidth reference lenses, which facilitates tracking of the disk and improves overall system cost and reliability.

Chapter 7 The Little Piece of Cortex

Contents

The previous chapters focused on holography for data storage and for correlators. Mass data storage is needed for AI approaches, and correlators form the basis for neural networks. However, neuromorphic engineering, which may be the most powerful form of neurally inspired computation, may also be the form which benefits the most from optics. Neuromorphic engineering involves emulating certain aspects of neurophysiology [11], as opposed to the overly-broad charicatures used by neural networks and AI. To date, most neuromorphic engineering has involved early sensory systems, such as artificial retinas and cochleas [11–17]. These examples have achieved tremendous dynamic ranges, with robust gain control, while consuming very low power. They consist mostly of sub-threshold, analog VLSI circuits, and they take advantage of the naturally exponential response of subthreshold transistors. This response facilitates the implementation of nicely behaving, saturating nonlinearities, which are common in nature. Efforts have been successful in early sensory systems because we know far more details about the early sensory systems than we do about higher systems, and early sensory systems lend themselves well to single-chip implementation. However, even a full artifical cochlea requires the integration of three chips [17], and vision chips typically compute only one function, such as simple motion or edge extraction. Asynchronous protocols have been developed to alleviate the bandwidth problem in attempting to put together multi-chip systems; however, this solution may not be applicable to the very large scales of integration that would be necessary for implementation of higher level systems, such as early visual cortex. Optical interconnects are necessary in order to achieve the connectivity necessary for implementation of higher systems, as well as to facilitate the implementation of multi-chip versions of early sensory systems.

The design of optical interconnects is a large field, and I will not attempt to elaborate too much on it here. My main purpose is to demonstrate how some of the concepts from this thesis, particularly shift multiplexing, can be used to implement interconnects, and to show how one might go about constructing a neuromorphic system to implement the functionality of the early visual system.

7.1 Primate Vision and Cortex

Information in the primate visual system is transduced at the retina, where initial processing begins. The retina has four different receptor types: three different cone types and one rod type. The cones have three different pigment types for color vision; they are specialized for day vision, having high temporal and spatial resolution with low gain that saturates only in intense light. The rods are specialized for low-light level vision, with very high gain but poor temporal and spatial resolution. Rods saturate in normal daylight. The output from these receptor cells are processed by a two-layer neural circuit consisting of feed-forward bipolar cells and lateral connections by horizontal and amacrine cells. The output cells of the retina are the retinal ganglion cells. Ganglion cells have on-center or off-center response properties, and are classified as either magnocellular or parvocellular. The magnocellular cells respond to motion in the visual field, while the parvocellular cells are more concerned with the fine details of the visual scene, including color. This segregation of the motion signal from the "form" signal is consistent throughout the visual system [57].

After passing through the lateral geniculate nucleus of the thalamus, the information goes to the visual cortex. The first area of cortex to receive visual input is V1. From here, the information passes through many layers of heirarchy, including areas V2, V3, V4, MT and IT [58,59]. While neuromorphic engineering has already made strides in emulating the functionality of the retina, it has, to date, made no real attempts at simulating cortex. The immense connectivity of cortex mandates the use of optics in any attempt at emulation. I will therefore concentrate on describing cortex and how one might use cortex as a model for developing an optoelectronic system.

The neocortex of man is a folded sheet about 2000 cm^2 in area and 3-4mm thick [60,61]. There are about 10^{10} neurons making a total of about 10^{13} synapses. One of the most striking aspects of cortical architecture is that while each area of cortex seems to serve a different function, from speech to vision to motor control and higher cognitive functions, all of cortex seems to have the same general structure. This has prompted Rodney Douglas to hypothesize a "canonical cortical circuit" [62]. The idea is that the same basic microcircuitry exists throughout cortex, with each area representing only minor changes to the same basic plan. If this is true, then an understanding of how one area of cortex works can rapidly lead to the understanding of other, seemingly disparate areas. Development of a neuromorphic cortex may then lead rapidly to applications in many areas, from artificial vision and speech to motor control and perhaps even higher levels of cognition. It makes sense to concentrate initial efforts in vision since more is known about the visual system than any other modality.

Cortex is horizontally organized into six layers and vertically organized into groups of neurons linked synaptically across the layers. These groups of neurons lie in a cylindrical volume oriented perpendicular to the surface of the cortex and are often refered to as "cortical columns." Each column represents a basic functional unit which is approximately 30μm across and contains 80-150 neurons. Each neuron in a column responds strongly to roughly similar stimuli. The horizontal layers are numbered I-VI, with layer I being closest to the skull (see figure 7.1). Layer IV is the main input layer, with some inputs also entering at layers II and III. Layers V and VI output to subcortical structures, with layer VI outputting mostly to thalamus. It is interesting to note that the thalamus sends and receives input from all areas of cortex,

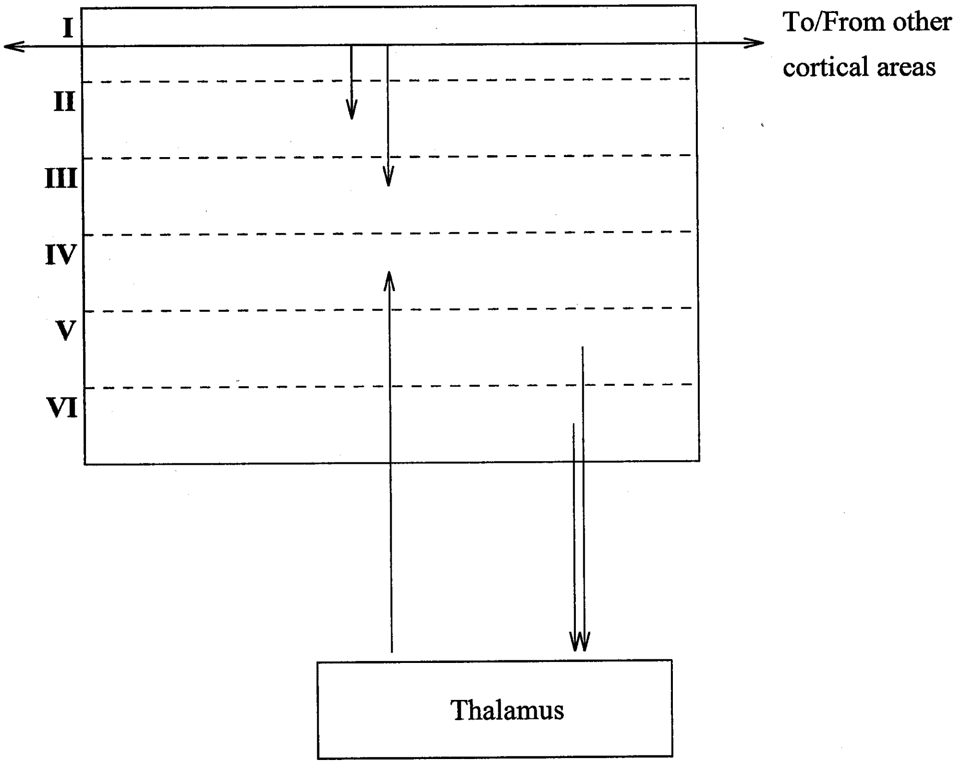

Figure 7.1: The six layered structure of cortex.

but it receives an order of magnitude more input fibers than it sends as output fibers. This massive feedback, and indeed the precise role of the thalamus itself, is not well understood. Layers II and III send output to other cortical areas, and layer I is composed mostly of fibers carrying information between cortical areas. There are very few cell bodies in layer I.

