PHOTOREFRACTIVE 3-D DISKS FOR OPTICAL DATA STORAGE AND ARTIFICIAL NEURAL NETWORKS

Thesis by

Hsin-Yu Sidney Li

In Partial Fulfillment of the Requirements

for the Degree of

Doctorate of Philosophy

California Institute of Technology

Pasadena, California

1994

(Submitted September 15, 1994)

Acknowledgement

There are so many people who have helped me one way or another to reach this stage of life, that it would take a separate chapter to properly thank them all. I can only try to acknowledge some of the people who have helped and influenced me these past five years. There are bound to be people who I fail to mention here. My apologies to them all.

First I would like to thank my advisor, Prof. Demetri Psaltis, who has been both an inspiring teacher and friend. It has been a great privilege to study with him at Caltech. He has always been kind and understanding to the failures, and generous and encouraging to even the smallest accomplishments. If I have achieved anything in the past five years, I owe it to him.

I would like to thank my old friend and teacher, Ken Hsu who first introduced me to Prof. Psaltis and helped me settle down in Caltech during my first year here; and David Brady, who together with Ken initiated me in the mysteries of optical experiments.

I would like to thank Yong Qiao, with whom I worked with on the face-recognition experiment. It was great working with him. Thanks also to Xin An, who picked up quickly where we had left off in the real time training of the face-recognition system; to Robert Denkewalter for helping us test the 1st generation of the face-recognition system; and to Annette Grot and Jiafu Luo for helping us make the spatial filter used in the face-recognition experiment.

I would like to thank all my other friends and colleagues at Caltech, past and present: Robert Snapp, Scott Hudson, Mark Neifeld, John Hong, Claire Gu, Seiji Kobayashi, Alan Yamamura, Chuanyi Ji,

Cheol-Hoon Park, Subrata Rakshit, Francis Ho, Charlie Stirk, Steve Lin, Kevin Curtis, David Marx, Geoffrey Burr, Jean-Jacques Drolet, Ernest Chuang, Michael Levene, Allen Pu, and Yayun Liu. I thank them for their friendship, support, and their intellectual stimulus. Many of them also helped in testing the face-recognition experiment. My greatest gratitude to all of them for their patience.

I would like to thank Helen Carrier, Linda Dozsa, and Su McKinley for helping me in the miscellaneous chores needed to get work going in the lab.

I would like to thank my parents, Chin-Fung Li and Pei-Lan Chen, who have been supportive in all ways, not only in these past few years, but since I was born. They laid the foundation of what I am today, and nothing else would have been possible if they had not spent so much time on me. I would like to thank my sister Sherry and my brother-in-law Yeou-Fong for help and advice throughout the year, and for lending me their lap-top computer on which half of this thesis was written. To my uncles, aunts, and cousins, I am grateful to them for taking care of my parents in Taiwan and keeping them company when I could not be there.

Finally I would like thank my wife, Shu-Fen Wang, for putting up with my temper and negligence when I was busy, and my snoring at night when I was tired, and for taking care of me for these last three years. I wouldn't have been able to finish this thesis without her.

Abstract

This thesis is on the application of 3-D photorefractive crystals disks for holographic optical data storage and optical neural networks.

Chapter 1 gives some introductory background and motivation for the materials given in this thesis. In Chapter 2, the coupled-mode analysis and Born's approximation in anisotropic crystals is reviewed. The results are similar to that of isotropic materials. However, there are approximations that are often neglected in the literature.

Chapter 3 starts with the description of the holographic 3-D disk for data storage, and analyzes the various alignment errors and tolerance problems for a 3-D disk system. Of particular interest is the effects in image reconstruction caused by rotational angle error. An optimum configuration is found that minimizes this error.

Chapter 4 examines the data storage density of 3-D disks and volume holographic storage systems that utilize wavelength/angle and spatial multiplexing. The maximum storage density and the geometry that achieves this density is derived.

Chapter 5 discusses the diffraction efficiency of 3-D disks fabricated with photorefractive crystals. Practical geometries and crystal orientations for achieving maximum uniform diffraction efficiency are given and compared to the maximum obtainable diffraction efficiencies using arbitrary cut crystals. Experimental results are shown.

Also derived in this chapter are the double grating effect from crystal anisotropy, and the optimum configuration for getting maximum diffraction efficiency using the 90 degree recording geometry. The Kuhktarev band-transport model of the photorefractive effect is examined briefly with emphasis on the anisotropy of the material. The proper expression for the permittivity term in the space-charge field formula is derived.

Chapter 6 gives an example of an optical neural network that uses photorefractive crystals. It is the real time face-recognition system. The setup and experiments are described. Some properties of volume holographic correlators are given in the Appendix.

Table of Contents

Chapter 1
Introduction

Holography and volume holography have been known since Gabor first invented the subject in 1948 [1]. Using photorefractive crystals as a recording medium for volume holograms began once people realized that "optical damage" in photorefractive crystals could be used for real-time holography [2–4].

It has long been known that photorefractive crystals can record volume holograms in real time (i.e., without the need for any processing, chemical or otherwise). Much study has been done on the properties of photorefractive crystals as a holographic recording material since the early sixties [5–11]. In recent years there has been increasing interest in recording volume holograms in photorefractive crystals for optical data storage and artificial neural networks. As with photorefractive crystals, neither of these subjects are new. However, it is fair to say that both have undergone a revival in recent years.

The idea of optical data storage using holograms also began in the 60's and early 70's. Despite the effort and results of early researchers, no practical system emerged, partly because of lack of proper components (e.g., lasers, spatial light modulators, etc.) and partly because of competition from magnetic recording. Although it was realized early that optical holographic storage could provide very high storage density, the difficulties for making a practical system far outweighed the needs at that time, which were modest by todays standard and could be better addressed by other emerging technologies.

Artificial neural networks suffered a similar fate. Early studies had shown interesting results, but they failed to live up to the great expectations people had in them. One of the problems with neural networks then (and today) was that it was often difficult or impractical to scale up the networks. Even when

larger networks could be built (usually with much effort), it was difficult to predict their behavior based on studies of smaller networks. Many ideas that seemed promising on a small scale did not work satisfactory when tested on larger problems. Because of the limited capacity of computers at that time, simulations and experiments with larger networks and real-life problems were difficult. Another problem was the lack of adequate algorithms (e.g., back error propagation [12]) for training multilayer networks. The realization of the limitations of single layer networks [13] contrasted sharply with the initial enthusiasm and promises. The disappointment was all the greater because of this, and the field gradually faded into the background.

At the same time, digital electronic computers and computing were advancing quickly. Although traditional artificial intelligence (AI) also failed to deliver true "mechanical intelligence," electronic computers were extremely useful in many other areas which did not require such "intelligence." For the next two decades the studies of neural networks and holographic data storage practically disappeared from the mainstream of research activities.

Since the days of these initial efforts in optical data storage and neural networks, great advances have been made in computer and their supporting technology, as evidenced by the situation today. Interestingly enough, these advances have only increased the demand for more computing power and more data storage capacity. The combination of electronics, semiconductor, digital computing, and magnetic storage had proven superior to optical data storage and neural networks in the early days, but now they, too, were hard pressed to answer the demands for more computing power and more data storage.

As the tasks which computers were called upon to handle became more and more complicated, programming became more difficult and error prone. Many problems that require certain amount of intelligence have still not been solved by traditional AI, and in the middle 80's people again turned to neural networks [14–18]. With the revival of artificial neural networks, some of the problems encountered in the early studies have reemerged. In particular, although present

day computers are orders of magnitude more powerful than their predecessors, the large number of interconnections is still difficult to simulate by software or build as hardware. A new development in implementing artificial neural networks is the combination of optics and neural networks. In particular, using holograms as the interconnection weight between neurons. Although similar ideas had been conceived in the 60's [6,19–20], it was not until much later that the study of optical neural networks became a field of its own [21–27]. Photorefractive crystals became the natural candidate for this application, not only because of its capacity for storing huge number of interconnections as holograms, but more importantly because holograms stored in them can be recorded and modified in real time.

With the demand for more data storage, optical data storage has reappeared as compact discs (CD) and magneto-optic disks (MO). But the storage capacities of these storage media are also approaching their limits. As was realized in the earlier effort, holographic data storage has the potential of higher storage densities than conventional CDs. Another element in the growing interest in holographic data storage is the speed at which the data may be read out. The requirement of data storage lies not only in capacity but also accessibility. As the speed of computers increase, the rate at which data can be transferred from hard disk to RAM has become a serious bottleneck. In this respect, holographic memories also satisfies the demand for high readout speeds with its capability of reading out whole pages of data in parallel.

Despite all the progress made since the 60's and 70's, the recording material is still a problem in holographic recording. Photopolymers and photorefractive crystals remain the most promising materials. What has changed in the intervening years is not only a better understanding of the material, but also improvements in supporting components such as lasers and spatial light modulators (SLM). In addition, new types of photorefractive crystals have been synthesized which improve the speed and efficiency of recording.

This thesis is on the application of photorefractive crystals in holographic

optical data storage and optical neural networks. The main topic is the analysis of one particular volume holographic data storage system — the 3-D holographic disk system [28]. Although we have in mind photorefractive crystals as the recording media, many of the results in this thesis apply to any volume holographic recording system.

The organization of this thesis is as follows: In Chapter 2, we start by reviewing two theories used for analyzing volume holograms: coupled mode analysis and Born's approximation applied to light diffraction from a volume hologram in an anisotropic (uniaxial) crystal.

In the first part of Chapter 2, the first-order Born's approximation is applied to the problem of volume holography. The effect of material anisotropy is shown to be minor, however some of the approximations needed are often neglected in the literature. In the second part, the coupled-mode theory of volume holography is examined, and is shown (in the un-depleted pump limit) to agree with the predictions of Born's approximation. The emphasis is again on the anisotropy of the recording material, and the results are shown to be similar to the results for isotropic materials.

In Chapter 3, we describe the 3-D holographic disk (HD) system, and apply the results in Chapter 2 to analyzing this system. We examine the various sources that cause alignment errors upon readout of the hologram. Of particular interest is the rotation alignment sensitivity of holographic disks. It will be shown how an optimum configuration may be set up so that the alignment sensitivity is minimized with respect to disk rotation. We also analyze how Bragg-matching in volume holograms affects image reconstruction when there is rotation misalignment.

In Chapter 4 the geometry for obtaining maximum storage density in volume holograms using angle or wavelength multiplexing is derived. It will be shown that the maximum storage density is obtained for image or Fourier plane holograms. A recording geometry that achieves both maximum storage density and minimum rotation alignment sensitivity is shown. We then discuss some of the issues that

need to be considered when designing a 3-D holographic disk for data storage.

In Chapter 5, we analyze the effect of crystal orientation on the diffraction efficiency of volume holograms stored in photorefractive crystals. This is especially important for the 3-D HD system since the crystal is anisotropic and needs to be rotated. We start with a short review of the Kukhtarev band-transport model, also with attention to the anisotropy of photorefractive crystals, and derive the proper expression for the permittivity used in the space-charge field formula.

Next, practical geometries and crystal orientations for achieving maximum uniform diffraction efficiency are given and compared to the maximum obtainable diffraction efficiencies from the crystals. An experimental 3-D disk system and its results are shown. Also derived in this chapter are the double grating effect due to crystal anisotropy, and the optimum configuration for getting maximum diffraction efficiency using the 90 degree recording geometry.

The last part of this thesis, Chapter 6, is on the real time face-recognition system. This system uses the real time recording property of photorefractive crystals, and demonstrates the capabilities of photorefractive crystals used as the interconnection of an optical neural network. The algorithm that is used for training the network is described, and experimental results are given that demonstrate the capabilities of the system. In the appendix, we analyze in some detail the volume holographic correlator, which is the basic building block of the network.

References

1. D. Gabor, "A New Microscopic Principle," *Nature*, **161**, 777-778 (1948).
2. A. Ashkin, G. D. Boyd, J. M. Dziedzie, R. G. Smith, A. A. Ballman, J.J. Leninstein, and K. Nassau, "Optically-induced Refractive Index Inhomogeneities in $LiNbO_3$," *Appl. Phys. Lett.*, **9**(1), 72–74 (1966).
3. F. S. Chen, "A Laser-induced Inhomogeneity of Refractive Indices in KTN," *J. Appl. Phys.*, **38**(8), 3418–3420 (1967).

4. F. S. Chen, "Optically-induced Change of Refractive Indices in $LiNbO_3$ and $LiTaO_3$," *J. Appl. Phys.*, **40**(8), 3389–3396 (1969).
5. F. S. Chen, J. T. LaMacchia, and D. B. Fraser, "Holographic Storage in Lithium Niobate," *Appl. Phys. Lett.*, **13**(7), 223-225 (1968).
6. P. J. Van Heerden, "Theory of Optical Information Storage in Solids," *Appl. Optics*, **2**, 393–400 (1963).
7. K. Bløtekjaer, "Limitations on Holographic Storage Capacity of Photochromic and Photorefractive Media," *Appl. Opt.* **18**, 57–67 (1979).
8. D. von der Linde and A.M. Glass, "Photorefractive Effects for Reversible Holographic Storage of Information," *Appl. Phys.*, **8**, 85–100 (1975).
9. J.-P. Huignard, J.-P. Herriau, and F. Micheron, "Coherent/Selective Erasure of Superimposed Volume Holograms in $LiNbO_3$," *Appl. Phys. Lett.*, **26**, 256–258 (1975).
10. L. d'Auria, J.-P. Huignard, C. Slezak, and E. Spitz, "Experimental Holographic Read-Write Memory using 3-D Storage," *Appl. Opt.*, *13*, 808–818 (1974).
11. J. B. Thaxter, "Electrical Control of Holographic Storage in Strontium Barium Niobate," *Appl. Phys. Lett.*, **15**(7), 210-212 (1969).
12. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations By Error Propagation," in *Parallel Distributed Processing, Vol 1*, Chapter 8. D. E. Rumelhart and J. L. McClelland, Eds. (MIT Press, Cambridge, 1986).
13. M. Marvin and S. Papert, *Perceptrons.* (MIT Press, Cambridge, 1986.)
14. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Natl. Acad. Sci. USA*, **79**, 2554–2558 (1982).
15. J. J. Hopfield, "Neurons with Graded Response Have Collective Computational Properties Like Those Of Two-State Neurons," *Proc. Ntl. Acad. Sci. USA*, **81**, 3088-3092 (1984).
16. J. A. Anderson, J. W. Silverstein, S. A. Ritz, and R. S. Jones, "Distinctive

Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model," *Psychological Review*, **84**, 413–451 (1977).

17. S. Grossberg, *Studies of Mind and Brain* (Reidel, Boston, 1982).
18. T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).
19. D. Gabor, "Associate Holographic Memories," *IBM J. Res. Devel.*, **13**(2), 156–159 (1969).
20. P. J. Van Heerden, "A New Optical Method Of Storing And Retrieving Information," *Appl. Optics*, **2**, 387–392 (1963).
21. D. Psaltis and N. Farhat, "Optical Information Processing Based on an Associative -Memory Model of Neural Nets with Thresholding and Feedback," *Opt. Lett.*, **10**, 98–100 (1985).
22. N. H. Farhat, D. Psaltis, A. Prat, S. Y. Shin, and S. Y. Lee, "Optical Implementation of Quadratic Associative Memory with Outer-Product Storage," *Opt. Lett.*, **13**, 693–695, (1988).
23. N. H. Farhat and D. Psaltis, "Optical Implementation of Associative Memory Based on Models of Neural Networks," in *Optical Signal Processing* Chap. 2.3, J. L. Horner, Ed. (Academic Press, San Diego, 1987).
24. L. S. Lee, H. M. Stoll, and M. C. Tackitt, "Continuous-time Optical Neural Associative Memory," *Opt. Lett.*, **14**(3), 162–164 (1989).
25. B. H. Soffer, G. J. Dunning, Y. Owechko, and E. Maron, "Associative Holographic Memory with Feedback using Phase-conjugate Mirrors," *Opt. Lett.*, **11**, 118–120 (1986).
26. D. Psaltis, D. Brady, and K. Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.*, **27**, 1752–1759 (1988).
27. D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in Artificial Neural Networks," *Nature*, **343**(25), 325–330 (1990).
28. Demetri Psaltis,"Parallel Optical Memories", *Byte*, **17**(9), 179–182 (1992).

Chapter 2
Diffraction From Anisotropic Materials

In this chapter, we review two theories describing diffraction from a volume hologram inside an anisotropic material. The first is Born's approximation, and the second is coupled-mode analysis. The coupled-mode analysis is the usual method for analyzing diffraction for volume holograms. For holographic optical storage systems, a large number of holograms is multiplexed within the same volume. In this case, the individual holograms are weak, and an alternative way of analyzing volume holograms is the first-order Born's approximation. As expected, the results from Born's approximation agree with that of coupled-mode analysis under the undepleted-pumping beam approximation.

In either model, the anisotropic aspect of the material is usually neglected. In this chapter, it will be shown that if certain approximations are made, then the usual results (derived under the assumption that the material is isotropic) are still applicable, provided minor changes are made. Although the results are essentially the same, the details are interesting, and are often neglected in the literature.

2.1. Born's Approximation

2.1.1. Born's Approximation in Isotropic Materials

We start by considering the first-order Born's Approximation [1] for isotropic media. In MKSA units Maxwell's equations are

$$\nabla \cdot \mathbf{B} = 0, \tag{2.1}$$

$$\nabla \cdot \mathbf{D} = \rho, \tag{2.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \tag{2.3}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}, \tag{2.4}$$

where

$$\mathbf{D} = \epsilon \mathbf{E} \tag{2.5}$$

and

$$\mathbf{B} = \mu \mathbf{H}. \tag{2.6}$$

We assume that ϵ and μ can have spatial variations.

From the four Maxwell's equations and the vector identity

$$\nabla^2 \mathbf{D} = -\nabla \times \nabla \times \mathbf{D} + \nabla(\nabla \cdot \mathbf{D}) = -\nabla \times \nabla \times \mathbf{D} + \nabla \rho, \tag{2.7}$$

we get the wave equation

$$\nabla^2 \mathbf{D} - \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2} = \left(\nabla \rho + \epsilon_0 \mu_0 \frac{\partial \mathbf{J}}{\partial t} \right) - \nabla \times \nabla \times (\mathbf{D} - \epsilon_0 \mathbf{E}) + \epsilon_0 \frac{\partial}{\partial t} \nabla \times (\mathbf{B} - \mu_0 \mathbf{H}), \tag{2.8}$$

where we use ϵ_0 and μ_0 to denote the unperturbed permittivity and permeability of the material [1], and $\mathbf{D} - \epsilon_0 \mathbf{E}$, etc., give the perturbation. The above equation is exact, when ϵ_0 and μ_0 are scalars; no approximations have been made up to this point. Assuming a time dependency of $e^{-j\omega t}$, Eq. (2.8) becomes [2]

$$(\nabla^2 + k^2)\mathbf{D} = (\nabla \rho - j\omega \epsilon_0 \mu_0 \mathbf{J}) - \nabla \times \nabla \times (\mathbf{D} - \epsilon_0 \mathbf{E}) - j\omega \epsilon_0 \nabla \times (\mathbf{B} - \mu_0 \mathbf{H}), \tag{2.9}$$

where $k^2 = \omega^2 \epsilon_0 \mu_0$. The right-hand side of Eq. (2.9) contain the "source" terms to the wave equation.

For the rest of this chapter, it will be assumed that μ_0 is a scalar and $\mathbf{B} - \mu_0 \mathbf{H} = 0$. We will also drop the $\nabla \rho - j\omega \epsilon_0 \mu_0 \mathbf{J}$ term. This can be done for the

---

[1] Not to be confused with the permittivity and permeability of vacuum.

[2] See for example Eq. (9.102) (page 420) of J. D. Jackson's *Classical Electrodynamics* (John Wiley & Sons, New York, 1962) [1].

following reasons: in photorefractive crystals, the total charge is neutral, and is practically static. Thus it does not contribute to the diffracted light that is at much higher frequencies. Similarly, the currents are small and slow varying.

We now assume that the perturbation is linear, i.e.,

$$\mathbf{D} = (\epsilon_0 + \Delta\epsilon)\mathbf{E}, \tag{2.10}$$

(where $\Delta\epsilon$ may be a tensor). In the (first-order) Born's approximation, we take

$$\mathbf{D} - \epsilon_0\mathbf{E} = \Delta\epsilon\mathbf{E} \approx \Delta\epsilon\mathbf{E}^{(0)} = \frac{\Delta\epsilon}{\epsilon_0}\mathbf{D}^{(0)}, \tag{2.11}$$

where the superscript (0) is used to denote the solution to the unperturbed incident wave equation (i.e., when $\Delta\epsilon = 0$).

The Green's function for the operator $(\nabla^2 + k^2)$ is [2]

$$G(\mathbf{x}, \mathbf{x}') = -\frac{1}{4\pi^2}\frac{e^{jk|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}. \tag{2.12}$$

We can formally write the solution to Eq. (2.9) as

$$\mathbf{D}(\mathbf{x}) \approx \mathbf{D}^{(0)}(\mathbf{x}) + \frac{1}{4\pi}\int d\mathbf{x}' \, \frac{e^{jk|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}\left\{\nabla\times\nabla\times\left[\frac{\Delta\epsilon}{\epsilon_0}\mathbf{D}^{(0)}(\mathbf{x}')\right]\right\}. \tag{2.13}$$

From physical considerations, we include only outward propagating waves and discard the converging wave solution. (Recall that the time dependency is $e^{-j\omega t}$.)

We now expand the spherical wave $e^{jk|\mathbf{x}|}/|\mathbf{x}|$ in terms of its plane wave components [3]

$$\frac{e^{jk|\mathbf{x}|}}{|\mathbf{x}|} = \frac{j}{2\pi}\int\int dk_x \, dk_y \sum_{\alpha=1}^{2}\frac{1}{|k_{z,\alpha}|}\cdot e^{j(k_x x + k_y y + k_{z,\alpha} z)}, \tag{2.14}$$

where $k_{z,\alpha} = \sqrt{k^2 - k_x^2 - k_y^2}$ for $\alpha = 1$ and $k_{z,\alpha} = -\sqrt{k^2 - k_x^2 - k_y^2}$ for $\alpha = 2$. The components with $\alpha = 1$ give plane waves traveling in the positive z direction, and those with $\alpha = 2$ give plane waves traveling in the negative z direction. In

[3] See for example reference [5]. A proof is given in Appendix A,

the $z > 0$ region, only the forward waves ($\alpha = 1$) exist, and in the $z > 0$ region, only the backward waves ($\alpha = 2$) exist. Using this, we can write Eq. (2.13) as

$$\begin{aligned}
&\mathbf{D}(\mathbf{x}) \\
\approx &\mathbf{D}^{(0)}(\mathbf{x}) + \frac{j}{8\pi^2} \sum_{\alpha=1}^{2} \int\int dk_x\, dk_y\, \frac{e^{j\mathbf{k}_\alpha \cdot \mathbf{x}}}{|k_{z,\alpha}|} \int d\mathbf{x}'\, e^{-j\mathbf{k}_\alpha \cdot \mathbf{x}'} \left\{ \nabla \times \nabla \times \left[\frac{\Delta\epsilon}{\epsilon_0} \mathbf{D}^{(0)}(\mathbf{x}') \right] \right\} \\
= &\mathbf{D}^{(0)}(\mathbf{x}) + \sum_{\alpha=1}^{2} \int\int \mathbf{A}(\mathbf{k}_\alpha)\, e^{j\mathbf{k}_\alpha \cdot \mathbf{x}}\, dk_x\, dk_y, \qquad (2.15)
\end{aligned}$$

where $\mathbf{k}_\alpha = (k_x,\ k_y,\ k_z, \alpha)$, and

$$\mathbf{A}(\mathbf{k}_\alpha) = \frac{j}{8\pi^2 |k_{z,\alpha}|} \cdot \int d\mathbf{x}'\, e^{-j\mathbf{k}_\alpha \cdot \mathbf{x}'} \left\{ \nabla \times \nabla \times \left[\frac{\Delta\epsilon}{\epsilon_0} \mathbf{D}^{(0)}(\mathbf{x}') \right] \right\}. \qquad (2.16)$$

Thus the diffracted wave is decomposed into its plane wave components, $\mathbf{A}(\mathbf{k}_\alpha)$. Note that $\mathbf{A}(\mathbf{k}_\alpha)$ is just the 3-D Fourier transform of $\nabla \times \nabla \times \left[\frac{\Delta\epsilon}{\epsilon_0} \mathbf{D}^{(0)}(\mathbf{x}') \right]$ multiplied by a constant.

2.1.2. Diffraction from an Isotropic Medium

We now consider the case where the incident wave is a plane wave. Let the (unperturbed) incident wave be [4]

$$\mathbf{D}^{(0)}(\mathbf{x}) = D_0 e^{j\mathbf{k}_2 \cdot \mathbf{x}}\, \mathbf{e}_2, \qquad (2.17)$$

where $\mathbf{e}_2$ is a unit vector that gives the polarization of the plane wave, and satisfies $\mathbf{e}_2 \cdot \mathbf{k}_2 = 0$. We first analyze the situation for a simple grating.

Assume that inside the region V there is a simple grating given by

$$\Delta\epsilon = \epsilon_0 \Delta\epsilon_r e^{-j\mathbf{K} \cdot \mathbf{x}} + c.c., \qquad (2.18)$$

---

[4] We will use the subscript 1 in the wave vectors and polarization vectors to denote the scattered, or diffracted wave (the signal beam) and subscript 2 to denote the incident wave (the reference beam).

where $\mathbf{K}$ is the grating vector and $\Delta\epsilon$ may be a tensor. (*c.c.* here means complex conjugate. We assume $\Delta\epsilon$ to be real.) In the following we will consider only the $e^{-j\mathbf{K}\cdot\mathbf{x}}$ term of $\Delta\epsilon$, since we can get the contribution from its complex conjugate by substituting $\mathbf{K}$ for $-\mathbf{K}$ and then use superposition to find the final answer. Let

$$\mathbf{k}_1 = -\mathbf{K} + \mathbf{k}_2. \tag{2.19}$$

Then

$$\frac{\Delta\epsilon}{\epsilon_0}\mathbf{D}^{(0)}(\mathbf{x}) = D_0(\Delta\epsilon_r\mathbf{e}_2)e^{j\mathbf{k}_1\cdot\mathbf{x}} \tag{2.20}$$

and

$$\nabla\times\nabla\times\left[\frac{\Delta\epsilon}{\epsilon_0}\mathbf{D}^{(0)}(\mathbf{x})\right] = -D_0\ \mathbf{k}_1\times\big(\mathbf{k}_1\times(\Delta\epsilon_r\mathbf{e}_2)\big)e^{j\mathbf{k}_1\cdot\mathbf{x}}. \tag{2.21}$$

Note that the expression in Eq. (2.21) is perpendicular to $\mathbf{k}_1$. If we let $\mathbf{v} = \mathbf{k}_1/k_1$ be the unit vector in the direction of $\mathbf{k}_1$, and $\mathbf{u}_1$ the unit vector in the direction of

$$-\mathbf{k}_1\times\big(\mathbf{k}_1\times(\Delta\epsilon_r\mathbf{e}_2)\big) = k_1^2[(\Delta\epsilon_r\mathbf{e}_2) - (\mathbf{v}\cdot\Delta\epsilon_r\mathbf{e}_2)\mathbf{v}], \tag{2.22}$$

then we have [5]

$$\nabla\times\nabla\times\left[\frac{\Delta\epsilon}{\epsilon_0}\mathbf{D}^{(0)}(\mathbf{x})\right] = k_1^2D_0\ (\mathbf{u}_1\cdot\Delta\epsilon_r\mathbf{e}_2)e^{j\mathbf{k}_1\cdot\mathbf{x}}\mathbf{u}_1. \tag{2.23}$$

Substituting this into Eq. (2.16), we have

$$\mathbf{A}(\mathbf{k}_\alpha) = \frac{jk_1^2D_0}{8\pi^2|k_{z,\alpha}|}\cdot\mathbf{u}_1(\mathbf{u}_1\cdot\Delta\epsilon_r\ \mathbf{e}_2)\int_V e^{-j(\mathbf{k}_\alpha-\mathbf{k}_2)\cdot\mathbf{x}'}\left(e^{-j\mathbf{K}\cdot\mathbf{x}'}\right)\ d\mathbf{x}', \tag{2.24}$$

where the integration is over the region V.

For a general $\Delta\epsilon$, we write it as a sum of simple gratings

$$\Delta\epsilon(\mathbf{x}') = \sum_a \epsilon_0\Delta\epsilon_a e^{-j\mathbf{K}_a\cdot\mathbf{x}'}. \tag{2.25}$$

---

[5] If $\mathbf{s} = \mathbf{w} - (\mathbf{w}\cdot\mathbf{v})\mathbf{v} = s\mathbf{u}$, where $|\mathbf{v}| = 1$ and $|\mathbf{u}| = 1$, then $s^2 = \mathbf{s}\cdot\mathbf{s} = \mathbf{s}\cdot\mathbf{w} = w^2 - (\mathbf{w}\cdot\mathbf{v})^2$ and $\mathbf{s} = s\mathbf{u} = (\mathbf{u}\cdot\mathbf{w})\mathbf{u}$.

In this case we have

$$\mathbf{A}(\mathbf{k}_\alpha) = \frac{jD_0}{8\pi^2|k_{\alpha z}|} \sum_a k_{a1}^2 \mathbf{u}_{a1} (\mathbf{u}_{a1} \cdot \Delta\epsilon_a \mathbf{e}_2) \int_V e^{-j(\mathbf{k}_\alpha - \mathbf{k}_2)\cdot\mathbf{x}'} \cdot e^{-j\mathbf{K}_a\cdot\mathbf{x}'} \, d\mathbf{x}', \tag{2.26}$$

where

$$\mathbf{k}_a = \mathbf{k}_2 - \mathbf{K}_a, \tag{2.27}$$

and $\mathbf{u}_a$ is the unit vector in the direction of $-\mathbf{k}_a \times (\mathbf{k}_a \times (\Delta\epsilon_a \mathbf{e}_2))$. The integral in Eq. (2.26) gives the amplitude of the plane wave component of the diffracted wave that has wave vector $\mathbf{k}_\alpha$.

In general the expression in Eq. (2.26) cannot be simplified further. However, if we assume that all $\mathbf{k}_a$ and $\mathbf{u}_a$ are approximately the same in the summation in Eq (2.25) and (2.26), then we can approximate

$$\mathbf{k}_a \approx \mathbf{k}_1 \tag{2.28}$$

$$\mathbf{u}_a \approx \mathbf{u}_1 \tag{2.29}$$

to be constants, and move them outside the summation. Eq. (2.26) then becomes

$$\mathbf{A}(\mathbf{k}_\alpha) \approx \frac{jk_1^2 D_0}{8\pi^2|k_{z,\alpha}|} \cdot \mathbf{u}_1 \big(\mathbf{u}_1 \cdot U(\mathbf{k}_\alpha)\mathbf{e}_2\big), \tag{2.30}$$

where

$$U(\mathbf{k}_\alpha) = \int_V d\mathbf{x}' \, e^{-j(\mathbf{k}_\alpha - \mathbf{k}_2)\cdot\mathbf{x}'} \, \frac{\Delta\epsilon(\mathbf{x}')}{\epsilon_0} \tag{2.31}$$

is the (tensor) 3-D Fourier transform of the perturbation $\Delta\epsilon/\epsilon_0$. Eq. (2.30) and (2.31) are exact (within the Born's approximation model) for a single grating, but are only approximately true when the assumptions in Eq.s (2.28) and (2.29) hold. For the applications we are interested in (namely, holographic data storage and optical neural networks), the conditions are true. For the general case, however, it is necessary to go back to Eq. (2.26).

We now apply these results to the case of a single grating inside an infinite slab. We take V to be the region between $z = -L/2$ to $L/2$, and use Eq.s (2.30)

and (2.31) (or Eq. (2.24)). The result is

$$\mathbf{D}(\mathbf{x}) = \mathbf{D}^{(0)}(\mathbf{x}) - j\left(\frac{k_1^2 D_0 L}{2k'_{1z}}\right)\mathrm{sinc}\left(\frac{L}{2}(k'_{1z}-k_{1z})\right)\,(\mathbf{u}_1\cdot\Delta\epsilon_r\mathbf{e}_2)\,e^{j\mathbf{k}'_1\cdot\mathbf{x}}\,\mathbf{u}_1$$
$$+ j\left(\frac{k_1^2 D_0 L}{2k'_{1z}}\right)\mathrm{sinc}\left(\frac{L}{2}(k'_{1z}+k_{1z})\right)\,(\mathbf{u}_1\cdot\Delta\epsilon_r\mathbf{e}_2)\,e^{j\mathbf{k}''_1\cdot\mathbf{x}}\,\mathbf{u}_1, \quad (2.32)$$

where

$$k'_{1z} = \sqrt{k^2 - k_{1x}^2 - k_{1y}^2}, \quad (2.33)$$

$$\mathbf{k}'_1 = (k_{1x}, k_{1y}, k'_{1z}), \quad (2.34)$$

and

$$\mathbf{k}''_1 = (k_{1x}, k_{1y}, -k'_{1z}). \quad (2.35)$$

The second term in Eq. (2.32) is a forward propagating plane wave (when $\mathbf{x}$ is in the $z > L/2$ region outside the crystal) while the third term is a backward propagating plane wave (when $\mathbf{x}$ is in the $z < -L/2$ region outside the crystal). The sinc function gives the familiar Bragg-mismatch factor. When k_{1z} (the z-component of $\mathbf{k}_1$) is close to k'_{1z} (the z component that satisfies the propagating condition, Eq. (2.33)), the backward propagating wave (the third term) is negligible. Earlier we had neglected the complex conjugate part of $\Delta\epsilon$ in Eq. (2.15). We can go back and repeat the process above to yield a similar expression for the contribution from the complex conjugate term in $\Delta\epsilon$. However, since we are assuming that $\mathbf{k}_1$ (Eq. (2.19)) is close to Bragg-match, the two terms we get from the complex conjugate term will be far from Bragg-match conditions, and are therefore negligible.

As mentioned before, $\mathbf{u}_1$ is perpendicular to $\mathbf{k}_1$, but *not* perpendicular to $\mathbf{k}'_1$. However, the amplitude of the forward propagating wave approaches zero when the z components of $\mathbf{k}'_1$ and $\mathbf{k}_1$ differ more than $1/L$, thus $\mathbf{u}_1$ will be approximately perpendicular to $\mathbf{k}'_1$.

The diffraction efficiency at Bragg-match angles (i.e., when $\Delta k_z = 0$) is from

Eq. (2.32) [6]

$$\eta = \left\{ \frac{k^2 L}{2k_{z0}} |\mathbf{u}_1 \cdot \Delta\epsilon_r \mathbf{e}_2| \right\}^2 = \left\{ \frac{kL}{2\cos\theta_1} |\mathbf{u}_1 \cdot \Delta\epsilon_r \mathbf{e}_2| \right\}^2 \tag{2.36}$$

where θ_1 is the angle between the z-axis (normal to the crystal surface) and $\mathbf{k}_1$.

On the other hand (for transmission type holograms), coupled-mode analysis gives us (at Bragg-matched angles) [3]

$$\eta = \left| \frac{\cos\theta_2}{\cos\theta_1} \right| \sin^2 \left(\frac{kL}{2\sqrt{\cos\theta_1 \cos\theta_2}} |\mathbf{e}_1 \cdot \Delta\epsilon_r \mathbf{e}_2| \right) \approx \left\{ \frac{kL}{2\cos\theta_1} |\mathbf{e}_1 \cdot \Delta\epsilon_r \mathbf{e}_2| \right\}^2, \tag{2.37}$$

for weak holograms, where θ_2 is the angle between $\mathbf{k}_2$ and the z-axis, and $\mathbf{e}_1$ is the polarization of the diffracted wave in the $\mathbf{k}_1$ direction. Comparison of Eq.s (2.36) and (2.37) show that the predictions of coupled-mode analysis for weak holograms are identical to the results derived using Born's approximation if we take $\mathbf{u}_1 = \mathbf{e}_1$. Noted that in coupled-mode analysis, we start out by postulating a polarization vector $\mathbf{e}_1$, whereas in Born's approximation, the polarization vector $\mathbf{u}_1$ comes out of the equations. In both cases, the Bragg-angle selectivity is given by the same sinc factor in Eq. (2.32). For reflection type holograms, the sine in Eq. (2.37) is replaced by a hyperbolic tangent, but for weak holograms, the result is the same.

As a second example, we consider the situation of a simple grating in the region $x = A/2$ to $x = -A/2$, $y = B/2$ to $y = -B/2$, and $z = L/2$ to $z = -L/2$. Using Eq. (2.24), we get

$$\mathbf{D}(\mathbf{x}) = \mathbf{D}^{(0)}(\mathbf{x}) + \frac{jk_1^2 D_0}{8\pi^2} (\mathbf{u}_1 \cdot \Delta\epsilon_r \mathbf{e}_2)\mathbf{u}_1 \cdot \int\int dk_x \, dk_y \sum_{\alpha=1}^{2} \left\{ \frac{ABL}{|k_{z,\alpha}|} \right.$$
$$\left. \operatorname{sinc}\left(\frac{A}{2}\Delta k_{x,\alpha}\right) \operatorname{sinc}\left(\frac{B}{2}\Delta k_{y,\alpha}\right) \operatorname{sinc}\left(\frac{L}{2}\Delta k_{z,\alpha}\right) \right\} e^{j\mathbf{k}_\alpha \cdot \mathbf{x}}, \tag{2.38}$$

where $\mathbf{k}_\alpha = (k_x, k_y, k_{z,\alpha})$, and

$$\Delta\mathbf{k}_\alpha = \mathbf{k}_\alpha - \mathbf{k}_1. \tag{2.39}$$

---

[6] We used $\epsilon = \epsilon_0 + \Delta\epsilon \, e^{-j\mathbf{K}\cdot\mathbf{x}} + c.c.$, whereas in some definitions, people use $\epsilon = \epsilon_0 + \frac{1}{2}\Delta\epsilon \, e^{-j\mathbf{K}\cdot\mathbf{x}} + c.c.$ instead.

Thus the diffracted wave given by the second term in Eq. (2.38) is a combination of plane waves centered around $\mathbf{k}_1$ where the "spread" in the spectrum is given by the sinc functions. In the limit where $A \to \infty$ and $B \to \infty$, the sinc's for x and y become delta functions, and the expression reduces to the expression in Eq. (2.32).

We can simplify the expression in Eq. (2.38) if we approximate the integral by integrating over only the central lobe of the sinc functions for k_x and k_y. We will assume that $\mathbf{k}_1$ has a positive z component, and consider only the $\alpha = 1$ components (i.e., plane waves that travels in the positive z direction) since for the $\alpha = 2$ components Δk_z is large. In the following, we will drop α from the notation. Note that the range of k_x and k_y are $2\pi/A$ and $2\pi/B$. From Eq. (2.39), we have

$$\Delta k_x = k_x - k_{1x} \tag{2.40}$$

$$\Delta k_y = k_y - k_{1y}. \tag{2.41}$$

We now change variables to Δk_x and Δk_y. We have

$$\begin{aligned}\sqrt{k^2 - k_x^2 - k_y^2} &= \left[k^2 - (k_{1x} - \Delta k_x)^2 - (k_{1y} - \Delta k_y)^2\right]^{1/2} \\ &\approx k'_{1z} + \frac{k_{1x}}{k'_{1z}}\Delta k_x + \frac{k_{1y}}{k'_{1z}}\Delta k_y,\end{aligned} \tag{2.42}$$

where k'_{1z} is given by Eq. (2.33).

The approximation above assumes that

$$k'^2_{1z} \gg 4\pi\left(\frac{k_{1x}}{A} + \frac{k_{1y}}{B}\right), \tag{2.43}$$

which is satisfied if we assume that $\mathbf{k}_1$ is approximately pointing along the positive z direction (paraxial approximation) and A and B are not too small. We also approximate $1/|k_z|$ by $1/|k_{1z}|$ and

$$\Delta k_z \approx k'_{1z} - k_{1z}. \tag{2.44}$$

Thus Δk_z gives the mismatch of $\mathbf{k}_1$ along the z direction. We can now write the integral in Eq. (2.38) (with only the $\alpha = 1$ term) as

$$\begin{aligned}
&\int\int dk_x\, dk_y \left\{ \frac{ABL}{|k_z|} \mathrm{sinc}\left(\frac{A}{2}\Delta k_x\right) \mathrm{sinc}\left(\frac{B}{2}\Delta k_y\right) \mathrm{sinc}\left(\frac{L}{2}\Delta k_z\right) \right\} e^{j\mathbf{k}\cdot\mathbf{x}} \\
\approx &\frac{ABL}{k_{1z}}\, e^{j\mathbf{k}_1'\mathbf{x}}\, \mathrm{sinc}\left(\frac{L}{2}(k_{1z}' - k_{1z})\right) \\
&\cdot \int d(\Delta k_x)\, \mathrm{sinc}\left(\frac{A}{2}\Delta k_x\right)\, \exp\left\{-j\left(x - \frac{k_{1x}}{k_{1z}'}z\right)\Delta k_x\right\} \\
&\cdot \int d(\Delta k_y)\, \mathrm{sinc}\left(\frac{B}{2}\Delta k_y\right)\, \exp\left\{-j\left(y - \frac{k_{1y}}{k_{1z}'}z\right)\Delta k_y\right\} \\
= &\frac{4\pi^2 L}{k_{1z}}\, \mathrm{sinc}\left(\frac{L}{2}(k_{1z}' - k_z)\right)\, e^{j\mathbf{k}_1'\mathbf{x}}. \\
&\mathrm{rect}\left(\frac{1}{A}\left(x - \frac{k_{1x}}{k_{1z}'}z\right)\right)\, \mathrm{rect}\left(\frac{1}{B}\left(y - \frac{k_{1y}}{k_{1z}'}z\right)\right).
\end{aligned} \tag{2.45}$$

Note that the rect functions are just the geometrical projections of the cross-section of the x–y plane along the $\mathbf{k}_1'$ direction at z (where the field is being measured). Eq. (2.45) is therefore the familiar Bragg-mismatched plane wave truncated by the geometrical projection of the finite x–y plane cross-section of the hologram. Since we have only assumed the paraxial approximation of $\mathbf{k}_1'$ along the z-axis, the above formula is valid also for the 90-degree recording geometry (this will be described in more detail in more detail in later Chapters). For example, we can have the incident wave (in the $\mathbf{k}_2$ direction) traveling approximately along the positive x direction, and the diffracted wave (in the $\mathbf{k}_1' \approx \mathbf{k}_1$ direction) traveling approximately along the positive z direction. From Eq. (2.45), the angle selectivity is determined by the thickness of the hologram L and the Bragg-mismatch $k_{1z}' - k_{1z}$ along the z direction.

2.1.3. Born's Approximation for Anisotropic Crystals

In the wave equation (Eq. (2.8)), it was assumed that both ϵ_0 and μ_0 were scalars. In the case of anisotropic crystals ϵ_0 (the unperturbed permittivity) is a tensor constant, and the wave equation, Eq. (2.8), is no longer true since in

general:

$$\nabla \times (\epsilon_0 \mathbf{E}) \neq \epsilon_0 \nabla \times \mathbf{E}. \tag{2.46}$$

Because of this, the x, y, z components of the fields are no longer separable and the scalar Green's function $e^{jk|\mathbf{x}-\mathbf{x}'|}/|\mathbf{x}-\mathbf{x}'|$ cannot be used. Nevertheless, the system is linear, and there exists a Green's function for the solution. For the anisotropic material, however, the Green's function is a tensor.

To derive Born's approximation for the general case of anisotropic crystals, we again start with Maxwell's equations

$$\nabla \cdot \mathbf{B} = 0, \tag{2.47}$$

$$\nabla \cdot \mathbf{D} = 0, \tag{2.48}$$

$$\nabla \times (\epsilon^{-1} \mathbf{D}) = j\omega\mu \mathbf{H}, \tag{2.49}$$

$$\nabla \times \mathbf{H} = -j\omega \mathbf{D}, \tag{2.50}$$

where we assume the $\epsilon^{-j\omega t}$ time factor, and assume that there is no current or charge. (As remarked earlier, the charge distribution in the material is very close to static and therefore does not contribute to the diffracted light field.) We assume that

$$\epsilon = \epsilon_0 + \Delta\epsilon, \tag{2.51}$$

where ϵ_0 is the unperturbed permittivity tensor. For small $\Delta\epsilon$ (also a tensor), we have [7]

$$\Delta(\epsilon^{-1}) = -\epsilon^{-1} \Delta\epsilon \, \epsilon^{-1}, \tag{2.52}$$

and we can write Eq.s (2.47) to (2.50) as

$$\nabla \cdot \mathbf{B} = 0, \tag{2.53}$$

$$\nabla \cdot \mathbf{D} = 0, \tag{2.54}$$

---

[7] We have $(\epsilon + \Delta\epsilon)^{-1} = [\epsilon(I + \epsilon^{-1}\Delta\epsilon)]^{-1} = (I + \epsilon^{-1}\Delta\epsilon)^{-1}\epsilon^{-1}$. But for small $\Delta\epsilon$, we have $(I + \epsilon^{-1}\Delta\epsilon)^{-1} \approx (I - \epsilon^{-1}\Delta\epsilon)$. Therefore $(\epsilon + \Delta\epsilon)^{-1} \approx \epsilon^{-1} - \epsilon^{-1}\Delta\epsilon\,\epsilon^{-1}$, and $\Delta(\epsilon^{-1}) \approx \epsilon^{-1}\Delta\epsilon\,\epsilon^{-1}$.

$$\nabla \times (\epsilon_0^{-1}\mathbf{D}) = j\omega\mu\mathbf{H} - \mathbf{M}, \tag{2.55}$$

$$\nabla \times \mathbf{H} = -j\omega\mathbf{D}, \tag{2.56}$$

where

$$\mathbf{M} = -\nabla \times (-\epsilon^{-1}\Delta\epsilon\, \epsilon^{-1}\mathbf{D}). \tag{2.57}$$

Born's (first-order) approximation is obtained when we approximate $\mathbf{D}$ by $\mathbf{D}^{(0)}$, the unperturbed solution to Eq.s (2.53)–(2.56). Thus we take

$$\begin{aligned}\mathbf{M} &\approx -\nabla \times (-\epsilon^{-1}\Delta\epsilon\, \epsilon^{-1}\mathbf{D}^{(0)}) \\ &= -\nabla \times (-\epsilon^{-1}\Delta\epsilon\, \mathbf{E}^{(0)}) \end{aligned} \tag{2.58},$$

and the problem is reduced to finding the solution for (the anisotropic) Maxwell's equations for a given magnetic current $\mathbf{M}$. Since the system is linear, if we can find the *dyadic* (or *tensor*) Green's function (impulse response) to the system, $G(\mathbf{x}, \mathbf{x}')$ (where G is a tensor), then in principle we can find the solution given any source $\mathbf{M}$ by superposition.

As noted earlier, for isotropic materials the problem could be *scalarized* since the components of $\mathbf{D}$ could be decoupled, and each component actually has the same (scalar) Green's function. In the general case of anisotropic materials, the components are coupled through the permittivity tensor ϵ, and we must use a dyadic (tensor) Green's function. Unfortunately, there is no simple expression for the dyadic Green's function for Maxwell's equations. It can be shown that the dyadic Green's function G can be expressed in terms of some scalar Green's function, g. However, no analytic solution to g is known [4]. Another approach is to express G directly in terms of plane wave components that are eigenmodes of the unperturbed anisotropic material [5–8]. This is the method that will be used below. We will consider only uniaxial crystals. Note that this method can also be used in the isotropic case, and would work even if we did not know that the Green's function for isotropic materials is the spherical wave.

First, recall that in a uniaxial crystal there are two propagating eigenmodes in any direction of propagation [9]: they are the extraordinary mode (e-mode)

where $H_c = 0$ (also called the TM mode), and the ordinary mode (o-mode) where $E_c = 0$ (also called the TE mode). Here E_c, etc., are the component of the fields along the direction of the optical axis. Usually the optical axis is the z axis. However, in this chapter, we assume an arbitrary crystal orientation and in this case z is not necessarily in the direction of the optical axis.

Given any transverse wave vector $\mathbf{k}_t = (k_x, k_y)$, we can find four plane wave modes having wave vector $\mathbf{k}_\alpha = (\mathbf{k}_t, k_\alpha)$ (i.e., $k_z = k_\alpha$, where k_α depends on $\mathbf{k}_t$) that satisfy Maxwell's equations. Each α here corresponds to an eigenmode plane wave: 2 traveling in the positive z direction ($\alpha = 1$: o-mode/TE, and $\alpha = 2$: e-mode/TM), and 2 traveling in the negative z direction ($\alpha = 3$: o-mode/TE, and $\alpha = 4$: e-mode/TM). [8] For any $\mathbf{E}$ that satisfies Maxwell's equations, we can then express $\mathbf{E}$ in terms of the plane wave solutions:

$$\mathbf{E}(\mathbf{x}) = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha z} f_\alpha(\mathbf{k}_t)\mathbf{e}_\alpha \right\} e^{j\mathbf{k}_t \cdot \mathbf{p}}, \tag{2.59}$$

where $\mathbf{p} = (x, y)$ and each k_α is a function of $\mathbf{k}_t$.

Now consider a point source $\mathbf{M} = \mathbf{u}_x \delta(\mathbf{x})$ (where $\mathbf{u}_x$ is the unit vector along the x-axis). Let the solution be $\mathbf{G}_x$, and write this in terms of the plane wave components

$$\mathbf{G}_x(\mathbf{x}) = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha z} a_\alpha(\mathbf{k}_t)\mathbf{e}_\alpha \right\} e^{j\mathbf{k}_t \cdot \mathbf{p}}, \tag{2.60}$$

where the a_α's are as yet unknown. Similarly, we have

$$\mathbf{G}_y(\mathbf{x}) = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha z} b_\alpha(\mathbf{k}_t)\mathbf{e}_\alpha \right\} e^{j\mathbf{k}_t \cdot \mathbf{p}} \tag{2.61}$$

and

$$\mathbf{G}_z(\mathbf{x}) = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha z} c_\alpha(\mathbf{k}_t)\mathbf{e}_\alpha \right\} e^{j\mathbf{k}_t \cdot \mathbf{p}} \tag{2.62}$$

as the solutions to point sources $\mathbf{M} = \mathbf{u}_y \delta(\mathbf{x})$ and $\mathbf{M} = \mathbf{u}_z \delta(\mathbf{x})$, respectively ($\mathbf{u}_y$ and $\mathbf{u}_z$ are the unit vectors in the direction of y and z). The a_α's, b_α's, and c_α's

---

[8] Of course if $\mathbf{k}_t$ is too large, the waves will not propagate.

are unknown at this point. In the next subsection, we will show how they can be found. For now, we will assume that they are known. We can then write the dyadic (tensor) Green's function G as

$$G(\mathbf{x},\mathbf{x}') = \{\mathbf{G}_x, \mathbf{G}_y, \mathbf{G}_z\} = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha(z-z')} \mathbf{e}_\alpha \mathbf{q}_\alpha^t \right\} e^{j\mathbf{k}_t\cdot(\mathbf{p}-\mathbf{p}')}, \quad (2.63)$$

where $\mathbf{q}_\alpha = (a_\alpha, b_\alpha, c_\alpha)$, and $\mathbf{q}_\alpha^t$ is the transpose of the column vector $\mathbf{q}_\alpha$ (the expression $\mathbf{e}_\alpha \mathbf{q}_\alpha^t$ is then a tensor). For arbitrary source $\mathbf{M}$, the solution is then

$$\mathbf{E}(\mathbf{x}) = \mathbf{E}^{(0)}(\mathbf{x}) + \int_V G(\mathbf{x},\mathbf{x}')\mathbf{M}(\mathbf{x}')\, d\mathbf{x}', \quad (2.64)$$

where integration is over the region V in which the grating is present (e.g., from $z = -L/2$ to $z = L/2$), and $\mathbf{E}^{(0)}$ is the unperturbed incident beam. The second term is then the diffracted wave from the grating.

It should be noted that in region $z > z'$, only the modes propagating in the positive z direction ($\alpha = 1, 2$) exist, whereas in the region $z < z'$, only the modes propagating in the negative z direction ($\alpha = 3, 4$) exist.

As before, we assume that the incident wave is of the form

$$\mathbf{E}^{(0)}(\mathbf{x}) = e^{j\mathbf{k}_{\alpha_2}\cdot\mathbf{x}} \mathbf{e}_{\alpha_2}, \quad (2.65)$$

and is an eigenmode. (Note that in anisotropic crystals, the polarization $\mathbf{e}_\alpha$ of the electric field is in general *not* perpendicular to the wave vector $\mathbf{k}_\alpha$). We also assume that the perturbation is of the form

$$\Delta\epsilon = \Delta\epsilon' e^{-j\mathbf{K}\cdot\mathbf{x}} + c.c., \quad (2.66)$$

and let

$$\mathbf{k}_{\alpha_1} = -\mathbf{K} + \mathbf{k}_{\alpha_2}. \quad (2.67)$$

For the rest of this section, we will write $\Delta\epsilon$ for $\Delta\epsilon'$.

As before, we will consider only the first term in Eq. (2.66) since it will be obvious later that the contribution from the second term will be far from Bragg-match conditions. We now have

$$\begin{aligned} \mathbf{M} &= -\nabla\times(\epsilon^{-1}\Delta\epsilon\, \mathbf{e}_{\alpha_2} e^{j\mathbf{k}_{\alpha_1}\cdot\mathbf{x}}) \\ &= -j\mathbf{k}_{\alpha_1} \times (\epsilon^{-1}\Delta\epsilon\, \mathbf{e}_{\alpha_2} e^{j\mathbf{k}_{\alpha_1}\cdot\mathbf{x}}). \end{aligned} \quad (2.68)$$

Let

$$\mathbf{b} = -j\mathbf{k}_{\alpha_1} \times (\epsilon^{-1}\Delta\epsilon\, \mathbf{e}_{\alpha_2}), \tag{2.69}$$

and write $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{b}e^{j\mathbf{k}_{t1}\cdot\mathbf{p}+jk_{\alpha_1}z}, \tag{2.70}$$

The diffracted wave is then

$$\mathbf{E}(\mathbf{x}) = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha z}\mathbf{e}_\alpha\, (\mathbf{q}_\alpha \cdot \mathbf{b}) e^{j\mathbf{k}_t\cdot\mathbf{p}} \int_V d\mathbf{x}'\; e^{j(\mathbf{k}_{t1}-\mathbf{k}_t)\cdot\mathbf{p}'+j(k_{\alpha_1}-k_\alpha)z'} \right\}. \tag{2.71}$$

For the case of the infinite slab where V is the region between $z = -L/2$ and $z = L/2$, the integral over $d\mathbf{x}'$ gives us

$$\int_V d\mathbf{x}'\; e^{j(\mathbf{k}_{t1}-\mathbf{k}_t)\cdot\mathbf{p}'+j(k_{\alpha_1}-k_\alpha)z'} = 4\pi^2 L\, \delta(\mathbf{k}_{t1}-\mathbf{k}_t)\, \mathrm{sinc}\left(\frac{L}{2}\Delta k_{\alpha,z}\right), \tag{2.72}$$

where $\Delta k_{\alpha,z} = k_{\alpha_1} - k_\alpha$ is the mismatch in the z direction (normal to the crystal surface). The expression for the diffracted wave in Eq. (2.71) then reduces to

$$\mathbf{E}(\mathbf{x}) = 4\pi^2 \sum_{\alpha=1}^{4} L\, \mathrm{sinc}\left(\frac{L}{2}\Delta k_{\alpha,z}\right) e^{j\mathbf{k}_{t1}\cdot\mathbf{p}+jk_\alpha z}\mathbf{e}_\alpha(\mathbf{q}_\alpha\cdot\mathbf{b}), \tag{2.73}$$

where k_α corresponds to the transverse wave vector $\mathbf{k}_t = \mathbf{k}_{t1}$. Note that $\mathbf{k} = (\mathbf{k}_{t1}, k_\alpha)$ corresponds to a propagating eigenmode and has the same transverse wave vectors as $\mathbf{k}_{\alpha_1} = (\mathbf{k}_{t1}, k_{\alpha_1})$, differing only in the z component (normal to the crystal surface). Thus the diffracted wave is in the form of four propagating eigenmode plane waves. As before, if $\mathbf{x}$ is in the region $z > L/2$ outside the crystal, then only the forward propagating components ($\alpha = 1$ and 2) are present; if $\mathbf{x}$ is in the region $z < -L/2$ outside the crystal, then only the backward propagating components ($\alpha = 3$ and 4) are present.

The angle-selectivity of each eigenmode is the familiar sinc function, and at Bragg-match angles (when $\Delta k_{\alpha,z} = 0$) the amplitude of each eigenmode plane wave is

$$4\pi^2 L\, e^{j\mathbf{k}_{t1}\cdot\mathbf{p}+jk_\alpha z}(\mathbf{q}_\alpha\cdot\mathbf{b}), \tag{2.74}$$

where by Eq (2.69), we have

$$\begin{aligned}\mathbf{q}_\alpha \cdot \mathbf{b} &= \mathbf{q}_\alpha \cdot [-j\mathbf{k}_{\alpha_1} \times (\epsilon^{-1} \cdot \Delta\epsilon\ \mathbf{e}_{\alpha 2})] \\ &= -j(\mathbf{q}_\alpha \times \mathbf{k}_{\alpha_1}) \cdot (\epsilon^{-1} \cdot \Delta\epsilon\ \mathbf{e}_{\alpha 2}). \end{aligned} \tag{2.75}$$

Now since ϵ is a symmetric tensor, so is ϵ^{-1}. [9] We may then write Eq. (2.75) as

$$\mathbf{q}_\alpha \cdot \mathbf{b} = -j(\epsilon^{-1}\mathbf{w}_\alpha) \cdot \Delta\epsilon\ \mathbf{e}_{\alpha 2}, \tag{2.76}$$

where

$$\mathbf{w}_\alpha = \mathbf{q}_\alpha \times \mathbf{k}_{\alpha_1} \tag{2.77}$$

is perpendicular to $\mathbf{k}_{\alpha_1}$. It remains now to find the $\mathbf{q}_\alpha$ vectors.

2.1.4. Solution to the $\mathbf{q}_\alpha$'s

To solve for the $\mathbf{q}_\alpha$ vectors, we will use a Reciprocity Theorem given in reference [7]. Let $\mathbf{E}_1, \mathbf{H}_1$ and $\mathbf{E}_2, \mathbf{H}_2$ be two solutions to the anisotropic Maxwell's equations (Eq.s (2.47) to (2.50)). We use $\mathbf{F}^\dagger(\mathbf{x}, t)$ to denote

$$\mathbf{F}^\dagger(\mathbf{x}, t) \stackrel{\text{def}}{=} \mathbf{F}(-\mathbf{x}, -t). \tag{2.78}$$

Thus the fields that have the superscript "†" have a time dependency of $e^{j\omega t}$ instead of $e^{-j\omega t}$. The reciprocity theorem states that (see Appendix B)

$$\oint_S d\mathbf{A} \cdot (\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2 \times \mathbf{H}_1^\dagger) = \int_V d\mathbf{x}\ (\mathbf{H}_1 \cdot \mathbf{M}_2^\dagger - \mathbf{H}_2^\dagger \cdot \mathbf{M}_1), \tag{2.79}$$

where integration is over the region V and its surface S.

Consider the case where the magnetic current source is

$$\mathbf{M} = \mathbf{P}\ \delta(\mathbf{x} - \mathbf{x}'), \tag{2.80}$$

---

[9] The property of symmetry for a tensor is independent of coordinate systems; i.e., if a tensor is symmetric in *any* particular coordinate system, then it is symmetric in *all* coordinate systems.

where $\mathbf{P} = (P_x, P_y, P_z)$ is an arbitrary constant vector, and $\mathbf{x}' = (x', y', z')$ is an arbitrary fixed point. Using the dyadic Green's function, we have solution

$$\mathbf{E}_1(\mathbf{x}) = G(\mathbf{x}, \mathbf{x}')\,\mathbf{P} = \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha(z-z')} \mathbf{e}_\alpha (\mathbf{q}_\alpha \cdot \mathbf{P}) \right\} e^{j\mathbf{k}_t \cdot (\mathbf{p}-\mathbf{p}')} \tag{2.81}$$

and

$$\begin{aligned} \mathbf{H}_1(\mathbf{x}) =& \frac{1}{j\omega\mu} \nabla \times \mathbf{E}_1(\mathbf{x}) \\ =& \frac{1}{\omega\mu} \int d\mathbf{k}_t \left\{ \sum_{\alpha=1}^{4} e^{jk_\alpha(z-z')} (\mathbf{k}_\alpha \times \mathbf{e}_\alpha)(\mathbf{q}_\alpha \cdot \mathbf{P}) \right\} e^{j\mathbf{k}_t \cdot (\mathbf{p}-\mathbf{p}')}. \end{aligned} \tag{2.82}$$

Note that $\mathbf{k}_\alpha \times \mathbf{e}_\alpha$ is perpendicular to $\mathbf{k}_\alpha$ (direction of propagation), but in general $\mathbf{e}_\alpha$ is *not* perpendicular to $\mathbf{k}_\alpha$. Let

$$\mathbf{E}_2(\mathbf{x}) = e^{j\mathbf{k}_{t0} \cdot \mathbf{p} + jk_\beta z} \mathbf{e}_\beta \quad \Longrightarrow \quad \mathbf{E}_2^\dagger(\mathbf{x}) = e^{-j\mathbf{k}_{t0} \cdot \mathbf{p} - jk_\beta z} \mathbf{e}_\beta, \tag{2.83}$$

where k_β corresponds to the transverse wave vector $\mathbf{k}_{t0}$ and $\mathbf{E}_2$ is the β^{th} propagating eigenmode plane wave with transverse wave vector $\mathbf{k}_{t0}$. We then have

$$\mathbf{H}_2(\mathbf{x}) = \frac{1}{\omega\mu} e^{j\mathbf{k}_{t0} \cdot \mathbf{p} + jk_\beta z} (\mathbf{k}_\beta \times \mathbf{e}_\beta) \quad \Longrightarrow \quad \mathbf{H}_2^\dagger(\mathbf{x}) = \frac{1}{\omega\mu} e^{-j\mathbf{k}_{t0} \cdot \mathbf{p} - jk_\beta z} (\mathbf{k}_\beta \times \mathbf{e}_\beta). \tag{2.84}$$

Since this is a plane wave solution, the corresponding magnetic source $\mathbf{M}_2$ is zero.

Consider the region between $z = z' + \Delta$ and $z = z' - \Delta$. In the limit where $\Delta \to 0$, the left-hand side of Eq. (2.79) gives us

$$\begin{aligned} &\int_{z=z'} (\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2^\dagger \times \mathbf{H}_1) \cdot \mathbf{u}_z \, d\mathbf{p} + \int_{z=z'} (\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2^\dagger \times \mathbf{H}_1) \cdot (-\mathbf{u}_z) \, d\mathbf{p} \\ =& \frac{4\pi^2}{\omega\mu} \sum_{\alpha=1}^{2} e^{-j\mathbf{k}_{t0} \cdot \mathbf{p}' - jk_\beta z'} \, \mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha](\mathbf{q}_\alpha \cdot \mathbf{P}) \\ &- \frac{4\pi^2}{\omega\mu} \sum_{\alpha=3}^{4} e^{-j\mathbf{k}_{t0} \cdot \mathbf{p}' - jk_\beta z'} \, \mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha](\mathbf{q}_\alpha \cdot \mathbf{P}). \end{aligned} \tag{2.85}$$

The right-hand side gives us

$$\begin{aligned} \int_V (-\mathbf{H}_2^\dagger \times \mathbf{M}_1) \, d\mathbf{x} &= -\mathbf{H}_2^\dagger(\mathbf{x}') \cdot \mathbf{P} \\ &= \frac{1}{\omega\mu} e^{-j\mathbf{k}_{t0} \cdot \mathbf{p} - jk_\beta z} (\mathbf{e}_\beta \times \mathbf{k}_\beta) \cdot \mathbf{P}. \end{aligned} \tag{2.86}$$

Substituting these results into Eq. (2.79), and using the fact $\mathbf{P}$ is arbitrary, we conclude that

$$\frac{1}{4\pi^2}(\mathbf{e}_\beta \times \mathbf{k}_\beta) = \sum_{\alpha=1}^{2} \mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha]\mathbf{q}_\alpha - \sum_{\alpha=3}^{4} \mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha]\mathbf{q}_\alpha, \quad (2.87)$$

where all the $\mathbf{e}_\alpha$, $\mathbf{k}_\alpha$, etc., are for the particular transverse wave vector $\mathbf{k}_{t0}$.

Define the 4×4 matrix

$$A = \{a_{\beta\alpha}\}_{\beta,\alpha=1}^{4}, \quad (2.88)$$

where

$$a_{\beta\alpha} = \begin{cases} +\mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha], & \text{if } \alpha = 1, 2; \\ -\mathbf{u}_z \cdot [\mathbf{e}_\beta \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) - (\mathbf{e}_\beta \times \mathbf{k}_\beta) \times \mathbf{e}_\alpha], & \text{if } \alpha = 3, 4. \end{cases} \quad (2.89)$$

Assume that A is invertible, and write its inverse A^{-1} as

$$A^{-1} = \{b_{\alpha\beta}\}_{\beta,\alpha=1}^{4}. \quad (2.90)$$

The solution to Eq. (2.87) is then

$$\mathbf{q}_\alpha = \frac{1}{4\pi^2} \sum_{\beta=1}^{4} b_{\alpha\beta}(\mathbf{e}_\beta \times \mathbf{k}_\beta). \quad (2.91)$$

Note that $\mathbf{e}_\beta \times \mathbf{k}_\beta$ is parallel to the direction of the magnetic field of the β^{th} eigenmode plane wave (for the transverse wave vector $\mathbf{k}_{t0}$).

The diagonal terms of the matrix A are

$$\begin{aligned} a_{\alpha\alpha} &= \pm 2\mathbf{u}_z \cdot [\mathbf{k}_\alpha - (\mathbf{e}_\alpha \cdot \mathbf{k}_\alpha)\mathbf{e}_\alpha] \\ &= \pm 2\mathbf{u}_z \cdot \big(\mathbf{e}_\alpha \times (\mathbf{e}_\alpha \times \mathbf{k}_\alpha)\big)|\mathbf{k}_\alpha| \sin\phi_\alpha, \end{aligned} \quad (2.92)$$

where we take plus for $\alpha = 1, 2$, and minus for $\alpha = 3, 4$. ϕ_α is the angle between $\mathbf{k}_\alpha$ and $\mathbf{e}_\alpha$. Note that $\mathbf{e}_\alpha \times (\mathbf{e}_\alpha \times \mathbf{k}_\alpha)$ is in the direction of the Poynting vector (which is not necessarily the same direction as $\mathbf{k}_\alpha$).

For $\alpha = 1, 3$, the modes are ordinary modes (TE), and $\mathbf{e}_\alpha \times \mathbf{k}_\alpha = 0$ (thus $\phi_\alpha = 90°$ and $\sin\phi_\alpha = 1$). We therefore have

$$a_{\alpha\alpha} = \pm 2\mathbf{u}_z \cdot \mathbf{k}_\alpha = \pm 2|\mathbf{k}_\alpha| \cos\theta_\alpha, \tag{2.93}$$

where θ_α is the angle between the z-axis and the Poynting vector (for ordinary waves, this is in the same direction as $\mathbf{k}_\alpha$). For $\alpha = 2, 4$, however, we have the extraordinary modes (TM), and in general $\mathbf{e}_\alpha \times \mathbf{k}_\alpha \neq 0$. In this case, we get

$$a_{\alpha\alpha} = \pm 2|\mathbf{k}_\alpha| \cos\theta_\alpha \sin\phi_\alpha. \tag{2.94}$$

Thus if the diagonal terms dominate, we have approximately

$$\begin{aligned} \mathbf{q}_\alpha &\approx \frac{1}{4\pi^2 a_{\alpha\alpha}} (\mathbf{e}_\alpha \times \mathbf{k}_\alpha) \\ &= \frac{\mathbf{e}_\alpha \times \mathbf{k}_\alpha}{8\pi^2 |\mathbf{k}_\alpha| \cos\theta_\alpha \sin\phi_\alpha}. \end{aligned} \tag{2.95}$$

Applying this to the results in Eq.s (2.74), (2.76), and (2.77), we get

$$\begin{aligned} \mathbf{w}_\alpha &= \mathbf{q}_\alpha \times \mathbf{k}_{\alpha_1} \\ &= \frac{(\mathbf{e}_\alpha \times \mathbf{k}_\alpha) \times \mathbf{k}_{\alpha_1}}{8\pi^2 |\mathbf{k}_\alpha| \cos\theta_\alpha \sin\phi_\alpha}. \end{aligned} \tag{2.96}$$

At exact Bragg-match angles, we have $\mathbf{k}_{\alpha_1} = \mathbf{k}_\alpha$, and

$$(\mathbf{e}_\alpha \times \mathbf{k}_\alpha) \times \mathbf{k}_\alpha = -|\mathbf{k}_\alpha|^2 \sin\phi_\alpha \, \mathbf{d}_\alpha, \tag{2.97}$$

where $\mathbf{d}_\alpha$ is the unit vector in the direction of the $\mathbf{D}$ field of the α^{th} eigenmode plane wave. [10] Using this result, we then have at Bragg-match angles

$$(\mathbf{q}_\alpha \times \mathbf{k}_{\alpha_1}) \cdot \epsilon^{-1} \Delta\epsilon \, \mathbf{e}_{\alpha_2} = -\frac{|\mathbf{k}_\alpha|}{8\pi^2 \cos\theta_\alpha} (\epsilon^{-1} \mathbf{d}_\alpha) \cdot \Delta\epsilon \, \mathbf{e}_{\alpha_2}. \tag{2.98}$$

---

[10] From Eq. (2.49), $\nabla \times \mathbf{E}_\alpha = j\mathbf{k}_\alpha \times \mathbf{e}_\alpha = j\omega\mu\mathbf{H}$, and from Eq. (2.50), $\nabla \times \mathbf{H}_\alpha = \frac{j}{\omega\mu}\mathbf{k}_\alpha \times (\mathbf{k}_\alpha \times \mathbf{e}_\alpha) = -j\omega\mathbf{D}_\alpha$. Therefore $(\mathbf{e}_\alpha \times \mathbf{k}_\alpha) \times \mathbf{k}_\alpha$ is in the same direction as $\mathbf{D}_\alpha$.

Now let

$$\epsilon_\alpha = \frac{1}{|\epsilon^{-1}\mathbf{d}_\alpha|} = \frac{1}{\sqrt{\mathbf{d}_\alpha^t \epsilon^{-2}\mathbf{d}_\alpha}}. \tag{2.99}$$

(This can be done since ϵ is symmetric.) Then ϵ_α is the permittivity seen "along the direction" of $\mathbf{d}_\alpha$ (at the optical frequency). [11] Since $\mathbf{d}_\alpha$ is in the direction of $\mathbf{D}_\alpha$, $\epsilon^{-1}\mathbf{d}_\alpha$ is in the direction of the electric field $\mathbf{E}_\alpha$. Let $\mathbf{e}_{\alpha_1}$ be the unit vector in the direction of $\mathbf{E}_\alpha$, then we have from Eq. (2.98)

$$(\mathbf{q}_\alpha \times \mathbf{k}_{\alpha_1}) \cdot \epsilon^{-1}\Delta\epsilon\, \mathbf{e}_{\alpha_2} = -\frac{|\mathbf{k}_\alpha|}{8\pi^2 \epsilon_\alpha \cos\theta_\alpha}(\mathbf{e}_{\alpha_1} \cdot \Delta\epsilon\, \mathbf{e}_{\alpha_2}), \tag{2.100}$$

and from Eq.s (2.74) and (2.75), the amplitude (electric field) of the scattered wave is then

$$E_1 = \frac{|\mathbf{k}_\alpha| L}{2\epsilon_\alpha \cos\theta_\alpha}|\mathbf{e}_{\alpha_1} \cdot \Delta\epsilon\, \mathbf{e}_{\alpha_2}|. \tag{2.101}$$

This is similar to the result we obtained for the isotropic case in Section 2.1.2. When the permittivity tensor ϵ becomes a scalar, the results are identical. These results are similar, but slightly different from the predictions of couple-wave analysis, which will be discussed in the next section.

In general, of course, the off-diagonal terms in A are non-zero and there is no simple solution to the inverse matrix A^{-1}. For the special cases where z (the normal of the crystal surface) is either parallel to the crystal axis, or perpendicular to the crystal axis, it turns out that the off-diagonal terms are exactly zero. These are of course the most commonly used cuts for recording reflection and transmission type holograms.

As remarked earlier, $\mathbf{e}_\beta \times \mathbf{k}_\beta$ is parallel to the direction of the magnetic field of the β^{th} eigenmode plane wave. Let $\mathbf{h}_\beta = \mathbf{k}_\beta \times \mathbf{e}_\beta$. It can be shown that for

---

[11] Strictly speaking, this is ambiguous since $\mathbf{E}$ is not parallel to $\mathbf{D}$. What we have here is that since $\mathbf{E} = E_\alpha \mathbf{e}_\alpha = \epsilon^{-1}\mathbf{D} = \epsilon^{-1} D_\alpha \mathbf{d}_\alpha$, therefore $E_\alpha^2 = \mathbf{E} \cdot \mathbf{E} = D_\alpha^2(\mathbf{d}_\alpha \cdot \epsilon^{-2}\mathbf{d}_\alpha)$, and $D_\alpha/E_\alpha = 1/\sqrt{\mathbf{d}_\alpha \cdot \epsilon^{-2}\mathbf{d}_\alpha}$.

these two special cases

$$\left\{\begin{array}{ll} (\mathbf{e}_1 \times \mathbf{h}_2)\cdot\mathbf{u}_z = 0, & (\mathbf{e}_2 \times \mathbf{h}_1)\cdot\mathbf{u}_z = 0, \\ (\mathbf{e}_1 \times \mathbf{h}_4)\cdot\mathbf{u}_z = 0, & (\mathbf{e}_4 \times \mathbf{h}_1)\cdot\mathbf{u}_z = 0, \\ (\mathbf{e}_3 \times \mathbf{h}_2)\cdot\mathbf{u}_z = 0, & (\mathbf{e}_2 \times \mathbf{h}_3)\cdot\mathbf{u}_z = 0, \\ (\mathbf{e}_3 \times \mathbf{h}_4)\cdot\mathbf{u}_z = 0, & (\mathbf{e}_4 \times \mathbf{h}_3)\cdot\mathbf{u}_z = 0, \end{array}\right\} \tag{2.102}$$

and

$$\left\{\begin{array}{l} [(\mathbf{e}_1 \times \mathbf{h}_3) + (\mathbf{e}_3 \times \mathbf{h}_1)]\cdot\mathbf{u}_z = 0, \\ [(\mathbf{e}_2 \times \mathbf{h}_4) + (\mathbf{e}_4 \times \mathbf{h}_2)]\cdot\mathbf{u}_z = 0. \end{array}\right\}. \tag{2.103}$$

Thus only the diagonal terms of the matrix A remain, and the solution is given by Eq. (2.95).

2.2 Coupled-Mode Analysis

Although it is possible to extend the treatment in Section 2.1 to higher order Born's approximation, coupled-mode analysis offers an alternative for analysis when the effect of the media on the incident beam can no longer be neglected.

The geometry of the coupled-mode analysis is shown in Figure 2.1. θ_1 (θ_2) is the angle between $\mathbf{k}_1$ ($\mathbf{k}_2$) and the positive z direction. However, $\mathbf{k}_1$, $\mathbf{k}_2$ and the z-axis do not necessarily lie in the same plane. As before we assume that $\mathbf{k}_2$ is the incident (reference) beam, and $\mathbf{k}_1$ is the diffracted (or signal) beam. The grating vector is $\mathbf{K} = \mathbf{k}_2 - \mathbf{k}_1$, and the time dependency is $e^{-j\omega t}$. The normal to the crystal surface is designated the z-axis, however this is not necessarily the crystal c-axis (i.e., the crystal slab has an arbitrary cut). θ_2 may be larger than 90°, in which case we have reflection type holograms.

The coupled-mode analysis starts out by assuming that

$$\mathbf{E} = \mathbf{e}_1 E_1(z) e^{j\mathbf{k}_1\cdot\mathbf{x}} + \mathbf{e}_2 E_2(z) e^{j\mathbf{k}_2\cdot\mathbf{x}}, \tag{2.104}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are polarization vectors. In the coupled-mode analysis, it is assume that $\mathbf{k}_2$ gives an exact propagating wave, whereas $\mathbf{k}_1$ does not. For

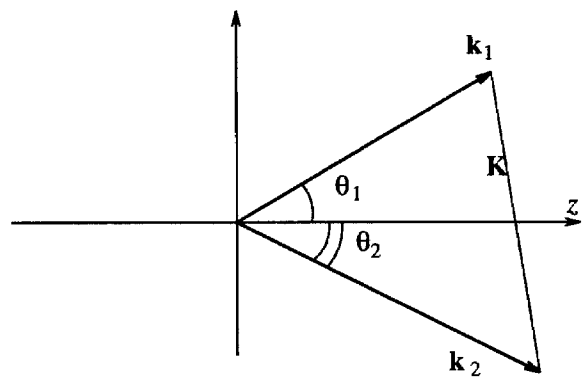


Figure 2.1. Geometry for coupled-mode analysis

anisotropic crystals, the $\mathbf{e}_i$'s are assumed to be eigen-polarizations. But this is somewhat ambiguous for $\mathbf{e}_1$ since $\mathbf{k}_1$ does *not* correspond exactly to a propagating wave. We will assume that $\mathbf{k}_1$ is "close" to a propagating wave, and that $\mathbf{e}_1$ is "close" to the corresponding eigen-polarization.

In isotropic crystals, the polarization vectors are perpendicular to the $\mathbf{k}$ vectors, so that $\mathbf{k}_i \cdot \mathbf{e}_i = 0$ (for $i = 1, 2$). In anisotropic crystals, however, this is no longer true, and instead we have

$$\mathbf{k}_i \cdot \epsilon \mathbf{e}_i = 0 \quad (i = 1, 2), \tag{2.105}$$

where ϵ is the permittivity tensor. In a source-less media, ρ and $\mathbf{J}$ are zero, and we have

$$\nabla \times \nabla \times \mathbf{E} = \omega^2 \mu \epsilon \mathbf{E}. \tag{2.106}$$

Now

$$\begin{aligned}
&\nabla \times \nabla \times \mathbf{E} \\
&= -\mathbf{k}_1 \times (\mathbf{k}_1 \times \mathbf{e}_1) E_1 e^{j\mathbf{k}_1 \cdot \mathbf{x}} + j \left\{ \mathbf{u}_z \times (\mathbf{k}_1 \times \mathbf{e}_1) + \mathbf{k}_1 \times (\mathbf{u}_z \times \mathbf{e}_1) \right\} \frac{dE_1}{dz} e^{j\mathbf{k}_1 \cdot \mathbf{x}} \\
&\quad + \frac{d^2 E_1}{dz^2} \left(\mathbf{u}_z \times (\mathbf{u}_z \times \mathbf{e}_1) \right) e^{j\mathbf{k}_1 \cdot \mathbf{x}} \\
&\quad - \mathbf{k}_2 \times (\mathbf{k}_2 \times \mathbf{e}_2) E_2 e^{j\mathbf{k}_2 \cdot \mathbf{x}} + j \left\{ \mathbf{u}_z \times (\mathbf{k}_2 \times \mathbf{e}_2) + \mathbf{k}_2 \times (\mathbf{u}_z \times \mathbf{e}_2) \right\} \frac{dE_2}{dz} e^{j\mathbf{k}_2 \cdot \mathbf{x}} \\
&\quad + \frac{d^2 E_2}{dz^2} \left(\mathbf{u}_z \times (\mathbf{u}_z \times \mathbf{e}_2) \right) e^{j\mathbf{k}_2 \cdot \mathbf{x}},
\end{aligned} \tag{2.107}$$

where $\mathbf{u}_z$ is the unit vector in the positive z direction. The second derivative terms ($\frac{d^2}{dz^2}$) in the above equation are dropped under the assumption that the E_i's are slow varying functions of z.

Since $\mathbf{e}_2$ corresponds to an eigen-polarization, it is a solution to the wave equation Eq. (2.106) when there is no perturbation. Thus we have

$$-\mathbf{k}_2 \times (\mathbf{k}_2 \times \mathbf{e}_2) = \omega^2 \mu \epsilon_0 \mathbf{e}_2, \tag{2.108}$$

where ϵ_0 is the unperturbed constant permittivity tensor. We assume that the perturbation is of the form [12]

$$\epsilon = \epsilon_0 + \Delta\epsilon\, e^{j\mathbf{K}\cdot\mathbf{x}} + c.c., \tag{2.109}$$

where $c.c.$ stands for complex conjugate, and $\Delta\epsilon$ is also a tensor. We then have

$$\begin{aligned}\omega^2 \mu \epsilon \mathbf{E} =& \omega^2 \mu \epsilon_0 \mathbf{E} + \omega^2 \mu \Delta\epsilon\, \mathbf{e}_1 E_1 e^{j\mathbf{k}_2\cdot\mathbf{x}} \\ &+ \omega^2 \mu \Delta\epsilon\, {}^*\mathbf{e}_2 E_2 e^{j\mathbf{k}_1\cdot\mathbf{x}} + (higher\ \ order\ \ terms).\end{aligned} \tag{2.110}$$

Substituting Eq.s (2.107), (2.108) and (2.210) into Eq. (2.106), matching the terms of $e^{j\mathbf{k}_i\cdot\mathbf{x}}$, and then taking the dot product with $\mathbf{e}_i$, we arrive at the coupled-wave equations

$$c_1 \frac{dE_1}{dz} = j\xi E_1 + j\Gamma_2 E_2 \tag{2.111}$$

$$c_2 \frac{dE_2}{dz} = j\Gamma_1 E_1, \tag{2.112}$$

where

$$\begin{aligned} c_i &= (\mathbf{k}_i \times \mathbf{e}_i)\cdot(\mathbf{u}_z \times \mathbf{e}_i) \\ &= 2k_i \cos\theta_i - 2(\mathbf{k}_i \cdot \mathbf{e}_i)(\mathbf{u}_z \cdot \mathbf{e}_i) \qquad (i = 1, 2), \end{aligned} \tag{2.113}$$

$$\Gamma_1 = \omega^2 \mu (\mathbf{e}_2 \cdot \Delta\epsilon\, \mathbf{e}_1), \tag{2.114}$$

$$\Gamma_2 = \omega^2 \mu (\mathbf{e}_1 \cdot \Delta\epsilon\, {}^*\mathbf{e}_2), \tag{2.115}$$

and

---

[12] See comment in footnote 6.

$$\xi = \omega^2\mu(\mathbf{e}_1 \cdot \epsilon_0 \mathbf{e}_1) + \mathbf{e}_1 \cdot [\mathbf{k}_1 \times (\mathbf{k}_1 \times \mathbf{e}_1)]$$
$$= \omega^2\mu(\mathbf{e}_1 \cdot \epsilon_0 \mathbf{e}_1) - |\mathbf{k}_1 \times \mathbf{e}_1|^2. \tag{2.116}$$

Note that in isotropic crystals, $\mathbf{k}_i \cdot \mathbf{e}_i = 0$ in the c_i factors. For anisotropic crystals, this is true only for the ordinary mode. In practice, however, the imaginary part of c_i is usually small and can be neglected. For an uniaxial crystal, it can be shown that for extraordinary modes, the cosine of the angle between $\mathbf{k}_i$ and $\mathbf{e}_i$ is

$$\cos(\mathbf{s_i}, \mathbf{e_i}) = \frac{\mathbf{k}_i \cdot \mathbf{e}_i}{k_i} = \frac{\sin\phi_i \cos\phi_i \left(\frac{1}{n_o^2} - \frac{1}{n_e^2}\right)}{\sqrt{\frac{\cos_i^2\phi}{n_o^4} + \frac{\sin_i^2\phi}{n_e^4}}} \approx \frac{\Delta n}{n} \sin 2\phi_i, \tag{2.117}$$

where ϕ_i is the angle between the c-axis and k_i, and $\mathbf{s}_i = \mathbf{k}_i/k_i$. (For ordinary mode, $\mathbf{k}_i \cdot \mathbf{e}_i$ is zero). Note that $\cos(\mathbf{s_i}, \mathbf{e_i}) = \mathbf{k}_i \cdot \mathbf{e}_i/k_i$ is zero when $\mathbf{k}_i$ is perpendicular or parallel to the c-axis, and reaches maximum around $\phi_i = 45°$. For $LiNbO_3$, $\Delta n/n$ is about 4%, while for $BaTiO_3$, it is about 3%.

At Bragg-match angles, $\mathbf{k}_1$ is a propagating wave, and using the result in Eq. (2.108) with the subscripts changed to 1, we get $\xi = 0$. The solution for $\xi = 0$ is well known. For transmission type holograms, the boundary conditions are $E_1(0) = 0$ and $E_2(0) = E_{20}$, and we have at $z = L$

$$\frac{E_1(L)}{E_{20}} = j\sqrt{\frac{c_2\Gamma_2}{c_1\Gamma_1}} \sin\left(\sqrt{\frac{\Gamma_1\Gamma_2}{c_1c_2}}\, L\right) \approx j\,\frac{\Gamma_2}{c_1}\, L, \tag{2.118}$$

for weak holograms. For reflection type holograms, the boundary conditions are $E_1(0) = 0$ and $E_2(L) = E_{20}$, and we have at $z = L$

$$\frac{E_1(L)}{E_{20}} = j\sqrt{-\frac{c_2\Gamma_2}{c_1\Gamma_1}} \tanh\left(\sqrt{-\frac{\Gamma_1\Gamma_2}{c_1c_2}}\, L\right) \approx j\,\frac{\Gamma_2}{c_1}\, L, \tag{2.119}$$

for weak holograms (note that for reflection type holograms, $\theta_2 > 90°$ and $c_2 < 0$). If we neglect the second term in c_1 and assume that $\Delta\epsilon$ is real, then we have (from either Eq. (2.118) or (2.119))

$$\frac{E_1(L)}{E_{20}} \approx j\,\frac{k_1(\mathbf{e}_1 \cdot \Delta\epsilon\, \mathbf{e}_2)}{2\epsilon_1 \cos\theta_1}\, L, \tag{2.120}$$

where $\omega^2\mu/k_1 = k_1/\epsilon_1$, and ϵ_1 is the permittivity associated with the wave vector k_1.

It should be noted that if we define the diffraction efficiency as the ratio of intensities (in power per unit area), then Eq. (2.118) gives us (assuming that the second term in c_1 is zero, etc.)

$$\eta = \left|\frac{\cos\theta_2}{\cos\theta_1}\right| \sin^2\left(\frac{k_1(\mathbf{e}_1 \cdot \Delta\epsilon\, \mathbf{e}_2)L}{2\epsilon_1\sqrt{\cos\theta_1 \cos\theta_2}}\right), \tag{2.121}$$

while Eq. (2.119) gives us

$$\eta = \left|\frac{\cos\theta_2}{\cos\theta_1}\right| \tanh^2\left(\frac{k_1(\mathbf{e}_1 \cdot \Delta\epsilon\, \mathbf{e}_2)L}{2\epsilon_1\sqrt{\cos\theta_1 \cos\theta_2}}\right), \tag{2.122}$$

If $\theta_1 = \theta_2$ or $\theta_1 = \pi - \theta_2$, then the factor $|\cos\theta_2/\cos\theta_1|$ is one. But in general it is not.

For the non-Bragg-matched situation $\xi \neq 0$. If we assume that $E_2 \approx E_{20}$ is approximately a constant, then Eq. (2.111) gives us

$$\frac{E_1(z)}{E_{20}} \approx -j\,\frac{\Gamma_2 z}{c_1}\,\mathrm{sinc}\left(\frac{\xi z}{2c_1}\right) e^{j\xi z/2c_1}. \tag{2.123}$$

The angle selectivity is therefore determined by the factor

$$\frac{\xi z}{2c_1} = \frac{\omega^2\mu(\mathbf{e}_1 \cdot \epsilon_0\mathbf{e}_1) - |\mathbf{e}_1 \times \mathbf{k}_1|^2}{4k_1\cos\theta_1 - 4(\mathbf{k}_1 \cdot \mathbf{e}_1)(\mathbf{u}_z \cdot \mathbf{e}_1)}. \tag{2.124}$$

For isotropic materials, this becomes

$$\frac{\xi z}{2c_1} = \frac{k_0^2 - k_1^2}{4k_1\cos\theta_1}\, z \approx \frac{\Delta k}{2\cos\theta_1} z \approx \frac{1}{2}\Delta k_z\, z, \tag{2.125}$$

where $k_0^2 = \omega^2\mu\epsilon_0$ is the wavenumber for the propagating wave, $\Delta k = k_0 - k_1$, and Δk_z is the mismatch of $\mathbf{k}_1$ in the z direction (See Figure 2.2). The width of the Bragg-selectivity is determined by setting $\xi z/2c_1 = \pi$.

If we compare the result in Eq. (2.120) with the corresponding result obtained using Born's Approximation, they are seemingly the same. There is, however, a subtle difference. In the Born's approximation (Section 2.1.3), the ϵ_1 in

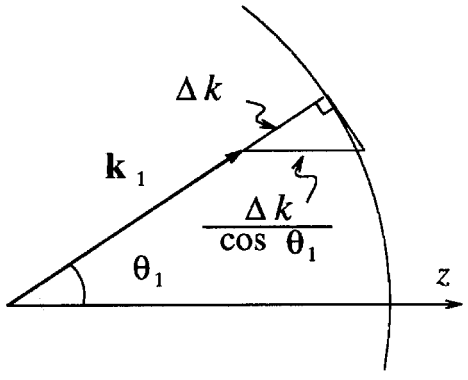


Figure 2.2. Bragg-mismatch of diffracted wave-vector

Eq. (2.120) was given by [13]

$$\epsilon_1 = \frac{1}{\sqrt{\mathbf{d}_1 \cdot \epsilon_0^{-2} \mathbf{d}_1}}, \tag{2.126}$$

where ϵ_0 is the permittivity tensor, and $\mathbf{d}_1$ is the unit vector in the direction of $\mathbf{D}$. As explained in the previous section, ϵ_1 is the ratio of the magnitude of $\mathbf{D}_1/\mathbf{E}_1$ (note that in general $\mathbf{D}_1$ and $\mathbf{E}_1$ are not parallel). In the coupled-mode analysis, however, ϵ_1 is given by

$$\epsilon_1 = \frac{k_1^2}{\omega^2 \mu}, \tag{2.127}$$

and is the effective permittivity for obtaining the k-vector of the eigenmode. To see the difference between these two, let ϕ_1 be the angle between the c-axis and $\mathbf{k}_1$. Then for the extraordinary wave, Eq. (2.126) (for Born's Approximation) gives us

$$\frac{1}{\epsilon_1^2} = \frac{\cos^2 \phi_1}{\epsilon_x^2} + \frac{\sin^2 \phi_1}{\epsilon_z^2}, \tag{2.128}$$

where the permittivity tensor is $\epsilon_0 = \mathrm{diag}\{\epsilon_x, \epsilon_x, \epsilon_z\}$. On the other hand, it can be shown that for the coupled-mode result [14]

$$\frac{1}{\epsilon_1} = \frac{\cos^2 \phi_1}{\epsilon_x} + \frac{\sin^2 \phi_1}{\epsilon_z}. \tag{2.129}$$

---

[13] See Eq. (2.99). The notations have been modified to be consistent with those used in this section.

[14] See for example Yariv & Yeh's *Optical Waves in Crystals*, Eq. (4.6-4) on page 87 [9].

For the ordinary waves, there is of course no difference, and for isotropic materials, $\epsilon_x = \epsilon_z$, and the two results are identical. For general uniaxial crystals, however, the definitions are, although the difference is practically zero. To see why this is true, we write Eq.s (2.128) and (2.129) as

$$\frac{1}{n_1^4} = \frac{\cos^2\phi_1}{n_o^4} + \frac{\sin^2\phi_1}{n_e^4} \tag{2.130}$$

and

$$\frac{1}{n_1^2} = \frac{\cos^2\phi_1}{n_o^2} + \frac{\sin^2\phi_1}{n_e^2}, \tag{2.131}$$

where n_o and n_e are the ordinary and extraordinary index of refraction (note that $\epsilon = \epsilon_0 n^2$). If we assume that $\Delta n = n_e - n_o$ is small compared to n_o and n_e, then to first order, both Eq.s (2.130) and (2.131) give us

$$n_1^2 = n_o^2(1 + 2n_o\Delta n \ \sin\phi_1). \tag{2.132}$$

As mentioned earlier, for $LiNbO_3$, $\Delta n/n_o$ is approximately 4%, and for $BaTiO_3$, it is about 3%. The two formulas give essentially the same results. [15]

Finally, there is one last approximation that was made in asserting the diffraction efficiency in Eq.s (2.121) and (2.122). For isotropic materials, the index of refraction does not depend on direction, and therefore the magnitude of the Poynting vector is just proportional to the square of the electric field. However, since the magnitude of the Poynting vector is

$$S = |\mathbf{E} \times \mathbf{H}| = \sqrt{\frac{\epsilon_0}{\mu}}\, nE^2, \tag{2.133}$$

where μ is the permeability of the material (assumed to be a scalar), and since for anisotropic materials the index depends on direction, the expression for η in Eq.s (2.121) and (2.122) should actually be multiplied by a factor of n_1/n_2, where n_i is the effective index of refraction for a plane wave propagating in the

---

[15] Nevertheless, the difference is fundamental, and would show up if $\Delta n/n$ were large.

$\mathbf{k}_i$ direction. In practice, the factor n_1/n_2 is different from 1 by at most $\Delta n/n$, which is only 3% to 4% for $LiNbO_3$ and $BaTiO_3$.

In summary, the results derived from coupled-mode analysis for anisotropic crystals are essentially the same as the results for isotropic materials (assuming that $\mathbf{k}_i$ and $\mathbf{e}_i$ are approximately perpendicular to each other).

In the limit of weak holograms, the results from coupled-mode analysis are similar, but slightly different from the results derived from Born's Approximation. In practice, however, the difference for anisotropic crystals such as $LiNbO_3$ and $BaTiO_3$ is negligible, and for isotropic materials, the difference disappears.

Appendix

A. Plane Wave Representation of Spherical Waves

The plane wave representation of a spherical wave, as given in Eq. (2.14), was obtained in reference [4] indirectly through arguments on radiation fields, etc. In this Appendix, we will derive it in a more straight forward fashion.

Consider the expression

$$\phi = -\frac{1}{4\pi r}e^{jkr} = \sum_{\alpha=1}^{2}\int\int dk_x dk_y\ A_\alpha(k_x,k_y)e^{j(k_x x+k_y y+k_{\alpha,z}z)}, \tag{2.134}$$

where $r=\sqrt{x^2+y^2+z^2}$, and

$$k_{\alpha,z} = \begin{cases} \sqrt{k^2-k_x^2-k_y^2}, & \text{if } \alpha=1; \\ -\sqrt{k^2-k_x^2-k_y^2} & \text{if } \alpha=2. \end{cases} \tag{2.135}$$

The component with $\alpha=1$ gives us the forward propagating wave (which exists only for $z>0$), and $\alpha=2$ gives us the backward propagating wave (which exists only for $z<0$). It is easy to show that ϕ satisfies

$$\nabla^2\phi + k^2\phi = \delta(\mathbf{x}), \tag{2.136}$$

Let

$$\psi = e^{-j\mathbf{k}_0\cdot\mathbf{x}} = e^{j(k_{0x}x+k_{0y}y+k_{0z}z)}, \tag{2.137}$$

where $k_{0x}^2 + k_{0y}^2 + k_{0z}^2 = k^2$. We will assume that $k_{0z} = \sqrt{k^2 - k_{0x}^2 - k_{0y}^2} > 0$. From Green's theorem,

$$\int_V (\psi\nabla^2\phi - \phi\nabla^2\psi)d\mathbf{x} = \oint_S (\psi\nabla\phi - \phi\nabla\psi)\cdot\mathbf{n}\, dA, \tag{2.138}$$

where S is the boundary surface of V and $\mathbf{n}$ is the unit vector normal and pointing outwards from the surface S. From Eq.s (2.135)–(2.137) we have

$$\psi\nabla^2\phi - \phi\nabla^2\psi = \psi\delta(\mathbf{x}), \tag{2.139}$$

so that the left-hand side of Eq. (2.138) is 1 if V includes the origin. On the other hand, we have

$$\psi\nabla\phi - \phi\nabla\psi = \sum_{\alpha=1}^{2} j\int\int dk_x dk_y\, A_\alpha(k_x,k_y)(\mathbf{k}_\alpha + \mathbf{k}_0)e^{j(\mathbf{k}_\alpha-\mathbf{k}_0)\cdot\mathbf{x}}. \tag{2.140}$$

We now take V to be the region between $z = -\Delta$ and $z = \Delta$. In the limit where $\Delta \to 0$, the right-hand side of Eq. (2.138) becomes

$$\begin{aligned}
&\oint_S (\psi\nabla\phi - \phi\nabla\psi)\cdot\mathbf{n}\, dA \\
=&\int\int dk_x dk_y\, A_1(k_x,k_y)\left\{\int\int_{z=0+} dxdy\, j(k_{1,z}+k_{0z})e^{j(\mathbf{k}_1-\mathbf{k}_0)\cdot\mathbf{x}}\right\} \\
&-\int\int dk_x dk_y\, A_2(k_x,k_y)\left\{\int\int_{z=0-} dxdy\, j(k_{2,z}+k_{0z})e^{j(\mathbf{k}_2-\mathbf{k}_0)\cdot\mathbf{x}}\right\} \\
=&j8\pi^2 k_{0z}A_1(k_{0x},k_{0y}).
\end{aligned} \tag{2.141}$$

Note that in the region $z > 0$ only the $\alpha = 1$ component exists, and in the region $z < 0$ only the $\alpha = 2$ component exists. We are assuming here that $k_{0z} = \sqrt{k^2 - k_{0x}^2 - k_{0y}^2} > 0$. If we take $k_{0z} = -\sqrt{k^2 - k_{0x}^2 - k_{0y}^2} < 0$ instead, then the above expression reduces to $j8\pi^2 k_{0z}A_2(k_{0x},k_{0y})$. Since the left-hand side of Eq. (2.138) is 1, we have

$$A_\alpha(k_x,k_y) = \frac{1}{j8\pi^2 k_{0z}}, \tag{2.142}$$

for both $\alpha = 1$ and 2, and from Eq. (2.134), we get

$$\frac{e^{jkr}}{r} = \frac{j}{2\pi}\int\int \frac{1}{k_z} e^{j(k_x x + k_y y \pm k_z z)}\, dk_x dk_y, \tag{2.143}$$

where $r = \sqrt{x^2+y^2+z^2}$ and $k_z = \sqrt{k^2 - k_x^2 - k_y^2} > 0$. We take the plus sign if $z > 0$ and the minus sign if $z < 0$.