In primary visual cortex (V1), each column contains neurons that respond best to lines oriented in a particular direction, at a particular location in the visual field, being presented to a particular eye [63,64]. Groups of columns whose receptive fields occupy the same general area of the visual field are referred to as hypercolumns. These hypercolumns are arranged retinotopically within most visual areas. A given hypercolumn within V1 can be broken up into occular dominance bands; cells within a particular occular dominance band all respond to input from the same eye. The occular dominance bands are composed of individual columns, each of which responds to stimuli from the same area of the visual field, presented to the same eye, but the

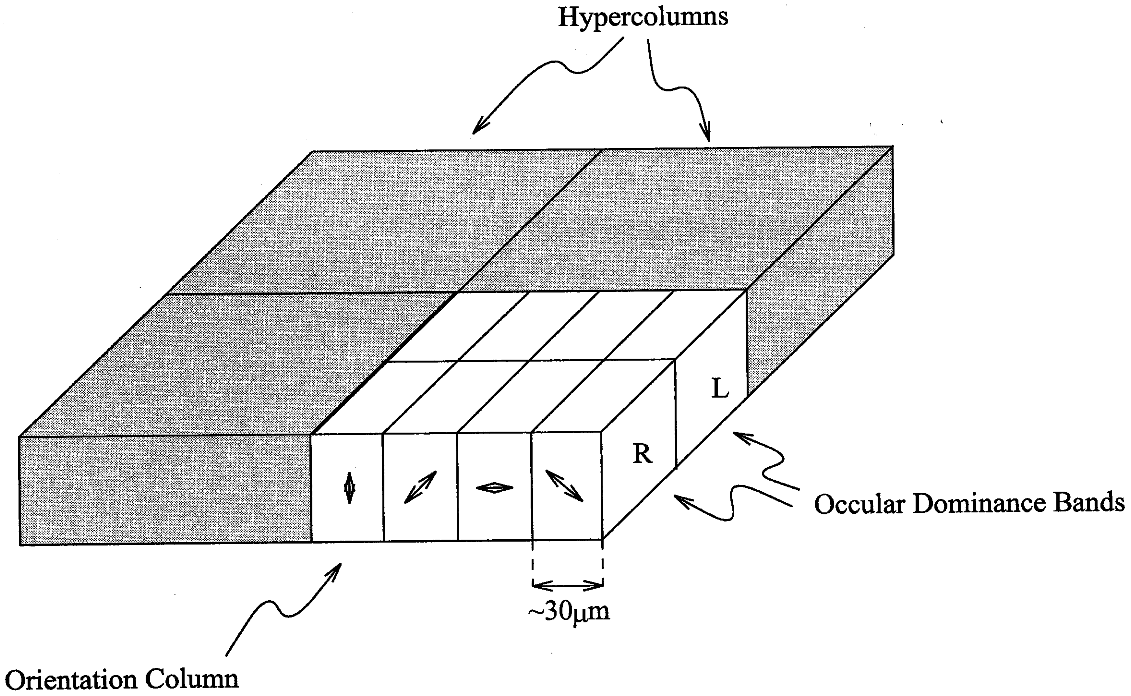

Figure 7.2: Hypercolumn, occular dominance bands, and orientation columns in V1.

orientation and size of the line segment that best stimulates a given column varies gradually from neighboring column to neighboring column (see figure 7.2).

Columns in cortex send and receive four major types of connections(see figure 7.3). The shortest range connections are typically on the order of 250μm, and are mostly inhibitory [65–67]. For example, columns within a hypercolumn are thought to inhibit their neighbors within the hypercolumn. However, very short connections, such as those that remain within the confines of a single column and its immediate neighbors, may be excitatory [68, 69]. These short range interactions may form a local, winner-take-all type of network, so that the column containing the best response to the stimulus in the local receptive field area will suppress all the other columns.

Pyramidal neurons also send out long range excitatory connections over horizontal distances of up to 3mm. These connections are "patchy" and terminate in discrete domains approximately 250μm apart [65, 70, 71]. In visual cortex, these long range connections may include excitation between columns with similar receptive field properties within different hypercolumns.

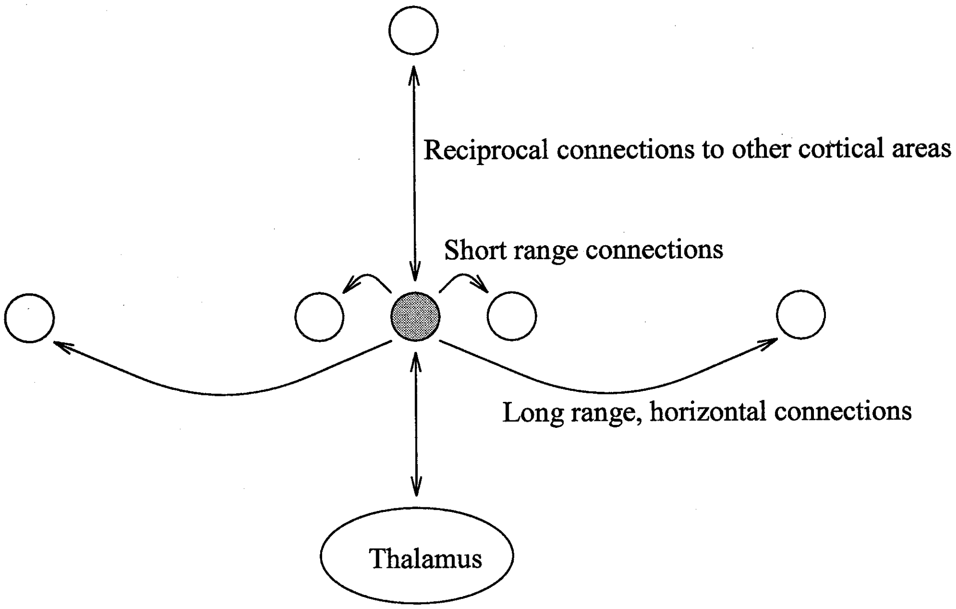
Reciprocal connections to other cortical areas
Short range connections
Long range, horizontal connections
Thalamus

Figure 7.3: The four major connection types for cortical columns.

In addition to the horizontal connections, there are also large feedforward and feedback connections to other cortical areas [72].

All areas of cortex have reciprocal connections with the thalamus [73]. The amount of feedback to thalamus, however, is typically an order of magnitude larger than the number of fibers in the feedforward path. The role of this massive feedback is not well understood.

All told, a typical pyramidal neuron in cortex receives between 1,000 and 10,000 connections. There are approximately 1.5 million retinal ganglion cells, and a similar number of cortical columns in area V1 [74]. From the computer vision perspective, simulating V1 alone is a massive undertaking. The benefits, however, would be tremendous. The primate visual system is capable of complex texture segmentation, figure-ground segmentation, saliency and top-down attentional mechanisms, brightness and color constancies, motion analysis, and depth analysis from stereoscopic and other clues. All of these abilities are greatly facilitated, and in some cases largely accomplished, by the computations that take place in V1. Moving beyond V1, to include simulations of higher cortical areas, will eventually serve to make an even more robust and powerful vision system.

7.2 Neuromorphic Cortex

It is clear from the above description that any attempt to emulate cortex will require massive amounts of interconnections. It is important to determine at what level to mimic biology. Given that the cortical microanatomy is still not well understood, it seems reasonable to assume a rougher scale simulation would be more appropriate. There have been numerous studies, both in neurophysiology and psychophysics, that relate to the functions of individual cortical columns. The classical, and to some extent the non-classical, receptive field properties of columns in visual cortex have been extensively studied, and new studies are revealing the nature of the nonlinear interactions between columns. Also, the anatomy of connections between columns have been fairly well elucidated. The level of the cortical column, then, would be the best level at which to attempt biological mimicry.