From this result, we have the 2-D Fourier transform pair:

$$\frac{1}{\sqrt{x^2+y^2+z^2}} e^{jk\sqrt{x^2+y^2+z^2}} \quad \Longleftrightarrow \quad \frac{j}{2\pi}\frac{1}{\sqrt{k^2-k_x^2-k_y^2}} e^{j\sqrt{k^2-k_x^2-k_y^2}|z|}, \tag{2.144}$$

B. Reciprocity Theorem

In this Appendix, we derive the reciprocity theorem used in subsection 2.1.4. in a slightly more general form than Eq. (2.79): [7]

$$\begin{aligned} &\oint_S d\mathbf{A} \cdot \{\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2 \times \mathbf{H}_1^\dagger\} \\ &= \int_V d\mathbf{x}\, (\mathbf{H}_1 \cdot \mathbf{M}_2^\dagger - \mathbf{H}_2^\dagger \cdot \mathbf{M}_1 + \mathbf{E}_2^\dagger \cdot \mathbf{J}_1 - \mathbf{E}_1 \cdot \mathbf{J}_2^\dagger). \end{aligned} \tag{2.145}$$

To derive this result, we start with Maxwell's equations. We have [16]

$$\nabla \times \mathbf{E}_1 = j\omega \mathbf{B}_1 - \mathbf{M}_1, \tag{2.146}$$

$$\nabla \times \mathbf{H}_1 = -j\omega \mathbf{D}_1 + \mathbf{J}_1, \tag{2.147}$$

$$-\nabla \times \mathbf{E}_2^\dagger = -j\omega \mathbf{B}_2^\dagger - \mathbf{M}_2^\dagger, \tag{2.148}$$

$$-\nabla \times \mathbf{H}_2^\dagger = j\omega \mathbf{D}_2^\dagger + \mathbf{J}_2^\dagger. \tag{2.149}$$

Then

$$\begin{aligned} \nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2^\dagger) &= \mathbf{H}_2^\dagger \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2^\dagger \\ &= -\mathbf{H}_2^\dagger \cdot (-j\omega \mathbf{B}_1) - \mathbf{E}_1 \cdot (j\omega \mathbf{D}_2^\dagger) - \mathbf{H}_2^\dagger \cdot \mathbf{M}_1 + \mathbf{E}_1 \cdot \mathbf{J}_2^\dagger \end{aligned} \tag{2.150}$$

---

[16] Recall that the superscript "†" is used to denote a reversal in position as well as time; i.e., $\mathbf{x} \to -\mathbf{x}$ and $t \to -t$. In this case, $\nabla\times \to -\nabla\times$ and $d/dt \to -d/dt$.

and

$$\nabla\cdot(\mathbf{E}_2^\dagger \times \mathbf{H}_1) = \mathbf{H}_1 \cdot \nabla\times\mathbf{E}_2^\dagger - \mathbf{E}_2^\dagger \cdot \nabla\times\mathbf{H}_1$$
$$= -\mathbf{H}_1 \cdot (j\omega\mathbf{B}_2) - \mathbf{E}_2^\dagger \cdot (-j\omega\mathbf{D}_1) + \mathbf{H}_1 \cdot \mathbf{M}_2^\dagger - \mathbf{E}_2^\dagger \cdot \mathbf{J}_1. \quad (2.151)$$

Since ϵ is symmetric, we have

$$\mathbf{E}_1 \cdot \mathbf{D}_2^\dagger = \mathbf{E}_1 \cdot (\epsilon\mathbf{E}_2^\dagger) = (\epsilon\mathbf{E}_1) \cdot \mathbf{E}_2^\dagger = \mathbf{D}_1 \cdot \mathbf{E}_2^\dagger. \quad (2.152)$$

We also assume that μ is symmetric (or a scalar), so that we have

$$\mathbf{H}_1 \cdot \mathbf{B}_2^\dagger = \mathbf{B}_1 \cdot \mathbf{H}_2^\dagger. \quad (2.153)$$

Using these results, we get

$$\nabla\cdot(\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2^\dagger \times \mathbf{H}_1) = -\mathbf{H}_2^\dagger \cdot \mathbf{M}_1 + \mathbf{E}_1 \cdot \mathbf{J}_2^\dagger + \mathbf{H}_1 \cdot \mathbf{M}_2^\dagger - \mathbf{E}_2^\dagger \cdot \mathbf{J}_1. \quad (2.154)$$

Now apply the divergence theorem to both sides of Eq. (2.154). We then have

$$\oint_S d\mathbf{A} \cdot (\mathbf{E}_1 \times \mathbf{H}_2^\dagger + \mathbf{E}_2^\dagger \times \mathbf{H}_1)$$
$$= \int_V d\mathbf{x}\, (-\mathbf{H}_2^\dagger \cdot \mathbf{M}_1 + \mathbf{E}_1 \cdot \mathbf{J}_2^\dagger + \mathbf{H}_1 \cdot \mathbf{M}_2^\dagger - \mathbf{E}_2^\dagger \cdot \mathbf{J}_1). \quad (2.155)$$

In the case where $\mathbf{J}_1$ and $\mathbf{J}_2$ are zero, Eq. (2.155) reduces to Eq. (2.79).

References

1. J. D. Jackson, *Classical Electrodynamics* (John Wiley & Sons, New York, 1962).
2. M. D. Greenberg, *Application of Green's Functions in Science and Engineering* (Prentice-Hall, Englewood Cliffs, 1971).
3. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell Syst. Tech. J.*, **48(9)**, 2909–2947 (1969).
4. W. S. Weiglhofer, "Green's Functions and Magnetized Ferrites," *Int. J. Electronics*, **73**(4), 763–771, 1992.

5. P. C. Clemmow, *The Plane Wave Spectrum Representation of Electromagnetics Fields*, *International Series of Monographs in Electromagnetic Waves*, vol. 12 (Pergamon Press, New York, 1966).

6. S. Barkeshli, "Electromagnetic Dyadic Green's Function for Multilayered Symmetric Gyroelectric Media," *Radio Science*, **28**(1), 23–26, Jan–Feb, 1993.

7. L. B. Felson and N. Marcuvitz, *Radiation and Scattering of Waves*, Chap. 8, Prentice-Hall, Englewood Cliffs, NJ, 1973.

8. H. Motz and H. Kogelnik, "Electromagnetic Radiation from Sources Embedded in an Infinite Anisotropic Medium and the Significance of the Poynting Vector," from *Electromagnetic Theory and Antennas*, Part I, E. C. Jordan, ed., *International Series of Monographs in Electromagnetic Waves*, vol. 6 (Pergamon Press, New York, 1963).

9. A. Yariv and P. Yeh, *Optical Waves in Crystals* (John Wiley & Sons, New York, 1984).

Chapter 3

Alignment Sensitivity of 3-D Holographic Disks

The holographic recording system considered in this thesis is the 3-D holographic disk. In this Chapter, we begin by describing the 3-D holographic disk system that combines spatial and angle or wavelength [1,2] multiplexing for recording data. The rest of the thesis will be mainly devoted to analyzing various issues associated with the 3-D disk system. After describing the 3-D disk system, the remainder of this chapter concentrates on the question of alignment sensitivities of 3-D disks.

3.1. Introduction

The theoretical upper limit on the storage density of volume holography is V/λ^3, where V is the volume of the hologram and λ is the operating wavelength of light. This limit is in the order of 10^{12} bits per cm^3, however in practical systems only 10^9–10^{10} bits per cm^3 is achievable due to the finite numerical aperture of the optical system that transfers the data into the optical system and the dynamic range of the crystal. For example, typically 10^3 holograms can are superimposed at the same location, each hologram consisting of $10^3 \times 10^3$ pixels, giving a total memory of 10^9 bits per location. To be competitive with magnetic and semiconductor memories, which are becoming cheaper and better all the time, it is necessary to further increase the capacity of holographic storage systems. This is done by recording on multiple locations; i.e., by spatial multiplexing. One

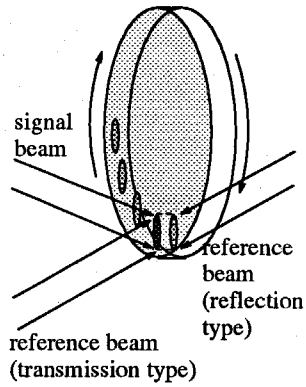


Figure 3.1. The 3-D holographic disk (HD).

of the simplest systems for performing spatial multiplexing is the 3-D holographic disk [3,4].

As in all spatial multiplexing schemes, the most crucial component of the 3-D holographic disk is the scanning mechanism that steers the readout mechanism to different locations of the disk. In the 3-D HD disk, spatial multiplexing is done in a disk configuration with rotation used to access different recording locations, as shown in Figure 3.1. Two light beams (a signal and a reference) interfere inside the photorefractive crystal to create a phase grating via the photorefractive effect. Multiple holograms are recorded at the same location by changing the reference beam angle (angle multiplexing) or by changing the wavelengths of the reference and signal beams (wavelength multiplexing). Because of the Bragg-matching requirement of volume holograms, individual holograms can be read out by changing the direction of the reference beam (for angle multiplexing), or the wavelength of the reference beam (for wavelength multiplexing).

In the holographic data storage system, high readout speed and capacity are achieved by reading out whole pages at a time. To convert this information to electronic signals for further processing by computers, etc., the hologram is imaged onto devices such as RAM chips with detectors or CCD cameras. For the data to be transferred correctly, it is necessary to position the reconstructed holographic image accurately. For data retrieval, even shifting the data page (the

reconstructed image) by one bit could be disastrous. Since we wish to store data at very high density (and therefore requiring small pixels), alignment errors can be a serious problem. It is therefore important to reduce alignment sensitivity for both spatial and angle/wavelength multiplexing.

Of course the conditions upon readout can not be precisely the same as during recording. Most of this alignment error comes from multiplexing. For example, because we use angle multiplexing, there is likely to be some error in the reference beam angle during readout, and this will cause the position of the reconstructed image to change. In addition, there are other sources of error such as disk wobbling from spinning the disk at high speed. For any spatial/angle/wavelength multiplexing holographic recording system, we can summarize all the possible sources of alignment errors into the following categories:

1. change in reference beam angle
2. change in reference beam wavelength
3. hologram rotation
4. translation (or shift) of hologram
5. tilt of hologram

The main sources of error, however, come from the changes used in the multiplexing scheme. Thus for a 3-D disk system that uses angle multiplexing, the main concerns are disk rotation and reference beam angle (and translation if we also scan in the radial direction). For a 3-D disk system that uses wavelength multiplexing, the concern is error in reference beam wavelength instead of reference beam angle. In this chapter, we examine the effect of these five sources of error on alignment sensitivity. For the main part of the paper we will assume that the reference beam is a plane wave.

The general holographic recording system that we consider is shown in Figure 3.2. We have some linear passive optical system $L1$ that maps each point source in the input plane into some wave-form (the impulse response of $L1$). At the holographic recording plane, this wave (the signal beam) interferes with a ref-

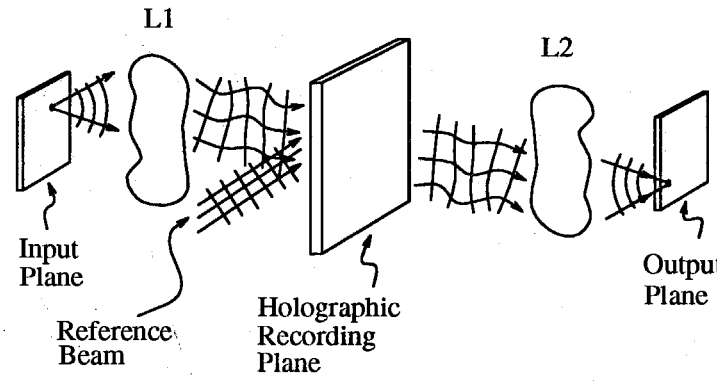


Figure 3.2. The general holographic recording system

erence beam to form an interference pattern that is recorded as the holograph. [1] Upon applying the same reference beam, the original wave-form is reconstructed and passes through another passive optical system $L2$ to be imaged to a point at the output plane. The mapping of the two optical systems is such that the image at the input plane will be reproduced at the output plane, possibly with scaling and/or inversion.

In this chapter, we analyze the effects of small changes in the holograph on image reconstruction. Strictly speaking, the treatment is for thin (or planar) holograms, but the results also hold for thick (volume) holograms within the Bragg-match regime. The effect of the Bragg-selectivity property of volume holograms will be examined briefly in each section, and then in more detail at the end of this chapter.

We will consider only two types of passive optical systems: (1) the Fourier-transform lens, where the impulse response is a plane wave (e.g., Fourier plane hologram recording systems), and (2) free space, where the impulse response is a spherical wave (e.g., image plane hologram recording systems). Almost all

---

[1] In this chapter, the term "holograph" will be used to denote the recorded interference pattern on the material, and the term "hologram" will be reserved to mean the reconstructed, or diffracted wave form emerging from the recording material when the reconstructing reference beam is applied.

practical imaging systems are equivalent to one of the two cases. The holograms are assumed to be either transmission type or reflection type holograms. We do not consider the 90-degree recording geometry here.

3.2. Angle Alignment Sensitivity

In this section we will assume that the reference beam is a plane wave, and consider the effect of error in the angle of the readout reference beam. The impulse response of the imaging system $L1$ is the signal beam, which can be either a plane wave or a spherical wave. We will assume that for case 1, the wave vectors of the reference beam, signal beam, and the normal to the (volume) holographic recording medium lie in the same plane. For case 2, we assume that the wave vector of the reference beam, the normal to the holographic recording medium, and the point source of the spherical wave signal beam lie in the same plane. We will refer to this condition as the co-planar geometry.

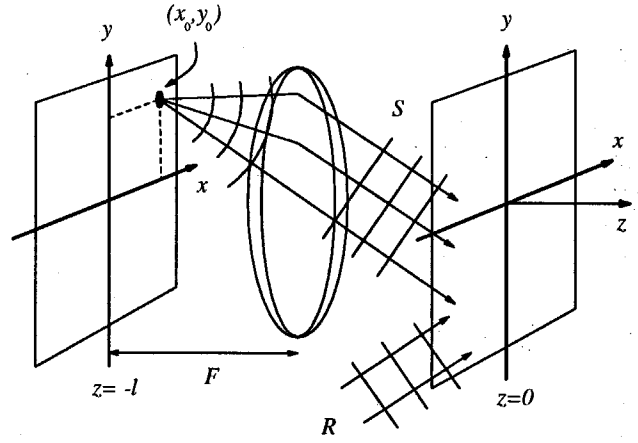


Figure 3.3. Recording system with plane wave reference beam and plane wave signal beam.

3.2.1. Plane Wave Signal Beam

We first examine the situation for case 1, shown in Figure 3.3. The input plane is at the plane at $z = -l$, and the image is in a region centered at a point

on the x axis (i.e., $y = 0$). We consider a particular point (x_0, y_0) of the input image. The signal beam coming from this point source is converted into a plane wave by the Fourier transform lens. It forms an interference pattern with the reference beam plane wave at the recording plane (at $z = 0$), which is at some distance away from the Fourier transform lens, but not necessarily at the Fourier plane. We will initially assume that the hologram is planar.

We assume that both reference and signal beams are propagating in the positive z direction, [2] and that the wave vector of the reference beam plane wave R lies in the x–z plane. The wave vector of the signal beam plane wave S lies close to the x–z plane (since the image lies near the x axis).

Let the signal S and reference R beams be

$$R = A\, e^{jk(u_x x + u_y y + u_z z)}, \tag{3.1}$$

$$S = B\, e^{jk(v_x x + v_y y + v_z z)}. \tag{3.2}$$

The wave vectors are then $k\mathbf{u}$ and $k\mathbf{v}$. Let the angle between $\mathbf{u}$ (reference beam) and the z axis be θ_R, and the angle between $\mathbf{v}$ (signal beam) and the z axis be θ_S. Under paraxial approximations, we have

$$v_x \approx \frac{x_0}{F} \approx \sin\theta_S, \tag{3.3}$$

$$v_y \approx \frac{y_0}{F} \approx 0, \tag{3.4}$$

$$v_z = \cos\theta_S, \tag{3.5}$$

$$u_x = -\sin\theta_R, \tag{3.6}$$

$$u_y = 0, \tag{3.7}$$

$$u_z = \cos\theta_R, \tag{3.8}$$

where F is the focal length of the Fourier transform lens. The grating vector $\mathbf{K}$ is

$$\mathbf{K} = k(\mathbf{u} - \mathbf{v}), \tag{3.9}$$

---

[2] This corresponds to a transmission type recording geometry. The results turn out to be the same for reflection type recording geometry.

where

$$K_x = k(u_x - v_x), \tag{3.10}$$

$$K_y = k(u_y - v_y). \tag{3.11}$$

When R changes angle in the x–z plane by $\Delta\theta_R$, the wave vector of R changes from $k\mathbf{u}$ to $k\mathbf{u}'$, where

$$\mathbf{u}' = \big(-\sin\theta_R - \cos\theta_R\ \Delta\theta_R,\ 0,\ \cos\theta_R - \sin\theta_R\ \Delta\theta_R \big). \tag{3.12}$$

The reconstructed hologram becomes at $z = 0$ (the holographic recording plane),

$$R'T\big|_{z=0} = C \exp\left\{ jk(v_x + u_z \Delta\theta_R)x + v_y y \right\}, \tag{3.13}$$

where $T = R^* S|_{z=0}$. Comparing this with the original signal beam, the wave vector of the reconstructed signal beam has changed to $k\mathbf{v}'$, where

$$\begin{aligned} \mathbf{v}' &= \big(\sin\theta_S + \cos\theta_S\ \Delta\theta_S,\ 0,\ \cos\theta_S - \sin\theta_S\ \Delta\theta_S \big) \\ &= \left(\sin\theta_S - \cos\theta_R\ \Delta\theta_R,\ 0,\ \cos\theta_S + \frac{\sin\theta_S\ \cos\theta_R}{\cos\theta_S}\ \Delta\theta_R \right), \end{aligned} \tag{3.14}$$

and $\Delta\theta_S$ is the angle by which the signal beam wave vector $k\mathbf{v}$ has rotated (in the x–z plane). Note that $|\mathbf{v}'| = 1$. The x, y components of $k\mathbf{v}'$ are the same as $k\mathbf{u}' - \mathbf{K}$, however there is a mismatch in the z component. The angle $\Delta\theta_S$ satisfies the relation

$$-\cos\theta_S\ \Delta\theta_S = \cos\theta_R\ \Delta\theta_R, \tag{3.15}$$

which at the input plane corresponds to a shift in the point source (x_0, y_0) by

$$\Delta x = F\cos\theta_S\ \Delta\theta_S = -F\cos\theta_R\ \Delta\theta_R. \tag{3.16}$$

For volume holograms, the above relation is still true within the Bragg-selectivity angle, which is given by [5] [3]

$$\Delta\theta_{R,Bragg} = \frac{2\lambda}{nL} \cdot \frac{\cos\theta_S}{\sin(\theta_R + \theta_S)}, \tag{3.17}$$

---

[3] This angle corresponds to the full width of the sinc function between the first nulls. See Eq. (2.32) or (2.123).

where L is the thickness of the hologram, λ is the wavelength, and n is the index of refraction of the recording material. Thus the point source at (x_0, y_0) will move by

$$\Delta x = \frac{2\cos\theta_R \cos\theta_S}{n \sin(\theta_R + \theta_S)} \cdot \frac{\lambda F}{L}, \tag{3.18}$$

before disappearing due to Bragg-mismatch. For the image of the point source at (x_0, y_0) to disappear before it moves by δ (the inter-pixel distance), we require that

$$\frac{g\lambda F}{L} < \delta, \tag{3.19}$$

where

$$g = \frac{2\cos\theta_R \cos\theta_S}{n \sin(\theta_R + \theta_S)}. \tag{3.20}$$

In the preceding analysis, the effect of refraction at the interface of the recording medium and air was ignored. To use the results above, we should first change to angles inside the recording medium. Roughly speaking, when seen from inside the recording media, distances from the object to the interface of the recording media will appear to be n times the actual distance (under paraxial approximations). Thus the focal length F of the lens should be replaced by nF and the quantity g in Eq. (3.20) should be changed to

$$g = \frac{2\cos\theta_R \cos\theta_S}{\sin(\theta_R + \theta_S)}, \tag{3.21}$$

where all angles are *inside* the recording media. Note that from Eq. (3.19), increasing the thickness of the recording media L increases the alignment sensitivity (i.e., the system becomes more sensitive to alignment errors).

The value of g is typically around 1. For example, if we take $n = 2.2$ (approximately the index of refraction of lithium niobate), $\theta_S = \theta_R = 30°$ outside the recording media, then $\theta_S = \theta_R = 13.1°$ inside the recording media, and g is approximately 4.3. If we have $F = 10$ cm, $\lambda = 500$ nm, and $L = 1$ cm, then from Eq. (3.19), δ should be larger than 21 μm for the image to disappear before moving by the inter-pixel distance due to change in the reference beam angle. If we take $\theta_S = 0$ and $\theta_R = 30°$ (outside the recording media) instead, then $g = 8.57$,

and δ has to be larger than 42 μm. The condition in Eq. (3.19) is usually satisfied. For example, a typical liquid crystal TV such as the Epson LCTV has pixel sizes of around 40 μm.

We now consider the implications of Eq. (3.19) in terms of storage density of holograms. [4] Under paraxial approximations, the spatial extent of the Fourier transform is

$$s = \frac{\lambda F}{\delta}. \tag{3.22}$$

Substituting this into Eq. (3.19), the condition becomes

$$gs < L. \tag{3.23}$$

Thus given the optimum recording size s (in terms of recording density), the thickness of the recording media L should be at least a factor of g larger than s in order for the image to disappear before it moves by the inter-pixel distance, δ, due to change in the reference beam angle.

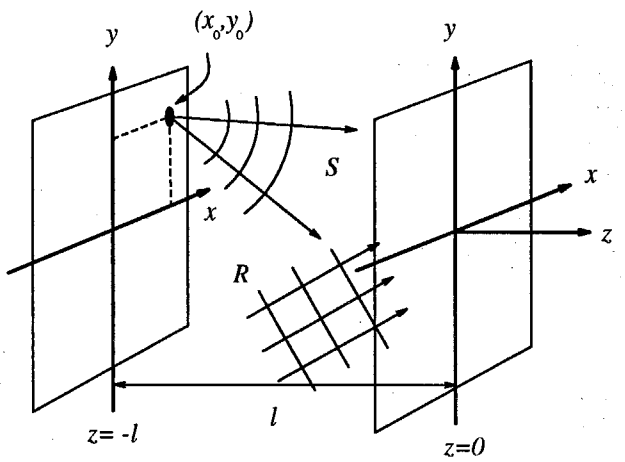


Figure 3.4. Recording system with plane wave reference beam and spherical wave signal beam.

[4] Please see Chapter 4 for more details.

3.2.2. Spherical Wave Signal Beam

We now turn to case 2 where the signal beam is a spherical wave, as shown in Figure 3.4. The reference beam is again a plane wave, and the signal beam is assumed to originate from the point source at $(x_0, y_0, -l)$, where $y_0 \approx 0$. The holographic recording plane is again at $z = 0$, so that the point source lies in the x–z plane. Under paraxial approximations, we have

$$R = A\, e^{jk(u_x x + u_y y + u_z z)}, \tag{3.24}$$

$$S = \frac{B}{l}\, e^{jkl}\, \exp\left\{ j\frac{k}{2l}\left[(x - x_0)^2 + (y - y_0)^2\right]\right\}, \tag{3.25}$$

where u_x, u_y, and u_z, are given by Eq. (3.6)–(3.8). The hologram at $z = 0$ then becomes

$$T = T_0 \exp\left\{ j\frac{k}{2l}\left[(x - a)^2 + (y - y_0)^2\right]\right\}, \tag{3.26}$$

where T_0 is a constant, and

$$a = x_0 - u_x l. \tag{3.27}$$

When the reference beam changes angle by $\Delta\theta_R$ (in the x–z plane), the wave vector becomes $k\mathbf{u}'$, where $\mathbf{u}'$ is given by Eq. (3.12). The reconstructed hologram (at $z = 0$) is then

$$S' = R'T = \frac{B'}{l}\, \exp\left\{ j\frac{k}{2l}\left[(x - x_0')^2 + (y - y_0)^2\right]\right\}, \tag{3.28}$$

where B' is some constant, and

$$x_0' = x_0 - l\, \cos\theta_R\, \Delta\theta_R. \tag{3.29}$$

Comparing Eq. (3.28) with (3.25), the reconstructed image of the point source (x_0, y_0) appears to have shifted in the x direction by

$$\Delta x = l\, \cos\theta_R\, \Delta\theta_R. \tag{3.30}$$

On the other hand, the Bragg-angle selectivity is given approximately by Eq. (3.17), where θ_S is the angle between the z axis and the line connecting the

point source $(x_0, y_0, -l)$ to the origin (the center of the holograph). [5] For the reconstructed image to disappear before it moves the distance δ, we require that

$$\frac{g\lambda l}{L} < \delta, \tag{3.31}$$

where g is given by Eq. (3.21). As before, all angles are inside the recording media. If we consider the effect of refraction, the angle θ_S is approximately

$$\theta_S = \tan^{-1}\left(\frac{x_0^2 + y_0^2}{nl}\right). \tag{3.32}$$

Since l is comparable to F, the conditions given by Eq. (3.19) and (3.31) are about the same.

In the limit where we record in the image plane, $l = 0$, the shift $\Delta x = 0$ to first order. [6]

3.2.3 Perpendicular Angle Changes

Up to now, we have assumed that the reference beam angle changes direction within the same plane formed by the reference and signal beams. We now consider the effect when the angle change is perpendicular to this plane.

If the angle changes by $\Delta\theta$, then for case 1, the image will shift by $\Delta x = F\Delta\theta$, where F is the focal length of the Fourier transform lens. For case 2, the shift will be approximately $\Delta x = l\Delta\theta$, where l is the distance from the image plane to the holograph. Since image reconstruction of volume holograms is not sensitive to angle changes in this direction, the reconstructed image moves a considerable amount before disappearing from Bragg-angle mismatch.

---

[5] The wave front of the signal beam spherical wave recorded near the origin — the center of the holograph — is approximately a plane wave traveling in the direction from $(x_0, 0, -l)$ to the origin.

[6] Strictly speaking the formulas in Eq. (3.25) are only valid when l is sufficiently large. Nevertheless the conclusion that Δx is zero when $l = 0$ is true.

3.3. Wavelength Alignment Sensitivity

In this section we consider the effect of wavelength errors in the reference beam on the reconstruction of holograms. We assume that the reference beam is a plane wave reference beam that changes in wavelength λ, but not in direction.

3.3.1. Plane wave signal beam

We assume as before the expressions in Eq.s (3.1)–(3.11). When the wavelength of the reference beam changes by $\Delta\lambda$, k changes by $\Delta k = -2\pi\Delta\lambda/\lambda^2$, and the wave vector of the reference beam changes to

$$k'\mathbf{u} = (k'u_x, k'u_y, k'u_z), \tag{3.33}$$

where $k' = k + \Delta k$. The wave vector of the reconstructed signal is then

$$k'\mathbf{v}' = k'\mathbf{u} - \mathbf{K} = (k'v'_x, k'v'_y, k'v'_z), \tag{3.34}$$

where $v'_x = v_x + \Delta v_x$, etc., and

$$\Delta v_x = \left(\frac{\Delta k}{k}\right)(u_x - v_x), \tag{3.35}$$

$$\Delta v_y = \left(\frac{\Delta k}{k}\right)(u_y - v_y), \tag{3.36}$$

$$\Delta v_z = \left(\frac{\Delta k}{k}\right)(u_z - v_z). \tag{3.37}$$

Note that here $|\mathbf{v}'|$ is in general not equal to 1. Following the same line of arguments as before, the reconstructed image of the original point source at (x_0, y_0) will appear to have moved in the x direction by

$$\begin{aligned}\Delta x &= F\Delta v_x \\ &= -\left(\frac{\Delta k}{k}\right)(\sin\theta_R + \sin\theta_S)\,F \\ &= \left(\frac{\Delta\lambda}{\lambda}\right)(\sin\theta_R + \sin\theta_S)\,F,\end{aligned} \tag{3.38}$$

since $v_x = \sin\theta_S = x_0/F$ by Eq. (3.3).

On the other hand, the mismatch of the z component of the wave vector $k'\mathbf{v}'$ as given by Eq.s (3.34)–(3.37) is

$$\begin{aligned}\Delta k_z &= k'\left(v'_z - \sqrt{1 - v'^2_x - v'^2_y}\right) \\ &\approx \left\{(u_z - v_z) + \frac{v_x}{v_z}(u_x - v_x)\right\}\Delta k \\ &= -\frac{\Delta k}{\cos\theta_S}\left[1 - \cos(\theta_S - \theta_R)\right] \\ &= \frac{2\pi n \Delta\lambda}{\lambda^2 \cos\theta_S}\left[1 - \cos(\theta_S - \theta_R)\right]. \end{aligned} \tag{3.39}$$

The wavelength-selectivity (full-width) is given by $\frac{L}{2}\Delta k_z = 2\pi$, and for the case where $\theta_R = 180°$ and $\theta_S = 0$ (reflection holograms), Eq. (3.39) gives us the familiar formula

$$\frac{\Delta\lambda}{\lambda} = \frac{\lambda}{nL}. \tag{3.40}$$

In general the Bragg-selectivity is given by

$$\Delta k = \frac{4\pi}{L}\,\frac{\cos\theta_S}{1 - \cos(\theta_S + \theta_R)}. \tag{3.41}$$

To avoid having the point source move by more than δ before disappearing due to Bragg-mismatch, we require that Eq. (3.19) hold, where g is now given by (after adjusting for refraction)

$$g = \frac{2\cos\theta_S\,(\sin\theta_R + \sin\theta_S)}{1 - \cos(\theta_S + \theta_R)}. \tag{3.42}$$

As before, all angles are inside the recording media. When recording wavelength multiplexed holograms, it is advantageous to record reflection type holograms with $\theta_R = 180°$ and $\theta_S \approx 0$, which gives us $g \approx 0$. In contrast, if we had taken $\theta_R = \theta_S = 30°$ (outside the recording media) and $n = 2.2$, then $g = 8.6$, which is twice that of angle multiplexing.

Thus compared to angle multiplexing, a system that uses only wavelength multiplexing is less sensitive to multiplexing error if we use the counter-propagating reflection geometry. For $\theta_R = 180°$, $g \approx x_0/F$ under paraxial approximations, and Eq. (3.19) becomes

$$\frac{\lambda x_0}{L} < \delta. \tag{3.43}$$

For wavelength multiplexing, g (for wavelength alignment sensitivity) is zero for the counter-propagating reflection geometry ($\theta_R = 180°$ and $\theta_S = 0$), while the angle selectivity for this geometry is at a minimum.

3.3.2. Spherical Wave Signal Beam

We now consider spherical wave signal beams. For spherical wave signal beams, we assume the same geometry as case 2 for angle multiplexing (Figure 3.4), except that here the wavelength rather than the direction of the reference beam is changing. As before, the hologram at the recording plane $z = 0$ is (from Eq.s (3.26) and (3.27))

$$T = T_0 \exp\left\{ j\frac{k}{2l}[(x - x_0 + u_x l)^2 + (y - y_0)^2] \right\} \tag{3.44}$$

for some constant T_0. Illuminating with the reference beam

$$R' = A\, e^{jk'\mathbf{u}\cdot\mathbf{x}}, \tag{3.45}$$

where $k' = k + \Delta k$, we get the reconstructed hologram at $z = 0$ to be

$$S' = R'T = C \exp\left\{ j\frac{k'}{2l'}[(x - x_0')^2 + (y - y_0)^2] \right\}, \tag{3.46}$$

where

$$l' = l + \Delta l = \left(1 + \frac{\Delta k}{k}\right), \tag{3.47}$$

$$x_0' = x_0 + \left(\frac{\Delta k}{k}\right) l u_x. \tag{3.48}$$

Thus the point source will appear to have moved from $(x_0, y_0, -l)$ to $(x_0', y_0, -l')$. It will not only have shifted by

$$\Delta x = \left(\frac{\Delta k}{k}\right) l u_x, \tag{3.49}$$

in the x–y plane, but also its depth will have changed by

$$\Delta l = \left(\frac{\Delta k}{k}\right) l. \tag{3.50}$$

This has the effect of defocusing, and the situation is therefore more complicated. If we require that $\Delta x < \delta$, for the change in the Δk given by Eq. (3.41) (where θ_R is the angle between the z axis and the line connecting the point source $(x_0, 0, -l)$ to the origin), then we again have the condition given by Eq. (3.19), where now

$$g = \frac{2 \sin\theta_R \ \sin\theta_S}{1 - \cos(\theta_S + \theta_R)}. \tag{3.51}$$

As in case 1, $g \approx 0$ for the reflection geometry where $\theta_R = 180°$ and $\theta_S = 0$. The concern for wavelength multiplexing therefore is primarily on the defocusing effect. However, since we are interested in high density storage, we would like to store at the image plane, where $l = 0$. This is exactly where the defocusing effect is zero.

3.4. Rotation Alignment Sensitivity: Plane Wave Reference Beam

We now consider the effect of spatial multiplexing in a 3-D disk system, which is done by rotating the disk. In this section we assume that the reference beam is a plane wave.

3.4.1. Plane Wave Signal Beam

We first consider case 1 where the signal beam is a plane wave, as shown in Figure 3.3. As in Section 3.2.1, we consider a point source at (x_0, y_0) on the input plane, and the signal and reference beams are plane waves given by Eq.s (3.1) and (3.2). Under paraxial approximations, we have

$$v_x = x_0/F \tag{3.52}$$

$$v_y = y_0/F \tag{3.53}$$

$$v_z = \sqrt{1 - v_x^2 - v_y^2}, \tag{3.54}$$

where F is the focal length of the Fourier transform lens. Following the same analysis as before, the holograph at $z = 0$ is

$$T = T_0 e^{j(K_x x + K_y y)}, \tag{3.55}$$

where K_x and K_y are given by Eq.s (3.10) and (3.11). When T is rotated about (x_c, y_c) by $\Delta\phi$, it becomes

$$T' = T_0' \exp\left\{j[(K_x - K_y\Delta\phi)x + (K_y + K_x\Delta\phi)y]\right\}, \tag{3.56}$$

for some constant T_0'. Illuminating with the reference beam R to T' reconstructs the hologram, which is

$$RT' = C \exp\left\{jk(v_x'x + v_y'y + v_z'z)\right\}, \tag{3.57}$$

where

$$v_x' = v_x - (v_y - u_y)\Delta\phi, \tag{3.58}$$

$$v_y' = v_y + (v_x - u_x)\Delta\phi. \tag{3.59}$$

Comparing this to the expression for S in Eq. (3.2) and using Eq.s (3.52) and (3.53), the reconstructed hologram appears to originate from a point source at

$$x_0' = x_0 - (y_0 - Fu_y)\Delta\phi \tag{3.60}$$

$$y_0' = y_0 + (x_0 - Fu_x)\Delta\phi. \tag{3.61}$$

This is just the original point source rotated about the center at

$$x_c' = Fu_x, \tag{3.62}$$

$$y_c' = Fu_y. \tag{3.63}$$

Thus upon rotating the holograph T by $\Delta\phi$, the image will appear to have rotated also by $\Delta\phi$ about the center $(x_c', y_c') = (Fu_x, Fu_y)$. This center is independent of the actual center of rotation (x_c, y_c) of the holograph and also independent of the distance l between the holograph and the Fourier transform lens.

Since the center of rotation of the (reconstructed) image (x_c', y_c') is different from the actual center of rotation of the disk (x_c, y_c), the radius of rotation of the image (R_I) could be larger or smaller than the radius of rotation of the holograph (R_H). In the present case, the radius of rotation of the image is

$$R_I = F\sqrt{u_x^2 + u_y^2}. \tag{3.64}$$

It is interesting to note that the apparent direction of motion of the reconstructed image is determined only by the reference beam angle, and is independent of the direction of motion of the holograph itself.

3.4.2. Spherical Wave Signal Beam

We now consider case 2 where the signal beam is a spherical wave, as shown in Figure 3.4. We assume the same situation as in Section 3.2.2, where R is given by Eq. (3.24) and S is given by Eq. (3.25). The holograph T is then

$$T = T_0 \exp\left\{ j\frac{k}{2l}\left[(x-a)^2 + (y-b)^2\right]\right\}, \tag{3.65}$$

where

$$a = x_0 + u_x l \tag{3.66}$$

$$b = y_0 + u_y l. \tag{3.67}$$

If we rotate T about (x_c, y_c) by a small angle $\Delta\phi$ (counter-clockwise), the point (x, y) moves to (x', y'), where

$$x' = x - (y - y_c)\Delta\phi \tag{3.68}$$

$$y' = y + (x - x_c)\Delta\phi. \tag{3.69}$$

The transparency function T then becomes

$$T' = T_0' \exp\left\{ j\frac{k}{2l}\left[(x-a')^2 + (y-b')^2\right]\right\}, \tag{3.70}$$

for some constant T_0', and

$$a' = a - (b - y_c)\Delta\phi \tag{3.71}$$

$$b' = b + (a - x_c)\Delta\phi. \tag{3.72}$$

Applying the original reference plane wave R to read out the hologram, we have at $z = 0$ (immediately after the transparency/holograph)

$$\begin{aligned} RT' &= AT_0' \exp\left\{ j\frac{k}{2l}\left[(x-a')^2 + (y-b')^2 + 2lu_x x + 2lu_y y\right]\right\} \\ &= C\exp\left\{ j\frac{k}{2l}\left[(x-a'+lu_x)^2 + (y-b'+lu_y)^2\right]\right\}, \end{aligned} \tag{3.73}$$

where C is some constant. Comparing this with the expression for the original signal beam (Eq. (3.25)), the reconstructed wave RT' is a spherical wave that appears to have originated from the point $(x_0', y_0', -l)$, where

$$x_0' = x_0 - (y_0 - y_c')\Delta\phi \tag{3.74}$$

$$y_0' = y_0 + (x_0 - x_c')\Delta\phi \tag{3.75}$$

and

$$x_c' = x_c - u_x l \tag{3.76}$$

$$y_c' = y_c - u_y l. \tag{3.77}$$

The original point source at (x_0, y_0) thus appears to have rotated by $\Delta\phi$ (counter-clockwise) about the point (x_c', y_c'), and for any image, when the holograph T rotates by $\Delta\phi$, the reconstructed image will appear to have rotated by the same angle about the center (x_c', y_c').

When $l = 0$, the hologram is recorded at the image plane, and (x_c', y_c') coincides with (x_c, y_c). [7] In this case $R_H = R_I$.

---

[7] Strictly speaking, the paraxial approximation that leads to Eq. (3.63) does not hold for $l = 0$. Nevertheless the conclusion above is valid.

In general, from Eq.s (3.76) and (3.77), the radius of rotation of the reconstructed image is

$$R_I = \left\{R_H^2 + (u_x^2 + u_y^2)l^2 - 2l(x_c u_x + y_c u_y)\right\}^{1/2}. \tag{3.78}$$

Comparing the R_I of the image plane hologram (which is equal to R_H) to the R_I of Fourier transform holograms, the ratio of the two is

$$\frac{R_I}{R_H} = \frac{F}{R_H}\sqrt{u_x^2 + u_y^2}. \tag{3.79}$$

3.4.3. Optimum Configuration

In either case — Fourier plane or image plane holograms — the reconstructed image will appear to rotate around some center by the same angle that the holograph rotates. The best that can be done then, is to make $R_I = 0$; i.e., have the reconstructed hologram rotate about the center of the image itself. From Eq.s (3.76) and (3.77), this can be done if we set

$$x_c = u_x l, \tag{3.80}$$

$$y_c = u_y l. \tag{3.81}$$

(It is assumed here that the center of the image is at $x = 0$, $y = 0$. If this is not true, then the conditions given above change, but the idea is the same.) A realization of such an arrangement is shown in Figure 3.5. In this optimum configuration, the pixels at the edge of the image will move the most, and in the worst case the radius of rotation is r, where r is the distance from the center of the image to the outermost pixel. This can be much less than the actual radius of rotation of the disk (R_H).

Of course, once l and (x_c, y_c) are fixed, there is only one reference beam angle that will give the optimum configuration. For angle multiplexing where the reference beam angle changes, we would set the center of the reference beam angle

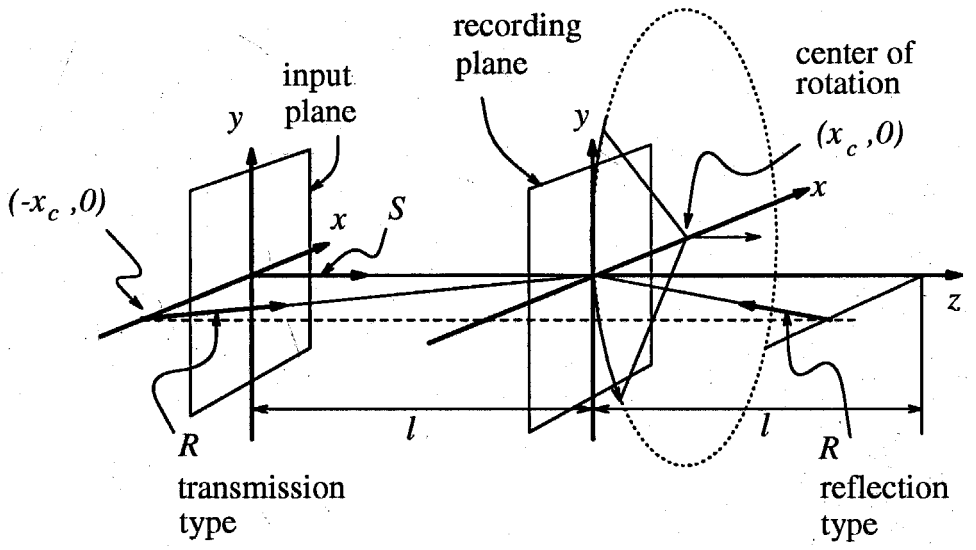


Figure 3.5. An optimum recording configuration system with minimum alignment sensitivity.

swing to be at the optimum angle. For wavelength multiplexing, the angle of the reference beam does not need to be changed, so this is not a problem.

In Figures 3.6, we show experimental data demonstrating the effect of holograph rotation. In Figure 3.6(a), a point source was recorded as an image plane hologram (Figure 3.7(a)). Upon rotation of the holograph, the image of the point moves horizontally in the same direction that the holograph is moving. The picture shown in Figure 3.6(a) was taken by making multiple exposures of the reconstructed image of a point source moving (due to disk rotation) at intervals of $\Delta\phi = 0.4°$. In the next experiment, the image of the point was recorded as a Fourier transform hologram (Figure 3.7(b)). Figure 3.6(b) shows the reconstructed image upon rotation of the holograph at $\Delta\phi = 0.4°$ intervals. Note that the point appears to move in a vertical direction even though the holograph itself is moving horizontally. In general, of course, the direction of motion will be neither horizontal nor vertical, but will depend on the distance between the image and the holograph.

In these experiments, the images were recorded as volume holograms on a

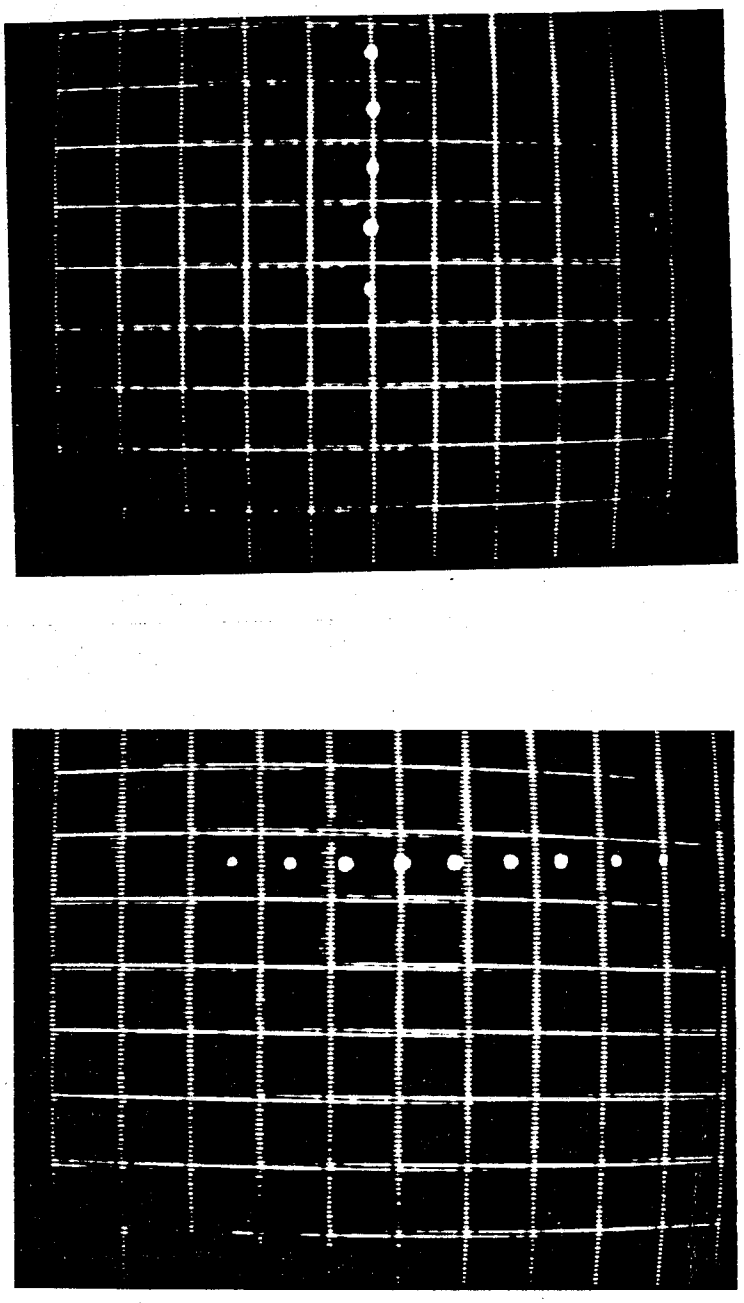


Figure 3.6. Experimental data:

(a) reconstruction of image plane hologram using the configuration in Figure 3.7(a). ($\Delta\phi = 0.4°$)

(b) reconstruction of Fourier plane hologram using the configuration in Figure 3.7(b). ($\Delta\phi = 0.4°$) sensitivity.