While the response properties of each cell in a column are similar, they are not identical. Multiple neurons are probably needed in order to perform the function of the column, and neurons within different layers perform different input and output functions, as described previously. However, given the variety and number of cell responses within a column, it is reasonable to assume each column should be modeled as more than one, but less than 100, functional units. A good compromise would be 10 units per column. These 10 units each represent slight variations of the average receptive field profile of neurons within a column. Biology uses such redundancy throughout the visual system to achieve tremendous robustness. This is important for our hardware implementation, because large, wafer scale integration can often have process yields which are significantly less than one [75]. Designing a system which is robust to failures of individual elements is therefore not only helpful, but necessary.

Given the huge fan-out and -in at each column, the best strategy is to take advantage of the modular architecture that is already present. Each unit can perform local computations, with all communication occurring optically.

7.2.1 The Learning Architecture

One question of importance in deciding the form of the optical interconnect is whether or not the weights of the interconnections, or synapses, will be stored electrically or as the strength of the optical connection itself. Although it is possible to train optical weights, experience has shown that it is not easy. It is difficult to develop effective training algorithms, because one cannot record a hologram without affecting all of the other hologram strengths. Additionally, the relative phase between the signal and reference beams needs to be maintained. Otherwise, when the system is attempting to strengthen a weight, it may actually be weakening it, and vice-versa.

Given the difficulties with real-time updating of holographic weights, and the ever-shrinking dimensions of DRAM, the best choice is to store information about the synapse strengths in a local DRAM cache. The idea is to time multiplex all communication. Each computational unit takes turns broadcasting its output according to a fixed clock cycle. When not broadcasting, each unit receives input from the currently active unit via a global optical bus. This input is multiplied by a weight, which is stored in a local DRAM, and accumulated until all inputs have been received and the output is computed. This strategy takes advantage of the high bandwidth of silicon, which can reach 1 GHz, in comparison with the slow rate of neural computation, which is around 100 Hz. A global optical bus takes advantage of the massive parallelism afforded by optics, while local DRAM uses the huge advances in silicon technology for a task that would be very difficult to implement in optics. Thus, this architecture uses each component in a way that fits its greatest strength.

The basic unit in silicon is represented in figure 7.4. All optical inputs and outputs will use a GaAs layer which is bonded to the silicon by the input and output vias, as discussed below. Each unit is connected to approximately 10,000 other units, with each unit taking turns broadcasting its output. To run at 100Hz requires a clock rate of 100Hz/unit×10,000 units = 1MHz. Analog communication would be too noisy at this rate, so a digital, pulse code must be used. Even with a pulse code, the system would still only need to run at 10MHz, well below the limit for optical communication

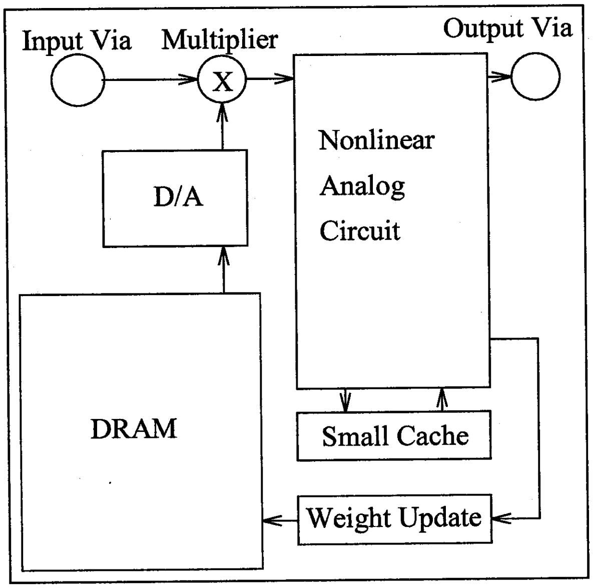

Figure 7.4: Silicon circuit for one unit of artificial cortex

(which can be as high as 10GHz).

With digital communications and local digital memory, digital logic would seem the obvious choice. However, much work has been done in simulation of neural circuits using subthreshold, analog circuitry. Sub-threshold circuitry also has a power advantage; given the enormous scale of a project that is designed to emulate cortex, power consumption is a very important concern. The trade-off in silicon area required for analog versus digital implementations is not obvious, and depends upon the specific function to be implemented. Digital circuits tend to require more transistors. However, the size of these transistors can be kept small; anolog designs can be sensitive to transistor mismatch and therefore can require large transistors in some parts of some circuits. An analog circuit would require quite a bit of digital to analog conversion, and analog to digital. However, this might not be as expensive in area as it first appears. Although a full digital to analog conversion needs to be made on the weights from memory, the update of these weights can be a simple threshold function to increment or decrement the weights, and therefore may only be a 1 bit analog to digital conversion. It may be possible to perform the mode conversions for the data input and output in the GaAs layer, if there is extra room, thus reducing the total area of silicon required.

Figure 7.4 also shows a small cache associated with the logic circuit. This represents a small working memory for whatever computation the circuit must do, and would probably take very little area. The size and form of cache necessary would again depend on the specifics of the circuit.

The total area of silicon required will most likely be dominated by the DRAM. The size of a DRAM cell will soon be down to 0.5μm on a side [76]. Assuming we need about 4Kbytes per unit, the total area for the DRAM will be less than 100μm on a side. Assuming 0.175μm technology, the total unit size should be approximately 200μm-400μm on a side, depending on the aggressiveness of the design. For 10,000,000 units to simulate all of striate cortex, the total silicon area required is from $60 \times 60 cm^2$ to $1.2 \times 1.2 m^2$. Obviously, even with wafer scale integration, such a system will need to occupy multiple chips.

Being a direct-bandgap semiconductor, silicon has no capacity for efficient light emitters. We therefore need to use GaAs to build our optical interface. Current technology involves flip-chip bonding of thinned GaAs wafers containing vertical cavity surface emitting lasers (VCSELs) and detectors to silicon substrates. The VCSELs can be made to emit out the top, or down through the bottom, passing through the silicon at transparent wavelengths. The yield for a wafer scale integration of GaAs flip chips would be far too low to be practical. Technology is under development, however, that would allow small pieces of GaAs to "find" their proper locations on the silicon using DNA assisted assembly [77]. Using these techniques, large scale integrations should be possible. Each piece of GaAs would contain the VCSEL emitter (our "axon") as well as a detector and the appropriate driver circuitry for the laser.

The optical architecture must be shift invariant, so that a single, or very few, gratings can be used to do all of the interconnects. Otherwise, the diffraction efficiency would be too small. The architecture should also be modular, so that higher level of cortex could eventually be integrated, complete with the massive feedback and feedforward connections. Figure 7.5 shows an architecture that meets these criteria. Each Si:GaAs wafer is connected to every other wafer via the diffractive element and