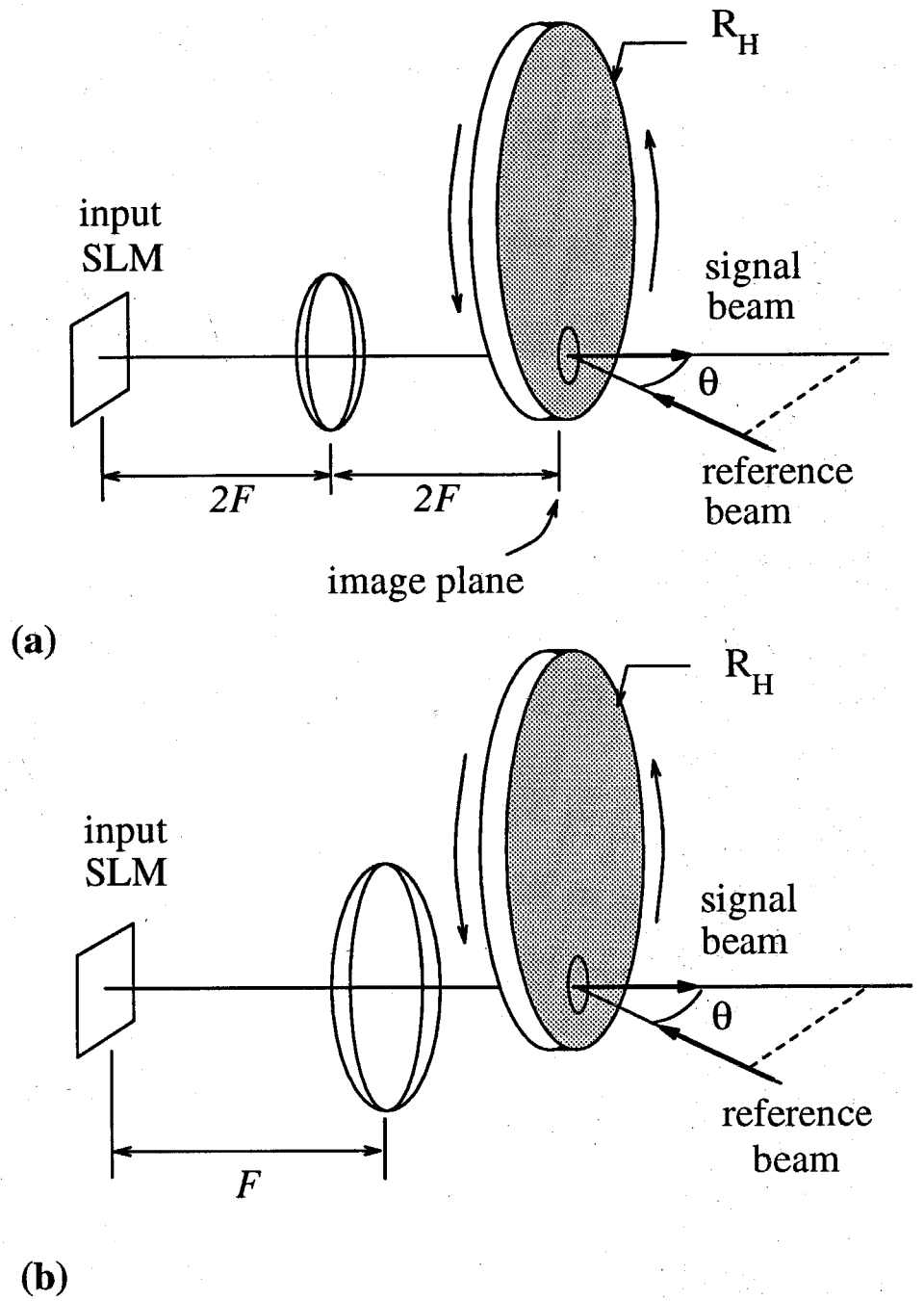


Figure 3.7. Recording geometries for the experiments in Figure 3.6.

(a) Recording geometry for image plane holograms. $R_H = 1.7$ cm and $\theta = 27°$ (outside the crystal).

(b) Recording geometry for Fourier plane holograms. $R_H = 1.7$ cm and $\theta = 27°$ (outside the crystal). The holograph is 4 cm before the Fourier transform plane.

5 mm thick lithium niobate crystal. As mentioned above, although the analysis here was carried out for planar holograms, within the Bragg-sensitivity angle the results still hold for volume holograms, as experiments have confirmed.

Figure 3.8 shows the effect of holograph rotation on the reconstruction of recorded images stored as (a) image plane holograms, (b) Fourier plane holograms, and (c) optimum configuration holograms. In all cases, the reconstructed holographic image rotates as the holograph (disk) rotates. For the optimum configuration, however, the reconstructed image rotates around the center of the image instead of moving out of the field of view.

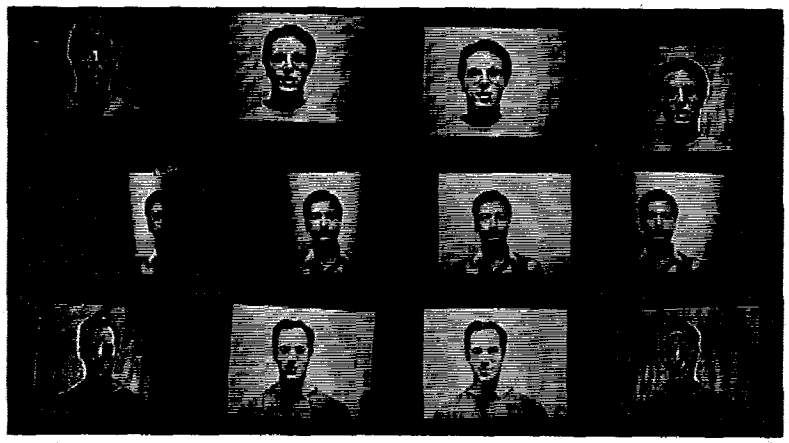


Figure 3.8. Reconstructed images from rotated holograph.

(a) 1st row: rotation of image plane hologram.

(b) 2nd row: rotation of Fourier plane hologram.

(c) 3rd row: rotation of hologram using optimum configuration for minimizing alignment sensitivity.

3.5. Rotation Alignment Sensitivity: Spherical Wave Reference Beam

In the previous section, we assume that the reference beam is a plane wave.

One might ask whether using some other wave form as the reference beam would make any difference in terms of rotational alignment sensitivity. The next simplest possible wave-form is the spherical wave. In this section, we apply the same method used in the previous section to the situation where we have spherical waves reference beams instead of plane waves reference beams. It will be shown that as before, the reconstructed image will rotate by the same angle as the disk rotates, but in general with a different radius.

Since the analysis is almost identical to the last section, some of the details will be omitted.

3.5.1. Plane Wave Signal Beam

The analysis for case 1 for a spherical wave reference beam is similar to case 2 when the reference beam is a plane wave; we just exchange the role of R and S and write

$$S = A\, e^{jk(v_x x + v_y y + v_z z)}, \tag{3.82}$$

$$R = \frac{B}{l}\, e^{jkl}\, \exp\left\{ j\frac{k}{2l}\left[(x-x_2)^2 + (y-y_2)^2\right]\right\}, \tag{3.83}$$

where v_x, etc., are given by Eq.s (3.52)–(3.54), and $(x_2, y_2, -l)$ is the point where the reference beam spherical wave originates.

The expressions for T and T' are then the same as the expressions given in Eq.s (3.65) and (3.70), with v_x replacing u_x, etc. The details of the rest of the derivation are similar and will not be repeated here. The conclusion is that the image will again rotate around some center (x'_c, y'_c), which in this case is given by

$$x'_c = \frac{F}{l}(x_c - x_2), \tag{3.84}$$

$$y'_c = \frac{F}{l}(y_c - y_2). \tag{3.85}$$

3.5.2. Spherical Wave Signal Beam

For case 2, we assume that the point source for the spherical wave reference beam is at $(x_2, y_x, -l_2)$ and consider a point source at $(x_1, y_1, -l_1)$ on the input plane. The holographic recording plane is again at $z = 0$. Taking the paraxial approximation, we have at $z = 0$

$$S = \frac{A}{l_2} e^{jkl_2} \exp\left\{ j\frac{k}{2l_2}\left[(x-x_2)^2 + (y-y_2)^2\right]\right\}, \tag{3.86}$$

$$R = \frac{B}{l_1} e^{jkl_1} \exp\left\{ j\frac{k}{2l_1}\left[(x-x_1)^2 + (y-y_1)^2\right]\right\}. \tag{3.87}$$

The holograph/transparency T is proportional to SR^*, and is of the same form as the expression in Eq. (3.65), with

$$\frac{1}{l} = \frac{1}{l_1} - \frac{1}{l_2}, \tag{3.88}$$

$$a = \frac{l}{l_1}x_1 - \frac{l}{l_2}x_2, \tag{3.89}$$

$$b = \frac{l}{l_1}y_1 - \frac{l}{l_2}y_2. \tag{3.90}$$

Rotating T by $\Delta\phi$, T becomes T', which is of the same form as the expression in Eq. (3.70), with

$$a' = a - (b - y_c)\Delta\phi, \tag{3.91}$$

$$b' = b + (a - x_c)\Delta\phi. \tag{3.92}$$

Applying R (Eq. (3.87)) to T', we get the reconstructed hologram $T'R$ as

$$T'R = C \exp\left\{ j\frac{k}{2l_1}\left[(x-a'')^2 + (y-b'')^2\right]\right\}, \tag{3.93}$$

where

$$a'' = x_1 - (y_1 - y_c')\Delta\phi, \tag{3.94}$$

$$b'' = y_1 + (x_1 - x_c')\Delta\phi, \tag{3.95}$$

and

$$x_c' = x_c + \frac{l_1}{l_2}(x_2 - x_c), \tag{3.96}$$

$$y_c' = y_c + \frac{l_1}{l_2}(y_2 - y_c). \tag{3.97}$$

Comparing these results to those in the previous section, we see that the reconstructed image rotates about the center (x_c', x_c'), where x_c' and y_c' are given by Eq.s (3.96) and (3.97).

3.5.3. Condition For Rotational Invariant Holograms

In view of the results on rotational sensitivity using plane wave reference beams and spherical wave reference beams, one might ask whether it is possible to design the system $L1$ (in Figure 3.2) and choose the reference beam wave form such that, upon rotation of the holograph T, the reconstructed image does not change (at least for small angles). We will now analyze this more carefully. Let $g(x, y)$ be the field distribution of the reference beam R at the recording plane $z = 0$, and let $h(x, y)$ be the field distribution at $z = 0$ of the impulse response of a point source at (x_0, y_0) on the input plane. The holograph is then

$$T(x, y) = g^*(x, y)h(x, y), \tag{3.98}$$

and upon rotation by an angle $\Delta\phi$, T becomes

$$\begin{aligned} &T\big(x - (y - y_c)\Delta\phi,\ y + (x - x_c)\Delta\phi\big) \\ \approx &g^*h - \frac{\partial}{\partial x}(g^*h)\,(y - y_c)\Delta\phi + \frac{\partial}{\partial y}(g^*h)\,(x - x_c)\Delta\phi. \end{aligned} \tag{3.99}$$

If we reconstruct the hologram by applying the reference beam g, we get

$$\begin{aligned} &g(x, y)T\big(x - (y - y_c)\Delta\phi,\ y + (x - x_c)\Delta\phi\big) \\ \approx &gg^*h - g\frac{\partial}{\partial x}(g^*h)\,(y - y_c)\Delta\phi + g\frac{\partial}{\partial y}(g^*h)\,(x - x_c)\Delta\phi. \\ = &|g|^2h\left\{1 - (y - y_c)\frac{\partial F}{\partial x}\Delta\phi + (x - x_c)\frac{\partial F}{\partial y}\Delta\phi\right\}, \end{aligned} \tag{3.100}$$

where

$$F = \ln T = \ln(g^*h). \tag{3.101}$$

For faithful reconstruction of h, we require that $|g|^2$ be approximately a constant and that the expression in Eq. (3.100) be approximately the same (within a proportionality constant) to the original impulse response $h(x,y)$. This requires that for all x and y,

$$-(y-y_c)\frac{\partial F}{\partial x}+(x-x_c)\frac{\partial F}{\partial y}=C, \tag{3.102}$$

where C is some constant.

In particular, for $x=x_c$, we have

$$(y-y_c)\frac{\partial F}{\partial x}=C \tag{3.103}$$

for all y. Therefore $C=0$, and Eq. (3.102) becomes

$$(y-y_c)\frac{\partial F}{\partial x}=(x-x_c)\frac{\partial F}{\partial y}. \tag{3.104}$$

We now change variables from x, y to u, v, where

$$u=\left(\frac{1}{2}y^2-y_cy\right)+\left(\frac{1}{2}x^2-x_cx\right), \tag{3.105}$$

$$v=\left(\frac{1}{2}y^2-y_cy\right)-\left(\frac{1}{2}x^2-x_cx\right), \tag{3.106}$$

and consider F as a function of u and v. From Eq. (3.104) we then get

$$\frac{\partial F}{\partial v}=0. \tag{3.107}$$

Thus F is a function of u only, and we have

$$\begin{aligned}F&=F\left(\left(\frac{1}{2}y^2-y_cy\right)+\left(\frac{1}{2}x^2-x_cx\right)\right)\\&=F\left(\frac{1}{2}(x-x_c)^2+\frac{1}{2}(y-y_c)^2-\frac{1}{2}(x_c^2+y_c^2)\right).\end{aligned} \tag{3.108}$$

This implies that T has circular symmetry about the center of rotation (x_c,y_c), which is what we would expect. Note that for the optimum configuration geometry discussed in the previous section, the holograph T (Eq. (3.65)) is circularly symmetric about the point (x_c,y_c) if the point source is at $x_0=0$, $y_0=0$ (from

Eq.s (3.66), (3.67) and (3.80), (3.81)). Thus the point $(0,0)$ does not move when the holograph rotates, while neighboring points rotate around it. There are of course impulse responses that satisfy the condition in Eq. (3.108). A trivial solution is to encode information as concentric rings of different intensity and radii. What we would really like, however, is an optical system where the impulse responses of all the points in a 2-D arrangement give rise to circularly symmetric holograms $T(x,y)$. It is not clear how such a system can be implemented.

Throughout Section 3.4 and 3.5, we have ignored the effect of refraction. To apply the results derived in these sections, the angles and lengths should first be converted to that seen from inside the recording media. As mentioned earlier, the distance from the object to the interface of the recording media will appear to be n times its actual distance, where n is the index of refraction of the recording material.

3.6. Translation And Tilt Alignment Sensitivity

We now turn to the effects of tilt and rotation of the holograph (by small angles) on image reconstruction. We assume in this section that the reference beam is a plane wave and assume the co-planar geometry described at the beginning of Section 3.2.

3.6.1. Translation Effects

We first consider the effect of transverse translation; i.e., when the holograph (the recording media) is translated by a small amount in the x-y plane (Figure 3.3 or 3.4).

In case 1, both the signal and reference beams are plane waves. We record the hologram and translate the holograph sideways (in the x–y plane), and then apply the reference plane wave. This will read out a plane wave in the same direction as the original recording plane wave, and therefore the position of the original

point source does not change when it passes through the Fourier Transform lens. If the holograph is not positioned exactly one focal length away from the lens, the phase of the reconstructed image will change, but the intensity pattern does not. For case 2, the signal beam is a spherical wave, and when the holograph moves by Δx, the reconstructed image will also appear to move by Δx.

Next we consider translation in the z direction (normal to the holograph). For case 1 (where the signal beam is a plane wave) the direction of the reconstructed plane wave signal beam again does not change, and as before the intensity pattern of the reconstructed image remains the same. For case 2 (where the signal beam is a spherical wave) the effect is the same as defocusing.

It should be noted that, whereas in the case of wavelength misalignment (Section 3.3), the change in depth depends on the distance l between the holograph and the image (Eq. (3.50)), in the case of translation errors it is independent of l. For high density storage, we usually try to record with the smallest possible pixels size (at the image plane). [8] However, to obtain small δ, it is necessary to increase the numerical aperture (or equivalently, reduce the $F/number$) of the imaging lens. This in turn decreases the range in which we have good focus.

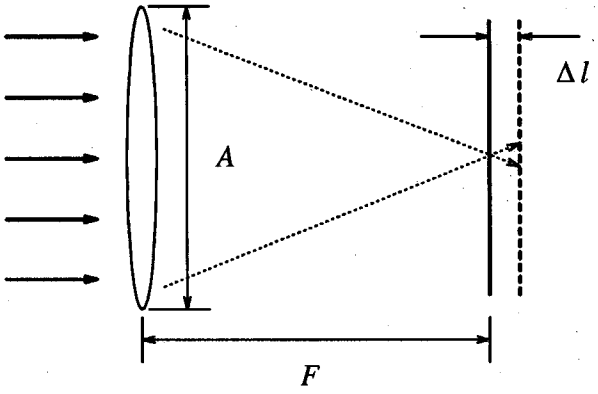


Figure 3.9. System for estimating depth of focus.

The range in which we have good focus can be found as follows [6]: consider

[8] This will be covered in more detail in Chapter 4.

the system shown in Figure 3.9. A lens of focal length F and aperture A is illuminated by a plane wave from the left. Because the finite aperture of the lens, at the back focal plane the image is not a perfect point, but instead has a width of $\delta = 2\lambda F/A$. This gives the resolution of the system. If we observe at a distance of Δl away from the back focus plane, then according to the Fresnel diffraction theory, the intensity distribution is proportional to

$$\begin{aligned} &\int e^{-j\frac{k}{2F}x^2} \cdot e^{j\frac{k}{2(F+\Delta l)}(x'-x)^2} \text{rect}\left(\frac{x}{A}\right) dx \\ \approx &e^{j\frac{\pi}{\lambda F}\left(1-\frac{\Delta l}{F}\right)x'^2} \int e^{-j\frac{2\pi}{\lambda F}\left(1-\frac{\Delta l}{F}\right)x'x} e^{j\frac{\pi\Delta l}{\lambda F^2}x^2} \text{rect}\left(\frac{x}{A}\right) dx. \end{aligned} \tag{3.109}$$

(For simplicity, we have taken only one dimension, and assumed a rectangular aperture instead of a circular one.) In the limit where Δl goes to zero, Eq. (3.109) gives us the familiar sinc function of width $\delta = 2\lambda F/A$. When Δl is not zero, we have the extra factor

$$\exp\left\{j\frac{\pi\Delta l}{\lambda F^2}x^2\right\} \tag{3.110}$$

in the formula. In order for this factor not to produce too much distortion, we require that Δl be less than

$$\Delta l \sim \frac{\lambda}{\pi}\left(\frac{F}{A}\right)^2 = \frac{1}{4\pi}\frac{\delta^2}{\lambda}, \tag{3.111}$$

where we have used the fact that $\delta = 2\lambda F/A$. [9]

The calculations made above implicitly assume that we are taking the paraxial approximation. This is of course not true when the pixel size δ becomes very small. Nevertheless, the exercise above indicates the issue here: namely that the depth of focus decreases rapidly as we go to finer resolutions.

---

[9] A similar estimate can be obtained heuristically by considering the angle of the highest spatial frequency, which is approximately λ/δ. If we take the geometrical projection of one pixel, which is of width δ, then the projected area grows to 2δ at a distance $\Delta l = \delta^2/2\lambda$. This is larger than the estimate given by Eq. (3.111).

3.6.2. Tilt Effects

The effect of tilting the holograph is similar to changing the reference beam angle. We can consider tilt as rotation of the hologram around some axis in the x-y plane. Suppose that the hologram tilts by an angle $\Delta\phi$. For case 1, the signal beam is a plane wave. Upon applying the readout reference beam to the holograph, the direction of the reconstructed signal beam and the original recording beam will differ by an angle of approximately $\Delta\phi$. The reconstructed image will therefore appear to shift by

$$\Delta x \approx F\Delta\phi, \tag{3.112}$$

where F is the focal length of the Fourier transform lens. Similarly for case 2, the reconstructed image will appear to shift by

$$\Delta x \approx l\Delta\phi, \tag{3.113}$$

where l is the distance from the holograph to the image plane. In the special case where we record in the image plane, l is zero, and the effect of tilt is a second-order effect.

As before, the angles here are inside the recording material. If we consider the effect of refraction, we should replace l and F by nl and nF in the expressions above. Although this may seem to make the sensitivity worse, the effect of refraction is such that the change of angle inside the recording material is roughly $1/n$ times the change in angle outside. The two effects therefore cancel.

3.7. Effect of Bragg-mismatch on Image reconstruction

The treatment in the preceding sections assumes that the holograms are planar. As mentioned earlier, volume holograms behave similar (in terms of image positions, but not intensity distributions) to planar holograms for small changes.

For reference beam angle/wavelength errors and tilt errors, if the error is large and in the direction of Bragg-selectivity, not only will the reconstructed image shift, but it is possible that the wrong hologram (page of data) will be read out.

Although changes caused by rotation misalignment are not as sensitive as angle or wavelength misalignments, there are still effects due to Bragg-mismatch, as evident by the pictures shown in Figure 3.8. These effects can be explained by considering the Bragg-mismatches that occur when the holograph rotates.

We start with case 1, where the signal beams (impulse responses) are plane waves. As explained earlier, each point in the input plane corresponds to a plane wave (Figure 3.3). In this section, we assume that the central plane wave is incident normally to the recording material (i.e., in Figure 3.3., the image is centered at the $x = 0$, $y = 0$). Under paraxial approximations, the wave vector of the signal beam plane wave corresponding to (x, y) is

$$\mathbf{k}_S = k\left(\frac{x}{F},\ \frac{y}{F},\ \sqrt{1-\frac{x^2+y^2}{F^2}}\right), \tag{3.114}$$

where F is the focal length of the Fourier transform lens. The wave vector of the reference beam plane wave is again given by

$$\mathbf{k}_R = k(0,\ -\sin\theta_R,\ \cos\theta_R), \tag{3.115}$$

and the grating vector is

$$\mathbf{K} = \mathbf{k}_R - \mathbf{k}_S = k\left(-\frac{x}{F},\ -\sin\theta_R - \frac{y}{F},\ \cos\theta_R - \sqrt{1-\frac{x^2+y^2}{F^2}}\right). \tag{3.116}$$

When the disk rotates along the z axis by $\Delta\phi$, $\mathbf{K}$ becomes (to first order of $\Delta\phi$)

$$\mathbf{K}' = k\Big(-\frac{x}{F} + (\sin\theta_R + \frac{y}{F})\Delta\phi,\ -\sin\theta_R - \frac{y}{F} - \frac{x}{F}\Delta\phi,$$
$$\cos\theta_R - \sqrt{1-\frac{x^2+y^2}{F^2}}\Big). \tag{3.117}$$

The wave vector of the reconstructed signal plane wave is then

$$\mathbf{k}'_S = \mathbf{k}_R + \mathbf{K}' = k\left(\frac{x}{F} - (\sin\theta_R + \frac{y}{F})\Delta\phi,\ \frac{y}{F} + \frac{x}{F}\Delta\phi,\ \sqrt{1-\frac{x^2+y^2}{F^2}}\right). \tag{3.118}$$

Thus the Bragg-mismatch along the z direction is (to first order of $\Delta\phi$) approximately [10]

$$\Delta k_z = \frac{kx}{F} \sin\theta_R \, \Delta\phi. \tag{3.119}$$

The Bragg-mismatch is given by the sinc function

$$\mathrm{sinc}\left(\frac{L}{2}\Delta k_z\right) = \mathrm{sinc}\left(\frac{kLx}{2F} \sin\theta_R \, \Delta\phi\right), \tag{3.120}$$

where L is the thickness of the recording material.

When $\Delta\phi = 0$, the sinc function is equal to 1, and the whole image is reconstructed as expected. As $\Delta\phi$ increase, the sinc function (as a function of x) becomes narrower, and the points further from the center (the larger x's) decreases in intensity more because of the sinc function. The center [11] of the image ($x = 0$) however, remains 1. Note that since the argument of the sinc function depends only on x, but not on y, the change in intensity occurs as strips parallel to the y axis. These conclusions are verified by the images shown in Figure 3.8(b). Note that as the disk rotates, the visible part of the image becomes narrower, and is centered around the head of the person. Eventually, of course, the holograph rotates outside the region of illumination, and the image disappears.

For case 2, where the signal beams (impulse responses) are spherical waves, the situation is slightly more complicated. For image plane holograms, we can apply the results above if we change the variable x to $\lambda F u_x$, where u_x is the spatial frequency component of the image in the x direction. The sinc factor then

---

[10] More accurately, it should be

$$\Delta k_z = \frac{kx}{F} \frac{\sin\theta_R \, \Delta\phi}{\sqrt{1-(x^2+y^2)/F^2}}.$$

However, for small $x^2 + y^2$, the expression in Eq. (3.119) is good enough.

[11] "Center" here refers to the center of the rotated image, not the original center.

becomes

$$\mathrm{sinc}\,(\pi L u_x \sin\theta_R\ \Delta\phi)\,. \tag{3.121}$$

In this case, the result of Bragg-mismatch is not a point-wise effect, but a modification in the spatial frequencies. For small $\Delta\phi$, the sinc factor cuts off the higher spatial. However, as the rotation $\Delta\phi$ becomes larger, the "spatial spectrum" (i.e., the Fourier transform of the image) starts to shift in the same way that the image shifts for case 1. At the same time, the width of the sinc function decreases. When $\Delta\phi$ is large enough, the central lobe of the sinc function becomes so narrow and shifts so much that the spatial spectrum has no DC component. In effect the reconstructed image appears to be edge enhanced, as shown in the sequence of pictures in Figure 3.8(a). From Eq. (3.120), the width of the central lobe is

$$w = \frac{2\lambda F}{L\sin\theta_R\ \Delta\phi}, \tag{3.122}$$

while the shift is

$$s = F\sin\theta_R\ \Delta\phi. \tag{3.123}$$

Thus we expect the edge enhancement effect to begin at $s = w/2$

$$\Delta\phi = \sqrt{\frac{\lambda}{L}}\,\frac{1}{\sin\theta_R}. \tag{3.124}$$

3.8. Discussions and Conclusions

Although we are interested primarily in 3-D holographic recording systems, and specifically the 3-D holographic disk systems, most of the results in this chapter apply equally well to any volume holographic recording system that employs spatial and angle/wavelength multiplexing, as well as 2-D holographic disks [7–9].

Holograph rotation, however, is unique to the 3-D (or 2-D) disk system. In this chapter, it has been shown that for a wide range of practical systems, the reconstructed holograms always rotate by the same angle as the holograph. However, the radius of rotation of the image is in general different from the radius

of rotation of the holograph. It is possible to design the system for minimum alignment sensitivity by placing the center of rotation at the center of the image.

Multiplexing error is a more serious problem for angle multiplexing than for wavelength multiplexing recorded in the counter-propagating reflection geometry. Wavelength multiplexing has the additional advantage that we can choose a fixed reference beam angle such that the rotation center of the image is at the center of the image. For angle multiplexing, the angle of the reference beam needs to be changed, and the optimum configuration is satisfied only at one angle (which we would set at the center of the angle swing). To minimize rotation alignment sensitivity, however, we need to record at angles off the counter-propagating reflection geometry for wavelength multiplexing. In this case, the alignment sensitivity is often worse for wavelength multiplexing. For example, as calculated in Section 3.2.1, for $\theta_S = 0$, $\theta_R = 30°$ (outside the recording material), and $n = 2.2$, the g factor (described in Section 3.2.1) is 5.6 for angle multiplexing and 17.4 for wavelength multiplexing.

For tilt alignment errors, image plane holograms are better than Fourier plane holograms. For translation alignment errors, the opposite is true.

For rotation alignment errors, image plane holograms are the worse, while the optimum configuration yields the best result. Although not immediately obvious from the results in Section 3.4, Fourier transform holograms give results that are comparable to the optimum configuration. This depends of course on how we store the image plane hologram and Fourier plane hologram. Since comparison is meaningful only after these are specified, we will postpone the discussion till Chapter 4 where we analyze the storage density of the 3-D disk (see subsections 4.3.4 and 4.3.5).

Overall, Fourier transform holograms are less sensitive to alignment errors. However, recording multiple Fourier transform holograms is also more difficult because of the dynamic range problem. The higher spatial frequencies (where most of the information is) are usually much weaker than the lower spatial frequencies. For good image reproduction, we would like all the spatial frequencies to be

recorded linearly. Since the intensities at higher spatial frequencies are weaker, the holograms recorded earlier tend to have their lower spatial frequencies erased more than their higher spatial frequencies. The result is that the reconstructed images are edge enhanced. Random phase diffusers may be used to compensate for this.

Recoring multiple holograms in the image plane is in comparison much easier, since the variation in intensity tends to be more distributed. On the other hand, image plane holograms are not only more susceptible to alignment errors, but also to material imperfections. For the Fourier transform hologram, the information of each pixel is distributed throughout the recording volume, whereas for image plane holograms the information is more localized. Localized imperfections in and on the material (such as scratches and dust, etc.) therefore affect the reconstructed image more seriously.

Further comparison between image plane and Fourier plane holograms will be given in Chapter 4 (Section 4.4) where we discuss the storage density and readout time of holograms.

Appendix

Conditions for Shift Invariant Holograms

The same reasoning used in Section 3.5.3. to analyze rotation invariant holographic recording systems may also be employed to analyzing the problem of shift invariant holographic recording systems. In this appendix, we examine the conditions necessary for the reconstructed image to be stationary (to first order) when the holograph T shifts by small amounts.

Consider a (planar) holograph T recorded in the same way as described in Section 3.5.3 (Eq. (3.98)). Let T be shifted by Δx and Δy in the x and y direction. T then becomes

$$T(x+\Delta x, y+\Delta y) \approx \left\{(g^* h) + \frac{\partial}{\partial x}(g^* h)\Delta x + \frac{\partial}{\partial y}(g^* h)\Delta y\right\}. \qquad (3.125)$$

Upon reconstruction by applying the reference beam g, we get

$$g(x,y)T(x+\Delta x, y+\Delta y) \approx \left\{ gg^*h + g\frac{\partial}{\partial x}(g^*h)\Delta x + g\frac{\partial}{\partial y}(g^*h)\Delta y \right\}. \quad (3.126)$$

To have faithful reproduction of the original h, we require

$$gg^* = |g|^2 \approx constant, \quad (3.127)$$

and to have shift invariance (to first order), we require that the expression in Eq. (3.126) be proportional to the original impulse response $h(x,y)$. Thus

$$g\frac{\partial}{\partial x}(g^*h) \approx C_x h(x,y), \quad (3.128)$$

$$g\frac{\partial}{\partial y}(g^*h) \approx C_y h(x,y), \quad (3.129)$$

where C_x and C_y are constants.

We may rewrite Equations (3.128) and (3.129) as

$$\left\{ |g|^2\frac{\partial}{\partial x} + |g|^2\left(\frac{1}{g^*}\frac{\partial g^*}{\partial x}\right) \right\} h = C_x h, \quad (3.130)$$

$$\left\{ |g|^2\frac{\partial}{\partial y} + |g|^2\left(\frac{1}{g^*}\frac{\partial g^*}{\partial y}\right) \right\} h = C_y h, \quad (3.131)$$

and think of this as an eigenvalue problem, where C_x and C_y are the eigenvalues, and the corresponding h is the eigenfunction or eigenmode of propagation. Any image would then be decomposed into these eigenmodes. If the optical system $L1$ in Figure 3.2 is designed such that these eigenmodes correspond to impulse responses, then the reconstructed image will have the same intensity distribution (but not necessarily the same phase distribution) as the original image.

Eqs. (3.130) and (3.131) can be solved if we assume that Eq. (3.127) is true. We then have

$$h(x,y) \approx \frac{A}{g^*(x,y)} \exp\left\{ \frac{C_x x + C_y y}{|g|^2} \right\}. \quad (3.132)$$

In the simplest case, both g (the reference beam) and h (the signal beam) are plane waves, which can be realized by recording Fourier transform holograms with a plane wave reference beam.

References

1. F. T. S. Yu, S. D. Wu, A. W. Mayers, and S. M. Rajan, "Wavelength Multiplexed Reflection Matched Spatial Filters Using $LiNbO_3$," *Opt. Comm.*, **81**(6), 343–347 (1991).
2. G. A. Rakuljic, V. Leyva, and A. Yariv, "Optical-data Storage by using Orthogonal Wavelength-Multiplexed Volume Holograms," *Opt. Lett.*, **17**(20), 1471–1473 (1992).
3. Demetri Psaltis,"Parallel Optical Memories", *Byte*, **17**(9), 179–182 (1992).
4. H.-Y. Li and D. Psaltis, "3-D Holographic Disks," submitted to *Applied Optics.*
5. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell Syst. Tech. J.*, **48(9)**, 2909–2947 (1969).
6. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill Books, New York, 1968).
7. M. A. Neifeld, S. Rakshit, A. A. Yamamura, S. Kobayashi, and D. Psaltis, "Optical Disk Implementation of Radial Basis Classifiers," *SPIE 1990 International Symposium on Optical and Optoelectronic Applied Science and Engineering*, SPIE 1347–02, San Diego, 1990.
8. D. Psaltis, M. A. Neifeld, A. Yamamura, "Image Correlators using Optical Memory Disks," *Opt. Lett.* **14** (9), 429–431 (1989).
9. M. A. Neifeld and D. Psaltis, "Optical Implementations of Radial Basis Classifiers," *Appl. Opt.* **32** (8), 1370–1379 (1993).

Chapter 4
Storage Density of 3-D Holographic Disks

In the 3-D disk system, data is stored and retrieved in parallel blocks or pages, each page consisting of approximately one million bits. As mentioned in Chapter 3, the actual storage density of a practical holographic storage system is often much less than the V/λ^3 upper limit. Various factors that limit the storage density include the numerical aperture of the lenses and dynamic range of recording material.

In this chapter, we will examine the storage capacity of the 3-D disk system. Although the system we have in mind is the 3-D holographic disk, the results derived in this chapter apply equally well to any volume holographic storage system that uses spatial and angle (or wavelength) multiplexing.

The storage density will be derived as a function of recording material thickness, pixel size of the spatial light modulator (SLM), page size (i.e., number of pixels of the SLM), and scanning parameters of the reference beam. The limitations considered in this chapter are "geometric" in nature, where we do not consider dynamic range. We will assume throughout this chapter that the image beam is at normal incidence to the recording material, and that we are recording either transmission type or reflection type holograms.

It will be shown that optimum storage density is approximately 100 bits/μm^2 to 190 bits/μm^2, depending on the resolution of the imaging system. Thus, a 3-D HD stores the equivalent of more than a hundred conventional 2-D disks of the same area.

4.1. Angle Multiplexed Holographic Disk

In this section we address the following question: What is the maximum

number of bits, N, that can be stored in a 3-D HD of area A using angle multiplexing? We will show that in order to maximize N we must properly select the thickness of the HD (L), the magnification of the optical system that transfers the data to the disk, and the angles of incidence for the reference beam. In what follows we derive these optimum parameters. The limits to storage capacity in this chapter are due to geometrical constraints. The dynamic range of the recording material imposes a limit on storage density independently. We will see that the capacity due to the geometric constraints is more restrictive than the material limitations in the 3-D HD system.

We can express N as follows:

$$N = N_S N_\theta N_p^2 \ . \tag{4.1}$$

In the above equation N_S is the number of separate locations on the disk where holograms are superimposed, N_θ is the number of holograms that are angularly multiplexed at the same location, and N_p^2 is the number of pixels in each stored hologram. We will derive an expression for each of the three quantities and then maximize their product with respect to the various parameters of the system.

4.1.1. Maximum Number of Angularly Multiplexed Holograms

We derive an expression for the maximum number of holograms, N_θ, that can be angularly multiplexed at a single location. We assume that data is stored by recording either reflection or transmission holograms. The reference beam is a plane wave whose incident angle is θ_R. The signal beam can be considered as a superposition of plane waves that spans a range of angles. We can calculate N_θ from the angular separation between adjacent holograms, which we take to be $\Delta\theta_R$, the full width of the Bragg angular selectivity of each hologram. An

approximate expression for $\Delta\theta_R$ is [1] [1]

$$\Delta\theta_R = \frac{8\lambda}{n\pi L}\frac{\cos\theta_S}{|\sin(\theta_R+\theta_S)|}, \tag{4.2}$$

where λ is the wavelength, L is the thickness of the hologram, n is the index, and θ_S is the incident angle of the central plane wave component of the signal beam (see Figure 4.1). For transmission holograms $0 < |\theta_R| < \pi/2$, and for reflection holograms $\pi/2 < |\theta_R| < \pi$. The signal beam is assumed to be in the range $0 < |\theta_S| < \pi/2$.

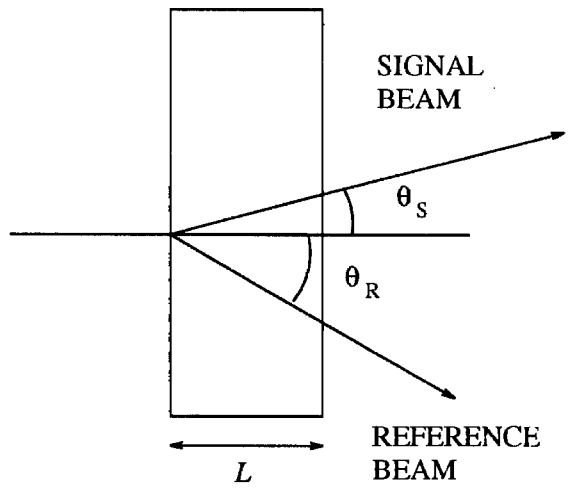


Figure 4.1. The recording geometry.

Eq. (4.2) is only an approximate estimate for the angular selectivity of the entire grating since different plane wave components have different $\Delta\theta_R$. However, Eq. (4.2) is commonly used for setting the angular separation between reference beam angles. The cross-talk resulting when holograms are angularly multiplexed in this way has been calculated recently [2].

To calculate the number of holograms that can fit into a range of reference beam angles θ_R spanning from θ_1 to θ_2 (each hologram being separating from its

---

[1] The criterion used here is slightly different from that used in Eq. (3.17), however the changes are minor. ($8/\pi = 2.5$ is slightly larger than the factor of 2 used in Eq. (3.17).

adjacent holograms by a corresponding $\Delta\theta_R$), we observe that

$$|\sin(\theta_R + \theta_S)|\Delta\theta_R = \frac{8\lambda}{n\pi L}\cos\theta_S, \tag{4.3}$$

which is valid for all possible θ_R angles. If we add together $N_\theta - 1$ such equations, one for each value of θ_R, and approximate the left-hand side of the summation by an integral, we obtain the following expression:

$$\int_{\theta_1}^{\theta_2} |\sin(\theta_R + \theta_S)| d\theta_R = \frac{8\lambda(N_\theta - 1)}{n\pi L}\cos\theta_S. \tag{4.4}$$

Solving for the number of reference angles, we get

$$N_\theta = 1 + \left(\frac{n\pi L}{8\lambda}\right)\frac{|\cos(\theta_S + \theta_1) - \cos(\theta_S + \theta_2)|}{\cos\theta_S}, \tag{4.5}$$

where it is assumed that either $0 < \theta_S + \theta_1 < \theta_S + \theta_2 < \pi$, or $-\pi/2 < \theta_S + \theta_1 < \theta_S + \theta_2 < 0$. Physically, this means that the reference beam is always to one side of the signal beam. The above calculations were carried out for angles inside the recording material. We can use Snell's law to convert to angles outside the recording material.

In the following, we will assume that the image beam has normal incidence ($\theta_S = 0$). In this case, Eq. (4.5) becomes

$$N_\theta = 1 + \left(\frac{n\pi L}{8\lambda}\right)|\cos\theta_1 - \cos\theta_2|, \tag{4.6}$$

where we have $0 < \theta_1 < \theta_2 < \pi/2$ for transmission holograms. We can increase N_θ by a factor of 2 by recording a second set of angularly multiplexed holograms in the range $-\theta_1$ to $-\theta_2$, since the number of holograms that can be angularly multiplexed in the same range of angles is equal to the expression in Eq. (4.6). It is also possible to simultaneously angularly multiplex reflection and transmission holograms, as shown in Figure 4.2. Therefore, the geometric limit on the total number of holograms that can be superimposed in the same location is four times the expression in Eq. (4.6) .

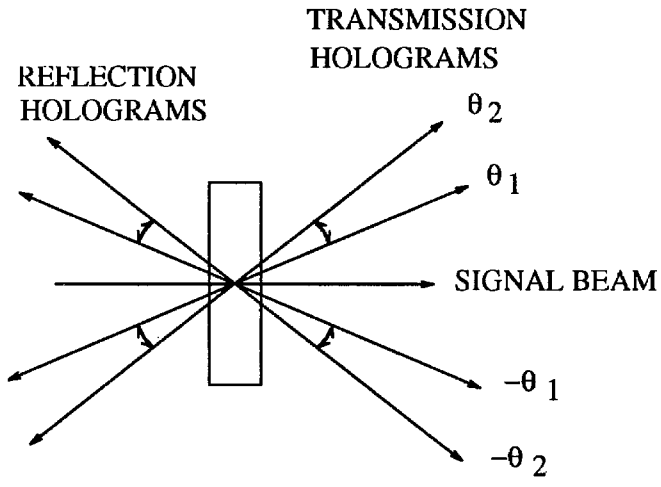


Figure 4.2. Angular multiplexing by reflection and transmission holograms from both sides of the signal beam.

4.1.2. Spatial Multiplexing

The number of non-overlapping spatial locations on a disk with area A is

$$N_S = \frac{A}{a} = \frac{A}{ww'}, \tag{4.7}$$

where $a = w \times w'$ is the area of each location. To find w and w' we need to consider the fact that the stored images can be in exact focus at only one plane in the volume of the recording material. As the thickness of the recording material increases, the area occupied by the defocused image at the surface of the hologram also increases. Moreover, the size of the area that is illuminated by the off-axis reference beam increases in one dimension as the recording material thickness and the angular sweep increase. We will derive expressions for w and w' with reference to the geometry of Figure 4. We assume that the images to be stored are at normal incidence and are focused at the middle of the recording material. We can calculate the extent of the defocused image on the surfaces by tracing the rays corresponding to the highest spatial frequency of the focused image. Let δ be the resolution or pixel spacing of the focused image. Then the maximum spatial frequency is approximately $1/\delta$, corresponding to a diffracted plane wave

traveling at an angle $\theta = \sin^{-1}(\lambda/n\delta)$. We use the ray optics approximation to trace this maximum spatial frequency component and obtain the size of the defocused image at the faces of the recording material:

$$w = N_p\delta + L\tan\theta = N_p\delta + \frac{L}{\sqrt{(n\delta/\lambda)^2 - 1}}. \qquad (4.8).$$

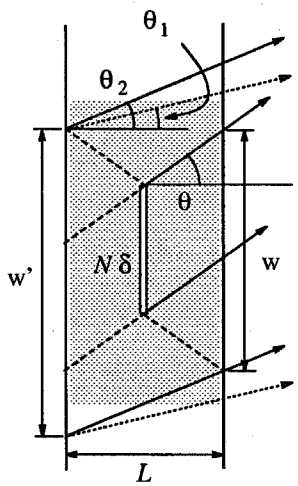


Figure 4.3. Angle multiplexing: extra area taken up by defocusing and reference beam angle change.

As shown in Figure 4.3, in order for the reference beam to fully illuminate the volume of the recording material that the signal beam occupies, it must illuminate a width larger than w in the direction of reference beam sweep. From the geometry of Figure 4.3, this width is

$$w' = w + L\tan\theta_2. \qquad (4.9).$$

The total area that must be devoted to each recording location is therefore

$$a = ww' = w(w + L\tan\theta_2), \qquad (4.10)$$

where w is given by Eq. (4.8).

4.1.3. Optimum N_p, θ_1, and δ

We can now write an expression for N, the total number of bits stored, by using Eqs. (4.7)–(4.10) and Eq. (4.1):

$$\begin{aligned} N &= AN_p^2 \frac{N_\theta}{ww'} \\ &= AN_p^2 \frac{1 + \frac{n\pi L}{8\lambda}(\cos\theta_1 - \cos\theta_2)}{\left[N_p\delta + \frac{L}{\sqrt{(n\delta/\lambda)^2 - 1}}\right]\left[N_p\delta + \frac{L}{\sqrt{(n\delta/\lambda)^2 - 1}} + L\tan\theta_2\right]} \end{aligned} \tag{4.11}$$

We wish to maximize the above expression by optimally selecting N_p, L, θ_1, θ_2, and δ, which are the parameters we can control.

First of all, we note that N decreases monotonically as θ_1 increases (in our analysis $0 < \theta_1 < \pi/2$), therefore $\theta_1 = 0$ is the optimum value. However, since the angular selectivity is very poor around $\theta_1 = 0$, in practice the minimum angle of the reference is set at $\theta_1 \approx 10°$ inside the recording material. Next we consider the optimum number of pixels, N_p. Taking the derivative of N with respect to N_p shows that N is a monotonically increasing function of N_p. This result confirms our intuition since the increase in the disk area required to store the holograms due to defocusing and angular multiplexing can be thought of as an "edge" effect. The use of larger images implies fewer recording locations on the same disk area, and therefore fewer edges. In practice N_p is limited by the number of pixels of the spatial light modulator (SLM) to approximately $N_p = 1,000$. For the rest of this section, we will consider θ_1 and N_p as given and fixed.

The determination for the three remaining variables (L, θ_2, and δ) is more difficult. We first consider the optimum pixel size δ. For a given L, N is maximized with respect to δ when w is minimized with respect to δ. To find the optimum δ it is convenient to write w as

$$w = \frac{\lambda N_p}{n}\left(y^{3/2} + \frac{c^{3/2}}{\sqrt{y^3 - 1}}\right), \tag{4.12}$$

where

$$y = \left(\frac{n\delta}{\lambda}\right)^{2/3} \tag{4.13}$$

and

$$c = \left(\frac{nL}{\lambda N_p}\right)^{2/3}. \tag{4.14}$$

To minimize w with respect to δ, we differentiate w with respect to y (since y increases monotonically with δ) and set the derivative to zero. This yields the following equation for y

$$y^3 = cy + 1. \tag{4.15}$$

The solution to this cubic equation is

$$y = \left\{\frac{1}{2} + \sqrt{\frac{1}{4} - \frac{c^3}{27}}\right\}^{1/3} + \left\{\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{c^3}{27}}\right\}^{1/3} \tag{4.16}$$

which can be evaluated for a given L to yield the optimum δ. Once y is determined we can solve for the optimum pixel spacing, δ_o, from the following equation:

$$\delta_o = \frac{\lambda}{n}\, y^{3/2}. \tag{4.17}$$

It can be shown that δ_o increases as L increases.

The above solution, however, may not always be realizable. In any practical system, the minimum δ (denoted as δ_{min}) is limited by the imaging system to a value larger than the wavelength λ. If we use an imaging lens of $F/number\ z$, the smallest resolvable spot is

$$\delta_{min} = \lambda\sqrt{4z^2 + 1}, \tag{4.18}$$

which corresponds to the highest spatial frequency plane wave traveling at an angle

$$\theta_i = \sin^{-1}\left(\frac{\lambda}{\delta}\right). \tag{4.19}$$

Inside the recording material, this becomes (from Snell's law)

$$\begin{aligned}\theta_r &= \sin^{-1}\left(\frac{1}{n}\sin\theta_i\right) \\ &= \sin^{-1}\left(\frac{\lambda}{n\delta_{min}}\right).\end{aligned} \tag{4.20}$$

Therefore the smallest resolvable spot size inside the recording material is also δ_{min} as given by Eq. (4.18).

δ_{min} is the lower bound for the size of δ. If L is too small, δ_o (from Eq. (4.16) and (4.17)) becomes less than δ_{min}. In that case, we set $\delta = \delta_{min}$. Since δ_o increases as L increases, we can use Eqs. (4.13)–(4.17) to find the smallest L for which δ_o is larger than δ_{min}:

$$L_{min} = \frac{\lambda N_p}{n}\left(\frac{y_{min}^3 - 1}{y_{min}}\right)^{2/3}, \tag{4.21}$$

where

$$y_{min} = \left(\frac{n\delta_{min}}{\lambda}\right)^{2/3}. \tag{4.22}$$

We will refer to the condition where the optimum δ is less than δ_{min} as the "thin" disk regime. If, on the other hand, the optimum pixel spacing is larger than the resolution limit of the lens, this corresponds to the "thick" regime. For example, for $n = 2.2$, $\lambda = 500$ nm, $N_p = 1000$, and using an $F/3$ imaging lens (i.e., $z = 3$), we get $\delta_{min} = 3.04$ μm and $L_{min} = 40.36$ mm. Note that L_{min} does not depend on θ_2.

To summarize, if $L > L_{min}$ (thick disk), we use $\delta = \delta_o$ (from Eq. (4.15)), and if $L < L_{min}$ (thin disk), we use $\delta = \delta_{min}$ (as given by Eq. (4.18)).

4.1.4. Optimum Thickness L

Our problem is now reduced to maximizing N with respect to the two remaining variables L and θ_2. We first treat θ_2 as fixed, and find the optimum L that maximizes N/A.

In the range $L < L_{min}$ we use δ_{min} as the optimum δ, and write N/A from Eq. (4.11) as

$$N/A = \frac{1}{\delta_{min}^2}\frac{1 + \alpha x}{(1 + \beta x)(1 + \gamma x)} \tag{4.23}$$

where

$$x = \frac{nL}{\lambda N_p}, \tag{4.24}$$

$$\alpha = \frac{\pi N_p}{8}(\cos\theta_1 - \cos\theta_2), \tag{4.25}$$

$$\beta = \frac{1}{y_{min}^{3/2}\sqrt{y_{min}^3 - 1}}, \tag{4.26}$$

and

$$\gamma = \beta + \frac{\tan\theta_2}{y_{min}^3}. \tag{4.27}$$

We can solve for the optimum L by differentiating the expression in Eq. (4.23) with respect to x. The maximum N/A turns out to be

$$N/A = \frac{1}{\delta_{min}^2}\frac{\alpha/\beta\gamma}{\left(\sqrt{\frac{1}{\beta}-\frac{1}{\alpha}} - \sqrt{\frac{1}{\gamma}-\frac{1}{\alpha}}\right)^2}, \tag{4.28}$$

which occurs at

$$L = L_o = \frac{\lambda N_p}{n}\left(-\frac{1}{\alpha} + \sqrt{\left(\frac{1}{\beta}-\frac{1}{\alpha}\right)\left(\frac{1}{\gamma}-\frac{1}{\alpha}\right)}\right), \tag{4.29}$$

assuming of course that $L_o < L_{min}$. If $L_o > L_{min}$ this means that the optimum thickness is outside the thin regime where the analysis used to derive L_o is valid. Within the thin regime the maximum thickness L occurs at the boundary since N/A is monotonically increasing with L for $L < L_{min}$. To obtain the overall optimum thickness L we must compare the maximum obtained from this regime (i.e., $L \leq L_{min}$) with the optimum thickness obtained from the thick regime ($L > L_{min}$) and finally select the thickness that yields the larger density N/A.