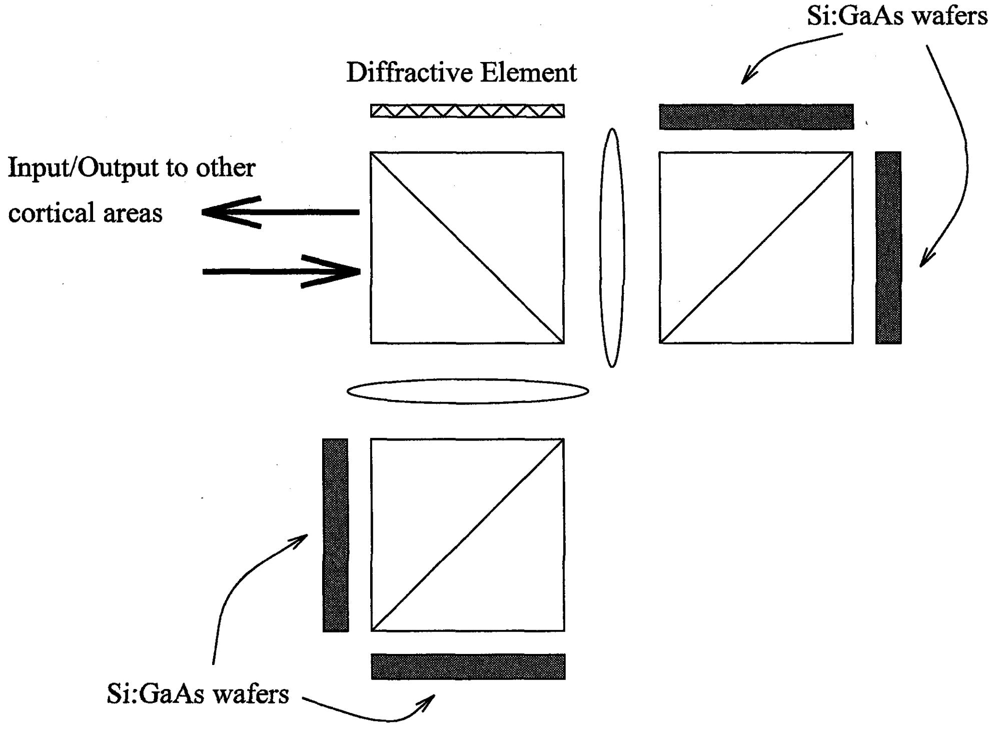

Figure 7.5: Modular architecture for artificial cortex with adaptable interconnection strengths.

four passes through a beamsplitter. Each wafer could be a part of a model striate cortex, or each wafer could represent a different cortical area. Although no single wafer is big enough to implement a cortical area in its entirety, implementing a small portion could be very fruitful, especially if it enables the implementation of several cortical areas at once. This cortical module is also able to output the combined activity of all the wafers, and to receive input, as shown in figure 7.5.

7.2.2 The Fixed Architecture

Although the above architecture can be very useful, it will not be very compact. Also, the integration of GaAs with Silicon is still under development, and large scale fabrication may prove to have yields that are too low, even for neural architectures. In some cases, however, one might be able to learn the appropriate synapse strengths off-line, through large scale, albeit very slow, computer simulations on large supercomputers. Or, if one were to build the above, adaptable architecture, one might then

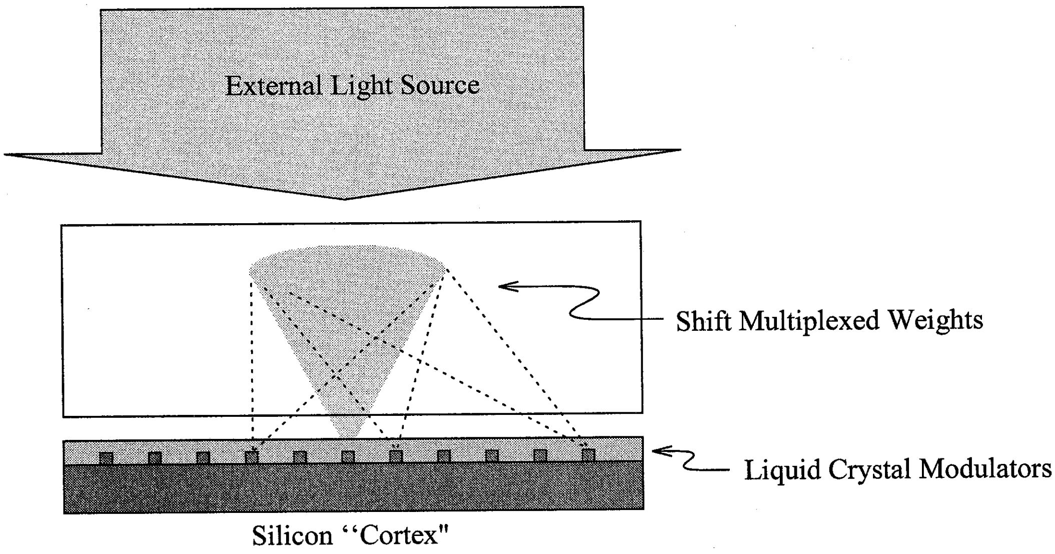

Figure 7.6: Fixed-weight architecture

read off the weights from the local DRAM and still desire to build a more compact implementation.

The Implementation Architecture

Figure 7.6 shows such an architecture. In this case, since we do not need adaptable weights, we can store the weights as analog connection strengths in a volume hologram. The entire system, in fact, can be analog. This drastically reduces the amount of silicon area required. In order to avoid the complexities of GaAs integration, liquid crystal modulators can be used. In this case, an external laser source is used to illuminate the wafer, and analog liquid crystal devices, which have been integrated with the silicon, are used to modulate the light. There is no need for time multiplexing, since each unit has its own output wights stored in the volume hologram. All units emit their output at all times. Since each emitter is a point source, the weights are shift-multiplexed in the volume hologram adjacent to the silicon wafer.

The silicon unit, shown in figure 7.7, is reduced to one liquid crystal pad, four detectors, and the nonlinear analog circuitry. While the exact number of detectors required might depend on the details of the design, four is a good estimate for two reasons. First, there are four basic types of connections as described above: local

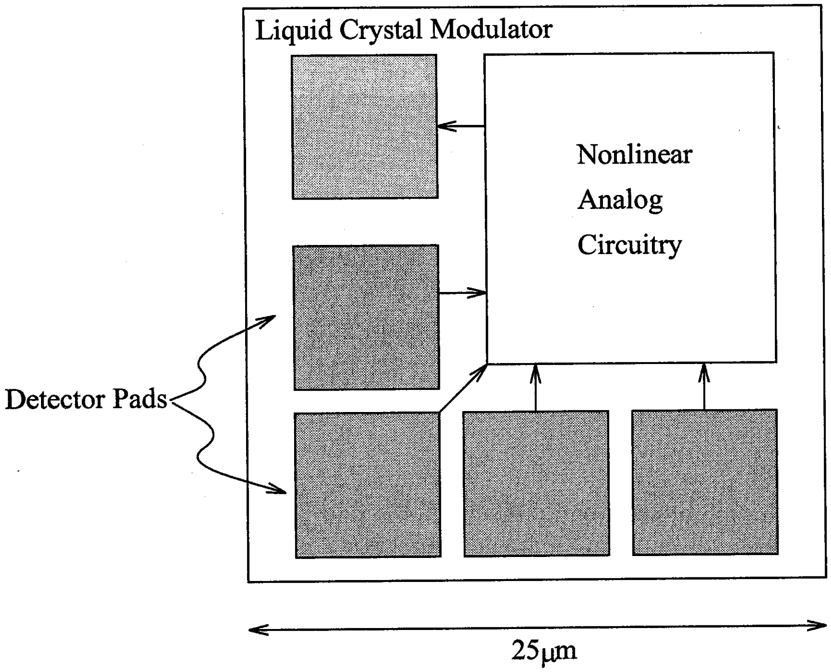

Figure 7.7: Circuit for one unit of the fixed architecture

inhibition, long range excitation, feedforward, and feedback connections. Secondly, there are four basic, signed operations we might want to implement: additive and multiplicative excitation, subtractive and divisive inhibition. Both arguments lead to the conclusion that four detectors will be required.