As an example, we continue with the previous example where $\delta_{min} = 3.04\ \mu$m (for an $F/3$ lens). If we take $\theta_1 = 10°$ and $\theta_2 = 20°$, we find L_o to be 16.74 mm, which is less than $L_{min} = 40.36$ mm. Therefore, the solution obtained from the thin regime is the valid optimum thickness. Note that as N_p increases, so does α, and therefore the expression in Eq. (4.28) increases. For large N_p, the maximum N/A as given by Eq. (4.28) increases approximately linearly with N_p, or the square root of the total number of pixels N_p^2.

For $L > L_{min}$, we can use Eq. (4.17) for δ and using Eqs. (4.12)–(4.15), Eq. (4.11) can be written as

$$N/A = \left(\frac{n}{\lambda}\right)^2 \frac{1 + \alpha c^{3/2}}{\left(y^{3/2} + \frac{c^{3/2}}{\sqrt{y^3 - 1}}\right)\left(y^{3/2} + \frac{c^{3/2}}{\sqrt{y^3 - 1}} + c^{3/2}\tan\theta_2\right)}. \tag{4.30}$$

The above expression may be evaluated numerically to find the value of L which maximizes N/A. We can also derive a relatively simple asymptotic expression (for large L) by observing that as $L \to \infty$, $y \to \sqrt{c}$. The asymptotic expression for N/A is

$$N/A \to \left(\frac{n}{\lambda}\right)^2 \frac{\pi N_p(\cos\theta_1 - \cos\theta_2)}{16\tan\theta_2}\sqrt{\frac{\lambda N_p}{nL}}. \tag{4.31}$$

The above expression predicts that the density will decrease as the disk thickness becomes very large. This is confirmed by the numerical results we present in the following section.

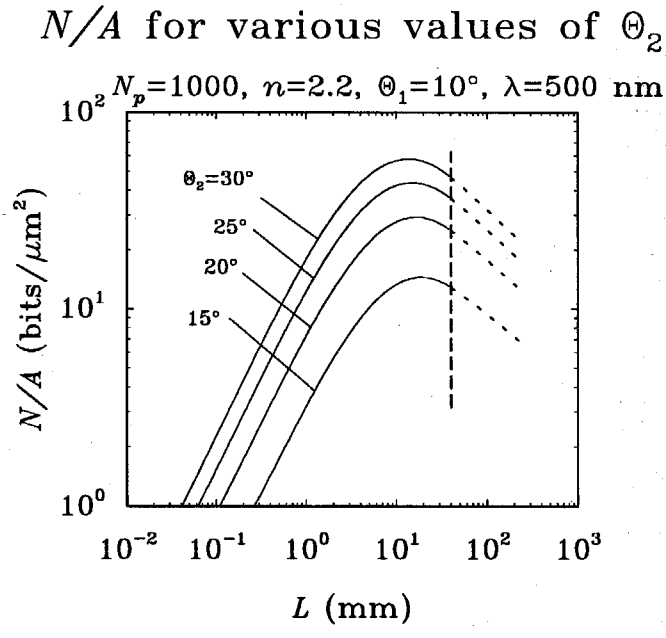


Figure 4.4. Angle multiplexing: N/A vs. L for various values of θ_2. We take $N_p = 1000$, $n = 2.2$, $\lambda = 500$ nm, and $\theta_1 = 10°$.

4.1.5. Optimum θ_2 and the Maximum Storage Density

The final step in the optimization of the storage density N/A, consists of optimally selecting θ_2. Since we cannot analytically derive the optimum angle, we resort to numerical methods. In Figure 4.4 we plot Eq. (4.23) in the thin regime (solid line) and Eq. (4.30) in the thick regime (dotted line) as a function of L for various values of θ_2 using the optimum value for δ. The vertical line indicates the transition from one regime to the other. The optimum values for L and θ are those that yield the maximum density. The parameters used in plotting Figure 4.4 are $\lambda = 500$ nm, $N_p = 1,000$, $n = 2.2$ (the index of refraction for $LiNbO_3$ crystals), and $\theta_1 = 10°$. $\theta_2 = 30°$ is the maximum value for which N/A is plotted since 27.04° is the largest angle that can be supported inside the recording material (due to Snell's law) without resorting to the use of index matching fluids.

From Figure 4.4, we see that the maximum N/A is obtained near $L = 1.5$ cm and increases monotonically with θ_2 for the parameters we selected. In this case, the optimum thickness is in the thin regime ($L_o < L_{min}$). Since it is not practical to use $\theta_2 = 30°$ inside the recording material (the critical angle is 27.04°), we get a realistic estimate for the achievable density by using $\theta_2 = 20°$. The corresponding angle swing outside the recording material is then 22.5° to 48.8° (total angular swing of 26.3°), which is practically achievable. The maximum density N/A is 29.3 bits/μm^2, which is obtained for a hologram thickness of $L = 16.74$ mm using $N_\theta = 1306$ angularly multiplexed holograms. This density can be increased by a factor of 4 (giving us $N/A = 117.2$ bits/μm^2) if we simultaneously record reflection and transmission holograms in the same reference angle range from both sides of the signal beam, as shown in Figure 3. The area for each recording location is $w \times w' = 4.3 \times 10.4\ mm^2$. Figure 4.5 is a plot of the optimum density and also the number of angularly multiplexed holograms, N_θ, as a function of L. For the thickness that yields maximum density, $N_\theta = 1,306$ holograms. Since more than 5,000 holograms have been recorded and faithfully reproduced in Lithium

Niobate [3], the geometric factors considered in this chapter limit the recording more severely than the material dynamic range. As another example, if we record only 100 holograms at each location, then the optimum thickness is a little over 1 mm and the corresponding storage density is about 8.8 bits/μm^2 (compared to about 30 bits/μm^2 for the optimum design). This density can be increased by a factor of 4 as we already described in Figure 4.2.

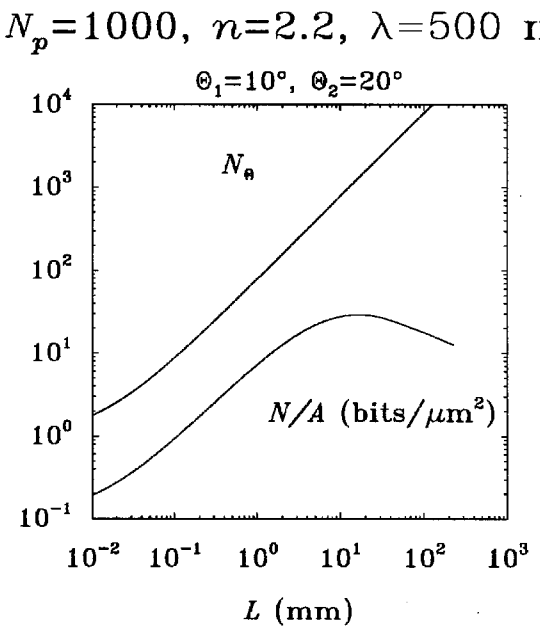


Figure 4.5. Angle multiplexing: optimum N/A (optimized with respect to δ) and N_p as functions of thickness L. We take $N_p = 1000$, $n = 2.2$, $\lambda = 500$ nm, $\theta_1 = 10°$, and $\theta_2 = 20°$.

4.2. Wavelength Multiplexing

Wavelength multiplexing [4,5] is an alternative method for multiplexing holograms in a single location on the HD. In this section we calculate the capacity of a wavelength multiplexed HD using a similar derivation as for angular multiplexing. The number of bits that can be stored is expressed as

$$N = N_S N_\lambda N_p^2, \tag{4.32}$$

where N_λ is the number of wavelength multiplexed holograms. We assume that the wavelength λ sweeps from λ_1 to λ_2, with $\lambda_1 < \lambda_2$.

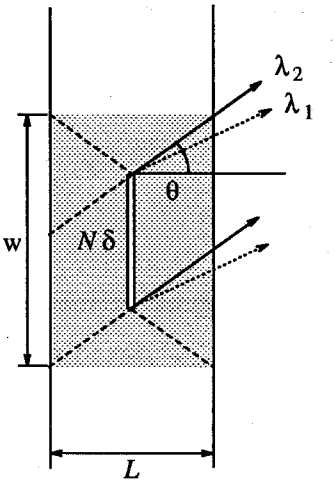


Figure 4.6. Wavelength multiplexing: extra area taken up because of defocusing and wavelength change.

For wavelength multiplexing, we again assume that the image is at normal incidence and focused at the middle of the recording material (Figure 4.6). We also assume that the reference beam is counter-propagating with the image beam and also at normal incidence. In this case, the problem of image defocusing at the crystal surface is the same, and we get Eq. (4.8) for the width w as before. However, since the reference beam is co-linear with the signal beam for all wavelengths, there is no extra width taken up by the $L \tan \theta_2$ term in the expression for w' in Eq. (4.9). On the other hand, as λ sweeps through λ_1 to λ_2, w changes. For any choice of δ, the largest w is for $\lambda = \lambda_2$. Therefore we have

$$N_S = \frac{A}{w^2}, \tag{4.33}$$

where

$$w = N_p \delta + \frac{L}{\sqrt{(n\delta/\lambda_2)^2 - 1}} \tag{4.34}$$

is a function of λ_2.

To find N_λ, we note that the half-width of the (frequency) selectivity $\Delta\nu$ is [1]

$$\Delta\nu = \frac{v_c}{nL}, \tag{4.35}$$

where v_c is the speed of light in vacuum. As λ sweeps over λ_1 to λ_2, the number of wavelength multiplexed holograms that can be stored is therefore

$$N_\lambda = 1 + \frac{\nu_1 - \nu_2}{2\Delta\nu} = 1 + \frac{nL}{2}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right), \tag{4.36}$$

where we take the separation between adjacent holograms to be the full width, $2\Delta\nu$. Using Eqs. (4.33), (4.34), and (4.36), we then have

$$N = AN_p^2 \frac{1 + \frac{nL}{2}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)}{\left[N_p\delta + \frac{L}{\sqrt{(n\delta/\lambda_2)^2 - 1}}\right]^2}. \tag{4.37}$$

4.2.1. Optimum N_p, λ_1, and δ

We now want to maximize N with respect to N_p, L, λ_1, λ_2, and δ. As before, N increases monotonically with N_p, which is limited by the SLM to about 1,000. N also increases as the minimum wavelength λ_1 decreases. This will be limited by the shortest usable wavelength we can get out of a tunable laser and/or the spectral sensitivity of the material. For the remainder of this section, we will assume that N_p and λ_1 are given and fixed.

The three remaining parameters λ_2, L, and δ are more complicated. We first take L and λ_2 as fixed, and find the optimum δ. Considering N/A as a function δ, we find as before that the maximum N/A is obtained when w is minimized with respect to δ. We then get the same set of equations as Eqs. (4.13)–(4.17), except with λ replaced by λ_2. We also have the same δ_{min} and L_{min} (with λ replaced by λ_2) conditions as given respectively by Eqs. (4.18) and (4.21). Note that both δ_{min} and L_{min} scale linearly with wavelength (since y_{min} depends only on n and z, the $F/number$ of the imaging lens). It should be emphasized, that

δ_{min} is the resolution of the imaging system using wavelength λ_2. The resolution of the system using λ_1 (which is less than λ_2) is of course better.

In summary, if $L > L_{min}$, we use $\delta = \delta_o$ (from Eq. (4.15)), otherwise we use $\delta = \delta_{min}$ (as defined in Eq. (4.18)); in these equations, λ is replaced by λ_2. As an example, for $\lambda_2 = 540$ nm and an $F/3$ imaging lens, we have $\delta_{min} = 3.28$ μm and $L_{min} = 43.59$ mm. For $\lambda_2 = 750$ nm, these become $\delta_{min} = 4.56$ μm and $L_{min} = 60.54$ mm.

4.2.2. Optimum L and λ_2

We now find the optimum thickness L that maximizes N/A. In the thin regime ($L < L_{min}$) we take $\delta = \delta_{min}$, and write N/A as

$$N/A = \frac{1}{\delta_{min}^2} \frac{1+\alpha x}{(1+\beta x)^2}, \tag{4.38}$$

where

$$x = \frac{nL}{\lambda_2 N_p}, \tag{4.39}$$

$$\alpha = \frac{N_p}{2}\left(\frac{\lambda_2}{\lambda_1} - 1\right), \tag{4.40}$$

and

$$\beta = \frac{1}{y_{min}^{3/2}\sqrt{y_{min}^3 - 1}}. \tag{4.41}$$

By differentiating the expression in Eq. (4.38) with respect to x, we find the maximum N/A to be

$$N/A = \frac{1}{\delta_{min}^2} \frac{\alpha^2}{4\beta(\alpha-\beta)}, \tag{4.42}$$

which occurs at

$$L = L_o = \frac{\lambda_2 N_p}{n}\left(-\frac{2}{\alpha} + \frac{1}{\beta}\right). \tag{4.43}$$

For example, for $\lambda_1 = 500$ nm, $\lambda_2 = 540$ nm ($\lambda_2/\lambda_1 = 1.08$), $n = 2.2$, and $N_p = 1000$, we get $L_o = 43.82$ mm. If λ_2 increases to 750 nm, L_o increase

to 60.88 mm. In both cases, L_o is larger than L_{min} (43.59 mm and 60.54 mm respectively). This means that there is no maximum in the thin regime and therefore in the range $L < L_{min}$, N/A is monotonically increasing with L. In this case, we would select the boundary value (L_{min}) for the best thickness obtainable from the thin regime. Notice that for wavelength multiplexed storage the optimum thickness of the disk can become quite large. Although we are not considering materials issues in this chapter, it should be pointed out that the useful thickness of the material in practice can be limited by absorption. In some materials (e.g., Lithium Niobate) it is possible to reduce the absorption by properly preparing the material (e.g., by adjusting the dopant and reduction/oxidation level). The reduced absorption will typically reduce the recording speed of the material for a given light intensity. Therefore, when materials considerations are included in the design process, this tradeoff between speed and density will emerge.

In the thick regime, $L > L_{min}$, we use Eqs. (4.15) and (4.17) (with λ replaced by λ_2) to obtain δ, and write N/A as

$$N/A = \left(\frac{n}{\lambda_2}\right)^2 \frac{1 + \alpha c^{3/2}}{\left(y^{3/2} + \frac{c^{3/2}}{\sqrt{y^3 - 1}}\right)^2}, \tag{4.44}$$

where α is given by Eq. (4.40). As before, as $L \to \infty$, $y \to \sqrt{c}$. For wavelength multiplexing, however, the asymptotic behavior of N/A is different. As $L \to \infty$, N/A saturates and approaches

$$N/A \to \frac{n^2 N_p}{8} \frac{1}{\lambda_2} \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right). \tag{4.45}$$

Thus for the range $L > L_{min}$, N/A also increases monotonically with L.

Note that the saturation value of N/A increases as N_p increases and λ_1 decreases. Also, for any choice of λ_2/λ_1, we have

$$\frac{1}{\lambda_2} \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \leq \frac{1}{4\lambda_1^2}, \tag{4.46}$$

with equality at

$$\frac{\lambda_2}{\lambda_1} = 2. \tag{4.47}$$

Thus, even if it is possible to have a light source with such a large range of wavelength tunability, the optimum setting for λ_2/λ_1 in order to obtain maximum saturation density is 2 (provided we use the same λ_1). For practical systems, λ_2/λ_1 is smaller than 2, and in this range the saturation value of N/A increases as λ_2/λ_1 increases. In the case where $\Delta\lambda = \lambda_2 - \lambda_1 \ll \lambda_1$ ($\lambda_2 \approx \lambda_1$), the saturation value given in Eq. (4.45) is approximately

$$N/A \approx \frac{n^2 N_p}{8} \frac{\Delta\lambda}{\lambda_1^3}, \tag{4.48}$$

which is proportional to $\Delta\lambda$.

In practice, the range of usable wavelengths is determined by the laser system. For instance, dye lasers can be tuned in the range from 370 nm to 890 nm, which gives us a λ_2/λ_1 of 2.40, in excess of the optimum $\lambda_2/\lambda_1 = 2$ requirement. It should be noted, however, that it is necessary to use several different dyes in order to obtain this range of wavelengths. For a typical broadband laser dye such as Coumarin 6, the range is from 510 nm to 550 nm, which only gives us a λ_2/λ_1 of 1.08. For Ti:Sapphire lasers, the range is 690 nm to 1025 nm, which gives us a λ_2/λ_1 of 1.48. [2]

As a specific example, consider the case where $N_p = 1000$, $n = 2.2$, and $\lambda_1 = 500$ nm. We plot N/A as a function of L (where N/A has been optimized with respect to δ) for various values of λ_2/λ_1. The result is shown in Figure 4.7. We see that N/A saturates for large L (around 5 cm) as expected, and the saturation value is largest for $\lambda_2/\lambda_1 = 2$. In Figure 4.8, we plot N/A and N_λ as functions of L for $\lambda_2/\lambda_1 = 1.08$ and $\lambda_2/\lambda_1 = 1.5$ using the same N_p, n, and λ_1. For $\lambda_2/\lambda_1 = 1.08$, N/A approaches 166.0 bits/μm^2, while for $\lambda_2/\lambda_1 = 1.5$, N/A approaches 537.8 bits/μm^2.

4.2.3. Storage Density and Optimum λ_2 for "Thin" Disks

The point where L causes N/A to reach saturation is of the order of 5 cm.

---

[2] Although this is the tunability of the laser in terms of the lasing media, it is necessary to change mirrors and output couplers to get this range.

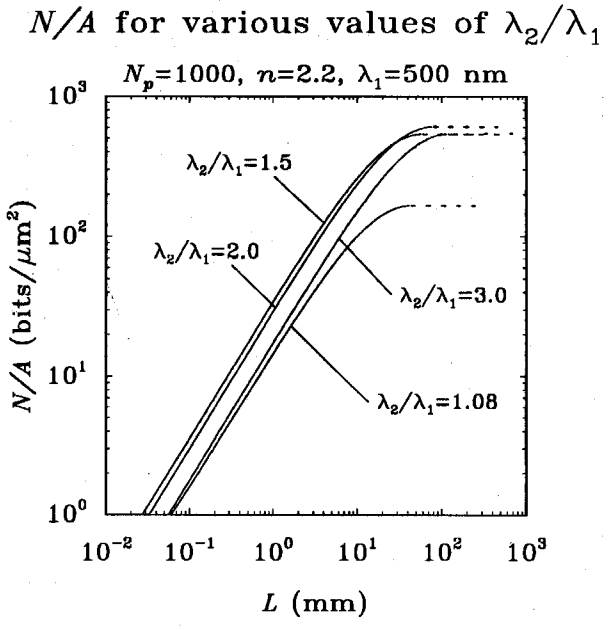


Figure 4.7. Wavelength multiplexing: optimum N/A (optimized with respect to δ) as a function of L for various values of λ_2/λ_1. We take $\lambda_1 = 500$ nm, $N_p = 1000$, and $n = 2.2$.

At this thickness, it becomes questionable about what we mean by a "disk." In practice it may be desirable or necessary (e.g., because of absorption) to keep the thickness small. In this case we are in the $L < L_{min}$ range (even though L_o may be larger than L_{min}), and N/A is given by Eq. (4.38). We can approximate Eq. (4.38) by

$$N/A \approx \frac{n}{2\delta_{min}^2}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right)L \propto \frac{1}{\lambda_2^2}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_2}\right), \tag{4.49}$$

if we assume that

$$\alpha x \gg 1 \gg \beta x. \tag{4.50}$$

In the previous example, $\alpha = 40$ for $\lambda_2/\lambda_1 = 1.08$, and $\alpha = 250$ $\lambda_2/\lambda_1 = 1.5$, while $\beta = 5.60 \times 10^{-3}$ for both cases, so the condition is satisfied.

If we limit x (and therefore the disk thickness L) to the range required by Eq. (4.50), the optimum λ_2 can be found by taking the derivative of the expression in Eq. (4.49) with respect to λ_2 and setting it to zero. In this case, it is easy to

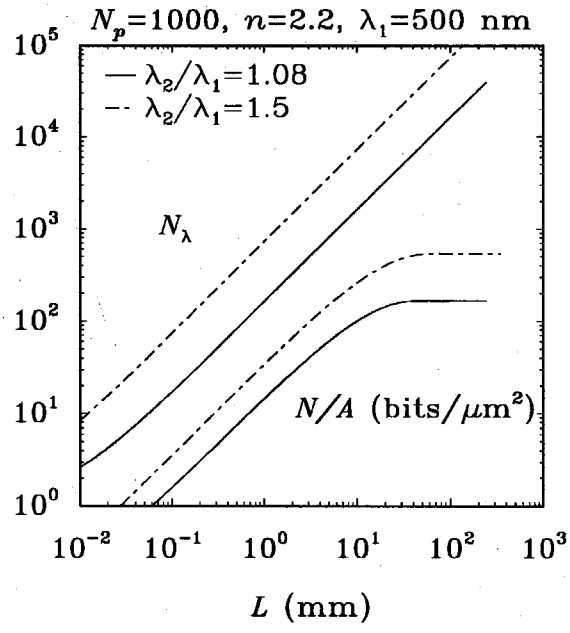


Figure 4.8. Wavelength multiplexing: optimum N/A and N_λ as functions of thickness L for $\lambda_2/\lambda_1 = 1.08$ and $\lambda_2/\lambda_1 = 1.5$. We take $\lambda_1 = 500$ nm, $N_p = 1000$, $n = 2.2$.

show that the maximum N/A occurs for $\lambda_2/\lambda_1 = 1.5$ (again assuming that we are using the same λ_1). This is very close to the range provided by Ti:Sapphire lasers. Therefore, in this case the density does not increase indefinitely with $\Delta\lambda$.

Finally we can also calculate the "knee" of the N/A curve, which we define as the point were the expression given by Eq. (4.48) reaches the saturation value. This is given by

$$L = L_K = \frac{nN_p\delta_{min}^2}{4\lambda_2} = \frac{4z^2+1}{4}\, n\lambda_2 N_p, \tag{4.51}$$

which is proportional to λ_2. For $\lambda_2/\lambda_1 = 1.08$, $L_K = 11.0$ mm, which gives us $N/A = 106.5$ bits/μm^2 and $N_\lambda = 1,794$. For $\lambda_2/\lambda_1 = 1.5$, $L_K = 15.3$ mm, which gives us $N/A = 34.46$ bits/μm^2 and $N_\lambda = 11,221$. In both cases, L_K is less than L_{min}, and the corresponding values of N/A is slightly over half the saturation values for N/A (i.e., approximately a 3 dB drop).

4.3. Design Considerations

Having derived the maximum storage densities for angle and wavelength multiplexing holograms, we now turn to various design considerations based on the results derived so far.

4.3.1. Angle vs. Wavelength Multiplexing

The values for the various parameters discussed so far in this chapter are summarized in Table 4.1, and the storage densities N/A are plotted in Figure 4.9 where we denote the densities of angle multiplexing and wavelength multiplexing by $(N/A)_\theta$ and $(N/A)_\lambda$, respectively.

Table 4.1. Value of Parameters used in Figure 4.9

(Assuming an $F/3$ imaging lens)

Parameters	**Angle multiplexing**	**Wavelength multiplexing**
Index of Refraction	$n = 2.2$	$n = 2.2$
Number of Pixels	$N_p^2 = 10^6$	$N_p^2 = 10^6$
Wavelength	$\lambda = 500$ nm	$\lambda_1 = 500$ nm, $\lambda_2 = 540$ nm
Angles	$\theta_1 = 10°$, $\theta_2 = 20°$,	—
Pixel Size	$\delta_{min} = 3.04$ μm	$\delta_{min} = 3.28$ μm
Critical Thickness	$L_{min} = 40.36$ mm	$L_{min} = 43.59$ mm
Optimum Thickness	$L_o = 16.74$ mm	$L_o \approx 30$ mm
Maximum Density	$N/A = 4 \times 29.3$ bits/μm^2	$N/A = 166.0$ bits/μm^2
Number of Holograms	$N_\theta = 4 \times 1306$	$N_\lambda \approx 5000$

In Figure 4.9, the curves for $(N/A)_\theta$ using just the angle range θ_1 to θ_2 (either as transmission or reflection holograms) is marked as ($\times 1$). We see that it is about a factor of 2 smaller than $(N/A)_\lambda$. However, if we angle-multiplex from

both sides of the signal beam, $(N/A)_\theta$ increases by a factor of 2 (denoted by the $\times 2$ curve in Figure 4.9). If we further record both reflection and transmission holograms (as in Figure 4.2), this increases by a factor of 4 (denoted by the $\times 4$ curve in Figure 4.9). In this case, $(N/A)_\theta$ becomes larger than $(N/A)_\lambda$ until L reaches about 12.5 mm, where both $(N/A)_\lambda$ and $(N/A)_\theta$ are about 115 bits/μm^2.

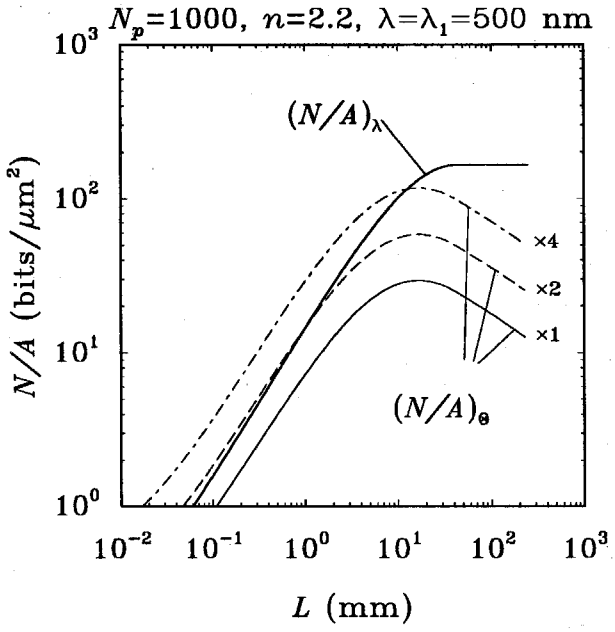


Figure 4.9. Comparison of angle multiplexing and wavelength multiplexing. We take $\lambda = \lambda_1 = 500$ nm, $\lambda_2/\lambda_1 = 1.08$, $n = 2.2$, $N_p = 1000$, $\theta_1 = 10°$, and $\theta_2 = 20°$. The density using angle multiplexing is denoted by $(N/A)_\theta$, and the density using wavelength multiplexing is denoted by $(N/A)_\lambda$.

4.3.2. Image Plane vs. Fourier Plane Holograms

One might ask whether it is possible to get higher density by recording in the Fourier plane instead of the image plane. It turns out that the storage density is the same. This is because the space-bandwidth-product is a constant. Specifically,

consider an image of extent $a = N_p\delta$, where δ is the pixel spacing and N_p is the number of pixels along one dimension. Let b be the extent of the Fourier transform of this image (by a lens of focal length F), and let $1/\delta'$ be the highest spatial frequency of the Fourier transform. Then within the paraxial approximations

$$N_p = \frac{a}{\delta} = \frac{b}{\delta'}. \tag{4.52}$$

This shows that recording in either the image plane or the Fourier plane will give the same minimum width.

If we record holograms at off-image or off-Fourier planes, the required width w increases. However, it is sometimes desirable to do this for purpose of noise, image quality, and alignment sensitivity. [3] The tradeoff between these requirements and storage density must be considered in the design of a practical system.

We next consider the problems involved in obtaining the small spatial extent necessary for maximum storage density. Image plane holograms require demagnifying the SLM in order to reach the optimum pixel size (typically, a factor of 20 to 30), which is at the diffraction limit. Note that we require not only high resolution, but also a fairly large area of view. Although possible to do, it requires very expensive optical systems. In addition, the space required for the demagnification system tends to be large. Fourier transform holograms, on the other hand, are relatively easy to shrink down in size. However, the space-band-width product requirement is the same, so the question of lens aberration still needs to be addressed. Also, as mentioned in Chapter 3, multiple Fourier transform holograms are also more difficult to record.

4.3.3. Storage Density vs. Alignment Sensitivity

In Chapter 3, it was shown that the condition for minimum rotational alignment sensitivity requires that the disk be placed some distance away from the

---

[3] Please see the discussion in Section 3.8.

image plane. This requirement is incompatible with the condition for maximizing the storage capacity, which requires the disk to be at the image plane or Fourier plane.

Although it turns out that the rotational alignment sensitivity of Fourier transform holograms is very close to the optimum configuration, [4] it is possible to simultaneously get both the minimum rotational alignment sensitivity and maximum storage density. Furthermore the configuration that we propose has several advantages over the usual image plane or Fourier plane recording configurations.

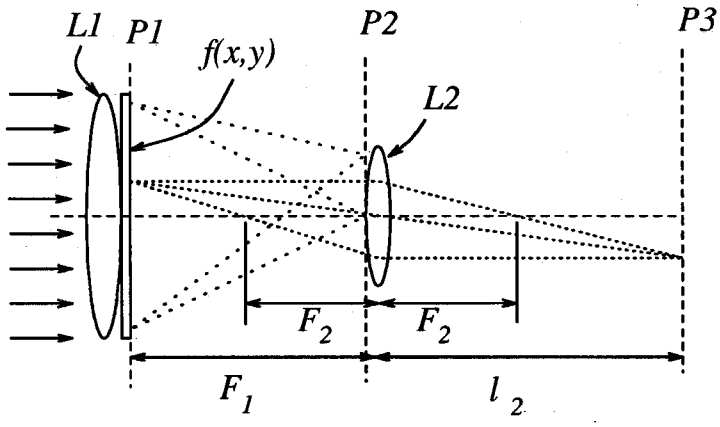


Figure 4.10. The Vander Lugt imaging system

The imaging system shown in Figure 4.10 was proposed by Vander Lugt for recording planar holograms [6]. The transparency or spatial light modulator (SLM) with transmittance $f(x, y)$ is placed at $P1$, immediately after the condenser lens $L1$. When illuminated with a plane wave, the condenser lens $L1$ provides a converging spherical wave that illuminates the SLM, and produces the Fourier transform of $f(x, y)$ (with an additional quadratic phase) at the back-focal plane ($P2$). The Fourier transform of $f(x, y)$ is then recorded as a hologram with a

[4] This will be discussed later in this section.

reference plane wave. A second lens $L2$ with focal length F_2 is placed immediately after the hologram at $P2$, and is used to form the (inverted) image of $f(x,y)$ at the image plane $P3$, where $P3$ is determined by the familiar imaging condition

$$\frac{1}{F_1} + \frac{1}{l_2} = \frac{1}{F_2}, \tag{4.53}$$

and F_1 is the focal length of the condenser lens $L1$.

On the other, point sources on the SLM at $P1$ are still converted to spherical wave impulse responses. Therefore it is possible to design the system according to Section 3.4.3 to set the center of rotation of the hologram at the center of the reconstructed image.

The minimum aperture required of $L2$ can be found as follows: let u_{max} be the highest spatial frequency of $f(x,y)$. Then the (spatial) extent of the Fourier transform at $L2$ (within paraxial approximations) is $a = 2u_{max}\lambda F_1$. Since all the information of the original image $f(x,y)$ is contained within this region, this is the minimum aperture required. If the pixel spacing of the SLM is δ, and N_p is the number of pixels along one dimension, then $u_{max} = 1/\delta$, and we can write a as

$$a = \frac{2\lambda F_1}{\delta} = 2\lambda N_p \left(\frac{F_1}{N_p \delta}\right) = \lambda N_p z, \tag{4.54}$$

where z is the F-number of the lens if the aperture of L_1 is the same as the size of the SLM, which is $N_p\delta$. Thus if the ratio $z = F_1/N_p\delta$ is kept fixed, then a is a constant even when the image size $N_p\delta$ changes (of course F_1 also needs to be changed).

It is interesting to compare this system (which we will refer to as the Vander Lugt system) with the conventional 4-F/Fourier transform system shown in Figure 4.11. If the focal length of the Fourier transform lens $L1'$ is also F_1, then the spatial extent of the Fourier transform is also $a = 2\lambda N_p(F_1/N_p\delta)$. But here the lens aperture of $L1'$ needs to be larger than $N_p\delta$ (the aperture of $L1$) to allow for all the spatial frequencies from $f(x,y)$ to pass through. Thus the F-number of $L1'$ is less than $L1$. In addition, the aberration corrections for $L1'$ are considerably more complicated than the condenser lens $L1$, which need only produce a

good converging spherical wave (a practical implementation would probably use an aspherical lens for $L1$). The requirements for $L2'$ are similar to $L1'$, and is more difficult than for $L2$ or $L1$. To increase the storage density, it is desirable to reduce a as much as possible. This requires that the F-number be as small as possible. Lenses with small (≈ 1) F-numbers are available but are expensive, especially if the aperture is large. Note that the space-bandwidth-product depends not only on the F-number, but also the aperture size.

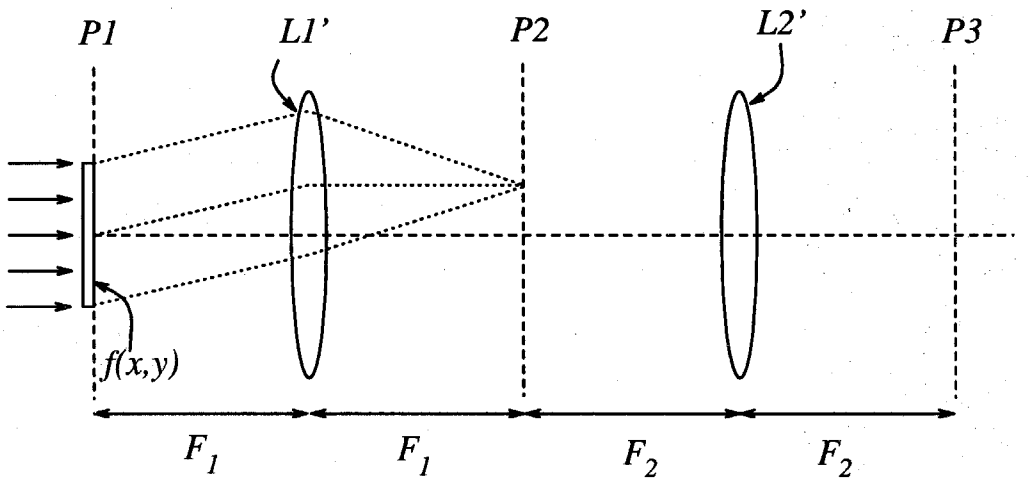


Figure 4.11. The conventional 4-F imaging system

In Vander Lugt's paper, the holograms were assumed to be planar. The system is of course also usable for volume holograms. Since the recorded pattern is essentially a Fourier transform hologram (except for the quadratic phase factor), the maximum storage density as analyzed in the previous sections is the same. However, since point sources at the input SLM give rise to spherical waves instead of plane waves (i.e., the impulse response of the system is a spherical wave), it is now also possible to adjust the reference beam angle to get minimum rotational alignment error without sacrificing the storage density.

As discussed above, since the space-bandwidth-product of the image and its Fourier transform is the same, the maximum achievable storage densities are

the same. To apply the formulas for Fourier transform holograms, we need only exchange the role of pixel spacing and spatial extent.

We will assume that the focal length of $L1$ (F_1) is fixed, and consider the pixel spacing δ as the variable. Given δ, the spatial extent is given by the expression in Eq. (4.54). The corresponding maximum spatial frequency of the Fourier transform is

$$u'_{max} = \frac{1}{\delta'} = \frac{N_p \delta}{\lambda F}, \tag{4.55}$$

where $\delta' = \lambda F / N_p \delta$ is the corresponding "pixel size" of the Fourier transform. If the F-number of $L1$ is sufficiently small, then we will always be in the "thick regime" discussed in Section 4.1.

We can now apply the results in Section 4.1 and 4.2 to find the conditions for maximum storage density by replacing all occurrences of δ by δ'.

The Vander Lugt system can be used so that the storage density in volume holograms is maximized, and the rotational alignment sensitivity is minimized. In addition the requirements on the lenses are modest, and the space needed (distance between the lenses, etc.) are small. In practice, however, there are some problems that need to be solved. First, as mentioned in Chapter 3, multiple Fourier transform holograms are in general more difficult to record because of the generally larger variation in intensity throughout the transform plane. Well designed phase diffusers can be used to compensate for this.

The second problem is the question of the reference beam. It is desirable to place the imaging lens $L2$ as close to the hologram as possible to reduce the requirement on the F-number (i.e., small aperture). If we record transmission type holograms, the reference beam will pass through the recording material and hit $L2$. Even with anti-reflection coating, the reflected/scattered light from $L2$ will add noise to the system. This can be a problem since the diffraction efficiencies of the holograms are low and the reference beam is much stronger than the reconstructed signal beam. If we record reflection type holograms, then the reference beam needs to pass through $L2$ to reach the recording material. This is not unsolvable, but the design of the optical system for reference beam steering (for

angle multiplexing) would be more complicated. For wavelength multiplexing, the situation is simpler, and one solution is to provide a point source at the focal plane of $L2$. In either cases, however, the lens aperture of $L2$ needs to be larger than the image beam in order for the reference beam to cover the entire signal beam inside the recording material. Because of these considerations, it is probably better to place $L2$ some distance away from the hologram, at the expense of a somewhat larger aperture.

The most serious problem with the Vander Lugt system, however, is the SLM. Most presently available (electrically or optically addressable) SLMs depend on liquid crystal (LC) technology. In the Vander Lugt system, the illumination is a spherical wave. This makes it very difficult to build the LC device with the appropriate LC layer spacing and polarization filters to get good contrast throughout the image. One solution is to place the SLM immediately *before* (instead of after) the condenser lens ($L1$), however this might require an additional lens with $L2$ to correct for aberrations. Although this makes the design of $L2$ more complicated, it might be worth the trouble if there is no simple solution to the SLM problem.

4.3.4. Alignment-Limited Maximum Readout Time

Having derived the configuration for minimum rotation alignment sensitivity (Chapter 3) and the configuration for maximum storage density (Chapter 4), we now address the problem of data transfer rate from a 3-D disk system designed for maximum storage density. We will assume that the disk is rotating at a constant speed. In this subsection, we first examine the limit in readout time due to misalignment from disk rotation. In the next subsection, we examine the limit due to detector noise.

Given the rotational speed of the disk, we want to calculate how long the reconstructed image will stay within half a pixel size. This will be the maximum amount of time we can use to read out the hologram if the disk is continuously

spinning. This time will of course depend on the way the holograms are stored (e.g., as Fourier transform holograms or image plane holograms). The comparison here is made under the assumption that in all cases (image plane, Fourier plane, or the optimum configuration shown in Chapter 3) the holograms are stored for maximum density.

For image plane holograms, the reconstructed image will rotate as the disk rotates. Let δ be the pixel spacing *at the image plane inside the crystal* and let R be the radius of rotation; i.e., the distance from the center of the hologram (in this case the image plane) to the disk rotation axis. If the disk rotates by an angle of $\Delta\phi$, then the image will rotate by the same angle and move a distance of $R\Delta\phi$. For this to be less than half the pixel-spacing, we require that

$$\Delta x = R\Delta\phi \leq \frac{1}{2}\delta. \tag{4.56}$$

If T is the period of rotation of the disk, then the time in which the reconstructed image remains within half a pixel spacing is

$$\tau = \frac{T}{2\pi}\Delta\phi \leq \frac{T}{2\pi}\cdot\frac{\delta}{2R} = \frac{T}{2\pi N_p}\cdot\left(\frac{N_p\delta}{2R}\right). \tag{4.57}$$

Note that τ is independent of image magnification (i.e., the actual size of the image at the readout plane). The value of the pixel-spacing δ is that which achieves maximum storage density, as explained earlier in this chapter.

For the optimum configuration (Chapter 3), the reconstructed image rotates around the center of the image at the same rate as the disk rotates. The center pixels of the image will therefore move very little, and the pixels that shifts the most will be the outermost pixels. Since the image size is $(N_p\delta)\times(N_p\delta)$, the most that the pixels will move when the disk rotates by $\Delta\phi$ is $\frac{1}{2}N_p\delta\ \Delta\phi$. For this to be less than $\delta/2$, we have

$$\Delta\phi \leq \frac{1}{N_p}, \tag{4.58}$$

and the maximum amount of time in which the hologram can be read out (when the disk is in continuous rotation) is

$$\tau = \frac{T}{2\pi N_p}. \tag{4.59}$$

It is interesting to note that this is independent of the pixel spacing δ, and in particular is true for the Vander Lugt recording configuration described in the previous subsection.

For Fourier plane holograms, we assume normal incidence for the signal beam. Let θ be the angle between the reference beam and the normal to the hologram surface. When the hologram rotates by $\Delta\phi$, the grating rotates also by the same amount. In this case the reconstructed wave vector tilts by an angle of $\sin\theta\,\Delta\phi$ and the reconstructed image of the pixel will therefore appear to move by

$$\Delta y = F\Delta\phi = F\sin\theta\,\Delta\phi, \tag{4.60}$$

where F is the focal length of the Fourier transform Lens for reconstructing the image from the hologram. On the other hand, the spatial extent of the Fourier transform hologram is $2\lambda F/\delta'$, where δ' is the pixel size of the reconstructed image. For optimum storage density, the spatial extent of the Fourier transform should be the same as $N_p\delta$, where δ is the (optimum) pixel size for obtaining maximum storage density in the case of image plane holograms; i.e.,

$$N_p\delta = 2\lambda F/\delta'. \tag{4.61}$$

The optimum pixel size for Fourier transform holograms is therefore

$$\delta' = \frac{2\lambda F}{N_p\delta}, \tag{4.62}$$

and for Δy to be less than have of this value, we get

$$\Delta\phi \leq \frac{\lambda}{N_p\delta\sin\theta}. \tag{4.63}$$

The maximum readout time is therefore

$$\tau = \frac{T}{2\pi N_p}\cdot\left(\frac{\lambda}{\delta\sin\theta}\right), \tag{4.64}$$

which is independent of the focal length F.

Note that although the displacements for the image plane hologram, Δx (Eq. (4.56)), and the displacements for the Fourier transform hologram, Δy

(Eq. (4.61)), are about the same (see also Eq. (3.78)), the pixels sizes which they are compared to (i.e., δ and δ') are very different. For Fourier transform holograms to have the same storage density as image plane holograms, the spatial extent of the Fourier transform hologram needs to be small. This implies that the pixel size of its reconstructed image is large. Thus although Δx and Δy are about the same, the pixel size of the image plane hologram δ is in general much smaller than δ'. Assuming the following parameters:

- Rotation speed: 3600 RPM, or $T = 1/60\ sec \approx 17\ msec.$
- Optimum pixel size: $\delta \approx 3\ \mu m.$ This is for the image plane hologram, assuming we use an $F/3$ lens. The corresponding optimum pixel size for the Fourier plane hologram is $\delta' = 33\ \mu m$, assuming $F = 10\ cm.$
- Radius of rotation: $R = 60\ mm.$
- Number of pixels per page: $N_p^2 = 1,000 \times 1,000.$
- Reference beam angle: $\theta = 15°$ (inside the holographic recording material).
- Wavelength: $\lambda = 500\ nm.$

The results are

- Image plane: 66 $nsec$
- Fourier plane: 1.7 μsec
- Optimum configuration: 2.7 μsec

4.3.5. Noise-Limited Minimum Readout Time

By noise-limited minimum readout time, we refer to the time it takes for the detector to accumulate enough photons from the light diffracted from the hologram, so that the detector signal rises significantly above noise level. As a typical example, we consider a CCD camera manufactured by DALSA. The spec-sheets quote a *noise equivalent exposure* of 45 pJ/cm^2 for a CCD that has a 16 μm pixel spacing. For a wavelength of $\lambda = 500\ nm$, this is approximately

4×10^{-19} J per photon. Thus each pixel requires about 26 photons to generate a signal equivalent to detector noise.

Let M as the number of required photons, η as the diffraction efficiency, N_p^2 as the number of pixels, and I_{inc} as the reference beam intensity. If the time it takes to accumulate M photons per pixel is τ, then

$$\frac{\eta I_{inc}}{N_p^2}\,\tau = \frac{Mhc}{\lambda}, \tag{4.65}$$

where h is Planck's constant and c is the speed of light. This gives us

$$\tau = \frac{MhcN_p^2}{\eta I_{inc}\lambda}, \tag{4.66}$$

and the maximum data transfer rate is

$$\frac{N_p^2}{\tau} = \frac{\eta I_{inc}\lambda}{Mhc}. \tag{4.67}$$

If we take

- Number of pixels per page: $N_p = 10^6$,
- Readout reference beam intensity: $I_{inc} = 10\ mW$,
- Minimum number of photons required: $M = 1,000$,
- Diffraction efficiency: $\eta = 10^{-7}$, [5]

then we get a transfer rate of 2.5 $Mbits/sec$ and a readout time of $\tau = 400\ msec$. If we are more optimistic and take $\eta = 10^{-6}$, we get a transfer rate of 25 $Mbits/sec$ and a readout time of $\tau = 40\ msec$.

4.4. Discussions and Conclusions

We have derived the optimum conditions for obtaining the maximum storage density of a 3-D HD disk using either angle multiplexing or wavelength multiplexing. Such optimally designed disks can store information with area densities

---

[5] If we take 1,000 holograms, then under symmetric read/write the individual hologram should have only 10^{-6} times the saturation diffraction efficiency [7].

more than 100 bits/μm^2 with disk thickness approximately 1 cm. However, the limits to storage density derived in this chapter are only due to the geometry of the system. The storage density can also be limited by noise (cross-talk, detector noise, media defects, etc.) and the limited dynamic range of the recording medium. These limits to N/A prove less restrictive than the geometric limits derived here. This is supported by recent experiments by Mok [8] where 1000 holograms were superimposed and reconstructed with extremely low probability of error in a lithium niobate crystal with 1 cm thickness. The parameters of this experiment were reasonably close to the optimum parameters given here.

We can increase the storage density further if we use lenses with smaller *F/numbers*. The practical limit is probably around 1, which is simple with the condenser lens. We had based most of the calculations in this chapter assuming an $F/3$ lens. If we had used an $F/1$ lens, then the density increases to 190 $bits/\mu m^2$: we assume as before

- Wavelength: $\lambda = 500\ nm$,
- Index of refraction: $n = 2.2$,
- Number of pixels: $N_p^2 = 1,000 \times 1,000$,
- Reference beam angle swing: 10° to 20° inside the crystal (this corresponds to 22.5° to 48.8° outside the crystal, a total angle swing of 26.3°).

For $F/1$ optics, the minimum resolvable spot size is $\delta_{min} = 1.12\ \mu m$ and $L_{min} = 5.16\ mm$. The optimum crystal thickness is $L_o = 3.21\ mm$, which gives us

- Optimum thickness: $L = 3.21\ mm$,
- Number of holograms: $N_\theta = 251$ (one "quadrant" only; with 4-side multiplexing, the number of holograms is 1004),
- Area per location: $a = 1.78 \times 2.95\ mm^2 = 5.27\ mm^2$.
- Storage density: $N/A = 47.68\ bits/\mu m^2$ (if we multiplex from four sides, as shown in Figure 4.2, then the density is 190.72 $bits/\mu m^2$).

The alignment-limited readout time (Section 4.4) for image plane holograms of course becomes shorter as we increase the storage density, since the pixel size δ decreases. However, it increases for Fourier plane hologram, while it does not change for the optimum configuration (which is of course always longer than that of the Fourier transform hologram).

From the discussions on the readout time in Section 4.3, the readout time limit due to rotation if we require fast, random access [6] is far too short for the detectors to pick up. One solution to this problem is to use pulse lasers, which have higher peak power than CW lasers. Taking $M = 1000$, $N_p^2 = 10^6$, a diffraction efficiency of 10^{-6}, and 4×10^{-19} J per photon, the total energy required to readout one page, which consists of 1 $Mbit$, is 0.4 mJ of laser light.

References

1. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell Syst. Tech. J.*, **48(9)**, 2909–2947 (1969).
2. C. Gu, J. Hong, I. McMichael, R. Saxena, and F. H. Mok, "Cross Talk Limited Storage Capacity of Volume Holographic Memory," *J. OSA. A*, **9**, 1–6 (1992).
3. F. H. Mok, "Angle-multiplexed Storage of 5,000 Holograms in Lithium Niobate," *Opt. Lett.*, **18**(11), 915–917 (1993).
4. F. T. S. Yu, S. D. Wu, A. W. Mayers, and S. M. Rajan, "Wavelength Multiplexed Reflection Matched Spatial Filters Using $LiNbO_3$," *Opt. Comm.*, **81**(6), 343–347 (1991).
5. G. A. Rakuljic, V. Leyva, and A. Yariv, "Optical-data Storage by using Orthogonal Wavelength-Multiplexed Volume Holograms," *Opt. Lett.*, **17**(20), 1471–1473 (1992).