In calculating the area of silicon per unit, we must estimate the size of the detectors and the liquid crytsal pad. Again, the actual circuitry will make a relatively minor contribution to the area for a 0.174 μm analog process. The size of the liquid crystal pad determines the numerical aperture of the output beam. Because we are using shift multiplexing, the numerical aperture of the input beams to the detectors cannot be larger than the numerical aperture of the output beam; the diameter of the hologram is limited by the output beam. Therefore, the detectors should not be made any smaller than the liquid crystal pad, since the input spot size cannot be made any smaller (see below). The minimum size for a liquid crystal pad, with a very aggressive design, is 5μm [78]. Therefore, each detector must also be 5μm on a side. The minimum unit size would then be approximately $25 \times 25 \mu m$. The entire striate cortex could therefore be implemented in a wafer that is less than 8cm on a side. This is very impressive,

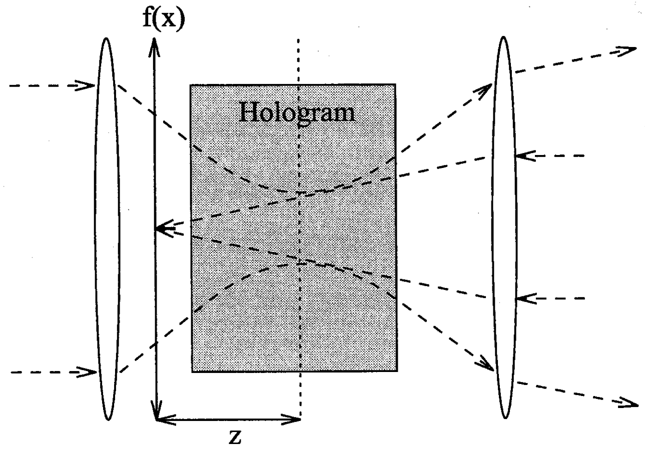

Figure 7.8: Architecture for recording shift multiplexed, fixed-weight interconnects

given that columns of primate cortex are about 30μm in diameter. Of course, primate cortex will still be far more powerful than our neuromorphic approximation.

As mentioned above, the optical interconnections will be shift-multiplexed, so that no lenses will be required in the system, and the volume hologram can be attached directly on top of the wafer. This brings about two questions: how to record the holograms, and what will the selectivity be?

The Recording Architecture

Figure 7.8 shows an architecture for recording the weights in the volume hologram. The biggest dificulty in recording the interconnect pattern is that the light from any emitter diverges into a cone that takes up only a fraction of the entire medium. The hologram must be contained only within this volume. If one were to attempt to record a regular, reflection geometry hologram, the light for detectors far from the emitter would not overlap, and so no connection would be made. In the architecture of figure 7.8, the interconnect pattern, $f(x)$, is placed after a focusing lens. This brings all the light from $f(x)$ into the small volume occupied by the emitter cone. The reference wave is a beam focused in the plane of $f(x)$, at the location where the emitter will be. The holograms will then be read out with the phase-conjugates of the reference waves, thus producing the phase conjugates of the interconnect mask, which will serve to connect the emitter to detectors in the same plane. The field at the focus of the transforming lens due to the mask is the Fourier transform of $f(x)$. The width of the Fourier transform is $2\lambda z/\Delta x$, where Δx is the detector size and z is the distance from the interconnect mask, $f(x)$. The width of the reference beam at z is approximately $2\lambda z/\Delta x'$, where $\Delta x'$ is the width of the emitter. For the best overlap between the reference and signal waves, $\Delta x = \Delta x'$. In other words, the detectors should be the same size as the emitters.

Shift Selectivity

The fixed architecture shown in figure 7.6 has no lenses in order to make it as compact as possible. This requires that the holographic interconnections are shift multiplexed with the holographic material virtually in direct contact with the silicon wafer. In order to compute the shift selectivity for the reflection geometry, we must use the refined, path integral theory from chapter 6. In this case, the path will be along the z axis. Because the numerical aperture is relatively small, we will assume that the effective change in angle, $\Delta\theta$, with shift δx is a constant across the lateral dimensions. However, because the source will be extremely close to the hologram, essentially in direct contact, it is impossible to approximate the change in angle with a single z_0. Instead, $\Delta\theta = \delta x/z$, were z is the distance from the emitter.

For the reflection geometry, the path integral becomes

$$A_d = \int e^{i2k(1-\cos(\Delta\theta))z} dz. \tag{7.1}$$

This can easily be seen from the k-sphere diagram in figure 7.9. The z component of the tilted reference beam is just $k\cos(\Delta\theta)$. The tip of the grating vector therefore moves away form the surface of the sphere by the distance $k(1-\cos\Delta\theta)$. But the curvature of the surface of the k-sphere moves away from the grating tip by the same amount, so that the total distance from the grating vector tip to the k-sphere surface becomes $2k(1-\cos\Delta\theta)$.

In chapter 6, the inhomogenous illumination of the spherical reference wave was

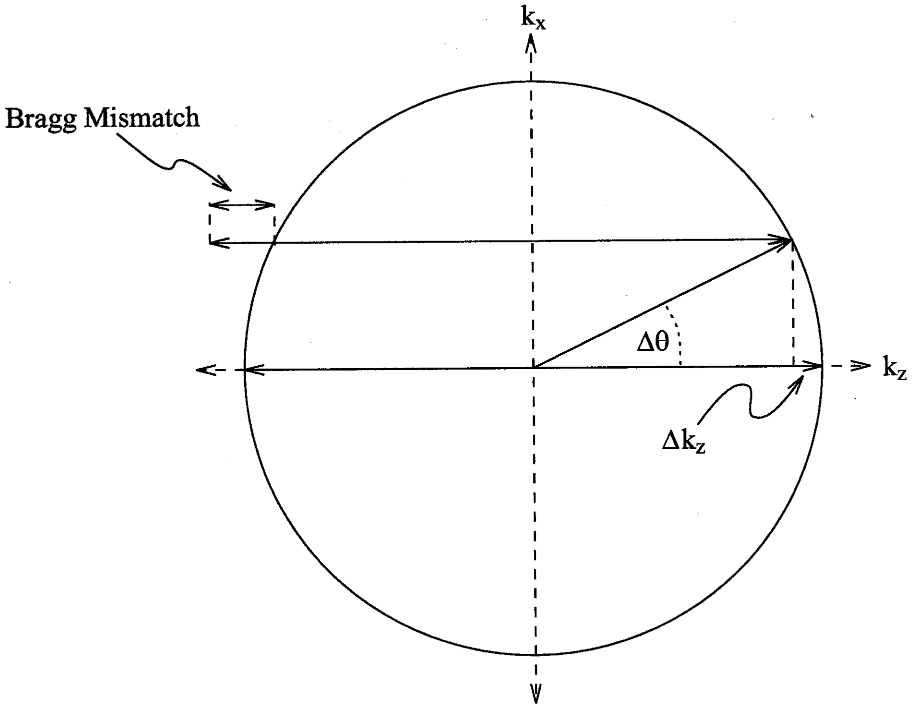

Figure 7.9: K-sphere for the reflection geometry

ignored. The recording rate for the hologram was assumed not to vary significantly, since the signal wave could be made strong and relatively homogeneous. For the present case, however, both the reference and signal waves can be strongly diverging. The writing of a hologram follows an exponential growth so that the strength of a hologram, A, is given approximately by

$$A = A_0(1 - e^{-t/\tau_r}) \tag{7.2}$$

where A_0 is the saturation grating strength and τ_r is the recording rate. The recording rate is proportional to the total intensity. For the case of counter-propagating spherical waves with a common focus at z_0, the strength of the recorded hologram can be calculated using

$$\frac{t}{\tau_r} = \frac{gI(L)}{I(z)} \tag{7.3}$$

where $I(L)$ is the intensity value at the face of the hologram opposite to the focal point of the sphercal waves and g is a constant that must be fit to the writing curve. The beam intensities can be solved assuming they are approximately gaussian. While this might not be the case for the liquid crystal implementation, it is of more general interest and matches more closely the experiment described below. It is a good enough approximation for our purposes here. The strength of the on-axis field for a gaussian beam, $E(z)$, is given by

$$E(z) = E_0 \frac{\omega_0}{\omega(z)} e^{-i[kz-\eta(z)]} \tag{7.4}$$

where

$$\begin{aligned} \omega_0 &= \frac{\lambda}{\sin\theta} && (7.5) \\ z_0 &= \frac{\pi \omega_0^2 n}{\lambda} && (7.6) \\ \omega(z) &= \sqrt{\omega_0^2 \left(1 + \frac{z^2}{z_0^2}\right)} && (7.7) \\ \eta(z) &= \arctan\left(\frac{z}{z_0}\right) && (7.8) \end{aligned}$$

Here, θ is the beam divergence angle and ω_0 is known as the beam waist because it is the width of the beam at its narrowest point. The coordinate z is measured from the beam waist. Figure 7.10 shows a plot of the relative hologram strength as a function of z for g=7.7.

We can now combine equations 7.1 and 7.4 to get

$$A_d = \int\limits_{L_{min}(\delta x)}^{L/2} (1 - e^{\frac{-|E(z'+L/2)|^2}{g|E(L)|^2}}) e^{i2k\left(1-\cos\left(\frac{\delta x}{z'+L/2}\right)\right)z'} dz' \tag{7.9}$$

where we have changed coordinates so that $z' = 0$ at the center of the hologram.

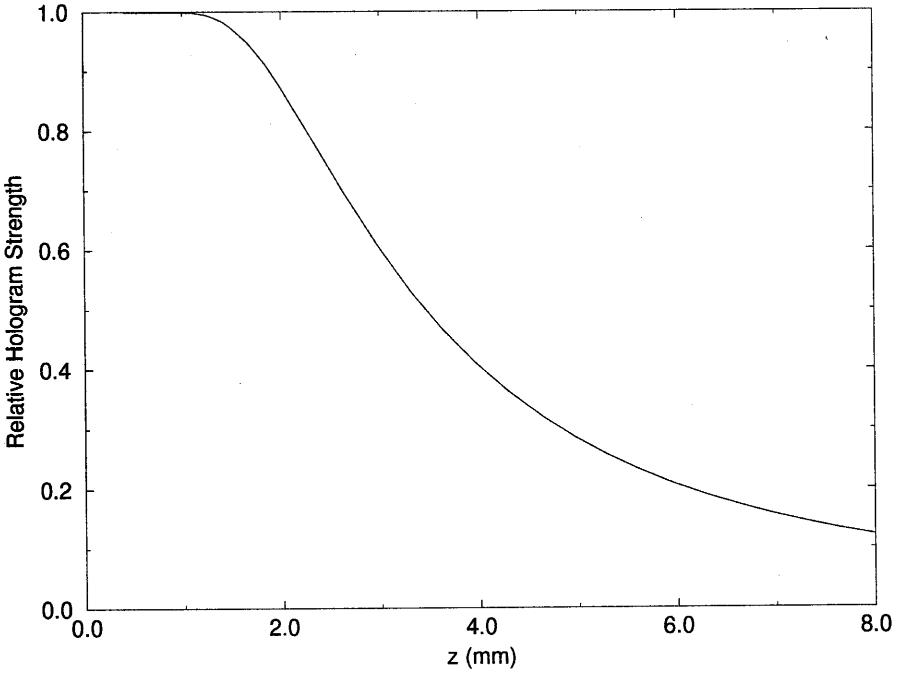

Figure 7.10: Hologram strength as a function of depth for the reflection geometry with a numerical aperture of 0.11 and g=7.7.

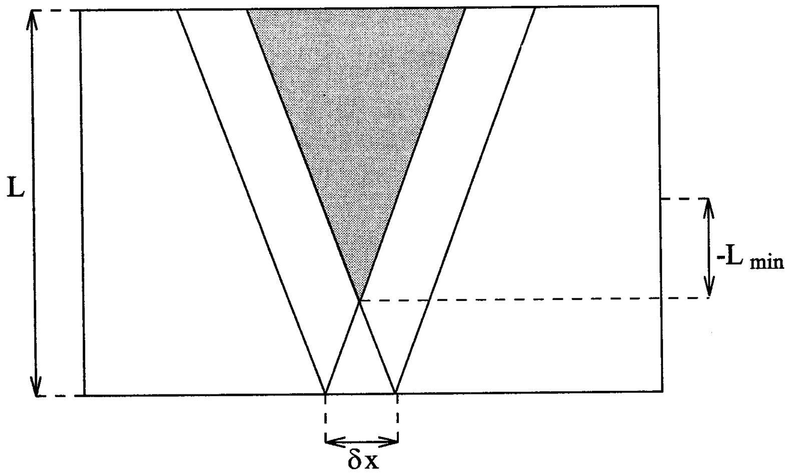

Figure 7.11: L_{min} begins where the new read-out beam first overlaps with the stored hologram.

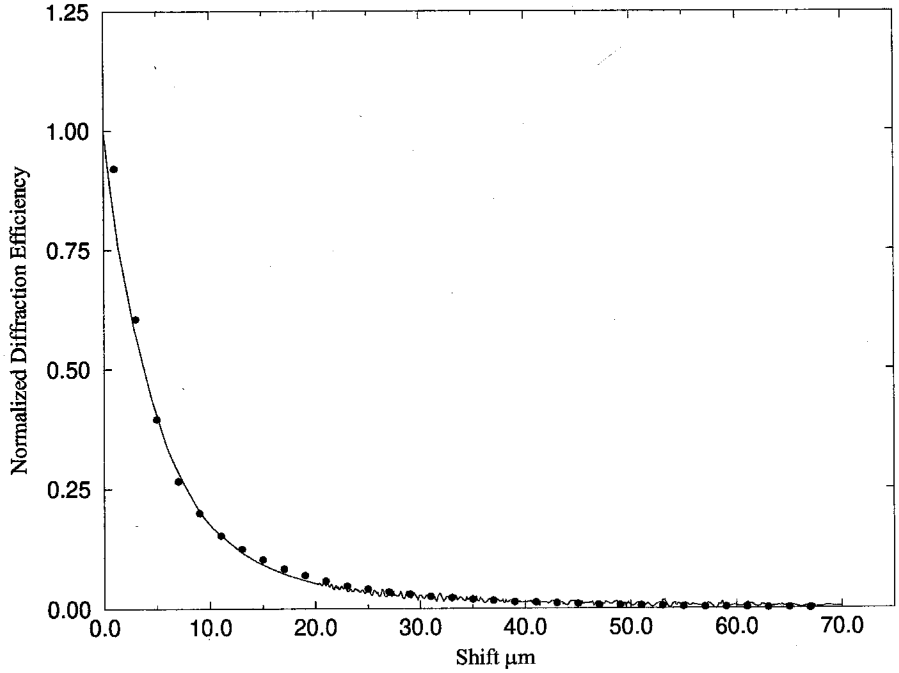

Figure 7.12: Selectivity curve for the reflection geometry. Recorded in 8mm thick $LiNbO_3$ with $z_0 = 0$. The numerical aperture was 0.11.

The lower limit, $L_{min}(\delta x)$, is a function of where the reference beam begins to overlap with the recorded hologram as δx varies. This is shown in figure 7.11 and is given by

$$L_{min}(\delta x) = -\frac{L}{2} + \frac{\delta x}{2\tan(\arcsin(\sin\theta/n))} \tag{7.10}$$

In order to test the validity of equation 7.9, I stored a hologram in the reflection geometry using counterpropagating spherical beams with numerical apertures of 0.11 in 8mm thick $LiNbO_3$. This represents, in the case of optical interconnects, one unit which has been connected to itself via the hologram. For more distant holograms, the angle between the reference and signal beams will be smaller, which will increase the shift selectivity. On the other hand, the overlap between the reference and signal beams will also decrease, making the effective length of the hologram shorter. The balance of these effects is hard to know without a more specific design. However, this experiment should at least provide us with a general feel for the amount of shift selectivity, as well as another confirmation of the theory. Figure 7.12 shows one half of the selectivity curve, as well as the matching theory. There is excellent agreement between theory and experiment. In this case, the parameter g was fit to a value of 7.7. The slight roughness to the theory curve in the wings was due to the working precision of the integration routine. The data was taken with a slow stage, and therefore only one side of the symmetric selectivity curve is shown. From the plot, it is clear that there will be some cross-talk between interconnect patterns for neighboring units of the little piece of cortex if they are only 25μm apart. This crosstalk may not be significant, however, given the robustenss that is inherent to neural architectures. Additionally, the strength of interconnections for neighboring units may not be totally independant; since their receptive field properties would be similar, it is likely that their interconnection patterns would be similar as well.