---

[6] If we rotate the disk at 3,600 RPM, this corresponds to a rotation period of approximately 17 $msec$, and on the average we would expect the access time to be half of this, at approximately 8.5 $msec$.

6. A. Vander Lugt, "Packing Density in Holographic Systems", *Appl. Opt.*, **14**(5), 1081–1087, 1975.

7. D. Brady, K. Hsu, and D. Psaltis, "Periodically Refreshed Multiply Exposed Photorefractive Holograms," *Opt. Lett.*, **15**(14), 817–819 (1990).

8. F. H. Mok, "Applications of Holographic Storage in Lithium Niobate," (presented at OSA 1992 Annual Meeting) *OSA 1992 Annual Meeting Technical Digest*, Vol. 23, **WE1**, 102, Sept. 1992.

9. D. Psaltis,"Parallel optical memories", *Byte*, **17**(9), 179–182 (1992).

10. H.-Y. Li and D. Psaltis, "3-D holographic disks," submitted to *Appl. Opt.*

11. F. H. Mok, D. Psaltis, and G. Burr, "Spatial and Angle Multiplexed Holographic Random Access Memory," SPIE vol. 1773c, San Diego, CA, July 1992.

Chapter 5

Crystal Orientation and Diffraction Efficiency of Photorefractive Crystals

In the two previous chapters, we discussed what might be called the *geometrical* aspects of 3-D holographic disk systems. We have ignored for the most part the question of dynamic range and diffraction efficiency. The diffraction efficiency obtainable from photorefractive crystals depends on the orientation of the crystal and the polarization of the incident light. This problem is of interest in general to holographic recording using photorefractive crystals, since it is always desirable to maximize the diffraction efficiency. It is particularly relevant to 3-D disks, since the orientation of the crystal changes as the disk rotates.

In this chapter, we analyze the diffraction efficiency of photorefractive crystals as a function of crystal orientation. We begin with a review of the Kukhtarev band-transport model [1] with special attention to the anisotropic aspect of photorefractive crystal.

Throughout this chapter we will use the usual coordinate system that diagonalizes the electrooptic tensor. Thus the z-axis will also be the crystal c-axis.

5.1. The Photorefractive Effect in Anisotropic Crystals

The saturation space-charge field in a photorefractive crystal in the absence of the photovoltaic effect [2] and external applied field is given by the well known formula [1,3,4]

$$E_{sc} = m \frac{\frac{k_B T}{e} K}{1 + \frac{\epsilon k_B T}{e^2 N_A} K^2} = m E_{sat}, \tag{5.1}$$

where m is the modulation depth, k_B is Boltzmann's constant, T is the temperature in Kelvins, N_A is the density of acceptor sites in the material, K is the magnitude of the grating vector, and ϵ is the permittivity. In the case of anisotropic crystals, however, ϵ is a tensor, and it is not immediately clear what value to use for ϵ. It turns out that the correct value is *not* given by either Eq. (2.128) or (2.129), but instead is given by

$$\epsilon = \epsilon_x \sin^2 \phi + \epsilon_z \cos^2 \phi, \tag{5.2}$$

where ϕ is the angle between the grating vector $\mathbf{K}$ and the c-axis. It should be noted that the permittivity we use for the photorefractive effect is the *low frequency* or *DC* permittivity, whereas the permittivity used for coupled-mode analysis are in the optical frequencies. For crystals such as lithium niobate, the difference between ϵ_x and ϵ_z is significant, differing in some cases by more than a factor of 2. For example, for lithium niobate, $\epsilon_z = 78$ and $\epsilon_x = 32$. The different formulas used to calculated ϵ can give very different answers.

In this section, an outline will be given for the derivation of the space-charge field based on the Kukhtarev band-transport model for anisotropic crystals. We will assume that there is only one species of charge carriers (the electron) and that there is only one species of trap sites. We also assume that the mobility μ of the crystal is a scalar, [1] but assume that the permittivity ϵ is a tensor. We will also ignore the effects of beam-coupling, and assume that the light intensity distribution is not affected by the grating.

According to the Kukhtarev model, the following equations describe the dynamics of the photorefractive effect when electrons are the only species of carriers

$$\frac{\partial N_D^+}{\partial t} = sI(N_D - N_D^+) - \gamma_D n_e N_D^+ \tag{5.3}$$

$$\mathbf{J} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e + p(N_D - N_D^+) I \mathbf{c} \tag{5.4}$$

---

[1] Not to be confused with the permeability. The assumption that μ is a scalar is not really true. However to simplify the analysis, we will assume in the treatment here that μ is a scalar.

$$\nabla \cdot \mathbf{J} = -e \frac{\partial}{\partial t} \left(N_D^+ - N_A - n_e \right) \tag{5.5}$$

$$\nabla \cdot (\epsilon \mathbf{E}) = e \left(N_D^+ - N_A - n_e \right) \tag{5.6}$$

where N_D is the density of trap sites (donor ions), N_D^+ is the density of the vacant trap sites, N_A is the density of the compensating ions (to maintain charge neutrality), I is the light intensity, s is the photon-absorption cross-section, γ_D is the recombination constant, $\mathbf{J}$ is the current density, μ is the electron mobility, e is the electron charge, n_e is the free electron density, k_B is Boltzmann's constant, T is the temperature in Kelvin, p is the photovoltaic constant, $\mathbf{c}$ is the unit vector in the direction of the photovoltaic field, and ϵ is the permittivity tensor.

We assume that the intensity grating is of the form

$$I = I_0 + I_1 e^{-i\mathbf{K}\cdot\mathbf{x}} + c.c., \tag{5.7}$$

where $\mathbf{K}$ is the grating vector. As mentioned earlier, we assume that beam-coupling effects are negligible, and take I as a constant of time.

Let $\mathbf{u}$ be the direction of the grating vector $\mathbf{K}$, and $\mathbf{c}$ be the direction of the photovoltaic field. We assume that

$$N_D^+ = D_0 + D_1 e^{-i\mathbf{K}\cdot\mathbf{x}} + c.c. \tag{5.8}$$

$$n_e = n_{e0} + n_1 e^{-i\mathbf{K}\cdot\mathbf{x}} + c.c. \tag{5.9}$$

$$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 e^{-i\mathbf{K}\cdot\mathbf{x}} + c.c. \tag{5.10}$$

Here $\mathbf{E}$ is not necessarily parallel to the grating vector $\mathbf{K}$. We also assume that

$$N_D \gg N_D^+ \gg n_{e0}, \tag{5.11}$$

and that the time response of n (the free electron density) is instantaneous compared to that of N_D ($dn_1/dt \ll dD_1/dt$).

Substituting Eqs. (5.7)–(5.10) into Eqs. (5.3)–(5.6), and collecting the zeroth and first-order terms, we get from Eq. (5.6)

$$D_0 \approx N_A = const. \tag{5.12}$$

Substituting this result into Eq. (5.3), we get

$$n_{e0} \approx \frac{sI_0(N_D - N_A)}{\gamma_D N_A} \approx \frac{sI_0 N_D}{\gamma_D N_A}, \tag{5.13}$$

which is a constant.

Now write the electric fields as

$$\mathbf{E}_1 = E_{1u}\mathbf{u} + E_{1v}\mathbf{v}, \tag{5.14}$$

where $\mathbf{v}$ is an unit vector perpendicular to $\mathbf{u}$. From Eq. (5.6), the first-order terms give us

$$-i\mathbf{K}\cdot\epsilon\mathbf{E}_1 = e(D_1 - n_1) \approx eD_1, \tag{5.15}$$

where it is assumed that $D_1 \gg n_1$. Using the fact that $\mathbf{K} = K\mathbf{u}$, the above equation gives us

$$\frac{ie}{K}D_1 = \epsilon' E_{1u} + \epsilon'' E_{1v}, \tag{5.16}$$

or

$$E_{1u} = \frac{ie}{\epsilon' K}D_1 - \frac{\epsilon''}{\epsilon'}E_{1v}, \tag{5.17}$$

where

$$\epsilon' = \mathbf{u}\cdot\epsilon\mathbf{u}, \tag{5.18}$$

and

$$\epsilon'' = \mathbf{u}\cdot\epsilon\mathbf{v}. \tag{5.19}$$

Note that the expression for ϵ' in Eq. (5.18) is exactly the value given by Eq. (5.2).

Next we substitute Eq. (5.4) into (5.5). we get

$$\begin{aligned}\nabla\cdot\mathbf{J} &= \mu e(n_e\nabla\cdot\mathbf{E} + \nabla n_e\cdot\mathbf{E}) + k_B T\mu\nabla^2 n_e + \alpha\nabla I\cdot\mathbf{c} \\ &= -e\frac{\partial}{\partial t}(N_D^+ - N_A - n_e).\end{aligned} \tag{5.20}$$

The first-order terms give us (assuming $dD_1/dt \gg dn_1/dt$)

$$\begin{aligned}\frac{1}{\mu K}\frac{d}{dt}(D_1 - n_1) \approx& \frac{1}{\mu K}\frac{dD_1}{dt} \\ =& \left[(E_d + iE_{0u})n_1 - \left(\frac{en_{e0}}{\epsilon' K} + i\frac{pI_0}{\mu e}(\mathbf{u}\cdot\mathbf{c})\right)D_1\right] \\ &+ i\frac{p(N_D - N_A)I_1}{\mu e}(\mathbf{u}\cdot\mathbf{c}) - i\left(\frac{\epsilon''}{\epsilon'}\right)n_{e0}E_{1v},\end{aligned} \tag{5.21}$$

where

$$E_d = \frac{k_B T K}{e}, \tag{5.22}$$

and

$$E_{0u} = \mathbf{E}_0 \cdot \mathbf{u}. \tag{5.23}$$

On the other hand, the first-order terms in Eq. (5.3) give us (by using Eqs. (5.11) and (5.13))

$$\begin{aligned}\frac{dD_1}{dt} &= -(sI_0 + \gamma_D n_{e0})D_1 + s(N_D - D_0)I_1 - \gamma_D D_0 n_1 \\ &\approx -n_{e0}\gamma_D \left(\frac{N_D}{N_D - N_A}\right) D_1 + s(N_D - N_A)I_1 - \gamma_D N_A n_1 \\ &= -\frac{1}{t_0} D_1 + s(N_D - N_A)I_1 - \gamma_D N_A n_1,\end{aligned} \tag{5.24}$$

where

$$t_0 = \frac{N_A}{sI_0 N_D} = \frac{1}{\gamma_D n_{e0}} \cdot \frac{N_D - N_A}{N_D}. \tag{5.25}$$

Eliminating n_1 from Eqs. (5.21) and (5.24), we get

$$\begin{aligned}(E_\mu + E_d + iE_{0u})\frac{dD_1}{dt} &= -\frac{1}{t_0}\left[E_N + E_d + i\left(E_{0u} + \frac{N_A}{N_D}E_{ph,u}\right)\right] D_1 \\ &+ [E_d + i(E_{0u} + E_{ph,u})]\left(\frac{N_D - N_A}{N_D}\right)\frac{mN_A}{t_0} \\ &- i\left(\frac{\epsilon''}{\epsilon'}\right)\frac{N_A}{t_0}\left(\frac{N_D - N_A}{N_D}\right) E_{1v},\end{aligned} \tag{5.26}$$

where

$$E_\mu = \frac{\gamma_D N_A}{\mu K}, \tag{5.27}$$

$$E_N = \frac{eN_A}{\epsilon' K}\left(\frac{N_D - N_A}{N_D}\right), \tag{5.28}$$

$$\mathbf{E}_{ph} = \frac{pI_0(N_D - N_A)}{\mu e n_{e0}}\mathbf{c}, \tag{5.29}$$

$$E_{ph,u} = \mathbf{E}_{ph} \cdot \mathbf{u}, \tag{5.30}$$

and

$$m = \frac{I_1}{I_0}. \tag{5.31}$$

Here, m is the modulation depth and $\mathbf{E}_{ph}$ is the photovoltaic field. We can write Eq. (5.26) as

$$\tau \frac{dD_1}{dt} = -D_1 + mfN_A - igN_A, \tag{5.32}$$

where we define

$$\tau = t_0 \frac{E_\mu + E_d + iE_{0u}}{E_N + E_d + i\left(E_{0u} + \frac{N_A}{N_D}E_{ph,u}\right)}, \tag{5.33}$$

$$f = \left(\frac{N_D - N_A}{N_D}\right) \frac{E_d + i(E_{0u} + E_{ph,u})}{E_N + E_d + i\left(E_{0u} + \frac{N_A}{N_D}E_{ph,u}\right)}, \tag{5.34}$$

and

$$g = \left(\frac{\epsilon''}{\epsilon'}\right)\left(\frac{N_D - N_A}{N_D}\right) \frac{E_{1v}}{E_N + E_d + i\left(E_{0u} + \frac{N_A}{N_D}E_{ph,u}\right)}. \tag{5.35}$$

This gives the dynamics of the space-charge field. Assuming that E_{1v} is known, we can solve for D_1, and then find E_{1u} using Eq. (5.17). Note that since $N_A/N_D \ll 1$ from Eq. (5.11) (typically 10^{-3}), the photovoltaic term $(N_A/N_D)E_{ph,u}$ in the denominators is negligible, and the factor $(N_D - N_A)/N_D$ is approximately 1. In particular the time constant τ is practically independent of the photovoltaic field. (It does, however, effect f, and hence the saturation field.) Intuitively, the presence of the photovoltaic current is equivalent to an applied field, even if the crystal is "short circuited" and there is no voltage drop across the crystal.

In practice, the crystal is finite, and we need to consider the boundary conditions. As the space-charge field develops inside the crystal, the electrons also start to migrate towards the surface of the crystal or towards the edge of illumination, and get trapped at the boundaries. The charge built-up creates an additional electric field. The case where the c-axis is parallel to the crystal surface and the grating vector, and where the charges accumulate at the boundaries of illumination has been treated in reference [4].

From Eq. (5.32), assuming that E_{1v} is a constant, the solution to D_1 is

$$D_1 = (mf - ig)N_A\left(1 - e^{-t/\tau}\right), \tag{5.36}$$

which gives us

$$E_{1u} = E_N(g + imf)\left(1 - e^{-t/\tau}\right) - \left(\frac{\epsilon''}{\epsilon'}\right) E_{1v}. \tag{5.37}$$

The value of E_{1v} should be determined by the boundary conditions and how the charges accumulate at the boundaries, etc. If we assume the "short circuit" condition in the direction perpendicular to the grating, then E_{1v} is zero, and g is zero. Eq. (5.37) then becomes (since $(N_D - N_A)/N_D \approx 1$)

$$E_{1u} = imfE_N\left(1 - e^{-t/\tau}\right), \tag{5.38}$$

and the space-charge field is parallel to the grating vector $\mathbf{K}$. When there is no applied field E_{0u} and no photovoltaic field $E_{ph,u}$ in the $\mathbf{K}$ direction, the expression in Eq. (5.38) becomes Eq. (5.1) as $t \to \infty$.

Typically, E_N is much larger than E_d and E_{0u}, and E_{1u} is approximately

$$E_{1u} = -m\left[(E_{0u} + E_{ph,u}) - iE_d\right] \cdot \left(1 - e^{-t/\tau}\right). \tag{5.39}$$

5.2. Diffraction from Photorefractive Crystals

In this section, we will apply the results from the previous section and the results from Chapter 2 to analyze the diffraction efficiency from photorefractive crystals. For simplicity, we will assume that there is no photovoltaic effect and no externally applied field. This approximation is good for crystals such as SBN and $BaTiO_3$. For lithium niobate, however, the photovoltaic effect is quite strong, and strictly speaking should not be neglected. As mentioned in the previous section, the effect is similar to having an applied field. In the analysis given here, we will ignore the photovoltaic effect and later discuss how it effects the diffraction efficiency. We will also assume that $E_{1v} = 0$. With these approximations and assumptions in mind, we now proceed to calculate the diffraction efficiency from photorefractive crystals.

From Chapter 2, the diffraction efficiency of a weak transmission- or reflection-type volume hologram is given approximately by

$$\eta = \frac{n_1}{n_2}\left|\frac{E_1}{E_{20}}\right|^2 \approx \frac{n_1}{n_2} \times \frac{(\omega^2\mu)^2(\mathbf{e}_2 \cdot \Delta\epsilon\, \mathbf{e}_1)^2}{4k_1^2 \cos^2\theta_1'} L^2, \tag{5.40}$$

where k_1 is the magnitude of the wave vector $\mathbf{k}_1$, $\Delta\epsilon$ is the change in the permittivity tensor, and L is the thickness of the crystal. θ_1' is the angle between $\mathbf{k}_1$ and the normal of the crystal (the normal is not necessarily the z-axis, which in this chapter is set to be the same as the c-axis). $\mathbf{e}_1$ and $\mathbf{e}_2$ are the unit vectors in the direction of the electric fields of the optical waves. As mentioned before, in general $\mathbf{e}_i$ is not perpendicular to $\mathbf{k}_i$. It is assumed that both the incident wave and the diffracted wave are eigenmodes; i.e., they are either ordinary or extraordinary waves.

For anisotropic crystals, the change in the permittivity tensor ϵ in the presence of an electric field is given by [2]

$$\begin{aligned} \Delta\epsilon &= -\epsilon_0[n^2(r\mathbf{E}^{sc})n^2] \\ &= -\frac{1}{\epsilon_0}[\epsilon(r\mathbf{E}^{sc})\epsilon], \end{aligned} \tag{5.41}$$

where ϵ_0 is the vacuum permittivity, r is the electrooptic coefficient tensor, $\mathbf{E}^{sc}$ is the space-charge field, and n is the index of refraction tensor. Note that both n and ϵ are symmetric tensors. We can now write $\mathbf{e}_2 \cdot \Delta\epsilon\, \mathbf{e}_2$ as

$$\mathbf{e}_2 \cdot \Delta\epsilon\, \mathbf{e}_2 = -\frac{1}{\epsilon_0}(\epsilon\mathbf{e}_2) \cdot (r\mathbf{E}^{sc})(\epsilon\mathbf{e}_2). \tag{5.42}$$

Since $\mathbf{e}_i$ is the direction of the electric field, $\epsilon\mathbf{e}_i$ is in the direction of the displacement vector $\mathbf{D}_i$, and is therefore perpendicular to the wave vector $\mathbf{k}_i$. Let $\mathbf{d}_i$ be

---

[2] We have $\Delta(1/n^2) = r\mathbf{E}^{sc}$ (where $1/n^2$ is a tensor) and $\epsilon(1/n^2) = I$ (the identity matrix). Therefore $(\Delta\epsilon\,)(1/n^2) + \epsilon\Delta(1/n^2) = 0$, and

$$\Delta\epsilon = -\epsilon\Delta(1/n^2)n^2 = -\epsilon_0[n^2\Delta(1/n^2)n^2],$$

since $\epsilon = \epsilon_0 n^2$.

the unit vector in the direction of $\mathbf{D}_i$, which is parallel to $\epsilon\mathbf{e}_i$. We then have

$$\epsilon\mathbf{e}_i = \epsilon_0 n_i^2 \mathbf{d}_i, \tag{5.43}$$

where (from Eq. (2.130)) n_i satisfies

$$n_i^{-2} = \sqrt{\mathbf{d}_i \cdot n^{-4}\mathbf{d}_i}. \tag{5.44}$$

On the other hand, we have

$$\frac{\omega^2\mu}{k_i} = \omega\sqrt{\frac{\mu}{\epsilon_0}}\,\frac{1}{n_i'}, \tag{5.45}$$

where (from Eq. (2.131)) n_i' satisfies

$$n_i'^{\,-1} = \sqrt{\mathbf{d}_i \cdot n^{-2}\mathbf{d}_i}. \tag{5.46}$$

As mentioned in Chapter 2, for ordinary waves $n_i = n_i' = n_o$, where n_o is the ordinary index of refraction. For extraordinary waves, if the difference between the ordinary index n_o and extraordinary index n_e is small, then both n_i and n_i' are approximately (from Eq. (2.132))

$$n_i \approx n_i' \approx n_o(1 + 2n_o\Delta n\,\sin\phi_i), \tag{5.47}$$

where $\Delta n = n_e - n_o$ and ϕ_i is the angle between $\mathbf{k}_i$ and the c-axis. We will therefore approximate n_i' by n_i for the remainder of this chapter.

With these results, we can now write the diffraction efficiency in Eq. (5.40) as

$$\eta = k_0^2 n_1^3 n_2^3 \left[\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1\right]^2 E_{sc}^2 \cdot \frac{L^2}{4\cos^2\theta_1'}, \tag{5.48}$$

where $k_0 = \omega\sqrt{\mu\epsilon_0}$, $\mathbf{u}$ is the unit vector in the direction of $\mathbf{E}^{sc}$, and E_{sc} is the magnitude of $\mathbf{E}^{sc}$.

If we assume that $E_{1v} = 0$ in Eq. (5.37), then the space-charge field is parallel to the grating vector $\mathbf{K}$, and the magnitude of the space-charge field is given by

$$E_{sc} = mE_{sat}\left(1 - e^{-t/\tau}\right) = \frac{\sqrt{I_1 I_2}}{I_1 + I_2}\,(\mathbf{e}_1 \cdot \mathbf{e}_2)\,E_{sat}\left(1 - e^{-t/\tau}\right), \tag{5.49}$$

where m is the modulation depth, I_i is the intensity of the $\mathbf{k}_i$ beam, τ is the time constant (which may be complex), and E_{sat} is the saturation space-charge-field. It is assumed here that the beams used for recording and the beams for reading out the hologram have the same polarizations. Although it is possible to record with beams that have polarizations different from that of the reading beams, this will create problems for simultaneously Bragg matching all the spatial frequencies. Similarly, we assume that only eigenmodes are used for recording and reading, so that we do not have the complication of "double gratings" [5]. This will be discussed in more detail in Section 5.4.

In the absence of applied fields and photovoltaic effect, the saturation space-charge-field E_{sat} is given by Eq. (5.1). From Eq. (5.2), the permittivity, ϵ', is given by

$$\epsilon' = \epsilon_0[\epsilon_x(u_x^2 + u_y^2) + \epsilon_z u_z^2]. \tag{5.50}$$

Here, ϵ_0 is the vacuum permittivity, and ϵ_x and ϵ_z are the dielectric constants perpendicular and parallel to the c-axis (z-axis). Note that these are the *DC* or *low frequency* dielectric constants, and not of optical frequencies.

Putting these results together, Eq. (5.48) becomes

$$\eta = \left\{k_0 n_1^{3/2} n_2^{3/2} [\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1] \cdot E_{sat}(\mathbf{e}_1 \cdot \mathbf{e}_2)\right\}^2 \cdot \frac{I_1 I_2}{(I_1 + I_2)^2} \cdot \frac{L^2}{4\cos^2\theta_1'} \left|1 - e^{-t/\tau}\right|^2, \tag{5.51}$$

where n_1 and n_2 are given by Eq. (5.44).

Our goal is to maximize the diffraction efficiency η. The diffraction efficiency of course increases as we increase the crystal thickness L and the modulation depth $m = I_1 I_2/(I_1 + I_2)$. These are independent of crystal orientation, and therefore will not be considered in the subsequent discussion. θ_1' is the angle between the $\mathbf{k}_1$ wave vector and the normal of the crystal, and is also independent of crystal orientation. In the following discussions, we will therefore also neglect the factor $L/\cos\theta_1'$. Note that for the same crystal thickness L, if the diffracted signal beam is at a larger angle θ_1', then the interaction length *along the direction of propagation* is $L/\cos\theta_1'$, and not L.

Another consideration is the surface area occupied by the signal beam, which is proportional to $1/\cos\theta_1'$. The storage density (in bits per unit surface area) is inversely proportional to the area occupied by the hologram, and is therefore proportional to $\cos\theta_1'$. On the other hand, from Eq. (5.51), diffraction efficiency is proportional to $(1/\cos\theta_1')^2$. The number of holograms that can be stored, however, is proportional to the *square root* of the diffraction efficiency, [3] and therefore to $1/\cos\theta_1'$. The effects therefore tend to cancel out.

With these considerations in mind, we will therefore concentrate on the quantity

$$G = k_0 n_1^{3/2} n_2^{3/2} (\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1) E_{sat} (\mathbf{e}_1 \cdot \mathbf{e}_2), \tag{5.52}$$

which is proportional to the square root of the diffraction efficiency, η. Our goal will be to maximize $|G|$ by choosing an optimal crystal orientation. We assume that the incident and diffracted waves are both eigenmodes (e- or o-modes). Note that the unit of G is $1/m$.

To calculate the value of G, we first write the *directions* of the wave vectors, $\mathbf{k}_1$ and $\mathbf{k}_2$, in spherical coordinates:

$$\frac{\mathbf{k}_i}{k_i} = \left(\sin\phi_i \cos\theta_i,\ \sin\phi_i \sin\theta_i,\ \cos\phi_i\right), \tag{5.53}$$

where ϕ_i is the angle between $\mathbf{k}_i$ and the z-axis. k_i is the magnitude of the wave vector, $\mathbf{k}_i$, along the direction of the unit vector on the right-hand side of Eq. (5.53).

To find the polarization vectors, we note that the extraordinary wave polarization is in the direction perpendicular to the wave vector $\mathbf{k}_i$, and lies in the same plane as the wave vector and the c-axis. The ordinary wave polarization vector is

---

[3] It can be shown that if we record multiple holograms on the same location, the diffraction efficiency of the individual holograms scale as $1/M^2$, where M is the number of holograms [6]. Given a minimum diffraction efficiency, the number of holograms M that can stored is therefore proportional to the square root of the diffraction efficiency.

perpendicular to the wave vector and the extraordinary wave polarization vector. Thus the ordinary mode polarization vector is given by

$$\mathbf{d}_i = \big(\sin\theta_i, \quad -\cos\theta_i, \quad 0\big), \tag{5.54}$$

and the extraordinary mode polarization vector is given by

$$\mathbf{d}_i = \big(-\cos\phi_i\ \cos\theta_i, \quad -\cos\phi_i\ \sin\theta_i, \quad \sin\phi_i\big). \tag{5.55}$$

We can then use Eq. (5.46) to find the values of the n_i's (or n_i''s), and from these, the k_i's (from Eq. (5.45)) and the wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$.

Next, we calculate the grating vector $\mathbf{K} = \mathbf{k}_2 - \mathbf{k}_1$, from which we determine the unit vector $\mathbf{u}$. We also need to calculate E_{sat}, as given by Eq. (5.1). Note that in Eq. (5.1), the permittivity ϵ (given by Eq. (5.50)) is not a constant, but depends on the direction ($\mathbf{u}$) of the grating vector, $\mathbf{K}$.

Assuming that r is known, we can finally substitute these values obtained above into Eq. (5.52) to get G. The expression is complicated, and for the general case no simple expression has been found for their maximum. However the value of G can be readily calculated numerically using a computer.

5.3. The Co-Planar Geometry

Although no simple expression for G is known (in terms of the θ_i and ϕ_i) for the general case, it can be reduced to a simple form for the special case where $\mathbf{k}_1$, $\mathbf{k}_2$ and z-axis are in the same plane. We will call this configuration the *co-planar* geometry. [4] The co-planar geometry is especially convenient for recording holograms on a 3-D disk in either the transmission or reflection geometry (where the c-axis is the disk rotation axis), since the eigenmode polarizations are either in the same plane as $\mathbf{k}_1$, $\mathbf{k}_2$ (o-mode) or perpendicular to it (e-mode), regardless

---

[4] This agrees with the definition given in Section 2.1, since in the discussion to follow, the z-axis is the normal of the crystal.

of the disk rotation angle. In terms of getting uniform diffraction efficiency as the disk rotates, choosing the c-axis as the rotation axis is a reasonable choice, since the c-axis is the axis with the highest symmetry.

In this section, we will apply the results derived in Section 5.2 to two particular categories of crystals that have special symmetry properties: those that belong to the $3m$ symmetry group, and those that belong to the $4mm$ symmetry group. These are of special interest because the majority of photorefractive crystals being used today (lithium niobate, barium titanate, SBN, etc.) belong to one of these two categories. Gallium arsanide, which belongs to crystals having the $\bar{4}3m$ symmetry group and exhibits some photorefractive effects, also has the $3mm$ symmetry. It can be shown, that under coordinate transformations (with the [111] direction as the new z-axis), the electrooptic tensor of the $\bar{4}3m$ symmetry group has the same form as that of the $3mm$ symmetry group.

5.3.1 Crystals From the $3m$ Symmetry Group: $LiNbO_3$

Lithium niobate is an example of a crystal having 3-fold symmetry about its c-axis. It belongs to the crystal group having the $3m$ symmetry [7]. In the contracted indices notation, the electrooptic coefficient tensor (in the coordinate system that diagonalizes the permittivity tensor) is given by [8]

$$r = \begin{pmatrix} 0 & -r_{22} & r_{13} \\ 0 & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{42} & 0 & 0 \\ -r_{22} & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\alpha & \beta \\ 0 & \alpha & \beta \\ 0 & 0 & \gamma \\ 0 & \delta & 0 \\ \delta & 0 & 0 \\ -\alpha & 0 & 0 \end{pmatrix}. \tag{5.56}$$

For lithium niobate, the coefficients are [9]: $\alpha = r_{22} = 6.8\ pm/V$, $\beta = r_{13} = 9.6\ pm/V$, $\gamma = r_{33} = 30.9\ pm/V$, and $\delta = r_{42} = 32.6\ pm/V$, with $n_o = 2.286$, $n_e = 2.200$, $\epsilon_x = 78$, and $\epsilon_z = 32$. In the calculations below, the wavelength is $500\ nm$.

Assuming the co-planar geometry, $\mathbf{k}_1$, $\mathbf{k}_2$ and the z-axis are in the same plane, and we have $\theta_1 = \theta_2 = \theta$. The grating vector $\mathbf{K}$ is then

$$\mathbf{K} = \mathbf{k}_2 - \mathbf{k}_1 = k_0 \begin{pmatrix} (n_2 \sin\phi_2 - n_1 \sin\phi_1)\cos\theta \\ (n_2 \sin\phi_2 - n_1 \sin\phi_1)\sin\theta \\ n_2 \cos\phi_2 - n_1 \cos\phi_1 \end{pmatrix}. \tag{5.57}$$

We now make the approximation that $n_1 \approx n_2 \approx n_a$, where $n_a = (n_e + n_o)/2$. For lithium niobate the difference between n_e and n_o is about 4%, so the approximation is fairly good [5]. We then have

$$\mathbf{K} = 2k_0 n_a \sin\frac{\phi_2 - \phi_1}{2}\mathbf{u}, \tag{5.58}$$

where

$$\mathbf{u} = (u_x, u_y, u_z) = \left(\cos\frac{\phi_2 + \phi_1}{2}\cos\theta,\ \cos\frac{\phi_2 + \phi_1}{2}\sin\theta,\ -\sin\frac{\phi_2 + \phi_1}{2}\right) \tag{5.59}$$

is the unit vector in the $\mathbf{K}$ direction. We then have (after expanding the contracted indices)

$$r\mathbf{u} = -\begin{pmatrix} -\alpha u_y + \beta u_z & -\alpha u_x & \delta u_x \\ -\alpha u_x & \alpha u_y + \beta u_z & \delta u_y \\ \delta u_x & \delta u_y & \gamma u_z \end{pmatrix}. \tag{5.60}$$

Depending on which polarizations we are interested in, we substituting either Eq. (5.54) or (5.55) for $\mathbf{d}_1$ and $\mathbf{d}_2$ in the expression for G (Eq. (5.52)), and after simplifying the expressions, we obtain

$$\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1 = A + B\sin 3\theta \tag{5.61}$$

if $\mathbf{d}_1$ and $\mathbf{d}_2$ are the same mode (i.e., both o-modes or both e-modes), or

$$\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1 = A + B\cos 3\theta \tag{5.62}$$

if $\mathbf{d}_1$ and $\mathbf{d}_2$ are different modes (i.e., one is o-modes and the other is e-modes). Note that we have the expected three-fold symmetry about the z-axis in the sine and cosine terms. The coefficients A and B for the various coupling modes are:

---

[5] From Eq. (5.44) or Eq. (5.46), it is easy to show that n_i and n_i' are between n_o and n_e.

1. $\mathbf{d}_1$ = *o-mode*, $\mathbf{d}_2$ = *o-mode: (o-o coupling)*

$$A = \beta \sin \frac{\phi_2 + \phi_1}{2} \tag{5.63}$$

$$B = -\alpha \cos \frac{\phi_2 + \phi_1}{2} \tag{6.64}$$

2. $\mathbf{d}_1$ = *e-mode*, $\mathbf{d}_2$ = *e-mode: (e-e coupling)*

$$A = \delta \sin(\phi_2 + \phi_1) \cos \frac{\phi_2 + \phi_1}{2} + \gamma \sin \phi_1 \ \sin \phi_2 \ \sin \frac{\phi_2 + \phi_1}{2} + \beta \cos \phi_1 \ \cos \phi_2 \ \sin \frac{\phi_2 + \phi_1}{2} \tag{5.65}$$

$$B = \alpha \cos \phi_1 \ \cos \phi_2 \ \cos \frac{\phi_2 + \phi_1}{2} \tag{5.66}$$

3. $\mathbf{d}_1$ = *o-mode*, $\mathbf{d}_2$ = *e-mode: (o-e coupling)*

$$A = 0 \tag{5.67}$$

$$B = \alpha \cos \phi_2 \ \cos \frac{\phi_2 + \phi_1}{2} \tag{5.68}$$

4. $\mathbf{d}_1$ = *e-mode*, $\mathbf{d}_2$ = *o-mode: (e-o coupling)*

$$A = 0 \tag{5.69}$$

$$B = \alpha \cos \phi_1 \ \cos \frac{\phi_2 + \phi_1}{2} \tag{5.70}$$

The coefficient A multiplied by the other factors in G give the average value of G, while B multiplied by the other factors in G gives the variation with respect to rotation angle θ. For e-o and o-e coupling, the average value of G is zero. This means that for these two coupling schemes, G will cross zero at some rotation angle, and therefore there is considerable variation in the diffraction efficiency as the disk rotates.

In Figure 5.1(a) to (d), we plot the average and variation of G in lithium niobate for the four various polarization couplings. [6] These values are plotted as

---

[6] From coupled-mode analysis (Chapter 2), the sign of G signifies the direction of two-wave mixing [8], or beam-coupling in the crystals. The sign of G is important if we are interested in optical gain from two-wave mixing. However, since we are primarily interested in optical storage here, and therefore diffraction efficiency, we are more concerned with the absolute value of G.

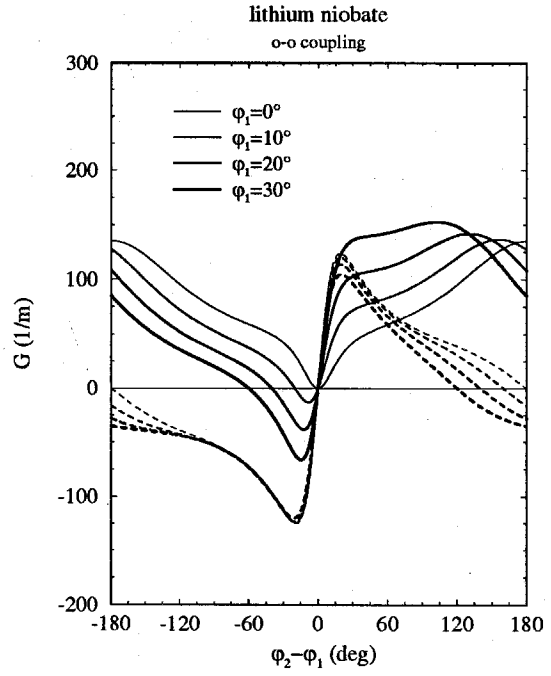
functions of $\phi_2 - \phi_1$, which is the angle between $\mathbf{k}_2$ and $\mathbf{k}_1$. Recall that $\mathbf{k}_1$, $\mathbf{k}_2$, and the z-axis lie in the same plane.

Note that the angles ϕ_1 and ϕ_2 are the angles *inside* the crystal. Since the largest angle we can have inside lithium niobate (due to refraction) is about 27°, the values of ϕ_1 in Figure 5.1 have been plotted only up to 30°. Similarly, certain values of $\phi_2 - \phi_1$ shown in Figure 5.1, can not be supported inside the crystal.

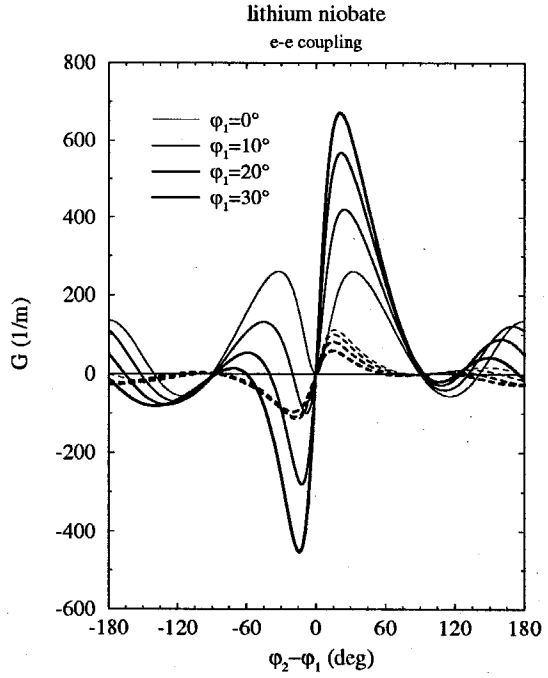
For c-o and o-e coupling, we have plotted only the variation, since the averages are zero. Strictly speaking, from Eq. (5.52), the values of G are zero for e-o and o-e coupling, since for these two cases, $\mathbf{e}_1 \cdot \mathbf{e}_2$ is zero. In Figures 5.1(c) and (d), the results were calculated by ignoring this, and taking $\mathbf{e}_1 \cdot \mathbf{e}_2$ to be 1 instead. The fact that $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$ means that if we use mixed polarizations, we can not form interference patterns, since the modulation depth (which is proportional to $\mathbf{e}_1 \cdot \mathbf{e}_2$) is zero. It is of course possible to record with one polarization, and then read out with another. However, the angle of the readout beam now changes because of crystal anisotropy. In order for the grating to be Bragg matched, the reference beam needs to be adjusted. For images however, it is often difficult to simultaneously Bragg match all the spatial frequencies. As a result, the reconstructed image is distorted. In any case, the results in Figures 5.1(c) and (d) have lower values than those of e-e and o-o coupling, and the averages are zero.

For o-o coupling, the polarization vectors (the direction of the electric fields) are perpendicular to the incident plane (the plane formed by $\mathbf{k}_1$ and $\mathbf{k}_1$). As shown in Figure 5.1(a), the absolute value of the average of G for small values of $\phi_2 - \phi_1$ is less than the variation. This implies that for transmission type holograms, we have large variation in diffraction efficiency as the disk rotates. For o-o coupling, it is therefore better to record in the reflection geometry. For example, if the image beam incidents normally on the crystal, we have $\phi_1 = 0$. In this case, we record in the reflection geometry where ϕ_2 is nearly 180°. In this configuration, the variation is also low.

For e-e coupling, the polarization vectors lie in the incident plane. As shown in Figure 5.1(b), the average is higher than for o-o coupling. For $\phi_1 = 0$ (signal

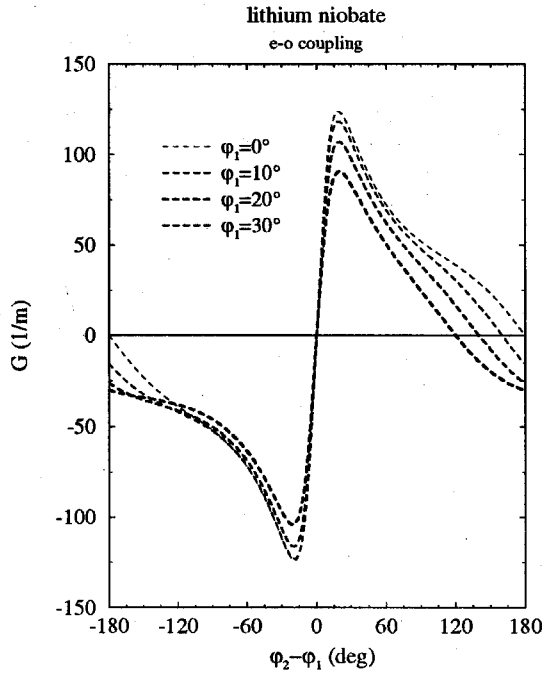


(a)

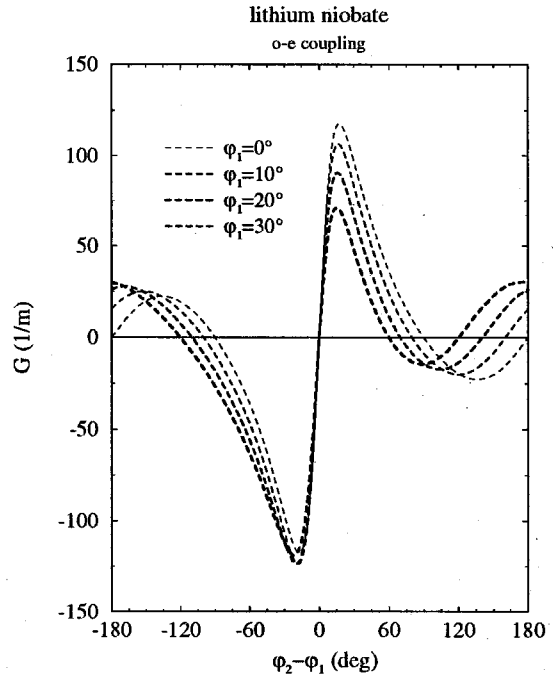


(b)

Figure 5.1. The average (solid lines) and variation (dashed lines) of G for: (a) o-o coupling, and (b) e-e coupling in lithium niobate.



(c)



(d)

Figure 5.1. (cont'd) The variation of G for: (c) e-o coupling, and (d) o-e coupling in lithium niobate. The average in both cases is zero.

beam at normal incidence), the average and variation are approximately the same for small $\phi_2 - \phi_1$ (transmission holograms). This means that there is large variation in η as the disk rotates. As we move to larger ϕ_1, the average increases while the variation decreases (for the same $\phi_2 - \phi_1$). However there is a large change in G as $\phi_2 - \phi_1$ changes (i.e., as the reference beam angle changes). This needs to be taken into account if we want to record multiple holograms with uniform diffraction efficiency.

For e-e coupling, recording reflection holograms is less favorable since the diffraction efficiency varies more over a smaller range of $\phi_2 - \phi_1$ angles compared to that of o-o coupling.

In conclusion, although the diffraction efficiency is higher with e-e coupling, a larger incident angle presents some problems, such as reflection from the crystal surface. Another problem is the larger variation in the diffraction efficiency as $\phi_2 - \phi_1$ changes. In comparison the "bandwidth" for o-o coupling with reflection holograms is larger. Thus it is more convenient to record reflection type holograms in lithium niobate 3-D disks.

5.3.2. Crystals From The *4mm* Symmetry Group: $BaTiO_3$ and SBN

$BaTiO_3$ and SBN have a 4-fold symmetry around the c-axis, and belong to the crystal group with the *4mm* symmetry. In the contracted indices notation, the electrooptic coefficient tensor (in the coordinate system that diagonalizes the permittivity tensor) is given by [8]

$$r = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{42} & 0 \\ r_{42} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \beta \\ 0 & 0 & \beta \\ 0 & 0 & \gamma \\ 0 & \delta & 0 \\ \delta & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.71}$$

Comparing this to the r coefficients for the *3m* crystals, the forms are identical if we take $\alpha = 0$. The results given in Eq. (5.61)–(5.70) (for the co-planar geometry)

are therefore still valid. But now since α is zero, B is zero, and this means that there is no variations with respect to change in θ, the rotation angle. In the context of 3-D disk geometry, this means that there is no variation in the diffraction efficiency with respect to disk rotation. Note also that there is no coupling between the o- and e-modes, since both A and B are zero for e-o or o-e coupling (Eqs. (5.67)–(5.70)).

For $BaTiO_3$, the coefficients are [8]: [7] $\beta = r_{13} = 8\ pm/V$, $\gamma = r_{33} = 23\ pm/V$, $\delta = r_{42} = 820\ pm/V$, with $n_o = 2.437$, $n_e = 2.365$, $\epsilon_x = 4300$, and $\epsilon_z = 106$. In the calculations below, the wavelength is 500 nm. We plot the values of G for e-e coupling and o-o coupling in Figures 5.2(a) and (b).

From these results, we see that it is preferable to use e-e coupling with transmission geometry in the case of barium titanate. The diffraction efficiencies are much larger than for lithium niobate. However, the variation in diffraction efficiency as $\phi_2 - \phi_1$ changes is larger.

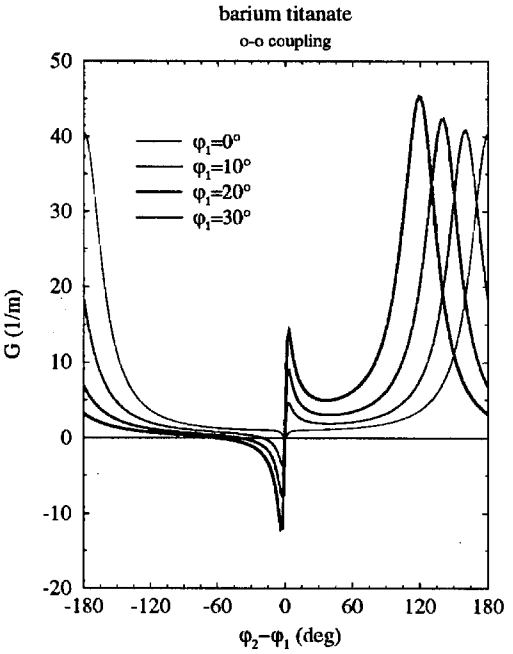
For SBN, the coefficients are [10]: $\beta = r_{13} = 55\ pm/V$, $\gamma = r_{33} = 224\ pm/V$, $\delta = r_{42} = 80\ pm/V$, with $n_o = 2.3$, $n_e = 2.27$, $\epsilon_x = 470$, and $\epsilon_z = 1100$. In the calculations below, the wavelength is 500 nm. We plot the values of G for e-e coupling and o-o coupling in Figures 5.3(a) and (b).

The result for SBN with e-e coupling is similar to the situation for barium titanate, but with lower diffraction efficiency. As with barium titanate, it is advisable to record in transmission geometry using e-e coupling. Again, the variation in diffraction efficiency with respect to $\phi_2 - \phi_1$ is larger than for lithium niobate.

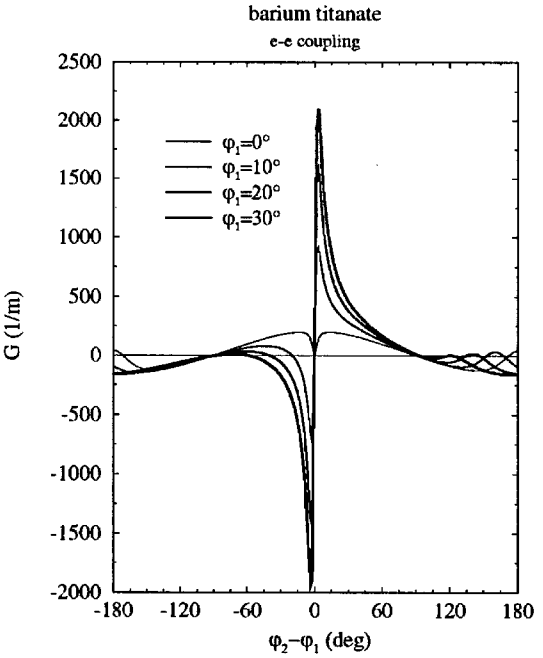
Note that for both crystals, the diffraction efficiency increases dramatically as ϕ_1 increases (i.e., the signal beam incidents at a more oblique angle). This

---

[7] Strictly speaking, the values for r, ϵ_x, and ϵ_z listed here are for high frequency [9]. For calculating the photorefractive effect, we should actually use the low frequency values. Thus the results shown here should not be trusted completely. Nevertheless, these results will give us an adequate estimate of the situation. In reference [19], the authors also used the high frequency values.