7.3 Conclusion

Neuromorphic engineering may be the most exciting development yet in the quest to mimic the computational power of the brain. While AI and neural networks could have developed without the involvment of optics, neuromorphic engineering needs optics in order to progress to larger scale systems. Shift multiplexing allows for very compact architectures and may eventually lead to the implementation of an artifical V1. Such a development would not only be a tremendous engineering achievement, it would help to further our understanding of how cortex works by creating a platform on which new computational theories of the brain could be tested.

Chapter 8 Conclusion

In this thesis, I have addressed the use of holography for optical correlators, optical memories, and optical interconnects.

The capacity of correlators for pattern recognition is a function of the geometry of the signal and reference waves. The optimal geometry is with the reference wave normal to the hologram, and the signal beam to be as oblique as possible. The best geometry is with the reference and signal beams 90° apart. Once the geometry is determined, the trade-off between shift-invariance and the number of templates required determines the final system characteristics. This trade-off is important for pattern recognition, and is dependant upon the task. For active vision systems, a few generic templates with large shift-invariance my be required to find objects of interest. The remaining correlators would require relatively little shift-invariance, since the object would be well centered in the visual field by the active vision system. The reflection geometry provides such a heterogenous system. However, it has a relatively low capacity according to the metric developed in chapter 3. The benefits of the heterogenous shift-invariance domains may outweigh considerations of total capacity in some applications.

Fresnel correlators were described in chapter 5. It is very easy to control the size of the shift-invariance domains for this system; one need only move the hologram into the Fresnel zone. Since this system does not rely on Bragg selectivity, it is feasible to use thin materials such as polymers. Some polymers, such as DuPont's HRF 150, are much more sensitive and have a larger dynamic range than photorefractive crystals. In this thesis, I demonstrated the basic principles behind the Fresnel correlator. The next step for this project is to develop a large system that takes advantage of the very large number of correlators that can be stored.

Shift multiplexed memories were described in chapter 6. I believe this technique has enormous commercial potential. The simple implementation allows for a robust and inexpensive design. One potential problem, however, is that as a disk spins past the reference beam, the reconstructed data page will shift across the imaging electronics. A possible way around this problem is the use of a pulsed laser with a fast-frame photodiode array. One could also try using large pixels. Although larger pixels will decrease the density somewhat, they will make tracking easier. Additionally, large pixels will make the system less sensitive to lens abberations in the signal arm. Both effects will help lead to a cheaper, more robust system at the cost of some loss in data density. Future work should analyze this trade-off against the practical constraints of a commercial product. The most important area for improvement in holographic data storage, however, is in the materials. New holographic materials that are sensitive enough for fast recording and have large dynamic ranges are necessary for commercial success.

Shift multiplexing can also be used for chip-to-chip interconnects and the construction of large neuromorphic systems. Chapter 7 only began to outline to first details of such a system; much work remains to be done in this area. For example, VCSEL technology needs to continue to improve, in order to provide lasers with stable polarization and wavelength. Dynamic range is also likely to be an issue for the fixed architecture described in chapter 7. Integration of GaAs and silicon is a very active area of research, and needs to continue in order to enable large scale integration of such systems. Finally, algorithms for vision need to be studied and developed. The better our understanding of primate vision is, the better we will be able to model it and develop sophisticated vision systems.

Bibliography

[1] W.S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[2] F. Rosenblatt. *Principles ofNeurodynamics.* Spartan, New York, 1962.

[3] *Adaptive Switching Circuits*, volume 4 of *1960 IRE WESCON convention record*, New York, 1960. IRE.

[4] B. Widrow. Generalization and information storage in networks of Adaline 'Neurons'. In M.C. Yovits, G.T. Jacobi, and G.D. Goldstein, editors, *Self-Organizing Systems 1962*, pages 435–461. Spartan, Washington, 1962.

[5] M.L. Minsky and S.A. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[6] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci., USA*, 79:2554–2558, 1982.

[7] P.J. Werbos. *Beyond regression: new tools for prediction and analysis in the bahavioral sciences.* PhD thesis, Harvard University, 1974.

[8] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagation errors. *Nature*, 323:533–536, 1986.

[9] D.E. Rumelhart, J.L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* MIT Press, Cambridge, MA, 1986.

[10] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley, New York, 1991.

[11] C. Mead. *Analog VLSI and Neural Systems.* Addison-Wesley, Reading, 1989.

[12] C.A. Mead and M.A. Mahowald. A silicon model of early visual processing. *Neural Networks*, 1:91–97, 1998.

[13] *A silicon retina with adaptive photoreceptors*, SPIE/SPSE Symposium on Electronic Science and Technology: From Neurons to Chips, Orlando, FL, April 1991.

[14] M. Mahowald. *VLSI analogs of neuronal visual processing: a synthesis of form and function.* PhD thesis, California Institute of Technology, 1994.

[15] T. Delbrück. Analog VLSI phototransduction by continous-time, adaptive, logarithmic photoreceptor circuits. CNS Memo 30, California Institute of Technology, Pasadena, CA, 1994.

[16] K.A. Boahen and A.G. Andreou. A contrast sensitive silicon retina with reciprocal synapses. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems 4*, pages 764–772. Morgan Kaufmann, San Mateo, CA, 1992.

[17] R. Sarpeshkar, R.F. Lyon, and C.A. Mead. A low-power wide-dynamic-range analog VLSI cochlea. *Analog Integrated Circuits and Signal Processing*, 16, 1998. In press.

[18] J.W. Goodman. *Introduction to Fourier Optics.* McGraw-Hill Publishing Co., 1968.

[19] P. J. van Heerden. "A new optical method of storing and retrieving information". *Applied Optics*, 2(4):387–392, 1963.

[20] E. N. Leith, A. Kozma, J. Upatnieks, J. Marks, and N. Massey. "Holographic data storage in three–dimensional media". *Applied Optics*, 5(8):1303–1311, 1966.

[21] E. G. Ramberg. "Holographic information storage". *RCA Review*, 33:5–53, 1972.

[22] D. Psaltis and F. H. Mok. "Holographic memories". *Scientific America*, 273(5):70–76, 1995.

[23] D. Psaltis, D. Brady, X. G. Gu, and S. Lin. "Holography in artificial neural networks". *Nature*, 343(6526), 1990.

[24] D. Psaltis, D. Brady, and K. Wagner. "Adaptive optical networks using photorefractive crystals". *Applied Optics*, 27(9):1752–1759, 1988.

[25] J.D. Jackson. *Classical Electrodynamics.* Wiley, New York, 1975.

[26] *Large-scale rapid access holographic memory*, volume 2514 of *SPIE Technical Digest Series*, 1995.

[27] X. An and D. Psaltis. Thermal fixing of 10,000 holograms in $LiNbO_3 : Fe$ using incremental fixing schedule. *submitted to Applied Optics*, 1997.

[28] D. Psaltis. "Class notes: EE133 Optical computing, Caltech", 1994.

[29] X. An, G. Burr, and D. Psaltis. 160,000-hologram system using $LiNbO_3 : Fe$. *submitted to Applied Optics*, 1997.

[30] H. S. Li. *"Photorefractive 3-D disks for optical data storage and artificial neural networks".* PhD thesis, California Institute of Technology, 1994.