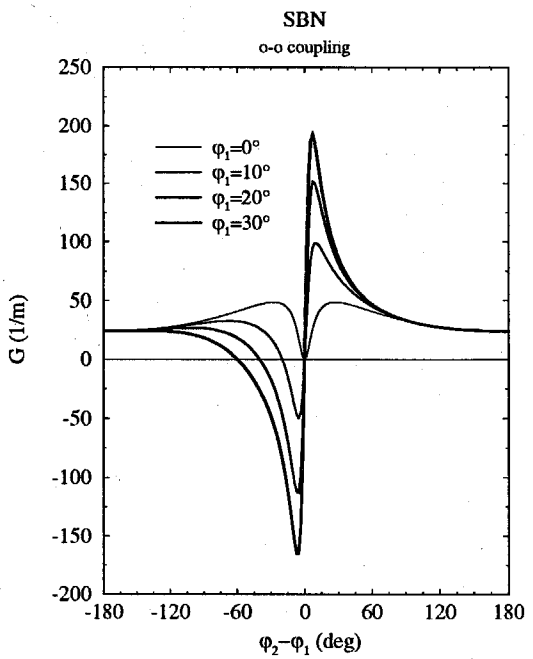


(a)

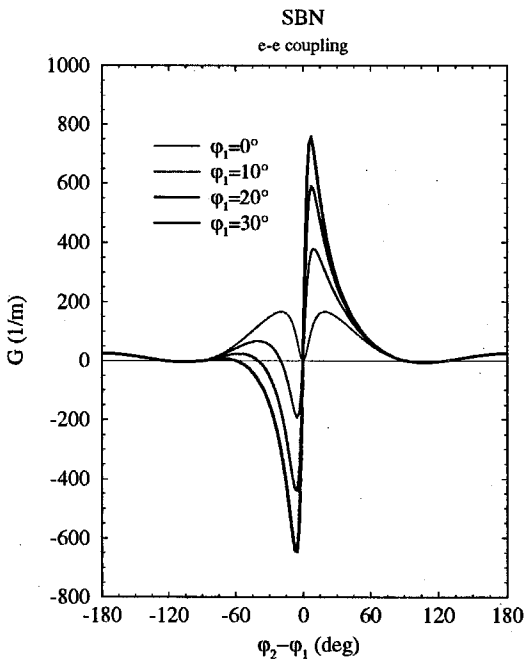


(b)

Figure 5.2. The values of G for: (a) o-o coupling, and (b) e-e coupling in barium titanate.



(a)



(b)

Figure 5.3. The values of G for: (a) o-o coupling, and (b) e-e coupling in SBN.

is due to the fact that the diffraction efficiency increases as the direction of the grating vector **K** is closer to the direction ofs the c-axis.

5.3.3. Demonstration of a 3-D Disk System

Having discussed some of the design considerations of the 3-D disk, we now present some results of holographic recording on an experimental 3-D disk.

The disk is an 0.01%-iron doped lithium niobate crystal (from *Deltronics*), with a thickness of 5 mm and a diameter of 1.5 inches. Lithium niobate was selected because it is commercially available in large sizes, and has good optical quality.

The holograms were recorded as near-image plane holograms using the reflection geometry, with the signal beam at normal incidence on the crystal surface, and the reference beam coming in from the other side of the crystal. The crystal was mounted on a rotation stage, and different recording locations along the rim of the crystal were accessed by rotating the crystal. Each recording location was approximately $5 \times 5\ mm^2$, and a total of 20 locations were recorded along the rim of the crystal. At each location, 100 holograms were recorded by angle multiplexing. A computer-controlled stepper motor was used to rotate the mirror that changed the reference beam angle. The reference beam passed through a 4-F imaging system so that the reference beam would always fall on the same location on the crystal as the angle changed. (This is the same as that used in the setup shown in Figure 6.1 in Chapter 6.)

The images used for recording were stored on a VCR, and a computer advanced the pictures frame-by-frame for recording. The VCR was connected to a liquid crystal TV (LCTV) taken from an *Epson* TV projector, and the LCTV was used as a transparency to bring the image onto the crystal. A recording schedule [6] was used for writing the holograms to ensure uniformity of the diffraction efficiencies in the individual holograms.

In Figure 5.4, we show the diffraction efficiency of 100 holograms recorded at

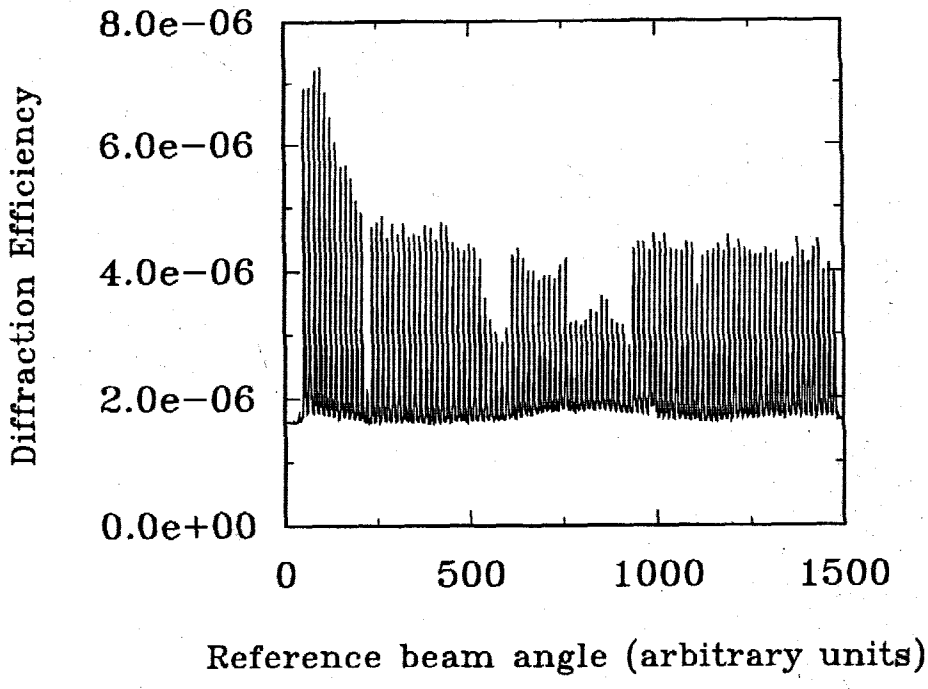


Figure 5.4. Diffraction efficiency of 100 angle multiplexed holograms.

Figure 5.5. Some of the reconstructed pictures from the 100 holograms.

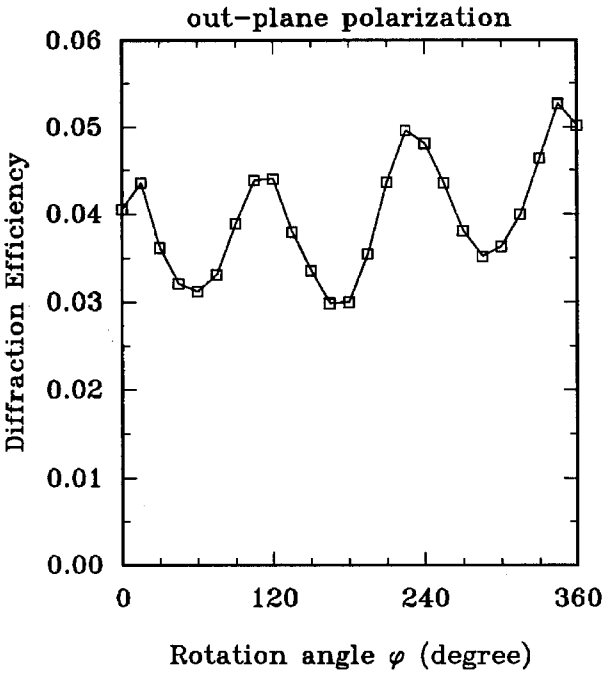


Figure 5.6. Diffraction efficiency as a function of rotation angle. Each hologram was exposed for 100 seconds during recording. The intensities of the recording beams are 20.5mW for the reference beam, and 2.5mW for the signal beam.

one location. The holograms were recorded in the reflection type geometry, and then thermally fixed [11]. [8] The diffraction efficiencies are shown as a function of reference beam angle. They are not completely uniform for two reasons. First, the images were taken from a segment of a cartoon, and the total intensity of the pictures changed from frame-to-frame. Another reason is that the estimate of the time constant was not accurate. It is known that the time constants for writing and erasing are not the same. In the experiment shown in Figure 5.4, the recording schedule was, for simplicity, calculated by assuming that the time constants were the same. This caused additional error in the uniformity.

In Figure 5.5, we show some of the reconstructed images from the 100 holo-

[8] Baked at 130° C for approximately 30 minutes.

grams recorded on the crystal. The reconstructed images were captured by a CCD camera and printed out from a *Sony* video printer. In Figure 5.6, we show the diffraction efficiency of holograms as the disk rotates. Note that as expected, we have the 3-fold symmetry due to the particular crystal property of lithium niobate.

5.4. Double Gratings

As mentioned briefly in Section 5.2, if we record holograms using polarizations that are not eigenmodes, we record "double gratings." This can be a problem when recording multiple holograms because of cross-talk. In this section, we examine this issue in more detail.

Because photorefractive crystals such as lithium niobate are anisotropic, plane waves that are not eigenmodes split into the ordinary wave (o-wave) and the extraordinary wave (e-wave). When this occurs, double gratings are written for each plane wave pair, each grating being associated with one of the eigenmodes (polarizations). This can be a problem for optical storage using angle multiplexing in volume holograms. In these systems, cross-talk between the various holograms is avoided by the angular Bragg selectivity of volume holograms. The width of the Bragg matching angle (which depends on the crystal thickness) dictates how far apart the reference beam angle needs to change from frame-to-frame to avoid cross-talk between the holograms.

In the usual setup for recording holograms in lithium niobate crystals, the c-axis lies on the plane of incidence (defined by the wave vectors of the signal and reference beams), and the optical (electric) field is polarized either in-plane (horizontal polarization) or perpendicular (perpendicular polarization) to the plane of incidence. When the signal and reference beams are plane waves, they are eigenmodes, and the beams do not split upon refraction. Therefore, a single grating is written, and the problem does not arise in this simple case. Cross-talk then depends only on the width of the Bragg matching angles.

For the 3-D disk, however, the disk rotates in order for the system to access different recording locations. The crystal orientation changes, and in general the incident beams are not eigenmode. If one or both of the recording beams (which are plane waves) are not eigenmodes, then double gratings are recorded instead of a single grating. Upon reconstruction of the holograms by scanning the reference beam angle, it is possible for each of the gratings to be Bragg matched at more than one angle. Thus there are three angles at which each hologram (consisting now of a double grating pair) can be potentially Bragg matched, and this results in additional cross-talk between the various holograms when we try to do angle multiplexing.

5.4.1. Theory

We describe an experiment that demonstrates the double grating effect. The geometry of our experiment is shown in Figure 5.7. The coordinates x, y, and z are the crystal axes. The c-axis (z) is tilted at an angle φ with respect to the plane of incidence. The direction of propagation of the signal beam is in the x direction, perpendicular to the surface of the crystal, the y-z plane. The reference beam is at an angle θ with respect to the signal beam. The recording beams are horizontally polarized, and have both e-wave and o-wave components. The angle between $\mathbf{k}_1$ and $\mathbf{k}_2$ is denoted as θ. This geometry was chosen only for convenience of analysis. The results are similar for other recording geometries.

Upon entering the crystal, the two eigenmodes refract according to Snell's law. The direction of the refracted beam depends on the index of the crystal, which can be found from the normal surfaces of the e- and o-waves. These are

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_o^2} = 1 \tag{5.72}$$

for the ordinary wave normal surface, and

$$\frac{x^2}{n_e^2} + \frac{y^2}{n_e^2} + \frac{z^2}{n_o^2} = 1 \tag{5.73}$$

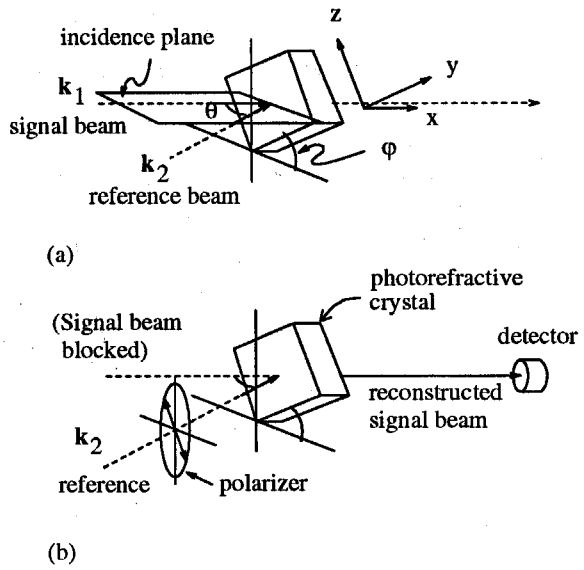


Figure 5.7. Geometry of recording and measurement. (a) Recording geometry.
(b) Setup for experiment in Fig. 5.12 and Fig. 5.13.

for the extraordinary wave normal surface.

The refracted beams remain within the incident plane since the signal beam is perpendicular to the crystal surface. Therefore we need only consider the intersections of the two normal surfaces with the incident plane. The intersections are two ellipses, as illustrated in Figure 5.8. They do not touch each other because the incident plane is at an angle φ with respect to the c-axis (z-axis). The equations for the two ellipses are readily found to be:

$$\frac{x^2}{n_o^2} + \frac{z'^2}{n_o^2} = 1 \tag{5.74}$$

for o-waves, and

$$\frac{x^2}{n_e^2} + \left(\frac{\sin^2\varphi}{n_e^2} + \frac{\cos^2\varphi}{n_o^2}\right)z'^2 = 1 \tag{5.75}$$

for e-waves. (x lies in the incident plane. z' is the direction on the incident plane perpendicular to x.)

Each beam splits into two eigenmodes according to Snell's law. The situation is shown in Figure 5.8. For our particular case, the two components for the signal

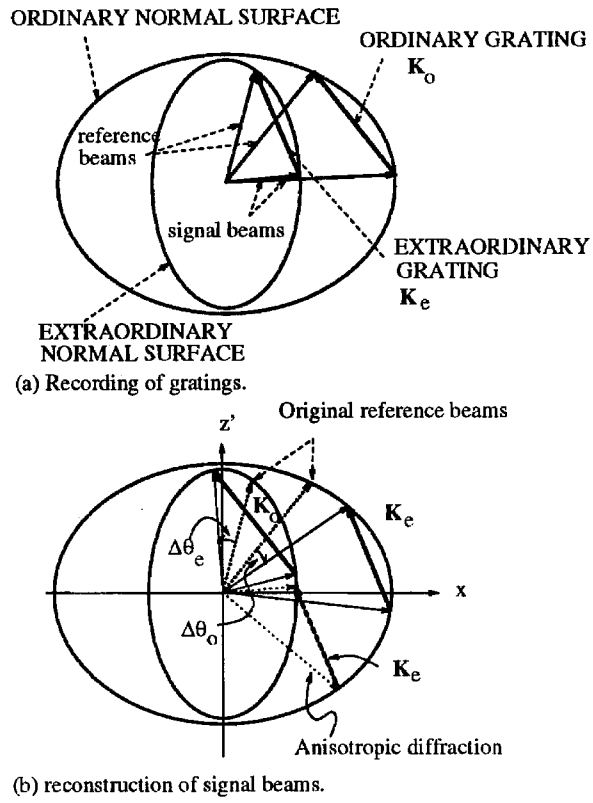


Figure 5.8. Normal surfaces and grating vectors.

beam are actually in the same direction since we have selected for convenience the direction of the signal beam to be along the crystal axis x. (In Figure 5.8, they are separated for clarity.) The two eigenmodes are perpendicular to each other in polarization, but it should be noted that the eigenmode polarizations are in general neither parallel nor perpendicular to the crystal axes. One of the reasons for choosing the signal beams to propagate along the direction of the crystal axis x, is because the eigenmode polarization directions are exactly along the other two crystal axes, y and z. Figure 5.9 shows the calculated slant angle, defined as the angle between the e-wave polarization and the direction normal to the plane of incidence, for several values of φ. The derivation of the slant angle is given in Appendix A.

The four refracted signal and reference beams inside the crystal form interference patterns which write gratings on the crystal via the photorefractive effect.

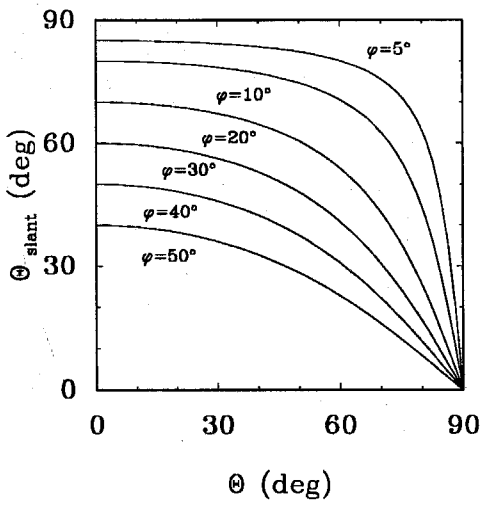


Figure 5.9. Slant angle between e-wave polarization and direction normal to incident plane.

The e-wave components form the grating $\mathbf{K}_e$, which we will call the extraordinary grating. The o-wave components form the grating $\mathbf{K}_o$, which we will call the ordinary grating. From the geometry of Figure 5.8, and the application of Snell's law, we can calculate the directions of the refracted beams, and hence the two gratings:

$$\mathbf{K}_e = \left(n_e\sqrt{1-\sin^2\theta\left(\frac{\sin^2\varphi}{n_e^2}+\frac{\cos^2\varphi}{n_o^2}\right)}-n_e,\ \sin\theta\right) \tag{5.76}$$

$$\mathbf{K}_o = \left(n_o\sqrt{1-\frac{\sin^2\theta}{n_o^2}}-n_o,\ \sin\theta\right). \tag{5.77}$$

In addition to the two gratings $\mathbf{K}_e$ and $\mathbf{K}_o$, we also get gratings from interference between the e-wave/o-wave component of the reference beam and the o-wave/e-wave component of the signal beam. However these two "cross gratings" are much weaker than $\mathbf{K}_o$ and $\mathbf{K}_e$ since the o-wave and e-wave modes are polarized orthogonally to each other for the same beam, and are almost perpendicular to each other for the signal and reference beams. Estimates of the modulation depths for $\mathbf{K}_e$, $\mathbf{K}_o$, and the cross gratings show that the modulation depths of the cross grating is less than 5% of that of $\mathbf{K}_e$ and $\mathbf{K}_o$. (In fact, it turns out that

the e-wave of the signal beam and the o-wave of the reference beam are exactly perpendicular to each other. See Appendix A.)

Once the gratings are recorded in the crystal, the signal beam can be reconstructed by the reference beam. The angle at which the signal can be reproduced is determined by the Bragg condition $\mathbf{k}_1 = \mathbf{k}_2 + \mathbf{K}$. Since we have two gratings, it is possible to satisfy the Bragg condition at more than one angle. In addition to the original recording angle (where $\mathbf{K}_e$ is read out by the e-wave and $\mathbf{K}_o$ is read out by the o-wave), we can get Bragg matching when $\mathbf{K}_e$ is read out by the o-wave and also when $\mathbf{K}_o$ is read out by the e-wave, as shown in Fig. 2(b). It is also possible that one of the two gratings can also satisfy the Bragg condition for anisotropic diffraction [12–15]. The anisotropic Bragg matching condition for $\mathbf{K}_e$ is also shown in Figure 5.8(b). We will not consider anisotropic diffraction for the rest this section. However, it can be treated in a way completely analogous to the isotropic case we will treat.

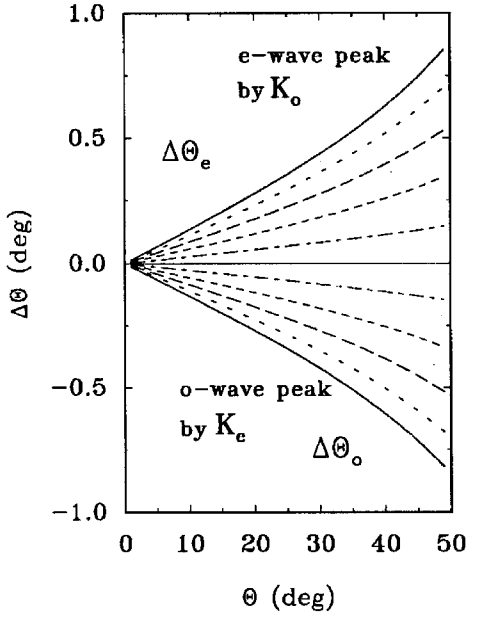


Figure 5.10. Deviation angle of the side-peaks from the recording reference beam angle as a function of θ for various values of φ.

The angles where the additional Bragg matching conditions occur are very

close to the original recording angle (the dotted lines in Fig. 5.8(b)). The deviation of the Bragg angle of the reconstructing beam from the original recording angle, denoted as $\Delta\theta_e$ (for e-wave) and $\Delta\theta_o$ (for o-wave), can be calculated as a function of θ and φ from the geometry of Fig. 5.8. The calculation is shown in Appendix B, and the result is plotted in Figure 5.10 for lithium niobate where $n_e = 2.208$ and $n_o = 2.286$. Note that $\Delta\theta$ depends only on the orientation of the crystal with respect to the writing beams (i.e., θ and ϕ). Inside the crystal, the Bragg angle is

$$\Delta\theta_{Bragg,inside} \approx \Lambda/L = \frac{\lambda}{2nL\sin\frac{\theta_{in}}{2}}, \tag{5.78}$$

where, θ_{in} here is the angle inside the crystal. Outside the crystal, because of Snell's law, we have

$$\Delta\theta_{Bragg} \approx \frac{\sqrt{n^2 - \sin^2\theta}}{\cos\theta}\Delta\theta_{Bragg,inside}, \tag{5.79}$$

where n is the index of refraction, and L is the interaction length. For lithium niobate ($n \approx 2.2$) with $L = 8mm$, and using $\lambda = 488nm$ and $\theta = 30°$, we get $\Delta\theta_{Bragg}$ about 0.01°. For $\varphi = 30°$ and $\theta = 30°$, we get from Figure 5.10 $\Delta\theta_e \approx \Delta\theta_o \approx 0.25°$, which is much larger than $\Delta\theta_{Bragg}$. This implies that in a system where we store images using angular multiplexing, the double grating will interfere with an image stored approximately 25 positions away.

5.4.2. Experiments

The experimental setup is shown in Figure 5.7(b), with $\theta = 34°$ and $\varphi = 30°$. Both signal and reference beams are polarized in the plane of incidence. We record a hologram in a lithium niobate crystal (grown by Deltronics; $20 \times 20 \times 8mm$, 0.01% Fe doping) using two plane waves, then block the signal beam. The detected intensity for the reconstructed beam (which has both e-wave and o-wave components) as a function of $\Delta\theta$ is shown in Figure 5.11. As expected, there are three peaks. In addition to one at the original angle ($\Delta\theta = 0°$), there are two side-peaks. The side-peaks are measured to be at $\Delta\theta_e = +0.36° \pm 0.01°$ and

$\Delta\theta_o = -0.35° \pm 0.01°$. The theoretical prediction is $\pm 0.31°$ for both $\Delta\theta_e$ and $\Delta\theta_o$. The central peak in Figure 5.11 has both e-wave and o-wave components, while the side-peak on the right is an e-wave, and the peak on the left is an o-wave.

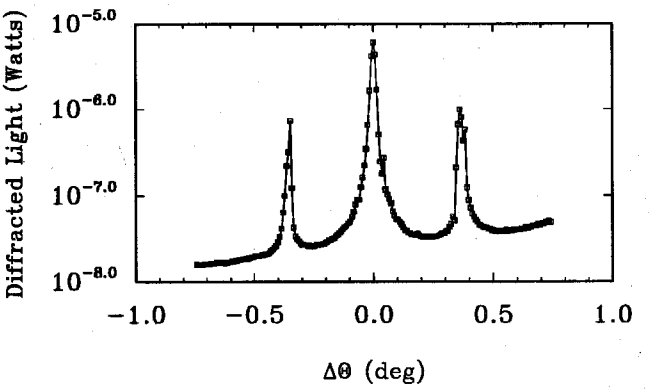


Figure 5.11. Measured intensity of reconstructed beam as a function of deviation of reference angle. Reference beam has both e-wave and o-wave components.

If we reconstruct the signal beam with a reference beam that has only e-wave components, we Bragg match $\mathbf{K}_e$ at one angle (the original recording angle), and $\mathbf{K}_o$ at a different angle. This gives us two peaks (both of which are e-waves), as shown in Figure 5.12(a). Similarly, we get two peaks when the reference beam has only o-wave components (Fig. 5.12(b)).

The relative heights of the two peaks shown in Fig. 5.12(a) and (b) can be estimated as follows. The diffraction efficiency of a thick phase grating is proportional to the square of the index change Δn of the grating (Chapter 2). This in turn is proportional to the modulation depth of the interference pattern for weak holograms. We can measure experimentally the e-wave components of the signal (I_{e1}) and reference beam (I_{e2}), and the o-wave components of the signal (I_{o1}) and reference beam (I_{o2}) that are transmitted through the crystal. From this we estimate the ratio of the modulation depths of $\mathbf{K}_e$ and $\mathbf{K}_o$ to be $\sqrt{I_{e1}I_{e2}/I_{o1}I_{o2}} = 2.94$, so the height of the two peaks (for both Fig. 6(a) and (b)) should have a ratio of $2.94^2 = 8.67$. (Strictly speaking, the diffraction efficiency

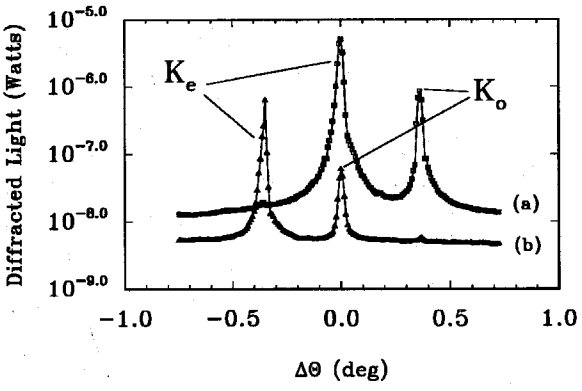


Figure 5.12. Measured intensity of reconstructed beam as a function of deviation of reference angle. (a) Reference beam has e-wave only. (b) Reference beam has o-wave only.

also depends on two-wave-mixing coupling coefficients [8] that are functions of θ and beam polarization. However $\Delta\theta$ is $< 0.5°$, so the change is negligible.) The actual measured ratios were 5.96 for the two e-wave peaks (Figure 5.12(a)) and 10.37 for the two o-wave peaks (Figure 5.12(b)).

The discrepancy between theory and measurement might be due to the fact that we do not actually have an infinite plane wave. Because of this, we have a spread of spatial frequencies centered around the $\mathbf{K}_e$ and $\mathbf{K}_o$ grating vectors (Figure 5.28). We can not Bragg match all these spatial frequencies simultaneously at the side-peaks, which is evident from the fact that at the side-peaks dark bands appear across the reconstructed image.

Instead of using plane waves, we can also record images. We recorded two images on a photorefractive crystal using the setup shown in Figure 5.13. The images 1 and 2 are photographic plate transparencies. The incident plane wave is split by the polarizing beamsplitter (PB) into horizontal and vertical components. Each illuminates one of the transparencies, and the two images are recombined by a second PB. The hologram is recorded with a reference beam that has both horizontal (e-wave) and vertical (o-wave) polarization. The first image consisted only of e-wave, and the second image consisted only of o-wave. After the holo-

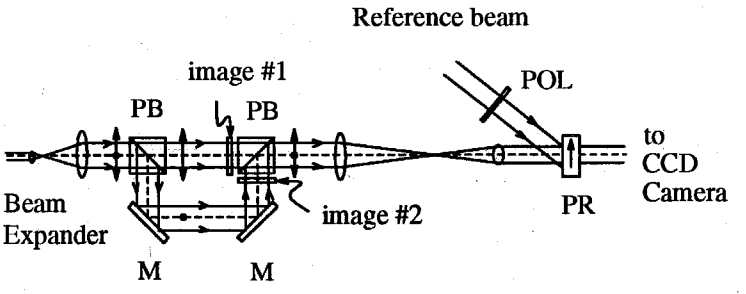
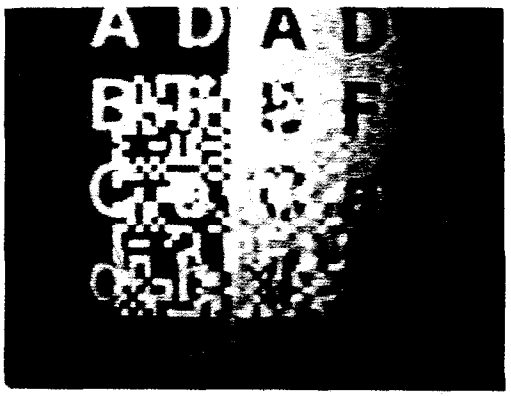


Figure 5.13. Setup for recording two images of orthogonal polarizations. Image 1 has only o-wave component. Image 2 has only e-wave component. (PB: polarizing beamsplitter. POL: polarizer. PR: photorefractive crystal. M: mirror.)

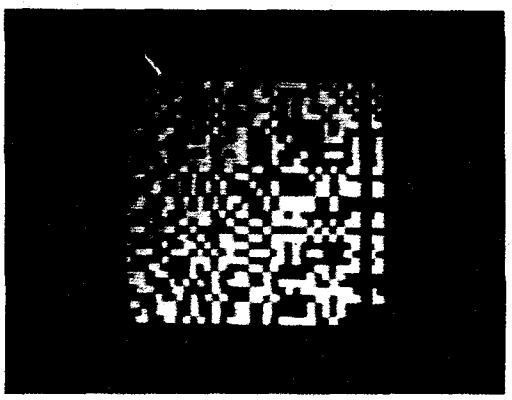
gram was recorded, we observed the image reconstructed by the same reference beam used for recording. As we change the polarization of the reference beam, images 1 and 2 appear and disappear in turn, since the former is recorded as $\mathbf{K}_e$ gratings, and the latter is recorded as $\mathbf{K}_o$ gratings. The results are shown in Figure 5.14. Figure 5.14(a) shows the reconstructed image when both o- and e-wave components are present in the reference beam. Figure 5.14(b) and Figure 5.14(c) shows that only one of the two images appears when the reference beam is o-wave or e-wave only. Note that there is some cross-talk: there is a trace of image 2 in image 1. This is probably because the polarizations were not exactly aligned with the c-axis of the crystal, so there was some o-wave component in the e-wave image of 1. The pictures were captured by a CCD camera, then taken off the monitor by a polaroid camera.

5.4.3. Design Considerations and Applications

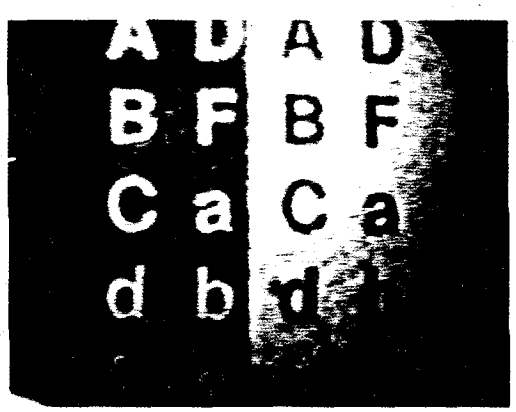
Double gratings are written in photorefractive crystals when the recording beams have both e-wave and o-wave components. As a result, it is possible to satisfy the Bragg matching condition at more than one reference beam angle. This can be a problem when we record multiple holograms by angle multiplex-



(a)



(b)



(c)

Figure 5.14. (a) Reconstructed image when both e- and o-wave are present in reference beam. (b) Reference beam has e-wave only. (c) Reference beam has o-wave only. The images were captured by a CCD camera, then taken off the monitor by a polaroid camera.

ing. When we read out one of the holograms recorded at a new reference beam angle, the new reference beam can coincide with one of the side-peaks of previous holograms, even though the new reference beam angle is separated from the other reference beam angles by more than several $\Delta\theta_{Bragg}$.

To avoid this problem, we might limit the range of θ and ϕ that we are working with, so that $\Delta\theta_e$ and $\Delta\theta_o$ are larger than the range of reference beam angles that we want to use. In this case, the side-peaks and central-peaks effectively coincide.

As an example, consider a crystal that has thickness of $8mm$, index $n = 2.2$, and $\theta = 30°$. We have $\Delta\theta_{Bragg} \approx 0.01°$ from Eq. (5.79). Taking twice this value as the separation between reference beam angles, storing 1000 holograms would require a range of θ of about 20°. We can calculate $\Delta\theta_e$ and $\Delta\theta_o$ for various values of φ at $\theta = 20°$, $\theta = 30°$, and $\theta = 40°$ (see Appendix B). The result is plotted in Figure 5.15. In lithium niobate we would like to work near $\varphi = 0$, where the c-axis is close to the incident plane, and the diffraction efficiencies are high. $\Delta\theta_e$ and $\Delta\theta_o$ in this region are about 1°, which means that we can place in about $0.5°/0.02° = 25$ holograms before we start hitting the side-peaks. In this example, we would have to record 25 adjacent, angularly multiplexed holograms, and then leave a blank of 25 slots for the cross-gratings. It is in principle possible to use the system in Figure 5.13, to record two separate images in two polarizations in such a way that the cross-gratings that appear in the "empty" slots would actually be new images. With this scheme, there is no loss in storage density, at the expense of considerable added complexity. As φ increases, the number of adjacent holograms that can be recorded before we must leave a gap decreases.

The second alternative is to squeeze the side-peaks as close as possible around the central-peak. As shown in Figure 5.15(b), around $\varphi \approx 44.1°$, all three peaks actually coincide. However, for the separations of the side-peaks and central-peaks to be much smaller than $\Delta\theta_{Bragg}$ throughout the whole working range of θ (about 20° to 40°), we see that we are limited to a range of φ (tilt) of only one or two degrees. This puts a rather strict constraint on system design. In general,

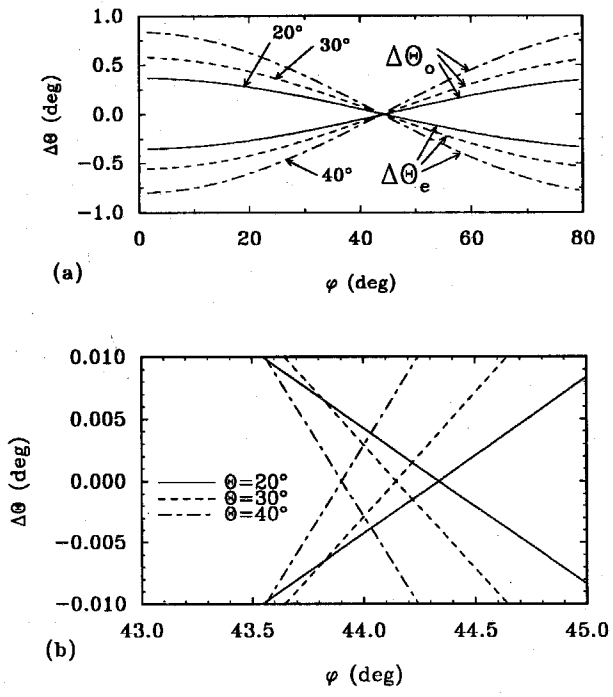


Figure 5.15. (a) Deviation angle of the side-peaks as a function of φ for $\theta = 20°$, $30°$, and $40°$. (b) Enlargement of the region where $\Delta\theta_e$ and $\Delta\theta_o$ are zero.

either we have to limit the number of holograms we can store, or limit the amount of tilting allowed. The example above shows that neither cases is very satisfactory. There is, however, one situation where the side-peak problem does not occur for the range of angles used in angle multiplexing. This is when we go to reflection type holograms, where $\theta > 90°$. In this case, calculations similar to those shown in Figure 5.10 and Figure 5.15 show that the o-wave peak disappears ($\Delta\theta_o \to \infty$) and the e-wave peak is far from the original reference beam angle. The result is plotted in Figure 5.16. For $\varphi = 20°$, $\Delta\theta_o$ is larger than $20°$ for $\theta > 110°$, enough room to store the 1000 holograms in the previous example without hitting the side-peaks.

Another possible way to avoid cross-talk due to double grating formation, is to have at least one of the recording beams (preferably both) being only e-wave or o-wave during recording, and using only one polarization during reconstruction of

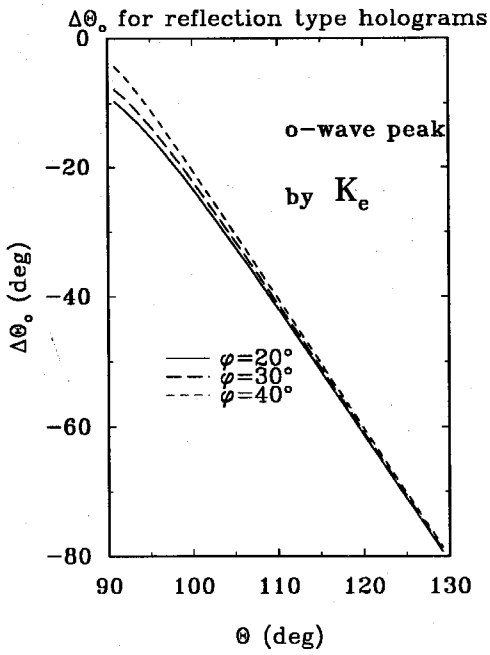


Figure 5.16. Deviation angle of the o-wave side-peak from the recording reference beam angle as a function of θ for $\theta = 20°$, $30°$, and $40°$.

the hologram. In a spatially multiplexed holographic memory, this means that we have to change the polarization as we scan the writing beams from one location to another. If the scanning is done by moving the crystal, we can accomplish this by using circularly polarized waves, and attaching a polarizer in front of the crystal (at the expense of losing half the light).

It should be pointed out that as long as the writing beams are not eigenmodes (e- and o-waves), double grating formation can occur, even when the crystal is not tilted with respect to the incident plane ($\varphi = 0$). In addition, if the signal beam is not a single plane wave but an image consisting of a spectrum of plane waves, then the condition $\varphi = 0$ cannot be satisfied for most of the plane wave components. The results of this section can be used to predict the expected cross-talk for this case as well, but the remedies that we outline above do not necessarily apply.

In some cases, the double grating effect might be of some use instead of a nuisance. As we have shown, by controlling the patterns that are e-wave and o-wave, it is possible to store two images (at the same reference beam angle) in

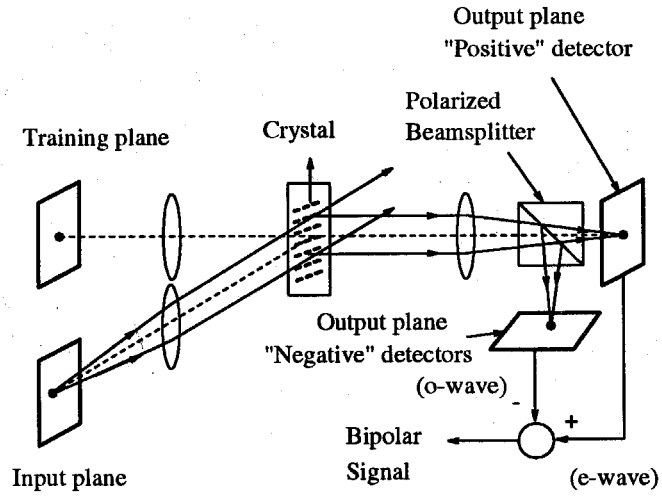


Figure 5.17. Possible setup for implementing bipolar input/output signals in a optical neural network. Signals at both the training and input plane have e-wave and o-wave components.

one single exposure. We can select which image to be readout by controlling the polarization at either the reference beam or reconstructed signal beam. An example of the usage of this double storage scheme is the representation of positive and negative numbers in optical neural net architectures (e.g., perceptron learning [16,17]) by the two orthogonal polarizations. This is shown in Figure 5.17. Here we take the e-wave polarization to represent the "positive" part of the connection, and the o-wave polarization to represent the "negative" part. For each connection from the input plane to the output plane, there is a $\mathbf{K}_e$ and a $\mathbf{K}_o$ grating. Upon readout, the signal from the input plane has both e-wave and o-wave components (this can be done by the same setup used in Figure 5.13). The e-wave component, however, will diffract only off the $\mathbf{K}_e$ grating, while the o-wave component will diffract only off the $\mathbf{K}_o$ grating. The two components in the diffracted light are detected separately, and the difference of the two intensities are taken (electronically) to be the bipolar signal.

5.5. Maximum Diffraction Efficiency

In Section 5.3, the c-axis was chosen as the rotation axis for the 3-D disk. Several questions that immediately come to mind are: how do they compare with that of the usual recording geometry for transmission type holograms (where the c-axis is parallel to the crystal surface)? And how do they compare to the *maximum* obtainable diffraction efficiency for arbitrary cut crystals? What is the geometry that gives us this diffraction efficiency?

In this section, we will try to answer these questions. We will consider only the *3mm* crystals here. The results for *4mm* crystals may be obtained by taking $\alpha = 0$.

5.5.1. Heuristic Argument

Before we present the results from the numerical calculations of the diffraction efficiency, it is instructive to first look at the problem from a more heuristic point of view. Given the electrooptic coefficient tensor r, we would like to know what geometry gives us strong diffraction efficiency.

In order to get large diffraction efficiency from a photorefractive crystal, the polarization and geometry should be such that the factor $\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1$ in Eq. (5.52) is as large as possible. To be precise, we should also consider the other coefficients in G also. However, the dominant variation comes from $\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1$ and $\mathbf{e}_1 \cdot \mathbf{e}_2$. In our estimate, we will concentrate only on $\mathbf{d}_2 \cdot (r\mathbf{u})\mathbf{d}_1$, and later compare this to the numerical results that include all the other factors.

As a crude estimate, we consider the non-zero coefficients of r one at a time while ignoring the others. If the crystal has a dominant coefficient, then the result should be a good indication of what is needed. For example, to make full use of the coefficient $r_{42} = r_{yzy}$ (the largest in lithium niobate), we need the grating vector $\mathbf{K}$ to be parallel to the y-axis, [9] and the polarization vectors $\mathbf{d}_1$

[9] Recall that $\mathbf{K}$ is parallel to $\mathbf{u}$.

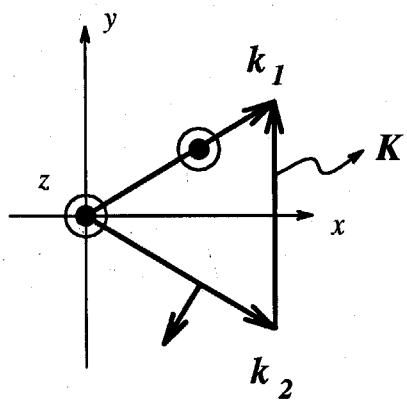
and $\mathbf{d}_2$ to be parallel to the y-axis and z-axis. We also require the wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ to be perpendicular to the polarization vectors, and satisfy the relation $\mathbf{k}_2 - \mathbf{k}_1 = \mathbf{K}$. After a little thought, it is obvious that the geometry that uses r_{42} most efficiently is the configuration shown in Figure 5.18(a), where the crystal is x-cut, and transmission type holograms are recorded. Note that in this case the grating vector $\mathbf{K}$ is *perpendicular* to the c-axis instead of parallel to it.

The other coefficients in the electrooptic coefficient tensor for *3mm* crystals may be explored in the same fashion, and the results are summarized in Figures 5.18(a) through (f). Because of the symmetry properties of the crystal, many of the coefficients are the same, and the results for those that are not listed may be obtained simply by exchanging the axes. For example, since $r_{42} = r_{yzy} = r_{51} = r_{xzx}$, we may exchange the x and y-axes.

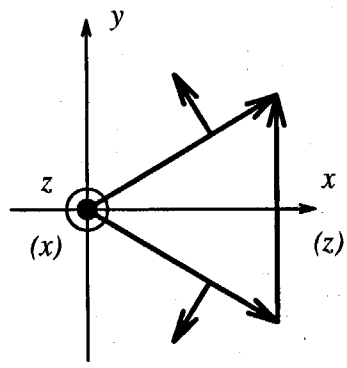
Note also that (by symmetry) $\mathbf{k}_1$ and $\mathbf{k}_2$ may be exchanged without affecting the value of G.

The largest coefficient for lithium niobate is r_{42}. Although the geometry shown in Figure 5.18(a) is expected to yield large diffraction efficiency, it is rarely used in practice for recording images. The main problem is that the polarization vectors are perpendicular to each other. As mentioned earlier in Section 5.3, this creates problems for recording and reconstructing holograms because of the difficulty in simultaneously Bragg-matching all spatial frequency components. The next largest coefficient for lithium niobate is r_{33}, which can be used in the configuration shown in Figure 5.18(b). This is of course the geometry most commonly used.

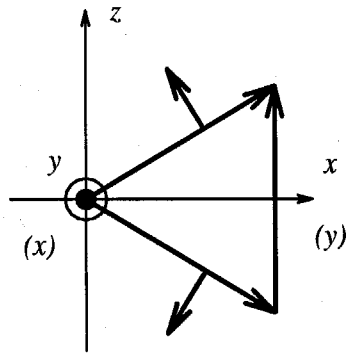
The configurations shown in Figure 5.18 are all suitable for recording transmission holograms, where the angle between the two wave vectors are small. To record reflection holograms (increasing the angle between the wave vectors in the figures to near 180 degrees), the geometries in Figure 5.18(c) and (e) are more efficient. Of course, we need to take into account the value of the r coefficients also. Although in lithium niobate r_{42} and r_{33} are larger than r_{13} (by a factor of about 4), it is wasted because of the angles necessary to have reflection geometry.



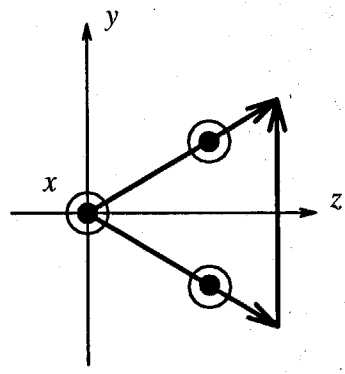
(a) $\mathbf{r}_{42}$



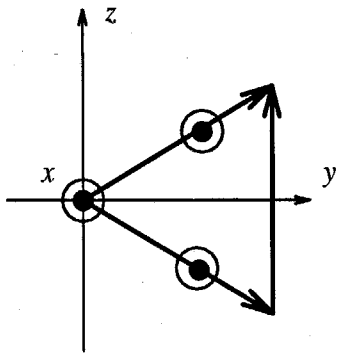
(d) $\mathbf{r}_{22}$



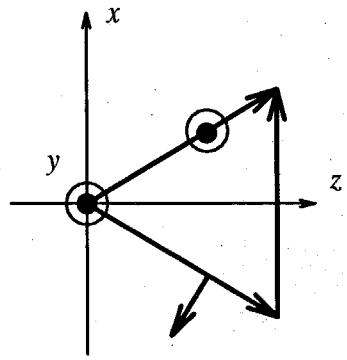
(b) $\mathbf{r}_{33}$



(e) $\mathbf{r}_{12}$



(c) $\mathbf{r}_{13}$



(f) $\mathbf{r}_{61}$

Figure 5.18. Best recording geometry for various electrooptic coefficients.

As a result, there is not much advantage in trying to use those instead of r_{33}. This is confirmed by the results shown in Figures 5.1(a)–(d). [10]

For barium titanate and SBN, the coefficient r_{22}, r_{12}, and r_{61} are zero. The dominant coefficient in barium titanate are $r_{42} = r_{51}$, and for SBN, it is r_{33}. The qualitative conclusions for lithium niobate hold more or less for these crystals also. However, the geometry that works best for reflection type holograms in barium titanate is not (c) (o-o coupling), but turns out to be a modification of (b) (e-e coupling). From Figure 5.2, the diffraction efficiency is larger if we take larger ϕ_1 and record in the reflection geometry (where $\phi_2 - \phi_1$ is close to 180 degrees). The reason is because r_{13} is much smaller than r_{42}.

5.5.2. Numerical Results

The G factor given by Eq. (5.52) can be calculated readily following the procedure listed in Section 5.3. It is not easy to find the maximum of G since there are many degrees of freedom ($\mathbf{k}_1$ and $\mathbf{k}_2$ can be in any direction) and also many implicit constraints (the triangle relation $\mathbf{k}_2 - \mathbf{k}_1 = \mathbf{K}$ needs to be satisfied, and the polarization vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ need to be eigenmodes).

Because of the complexity, no analytic solution was attempted, and an exhaustive search was done on computer to find the maximum G value. The search was done in several steps. We first picked a polarization coupling (e-e, o-o, e-o, or o-e), and for a fixed direction of $\mathbf{k}_1$, we allowed $\mathbf{k}_2$ to vary (in increments of 0.9° in θ and ϕ, where θ is the angle to the z-axis). The maximum G was then found, and the procedure was repeated for a different $\mathbf{k}_1$. We then compared the maximum G's associated with each $\mathbf{k}_1$, and picked out the largest value. The increments in $\mathbf{k}_1$ was set at 1° for the θ_1 angle, and 1.5° for ϕ_1 (Eq. (5.53)).