[31] Y.N. Denisyuk. Photographic reconstruction of the optical properties of an object in its own scatteredradiation field. *Sov. Phys. Dokl.*, 7:543, 1962.

[32] H.-Y. S. Li and D. Psaltis. Three-dimensinal holographic disks. *Appl. Opt.*, 33:3764–3774, 1994.

[33] F.M. Smits and L.E. Gallaher. Design considerations for a semipermanent optical memory. *The Bell System Technical Journal*, pages 1267–1278, 1967.

[34] A.L. Mikaelian, V.I. Bobrinev, S.M. Naumov, and L.Z. Sokolova. Design principles of holographic memory devices. *IEEE Journal of Quantum Electronics*, 6:193–198, 1972.

[35] P. Graf and M. Lang. Geometrical aspects of consistent holographic memory design. *Applied optics*, 11:1382–1388, 1972.

[36] B. Hill. Some aspects of a large capacity holographic memory. *Applied Optics*, 11:182–191, 1972.

[37] A. Vander Lugt. Design relationships for holographic memories. *Applied Optics*, 12:1675–1685, 1973.

[38] A. Vander Lugt. Packing density in holographic systems. *Applied Optics*, 14:1081–1087, 1975.

[39] *Capacity of optical correlators*, volume 825-22 of *SPIE*, San Diego, CA, Aug. 1987.

[40] F.T.S. Yu, S. Wu, A.W. Mayers, and S. Rajan. Wavelength multiplexed reflection matched spatial filters using $linbo_3$. *Optics Communications*, 81:343–47, 1991.

[41] G.A. Rakuljic, V. Leyva, and A. Yariv. Optical data storage by using orthogonal wavelength-multiplexed volume holograms. *Optics Letters*, 17:1471–1473, 1992.

[42] H. Yamamoto, K. Maeda, S. Ishizuka, and T. Kubota. Real-time measurement of wavelength selectivity of reflection holograms. *Applied Optics*, 31:7397–7399, 1992.

[43] G.A. Rakuljic and V. Leyva. Volume holographic narrow-band optical filter. *Optics Letters*, 18:459–461, 1993.

[44] J. Rosen, M. Segev, and A. Yariv. Wavelength-multiplexed computer-generated volume holography. *Optics Letters*, 18:744–746, 1993.

[45] S. Yin, F. Zhou, M. Wen, Z. Yang, J. Zhang, and F.T.S. Yu. Wavelength multiplexed holographic storage in a sensitive photorefractive crystal using a visible-light tunable diode laser. *Optics Communications*, 101:317–321, 1993.

[46] Z. Zhao, H. Zhou, S. Yin, and F.T.S. Yu. Wavelength-multiplexed holographic storage by using the minimum wavelength channel separation in a photorefractive crystal fiber. *Optics Communcation*, 103:59–62, 1993.

[47] F.T.S. Yu, S. Yin, and Z.H. Yang. Thick volume photorefractive crystal wavelength-multiplexed reflection-type matched filter. *Optical Memory and Neural Networks*, 3:207–215, 1994.

[48] H. Zhou, Z. Zhao, and F.T.S. Yu. Diffraction properties of a reflection photorefractive hologram. *Applied Optics*, 33:4345–4352, 1994.

[49] S.-S. Chen and H.J. Caulfield, editors. *Optical neural networks*, volume CR55 of *Critical Reviews of Optical Science and Technology*, Orlando, Florida, April 1994. SPIE.

[50] C.X.-G. Gu. *Optical neural networks using volume holograms.* PhD thesis, California Institute of Technology, Pasadena, CA, 1990.

[51] see http://www/ntt.co.jp.

[52] G. Burr. *Volume holographic storage using the* 90° *geometry.* PhD thesis, California Institute of Technology, Pasadena, CA, 1996.

[53] K. Wagner and D. Psaltis. Multilayer optical learning networks. *Appl. Opt.*, 26:5061–5076, 1987.

[54] L. Solymar and D.J. Cooke. *Volume Holography and Volume Gratings.* Academic, New York, 1991.

[55] H.C. Külich. A new approach to read volume holograms at different wavelengths. *Opt. Commun.*, 64:407–411, 1987.

[56] H.C. Külich. Reconstructing volume holograms without image field losses. *Appl. Opt.*, 30:2850–2857, 1991.

[57] J. H. Kandel, E. R.; Schwartz and T. M. Jessel. *Principles of Neural Science.* Elsevier, New York, 1991.

[58] R. Desimone and L. G. Ungerleider. Neural mechanisms of visual processing ir monkeys. In F. Boller and J. Grafman, editors, *Handbook of Neuropsycholo Vol. II*, pages 267–299. Elsevier, Amsterdam, 1989.

[59] A. D. Milner and M. A. Goodale. *The Visual Brain in Action.* Oxford University Press, Oxford, 1995.

[60] V.B. Mountcastle. The columnar organization of cortex. *Brain*, 120:701–722, 1997.

[61] H.A. Drury and D.C. Van Essen. Functional specializations in human cerebral-cortex analyzed using the visible-man surface atlas. *Human Brain Mapping*, 5:223–237, 1997.

[62] R. J. Douglas and K.A.C. Martin. Neocortex. In Gordon M. Shephard, editor, *The Synaptic Organization of the Brain*, pages 389–438. Oxford University Press, New York, 1990.

[63] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)*, 160:106–154, 1062.

[64] D.H. Hubel and T.N. Wiesel. Functional architecture of macaque monkey visual cortex (ferrier lecture). *Proc. R. Soc. Lond. [B]*, 198:1–59, 1077.

[65] J.S. Lund, T. Yoshioka, and J.B. Levitt. Comparison of intrinsic connectivity in different areas of macaque monkey cerebral-cortex. *Cereb. Cort.*, 3, 1993.

[66] J.S. Lund and C.Q. Wu. Local circuit neurons of macaque monkey striate cortex. 4. neurons of laminae 1-3a. *J. Comp. Neur.*, 384:109–126, 1997.

[67] J.B. Levitt, J.S. Lund, and T. Yoshioka. Anatomical substrates for early stages in cortical processing of visual information in the macaque monkey. *Beh. Bra. Res.*, 76:5–19, 1996.

[68] Z.F. Kisvarday and U.T. Eysel. Functional and structural topography of horizontal inhibitory connections in cat visual-cortex. *Euro. J. Neuro.*, 5:1559–1572, 1993.

[69] R. Niewenhuys. The neocortex - an overview of its evolutionary development, structural organization and synaptology. *Anat. Embryo.*, 90:307–337, 1994.

[70] M. Weliky, K. Kandler, D. Fitzpatrick, and L.C. Katz. Patterns of excitation and inhibition evoked by horizontal connections in visual-cortex share a common relationship to orientation columns. *Neuron*, 15:541–552, 1995.

[71] T. Yoshioka, J.B. Levitt, and J.S. Lund. Intrinsic lattice connections of macaque monkey visual cortical area-v4. *Journal of Neuroscience*, 12:2785–2802, 1992.

[72] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1990.

[73] E.G. Jones. *The Thalamus.* Plenum Press, New York, 1995.

[74] A. Peters. The organization of primary visual cortex in the macaque. In A. Peters and E.G. Jones., editors, *Cerebral Cortex Vol 10*, pages 1–36. Plenum Press, New York, 1991.

[75] Z. Koren and I. Koren. On the effect of floorplanning on the yield of large area integrated circuits. *IEEE-VLSI*, 5:3–14, 1997.

[76] K. Kim, C.-G. Hwang, and J.G. Lee. DRAM technology in the gigabit era. *IEEE Trans. Electron Devices*, 45:598–608, 1998.

[77] S. Esener. personal communication, Oct. 1997.

[78] J.J. Drolet. personal communication, Oct. 1997.