The whole procedure was done for e-e, e-o, and o-o coupling. It was not necessary to do the o-e coupling because from symmetry, it is equivalent to the

---

[10] The geometry for reflection type holograms occurs in the region in Figures 5.1(a)–(d) near $\phi_2 - \phi_1 = 180°$.

e-o coupling. The crystal symmetry properties were used to reduce the amount of search needed. For example, since lithium niobate has the 3-fold symmetry, it was only necessary to search through 120° of the θ_1 angle instead of the full 360°. (Actually, it is only necessary to search through 60°, since in addition to the 3-fold rotation symmetry, there is also a mirror symmetry.) Although the symmetry properties greatly reduced the amount of parameter space that needed to be covered, the whole process still required considerable computation time. [11]

In these calculations, it has been assumed that there is no photovoltaic effect and external applied field. As noted earlier, this is not always valid. Note that unless the E_{1v} component in Eq. (5.37) is non-zero, the space-charge field is still parallel to the grating vector $\mathbf{K}$. Thus the effect of ignoring the photovoltaic effect and external applied field on the diffraction efficiency is through the space-charge field E_{sc}.

The results for the numerical calculations will be presented by giving the directions of the wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ in θ and ϕ. The polarizations may be found according to the method outlined in Section 5.3.

The results are as follows: (the parameters used in the calculations are the same as those used in Section 5.3)

Lithium Niobate:

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

$$\theta_1 = 90°,\ \phi_1 = 138.6°,\ \theta_2 = 90°,\ \phi_2 = 118.8° \quad \Longrightarrow \quad G_{max} = 758\ m^{-1}.$$

In this case the wave vectors and polarization vectors all lie in the y–z plane, and the grating vector $\mathbf{K}$ is approximately at 39° with the c-axis. The wave vectors are 48.6° and 28.8° with the y-axis, and the angle between them is about 20°. (All these angles are inside the crystal.) Because we are neglecting

---

[11] The modulation depth factor $\mathbf{e}_1 \cdot \mathbf{e}_2$ in this case was not discarded in calculating the results for e-o coupling.

the photovoltaic effect, we expect that the maximum diffraction efficiency to occur when the direction of the grating vector $\mathbf{K}$ is closer to parallel to the c-axis. In this case, the angles ϕ_1 and ϕ_2 are closer to 90°.

2. e-o coupling: ($\mathbf{k}_1$ is e-mode and $\mathbf{k}_2$ is o-mode)

$$\theta_1 = 30.5°,\ \phi_1 = 27°,\ \theta_2 = 120.1°,\ \phi_2 = 0.41° \quad \Longrightarrow \quad G_{max} = 344\ m^{-1}.$$

For e-o coupling, the wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$ and the z-axis do not lie in the same plane. The wave vectors are almost perpendicular to each other. However, their polarization vectors are not since $\mathbf{k}_1$ is e-mode polarized while $\mathbf{k}_2$ is o-mode polarized. (They lie in the plane formed by the z-axis and $\mathbf{k}_1$.) It should be noted that the result here is degenerate: the $\mathbf{k}_2$ wave vector is also (approximately) an extraordinary wave. Therefore the configuration may also be considered as e-e coupling. It turns out that this is true also for barium titanate and SBN (to be presented below).

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

$$\theta_1 = 90°,\ \phi_1 = 88.5°,\ \theta_2 = 90°,\ \phi_2 = 59.7° \quad \Longrightarrow \quad G_{max} = 303\ m^{-1}.$$

For o-o coupling, the wave vectors again lie in the y–z plane, but the polarization vectors are in the x direction. The $\mathbf{K}$ vector is approximately 14° with the c-axis, and the angle between the wave vectors is about 29°.

Because of the 3-fold symmetry of *3mm* crystals, we obtain the same diffraction efficiency if we rotate around the z-axis by 120°. In addition, there are other mirror symmetry planes and inversion symmetries. Thus the same maximum diffraction efficiencies can be obtained if we change angles according to these symmetries.

From the results above, the maximum diffraction efficiency is obtained for e-e coupling with a crystal cut of approximately 50°. The polarizations are "in-plane." As mentioned before, since we have neglected the photovoltaic effect, the

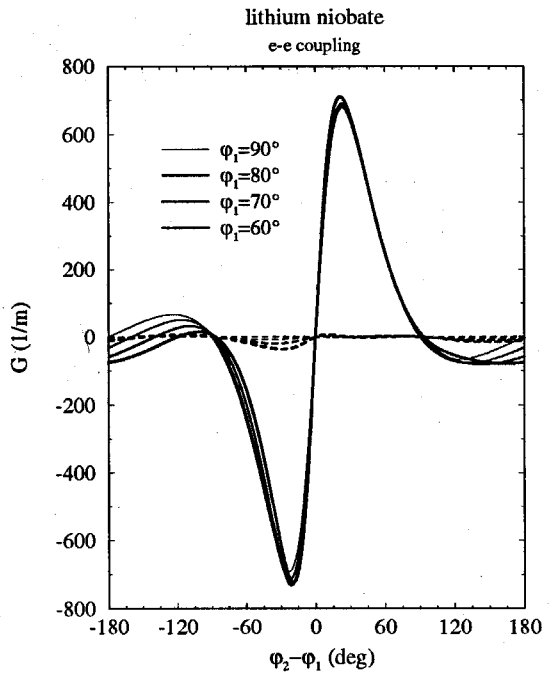


Figure 5.19. The average and variation of G near the conventional recording geometry in lithium niobate. The holograms are recorded in the transmission geometry using e-e coupling.

actual optimum cut should be somewhat smaller than 50°. For practical purposes, a 45° cut should be used.

As a comparison to the commonly used geometry, where the c-axis is parallel to the crystal surface, we plot the results shown in Figure 5.19 for lithium niobate. In this case, the signal beam angle ϕ_1 is approximately perpendicular to the c-axis. Note that the variation in G with respect to rotation around the z-axis is relatively small. Thus the difference between the x-cut and y-cut is not significant. Note, however, that the diffraction efficiency in this case is not significantly smaller than the maximum value calculated above.

The maximum G for lithium niobate is 758 m^{-1}. In the 3-D disk system using lithium niobate, the recommended recording configuration using o-o coupling and reflection type geometry gives us about $G = 130\ m^{-1}$. This is about 17% of the maximum G. Although this is a significant reduction in diffraction efficiency, one

benefit is the fact that fanning does not occur along the direction of the signal beam, since the maximum gain is not along this direction.

For *4mm* crystals, there is invariance (as far as diffraction efficiency is concerned) with respect to rotation around the c-axis. This can be seen by considering a coordinate transform for such a rotation. It can be shown that the electrooptic tensor and the permittivity tensors do not change upon such rotations. Thus in the results below, the same diffraction efficiencies can be obtained for any rotation angle around the c-axis.

Barium Titanate:

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

$$\phi_1 = 46.5^\circ,\ \phi_2 = 42.9^\circ,\ \theta_1 = \theta_2 \quad \Longrightarrow \quad G_{max} = 2365\ m^{-1}.$$

   In this case, the wave vectors and the c-axis lie in the same plane. The angle between the wave vectors are about 4°, and the angle between the grating vector $\mathbf{K}$ and the c-axis is about 45° [19]. The reason the optimum angle between the wave vector is so small is because the permittivity is much larger than for lithium niobate. Because of this, the denominator term in E_{sat} quickly becomes large when $\mathbf{K}$ increases. The same effect can be seen also from Figure 5.2.

2. e-o coupling: ($\mathbf{k}_1$ is e-mode and $\mathbf{k}_2$ is o-mode)

$$\phi_1 = 13.5^\circ,\ \phi_2 = 0.2^\circ,\ \theta_2 - \theta_1 = 270.4^\circ \quad \Longrightarrow \quad G_{max} = 196\ m^{-1}.$$

   The polarization vectors both lie approximately in the x–z plane, and the angle between them is about 13.5°. The wave vectors are both approximately parallel to the c-axis, so the grating vector is perpendicular to it. The angle between the wave vectors is about 14°.

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

$$\phi_1 = 82.5^\circ,\ \phi_2 = 97.8^\circ,\ \theta_2 = \theta_1 \quad \Longrightarrow \quad G_{max} = 153\ m^{-1}.$$

In this case we have the co-planar geometry again, and the wave vectors are approximately perpendicular to the c-axis. In this case, however, the angle between the wave vectors is slightly larger (about 15°).

The optimum crystal cut for barium titanate is therefore the well known 45° cut. In practice, however, the optical gain is so large for this crystal cut that there is serious fanning problem. For lithium niobate, the diffraction efficiency using the conventional geometry is not much smaller than that of the maximum value. For barium titanate, however, this is not true. It turns out that there is very significant increase in diffraction efficiency if the grating vector moves towards the 45° angle between the c-axis. The result is shown in Figure 5.20.

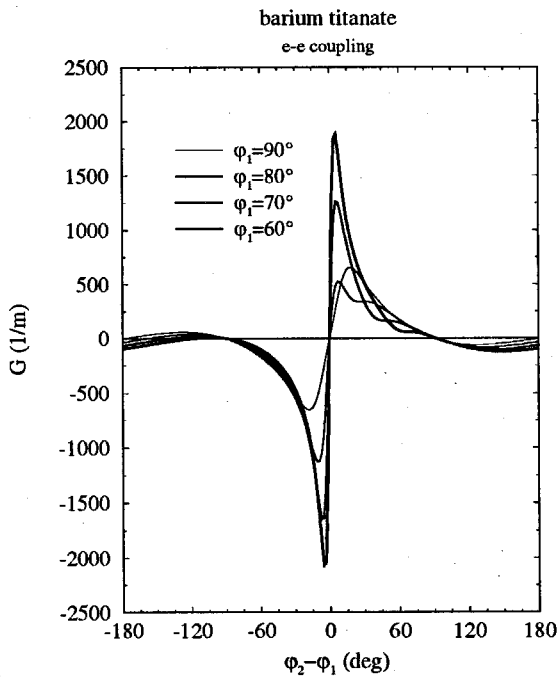


Figure 5.20. The values of G near the conventional recording geometry in barium titanate. The holograms are recorded in the transmission geometry using e-e coupling.

The results for SBN are summarized below:

SBN:

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

$$\phi_1 = 93^\circ,\ \phi_2 = 87.6^\circ,\ \theta_2 = \theta_1 \quad \Longrightarrow \quad G_{max} = 1066\ m^{-1}.$$

   We have the co-planar geometry here. The wave vectors are approximately perpendicular to the c-axis, and the angle between the wave vectors is about 5°. The reason for the small angle is again because of the large permittivity.

2. e-o coupling: ($\mathbf{k}_1$ is e-mode and $\mathbf{k}_2$ is o-mode)

$$\phi_1 = 18^\circ,\ \phi_2 = 0^\circ,\ \theta_2 - \theta_1 = 90.12 \quad \Longrightarrow \quad G_{max} = 166\ m^{-1}.$$

   The wave vectors in this case are almost parallel to the c-axis, with an angle of approximately 18° between them. The polarization vectors both lie in the x–z plane, and the angle between them is about 18°.

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

$$\phi_1 = 87^\circ,\ \phi_2 = 92.4^\circ,\ \theta_1 = \theta_2 \quad \Longrightarrow \quad G_{max} = 274\ m^{-1}.$$

   We have the co-planar geometry here. The wave vectors are approximately perpendicular to the c-axis, and the angle between the wave vectors is also about 5°.

5.5.3. Results for the 90 Degree Recording Geometry

There has been interest in recent years of recording holograms using the 90 degree geometry. It can be shown [18] that the expression for G is valid also for the 90 degree recording geometry if we take $\theta_1' = 0$ in Eq. (5.48).

From Figure 5.18, the configuration that work for reflection type holograms also work for the 90 degree geometry. These are the configurations of Figures 5.18(c) and (e). For lithium niobate, r_{13} is slightly more than 1/4 of the maximum coefficient, r_{42}, thus we expect a configuration similar to Figure 5.18(c)

to work well. This agrees with the result shown in Figure 5.1(a) (o-o coupling), where the maximum at $\phi_2 - \phi_1 = 90°$ is near $\phi_1 = 45°$. This is also confirmed by numerical result, where we have

Lithium Niobate:

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

   $$\theta_1 = 32°,\ \phi_1 = 52.5°,\ \theta_2 = 86.5°,\ \phi_2 = 127.1° \quad \Longrightarrow \quad G_{max,90} = 54.9\, m^{-1}.$$

2. e-o coupling: ($\mathbf{k}_1$ is e-mode and $\mathbf{k}_2$ is o-mode)

   $$\theta_1 = 0°,\ \phi_1 = 0°,\ \theta_2 = 90°,\ \phi_2 = 90° \quad \Longrightarrow \quad G_{max,90} = 133\ m^{-1}.$$

   ($\mathbf{k}_1$ is degenerate: it is also an ordinary wave. Thus this is also an o-o coupling.)

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

   $$\theta_1 = 90°,\ \phi_1 = 34.5°,\ \theta_2 = 90°,\ \phi_2 = 124.5° \quad \Longrightarrow \quad G_{max,90} = 192\ m^{-1}.$$

   (Thus for o-o coupling $\mathbf{k}_1$ and $\mathbf{k}_2$ are both in the y-z plane).

For lithium niobate, the best result is obtained by using o-o coupling, with a crystal cut at 34.5°. If we had used a 45°-cut crystal instead of the optimum 34.5°, we would have gotten $G = 179.8\ m^{-1}$, which is 94% of the maximum value for 90 degree geometries. In the calculations above, it was assumed that there is no photovoltaic effect. In reality the photovoltaic effect is significant in lithium niobate crystals. Because of this, we expect an increase in diffraction efficiency when $\mathbf{K}$ is closer to parallel to the c-axis. Thus the 45°-cut (which should make $\mathbf{K}$ exactly parallel to the c-axis) should work even better than indicated by the calculations shown above.

Barium Titanate

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

   $$\theta_1 = 0°,\ \phi_1 = 61.5°,\ \theta_2 = 73.2°,\ \phi_2 = 118.1° \quad \Longrightarrow \quad G_{max,90} = 63.73\, m^{-1}.$$

2. e-o coupling: ($\mathbf{k}_1$ is e-mode and $\mathbf{k}_2$ is o-mode)

$$\theta_1 = 0°,\ \phi_1 = 51°,\ \theta_2 = 37.9°,\ \phi_2 = 134.3° \quad \Longrightarrow \quad G_{max,90} = 97.29\ m^{-1}.$$

(In this case, however, none of the wave vectors are degenerate.)

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

$$\theta_1 = 0°,\ \phi_1 = 45°,\ \theta_2 = 0°, \phi_2 = 135° \quad \Longrightarrow \quad G_{max,90} = 56.18\ m^{-1}.$$

(Thus for o-o coupling $\mathbf{k}_1$ and $\mathbf{k}_2$ lie in the x-z plane.)

SBN:

1. e-e coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are e-mode)

$$\theta_1 = 0°,\ \phi_1 = 69°,\ \theta_2 = 81.6°,\ \phi_2 = 111° \quad \Longrightarrow \quad G_{max,90} = 60.74\ m^{-1}.$$

2. e-o coupling: ($\mathbf{k}_1$ is e-mode, $\mathbf{k}_2$ is o-mode)

$$\theta_1 = 0°,\ \phi_1 = 0°,\ \theta_2 = 270°,\ \phi_2 = 90° \quad \Longrightarrow \quad G_{max,90} = 33.57\ m^{-1}.$$

($\mathbf{k}_1$ is again degenerate: it is also an ordinary wave.)

3. o-o coupling: (both $\mathbf{k}_1$ and $\mathbf{k}_2$ are o-mode)

$$\theta_1 = 0°,\ \phi_1 = 75°,\ \theta_2 = 0°,\ \phi_2 = 165° \quad \Longrightarrow \quad G_{max,90} = 34.28\ m^{-1}.$$

Compared to the results for lithium niobate, the maximum G for barium titanate and SBN turns out to be *less* than that of lithium niobate. Part of the reason is because the dielectric constants of these crystals are larger, and therefore E_{sat} is smaller when the magnitude of the grating vector is larger. For $BaTiO_3$, the maximum G occurs for e-o coupling, while for SBN, the maximum G occurs for e-e coupling. Note, however, that these results require peculiar crystal cuts which are more difficult to fabricate. Another problem is getting the eigenmode polarizations correct.

For more reasonable orientations for SBN, consider the case where $\mathbf{k}_1$ is required to be parallel to the y-axis ($\theta_1 = \phi_1 = 90°$). Using e-e coupling and allowing $\mathbf{k}_2$ to vary (but perpendicular to $\mathbf{k}_1$) the best we can get is $G = 45.79\ m^{-1}$, which occurs at $\theta_2 = 0°$, $\phi_2 = 59.4°$. This is only 75% of the maximum (but is still larger, however, than the maximum G obtainable using o-o or o-e coupling).

For barium titanate, the situation is more awkward, since the maximum G occurs for o-e coupling. To obtain a simpler orientation, if we try having $\mathbf{k}_1$, $\mathbf{k}_2$ and z-axis lie on the same plane (i.e., the co-planar geometry), then it turns out that for e-o coupling, G is zero (Eq. (5.67) or (5.70)). If we use o-o or e-e coupling instead, the best we can get is 57.7% of the optimum G if we use o-o coupling, or 65.6% of the optimum G if we use e-e coupling.

It is interesting to compare the maximum diffraction efficiency for the 90 degree recording geometry with the results for the conventional transmission geometry. For lithium niobate, the diffraction efficiencies are only slightly lower than the diffraction efficiencies obtained for the reflection geometry using o-o coupling (about 17% of the maximum). For barium titanate and SBN, however, the results are significantly lower than that of the conventional recording geometry.

5.6. Discussions and Conclusions

In this chapter, we have analyzed the diffraction efficiency of photorefractive crystals for use in a 3-D holographic disk system. The c-axis was chosen to be the rotation axis because it is the highest symmetry axis. One might ask, however, whether using the c-axis as the rotation axis is the best choice. With transmission holograms, the gratings formed inside the crystal is almost perpendicular to the c-axis, and the diffraction efficiency is low. If we record in the reflection geometry, the gratings are closer to parallel to the c-axis. However, in this case, the magnitude of the grating vector is larger, and the K^2 term in the denominator in the expression for E_{sat} (Eq. (5.1)) causes the diffraction efficiency to be lower. If we use the x-axis or the y-axis as the rotation axis, the diffraction efficiency is much

better. However, now there is also larger variation in the diffraction efficiency as the disk rotates.

In Section 5.3, the criterion for a desirable recording geometry was one that was high in diffraction efficiency, but preferably changes little as the angle of the reference beam changes and the disk rotates. The c-axis is a natural choice because it has the highest axis of symmetry. From the point of view of storing as many holograms as possible, however, it might be argued that we should look at the *total* amount of diffraction efficiency. Thus even if there are some locations on the disk that yield lower diffraction efficiency, it is preferable as long as there are locations on the disk that have higher diffraction efficiency to compensate for the low efficiency areas. In this case, using the x or y-axes as rotation axis would be a better choice. The tradeoff of course, is that the control of the hologram recording schedule [6] to get uniform diffraction efficiency of the individual holograms is more complicated because of the larger variation with respect to disk rotation and reference beam angle change.

If we use the number of holograms as the figure of merit, then we are concerned with the total amount of diffraction efficiency (measured by summing or integrating the diffraction efficiency over all rotation angles). As shown in Section 5.5, the geometry that gives the highest diffraction efficiency turn out to be the co-planar geometry for all of the three crystals we have discussed. For obtaining the maximum total amount of diffraction efficiency, however, it is not obvious that the x or y-axes are the optimum choice. More work needs to be done to answer this question.

For lithium niobate, it has been shown that using the c-axis as rotation axis and recording reflection holograms with o-o coupling gives us a diffraction efficiency of about 17% of the maximum obtainable diffraction efficiency. Although there is a significant loss in diffraction efficiency, there are several advantages. One is that the problem of double gratings is easily avoided, since having the polarization vectors perpendicular to the incident plane (the plane that contains the signal and reference beams) will guarantee that we have o-o coupling regardless

of how the reference beam angle changes or how the disk rotates. The second advantage is that fanning does not occur in the direction of recording.

If the loss of the diffraction efficiency using the c-axis is not acceptable for implementing a 3-D holographic disk system, then we need to consider other crystal orientations. The disadvantage is that the variation in diffraction efficiency with respect to disk rotation and reference beam angles is larger. Polarization is also a problem, since we would like to avoid having double gratings. A solution might be to use sheet polarizers glued on to the crystal surface and use circularly polarized light, however such solutions do not work very well. An alternative to the problem of uniformity versus diffraction efficiency is to use not a single piece of crystal, but to put together several pieces into a single disk. This takes care of the problem with crystal orientation and light polarization, and also allows us to have larger 3-D disks. The disadvantage is that the crystal needs to be cut to the appropriate fan shapes, and then assembled into a single piece.

Appendix

A. Derivation of Slant Angle

Figure 5.21 shows the geometry used in our experiment. The signal beam $\mathbf{k}_1$ propagates along the +x direction. Its e-wave polarization is in the z direction, and its o-wave polarization is in the y direction. For the reference beam, $\mathbf{k}_2$, the eigenmode polarizations are $\mathbf{q}$ (e-wave) and $\mathbf{r}$ (o-wave). It can be shown that $\mathbf{k}_2$, $\mathbf{q}$ and $\mathbf{r}$ are perpendicular to each other. Also, $\mathbf{r}$ lies in the x-y plane. To see why this is true, consider rotating $\mathbf{k}_2$ about the z-axis until $\mathbf{k}_2$ lies in the x-z plane. $\mathbf{q}$ and $\mathbf{r}$ will now rotate with $\mathbf{k}_2$. But when $\mathbf{k}_2$ lies in the x-z plane, it is readily seen that $\mathbf{r}$ is parallel to the y-axis. This shows that $\mathbf{r}$ lies in the x-y plane. Given this fact, we see that $\mathbf{q}$ lies in the plane formed by the z-axis and $\mathbf{k}_2$. We are interested in the angle between $\mathbf{q}$ and $\mathbf{p}$, the direction perpendicular to the incident plane formed by $\mathbf{k}_2$ and $\mathbf{k}_1$.

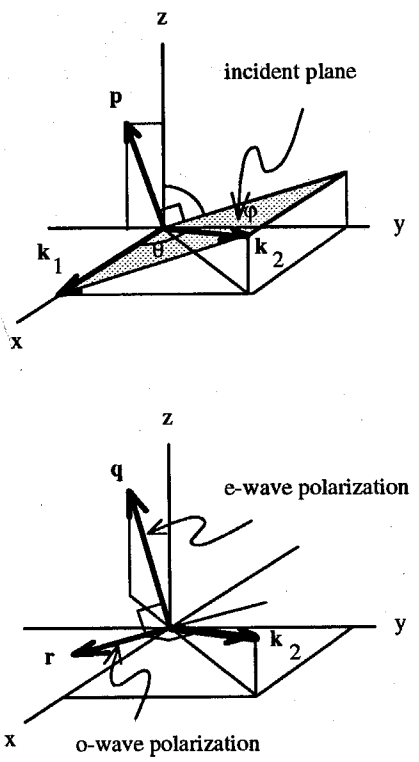


Figure 5.21. (a) Geometry of reference ($\mathbf{k}_2$) and signal ($\mathbf{k}_1$) beams with respect to crystal axes. (b) Reference beam and its two eigenmode polarization directions, $\mathbf{q}$ (e-wave) and $\mathbf{r}$ (o-wave).

Note that since $\mathbf{r}$ (the o-wave polarization of $\mathbf{k}_2$, the reference beam) lies in the x-y plane, and the e-wave polarization of $\mathbf{k}_1$ (the signal beam) is in the z direction, they are perpendicular to each other.

We have

$$\mathbf{p} = (0, -\cos\varphi, \sin\varphi) \tag{5.80}$$

Thus the slant angle θ_{slant} is

$$\theta_{slant} = \cos^{-1}(\mathbf{u} \cdot \mathbf{p}) = \cos^{-1}(-u_y \cos\varphi + u_z \sin\varphi) \tag{5.81}$$

where, $\mathbf{u}$ is the unit vector parallel to $\mathbf{q}$. It is easy to show that within a proportionality constant

$$q_x = -\sin\theta \ \cos\theta \ \cos\varphi \tag{5.82}$$

$$q_y = -\sin^2\theta\ \sin\varphi\ \cos\varphi \tag{5.83}$$

$$q_z = \cos^2\theta + \sin^2\theta\ \sin^2\varphi. \tag{5.84}$$

This is all we need to calculate the slant angle for various values of θ and φ.

B. Bragg Matching Angle

The problem of finding the Bragg matching angle, given a grating vector, can be stated as follows: given a normal surface (in 2-D) described by the ellipse Γ

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{5.85}$$

and a grating vector

$$\mathbf{K} = (K_x, K_y) \tag{5.86}$$

we want to find (x, y) such that both (x, y) and $(K_x + x, K_y + y)$ both lie on Γ.

We have

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{5.87}$$

and

$$\frac{(K_x + x)^2}{a^2} + \frac{(K_y + y)^2}{b^2} = 1 \tag{5.88}$$

which gives us

$$(\frac{2K_x}{a^2})x + (\frac{2K_y}{b^2})y + (\frac{K_x}{a})^2 + (\frac{K_y}{b})^2 = 0. \tag{5.89}$$

Let

$$u = \frac{x}{a}, \qquad v = \frac{y}{b} \tag{5.90}$$

and

$$A = \frac{K_x}{a}, \qquad B = \frac{K_y}{b}, \tag{5.91}$$

then

$$u^2 + v^2 = 1 \tag{5.92}$$

and

$$2Au + 2Bv + A^2 + B^2 = 0. \tag{5.93}$$

There are two solutions:

$$x = \frac{K_x}{2} \mp \frac{K_y}{2}\left(\frac{a}{b}\right)\sqrt{\frac{4}{A^2+B^2} - 1} \tag{5.94}$$

$$y = \frac{K_y}{2} \mp \frac{K_x}{2}\left(\frac{b}{a}\right)\sqrt{\frac{4}{A^2+B^2} - 1}. \tag{5.95}$$

From these, we can calculate the angle between (x, y) and the x-axis.

Given the recording geometry of the crystal, we can find $\mathbf{K}_e$ and $\mathbf{K}_o$ from θ and φ (Eq. (5.76) and Eq. (5.77)). The problem then becomes finding the Bragg matching angle for $\mathbf{K}_o$ on the e-wave normal surface, and finding the Bragg matching angle for $\mathbf{K}_e$ on the o-wave normal surface (see Figure 5.8). Once these are found, the corresponding incident angles using Snell's law can be found. We then subtract them from θ to obtain $\Delta\theta_e$ and $\Delta\theta_o$.

C. Traveling Gratings for Recording Multiple Holograms in Photorefractive Crystals

For angle multiplexed volume holograms, different holograms can be read out by varying the reference beam angle. The speed of accessing different locations depends on the time needed to change the reference beam angle. Using acoustooptic (AO) deflectors, we can achieve transit times of 1 to 10 μsec. However these devices also Doppler-shift the frequency of the deflected light, and unless the frequency of the signal beam is also changed to the same frequency as the reference beam, the interference pattern that forms inside the crystal will not be stationary, but will be "moving" or "traveling."

The theory of moving gratings inside photorefractive crystals has been developed by Valley [20] and Refregier *et al.* [21]. It is found that when the frequency of the running grating is matched with the characteristics of the crystal, it is possible to achieve a resonance effect and enhance the diffraction efficiency of the

holograms. However, although the saturation diffraction efficiency is enhanced, the time constant of growth is also affected. For multiple holograms, it turns out that the effects cancel each other, and do not increase the diffraction efficiency of individual holograms. In this appendix, we will give the derivation of these results.

Consider the interference pattern formed by two plane waves with frequency difference of ω. The pattern is of the form

$$I = I_0 + I_1 e^{-i(Kx-\omega t)} + c.c., \tag{5.96}$$

where K is the grating vector. The interference grating moves with velocity $v = \omega/K$ (in the $+x$ direction). We will assume that $E_{1v} = 0$. Following the same line of argument as described in Section 5.1, the space-charge field E_1 inside the photorefractive crystal due to the interference pattern in Eq. (5.96) is governed by (cf. Eqs. (5.17) and (5.32))

$$\tau \frac{dE_1}{dt} = -E_1 + imf' E_N \, e^{-i\omega t}, \tag{5.97}$$

where m is given by Eq. (5.31), τ is given by Eq. (5,33), E_N is given by Eq. (5.28), and f' is given by

$$f' = \frac{E_d + i(E_{0u} + E_{ph,u})}{E_N + E_d + i\left(E_{0u} + \frac{N_A}{N_D} E_{ph,u}\right)}, \tag{5.98}$$

(cf. Eq. (5.34)). τ is the time constant of the grating development when $\omega = 0$. In general, it is a complex number of the form

$$\frac{1}{\tau} = \frac{1}{\tau_g} + i\omega_g, \tag{5.99}$$

where τ_g and ω_g are real numbers. As described in this Chapter, the induced change in index of refraction is proportional to the space-charge field, and thus we get a phase grating in the photorefractive crystal. The diffraction efficiency can then be calculated, and for weak holograms it is proportional to the square of the magnitude of E_1.

Assuming zero initial condition (when we first record a hologram), Eq. (5.97) gives us

$$E_1 = imE'_N f\left(1 - e^{t/\tau'}\right) e^{-i\omega t} \quad , \tag{5.100}$$

where

$$\tau' = \frac{\tau}{1 - i\omega\tau} \quad , \tag{5.101}$$

and

$$E'_N = \frac{E_N}{1 - i\omega\tau} \quad . \tag{5.102}$$

The magnitude of the steady-state E_1 is modified by a factor of

$$\left|\frac{1}{1 - i\omega\tau}\right| = \sqrt{\frac{1 + \omega_g^2\tau_g^2}{1 + (\omega - \omega_g)^2\tau_g^2}} \quad , \tag{5.103}$$

while the time constant has changed to give

$$\frac{1}{\tau'} = \frac{1}{\tau_g} + i(\omega_g - \omega). \tag{5.104}$$

For a decaying hologram, the time constant is just τ_g.

For a single hologram, the quantity in Eq. (5.103) enhances the saturation diffraction efficiency by a factor of

$$\sqrt{1 + \omega_g^2\tau_g^2} \tag{5.105}$$

when $\omega = \omega_g$ (resonance). For multiple holograms, however, the final diffraction efficiency depends not only on the saturation space-charge field magnitude, but also on the rise time of the space-charge field.

For multiple holograms, we want to adjust the exposure times of the holograms such that all the holograms come out to have the same strength at the end of the recording. The exposure schedule can be worked out as follows. Assume that at a certain point, the previously recorded holograms all have the same strength $mE_N fA$. The next hologram should be written such that

$$\left|\frac{1}{1 - i\omega\tau}\right| \left|1 - e^{-t/\tau_g} e^{-i(\omega_g - \omega)t}\right| = Ae^{-t/\tau_g}. \tag{5.106}$$

Solving for t gives the recording time needed for the next hologram. In this way, given an initial exposure time, all subsequent exposure times can be found. Eq. (5.106) can not be solved analytically, but is not difficult to find the solution numerically once the various parameters are given.

One case where Eq. (5.106) can be solved exactly is when $\omega = \omega_g = 0$ [6]. In this case, it can be shown that the magnitude of E_1 drops as $mE_N f/N$, where N is the number of holograms recorded. For $\omega_g \neq 0$, but $\omega = 0$ (i.e., complex time constant without traveling gratings), Eq. (5.106) becomes

$$\left|1 - e^{-t/\tau_g} e^{-i\omega_g t}\right| = Ae^{-t/\tau_g}, \tag{5.107}$$

and the asymptotic behavior can be shown to be

$$|E_{1,N}| \rightarrow \frac{\sqrt{1+\omega_g^2\tau_g^2}}{N} mE_N f, \tag{5.108}$$

where the subscript '$1, N$' denotes the N-th hologram recorded. The magnitude of E_1 turns out to be enhanced by the factor in Eq. (5.105). Intuitively, the reason for this is that the slope of $|E_1/mE_N f|$ at $t = 0$ is the quantity shown in Eq. (5.113), whereas the time constant for decay is always τ_g. When the slope at $t = 0$ increases, it takes a shorter time for the new hologram to reach the value of the previously recorded holograms.

When $\omega \neq 0$ (traveling gratings), the inverse of the time constant changes from Eq. (5.98) to Eq. (5.104). If we ignore the $|1 - i\omega\tau|^{-1}$ factor in Eq. (5.106), then the enhancement factor would change from $\sqrt{1+\omega_g^2\tau_g^2}$ to

$$\sqrt{1+\tau_g^2(\omega_g - \omega)^2}. \tag{5.109}$$

However, because of the $|1-i\omega\tau|^{-1}$ factor, the enhancement for recording multiple holograms is actually

$$\left|\frac{1}{1-i\omega\tau}\right| \sqrt{1+\tau_g^2(\omega_g-\omega)^2} = \sqrt{1+\tau_g^2\omega_g^2}, \tag{5.110}$$

which is the same as for $\omega = 0$. Intuitively, although the saturation value of E_1 has changed, the time constant has also changed, and the slope of $E_1(t)$ at $t = 0$ is a constant regardless of the value of ω.

The preceding results show that for recording multiple holograms, using traveling gratings does not help increase the diffraction efficiency. In fact, the asymptotic value of diffraction efficiency for large number of holograms is the same. On the other hand, this means that if for some reason the intensity pattern (grating) were drifting at a steady rate, the rise time of the holograms would not change, and thus the recording results would not be affected.

Theoretically, although it is true that the asymptotic value of diffraction efficiency for "large number of holograms" is the same, in practice the reduction in the saturation diffraction efficiency by the factor in Eq. (5.103) is very severe. For example, a typical AO deflector would work at 500 MHz, whereas the time constant (which depends on the light intensity) of a photorefractive crystal such as $BaTiO_3$ or SBN would typically be greater than milliseconds. In this case, $\omega_g - \omega$ would be dominated by the the acoustic frequency, and we would have $(\omega_g - \omega)\tau_g \approx 5 \times 10^5$ at best, whereas $\tau_g \omega_g$ would be about 1. From Eq. (5.103), the saturation value would be reduced by a factor of 5×10^5. This would not be practical for recording multiple holograms unless we record more than 1 million holograms. For time constants on the order of seconds, the saturation value would be reduced by a factor of 5×10^8. The moral to all this is that if we use AO deflectors, we still need to compensate for the frequency differences.

References

1. N. V. Kukhtarev, V. B. Markov, S. G. Odulov, M. S. Soskin and V. L. Vinetskii, "Holographic Storage in Electrooptic Crystals, PartI: Steady State, Part II: Beam Coupling Light Amplification," *Ferroelectrics*, **22**, 949–964 (1979).
2. A. M. Glass, D. von der Linde, and T. J. Negran, "High-voltage Bulk Photo-

voltaic Effect and the Photorefractive Process in $LiNb_3$," *Appl. Phys. Lett.* **25**(4), 233–235 (1974).

3. M. Cronin-Golomb, "Large Nonlinearities in Four-wave Mixing in Photorefractive Crystals and Applications in Passive Optical Phase Conjugation," Ph.D. Thesis, California Institute of Technology, 1983.
4. C. Gu, J. Hong, H.-Y. Li, D. Psaltis, P. Yeh, "Dynamics of Grating Formation in Photorefractive Media," *J. of Appl. Phys.*, **69**(3), 1167–1172 (1991).
5. H.-Y. Li and D. Psaltis, "Double Grating Formation In Anisotropic Photorefractive Crystals," *J. Appl. Phys.*, **71**(3), 1394–1400 (1992).
6. D. Brady, K. Hsu, and D. Psaltis, "Periodically Refreshed Multiply Exposed Photorefractive Holograms," *Opt. Lett.*, **15**(14), 817–819 (1990).
7. N. W. Aschroft and N. D. Mermin, *Solid State Physics* (Holt, Rhinhart, & Winston Inc., New York, 1976).
8. A. Yariv, *Quantum Electronics*, 3rd ed. (John Wiley & Sons, New York, 1989).
9. A. Yariv and P. Yeh, *Optical Waves in Crystals* (John Wiley & Sons, New York, 1989).
10. S. R. Montgomery, J. Yarrison-Rice, D. O. Pederson, G. J. Salamo, M. J. Miller, W. W. Clark III, G. L. Wood, E. J. Sharp, and R. R. Neurgaonkar, "Self-pumped Phase Conjugation in the Red in Photorefractive $Ba_{0.5}Sr_{1.5}K_{0.25}Na_{0.75}Nb_5O_{15}$ and $Sr_{0.6}Ba_{0.4}Nb_2O_6$ with Cerium in 9-fold Coordinated Sites," *JOSA B*, **5**(8), 1775–1780 (1988).
11. D. L. Staebler, W. Burke, W. Phillips, and J. J. Amodei, "Multiple Storage and Erasure of Fixed Holograms in Fe-doped $LiNbO_3$," *Applied Physics Letters*, **26**(4), 182–184 (1975).
12. S. I. Stepanov, M. P. Petrov, and A. A. Kamshilin, "Optical Diffraction with Polarization-plane Rotation in a volume Hologram in an Electrooptic Crystal," *Sov. Tech. Phys. Lett.*, **3(9)**, 345–346, (1977).
13. E. Voit, C. Zaldo, and P. Günter, "Optically Induced Variable Light Deflection by Anisotropic Bragg Diffraction in Photorefractive $KNbO_3$," *Opt. Lett.*,

11(5), 309–311, (1986).

14. F. Vachass and T. Y. Chang, "Cross-polarization Two-beam Coupling in Optically Active Photorefractive Media," *J. Opt. Soc. Am. B*, **6(9)**, 1683–1692, (1989).
15. P. Gunter and E. Voit, "Anisotropic Bragg Diffraction in Photorefractive Crystals," *Ferroelectrics*, **78**, 51–61, (1988).
16. D. Psaltis, D. Brady, and K. Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.*, **27**(9), 1752–1759 (1988).
17. J. H. Hong, S. Campbell, and P. Yeh, "Optical-pattern Classifier with Perceptron Learning," *Appl. Opt.*, **29(20)**, 3019–3025, (1990).
18. L. Solymar and D. J. Cooke, *Volume Holography and Volume Gratings* (Academic Press, New York, 1981).
19. J. E. Ford, Y. Fainman, and S. H. Lee, "Enhanced Photorefractive Performance from 45°-cut $BaTiO_3$," *Appl. Opt.*, **28**(22), 4808–4815 (1989).
20. G. C. Valley, "2-Wave Mixing with an Applied Field and a Moving Grating," *J. Opt. Soc. Am. B*, **1**(6), 868–873 (1984).
21. P. Refregier, L. Solymar, H. Rajbenbach, and J. P. Huignard, "2-Beam Coupling in Photorefractive $Bi_{12}SiO_{20}$ Crystals with Moving Grating — Theory and Experiments," *J. Appl. Phys.*, **58**(1), 45–57 (1985).

Chapter 6
The Real Time Face-Recognition System

In this chapter an optical network is described that can recognize at standard video rate the identity of faces it has been trained to recognize. The system uses photorefractive crystals (lithium niobate) as the interconnecting weights (or synapses). It shows how the real time recording property of the crystal may be used to implement and train large number of interconnections or neural networks.

Such a system may be implemented with a 3-D disk system, where we allocate each location to a person. By rotating the disk, the system can attempt to recognize and identify the input face with the faces stored on the disk. Such a system is shown in Figure 6.1, where a 3-D disk is used to implement the first layer of the network and a 2-D disk is used to implement the second layer.

In this chapter, we will concentrate only on one location. The problems associated with rotating the 3-D to the correct angle for readout have already been discussed in Chapter 3.

The implementation of the interconnection weights in optical neural networks uses the holographic optical correlator. In the appendix, we examine in more detail the volume holographic correlator and determine the factors that affect the shift invariance property.

6.1. Introduction

The optical face recognition system that will described in this chapter is basically a two layer feedforward network. The adaptable interconnections of the network are implemented with holograms stored in a photorefractive crystal. The optical system is the standard holographic multi-layer architecture [1–6].

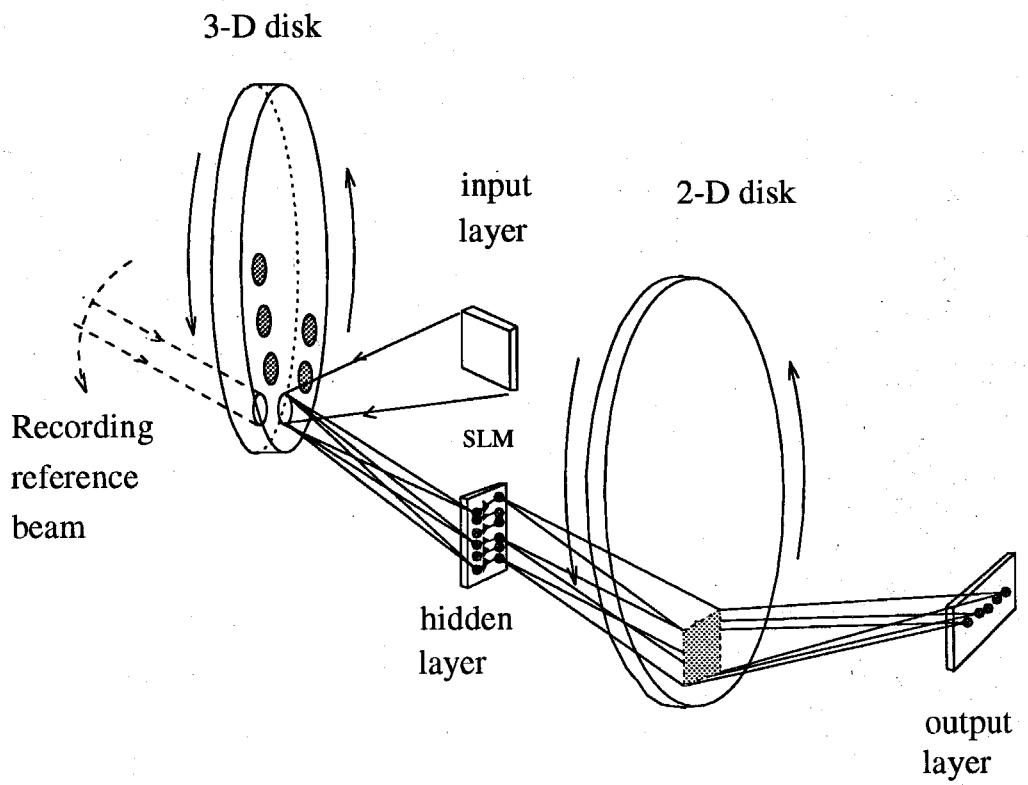


Figure 6.1. Optical Network implemented with a 3-D disk.

The second layer has fixed weights and a simple ad-hoc procedure is used to train the network. Choosing a training algorithm that is well suited to the optical implementation is the most crucial step carrying out a successful experiment. The back error propagation (BEP) algorithm [7] and its variants are the most popular procedures for training multi-layer optical networks [1,3,4,5]. Back propagation is an example of a learning algorithm that yields *distributed* representations in the hidden layers of a network. In a distributed representation a large portion (typically half) of the hidden units responds when the input is one of the training samples.

In contrast, in a *local* learning algorithm each hidden unit is trained to respond to only a small number of training examples. The Radial basis function (RBF) classifier is an example of a commonly used local learning algorithm. An optical RBF system has been recently demonstrated [8]. The advantage of local algorithms is that the training process is relatively easy. If an input training sam-

ple does not cause any of the existing hidden units to respond sufficiently, a new hidden unit is added and devoted to the new sample. The disadvantage of local algorithms is the large network size that is typically obtained. The disadvantage of distributed representation learning algorithms is that the training is difficult, typically requiring a large number of training cycles.

In selecting an algorithm for training an optical neural network, we can argue that distributed algorithms are well suited for optics because the computational speed of optics can be effectively used to speed up the training. However, the optical implementation of algorithms such as BEP requires a dynamic holographic medium that can be accurately controlled. In the experiment described in this chapter we use photorefractive crystals to implement the adaptive interconnections. When a new hologram is recorded in a photorefractive crystal the previously recorded signal is partially erased. This "weight decay" in effect limits the number of cycles a training algorithm can run on an optical system, since earlier exposures are erased as the training progresses. Dynamic copying [9–12] can overcome this problem by restoring the strength of the hologram through feedback. However, dynamic copying is still at the early stages of development and it is premature to construct a large scale network using this approach. Another way for bypassing the weight decay problem is to use local algorithms since they do not require long training sequences. In this case the large storage capacity of 3-D holograms can be used to synthesize the large networks that are required.

The algorithm used for training the face recognition is a hybrid. It has features of local algorithms in that each hidden unit is trained separately and the training method is not iterative. On the other hand, the representations that result are distributed. The distributed representation is crucial for two reasons. First, when the optical network was trained with purely local representations, it was found that the system became extremely susceptible to noise and the performance deteriorated very rapidly as the number of hidden units increased. This is because in a purely local representation, only one hidden unit is on at a time. Since the output is formed as a linear combination of all the hidden units, a

small amount of noise from each hidden unit will ultimately overwhelm the signal term as more hidden units are added. Poor generalization performance is the second reason to avoid purely local representations. By switching to distributed representations, the system performs much better when presented with images it had never seen before.

In the following, we first describe the optical architecture and the overall experimental setup, and then the training algorithm and the details of the training procedure. In the last section, we describe the performance obtained with the network.

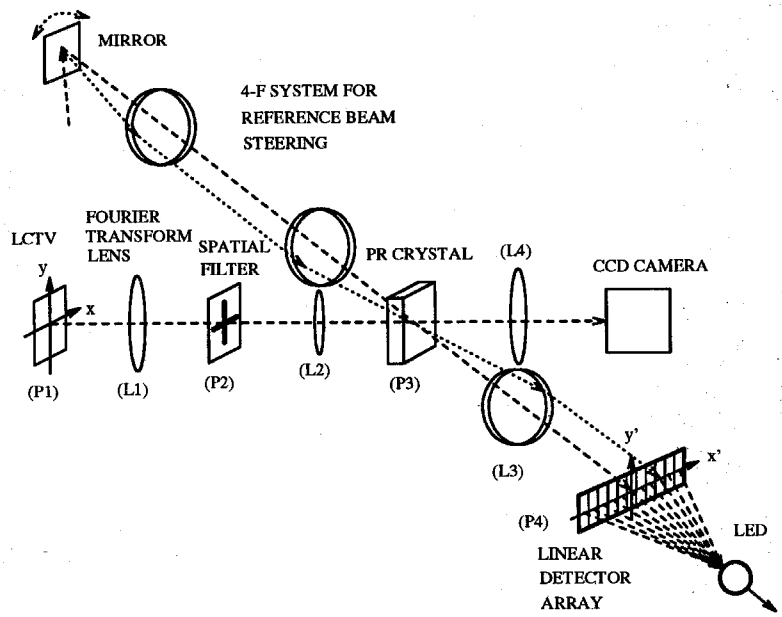


Figure 6.2. Optical setup of the face-recognition system.

6.2. Experimental Apparatus

The optical setup is shown in Figure 6.2. It is a 2-layer network with an optical pre-processing stage that performs edge enhancement. The input device to the network is a liquid crystal TV (LCTV) that has 320 by 220 pixels resolution and 2×2.5 cm^2 clear aperture. This device was extracted from an EPSON television projector. The LCTV is illuminated with collimated light from an Argon

laser ($\lambda = 488$nm). Lens L1 produces the Fourier transform of the input image at plane P2. A spatial filter is placed at P2 to perform two functions. First, it blocks the higher diffracted orders that result from the pixelation of the LCTV. The removal of the higher orders gives a smoother, less noisy image but it reduces the light efficiency of the LCTV. The second function of the spatial filter is to block the low frequency components of the input image. This enhances the edges of the input image and dramatically improves the ability of the system to discriminate between inputs from different classes. A photograph of the spatial filter is shown in Figure 6.3. It consists of a cross-hair and a DC block for high pass filtering. The purpose of the cross-hair is to remove the diffraction pattern at P2 that comes from the boundary of the LCTV image. This boundary, when edge enhanced, yields a very strong rectangle that is common to all inputs and makes discrimination difficult. The diameter of the DC block is 260 microns. Given the wavelength of light and the focal length of L1 ($F_{L1} = 50$ cm), the cutoff frequency is approximately .533 lines/mm. Roughly speaking, features in the input plane that are smaller than 1.9 mm are highlighted in the edge enhanced image. An iris (not shown in Figure 6.3) is used to block the higher orders not blocked by the cross-hair. An example of an image of a face and the edge enhanced version of it that was produced by the optical system is shown in Figure 6.4.

Lens L2 images with magnification 1 plane P2 onto plane P3, the plane of the hologram. The size of the spectrum on the hologram is approximately 5 mm in diameter. The hologram is formed by introducing a plane wave reference. The angle between the signal and reference beam varies from 29 to 31 degrees, outside the crystal. The reference beam is reflected off a mirror mounted on a computer controlled rotation stage. The rotating mirror is imaged onto the crystal with a unit magnification 4-f system. This allows the angle of the reference beam to be scanned without moving the position of the reference beam on the crystal. The crystal is an iron doped $LiNbO_3$, with doping level 0.01%. The c-axis of the crystal is in the horizontal direction in Figure 6.2. The crystal dimensions are $20 \times 20 \times 8$ mm^3.

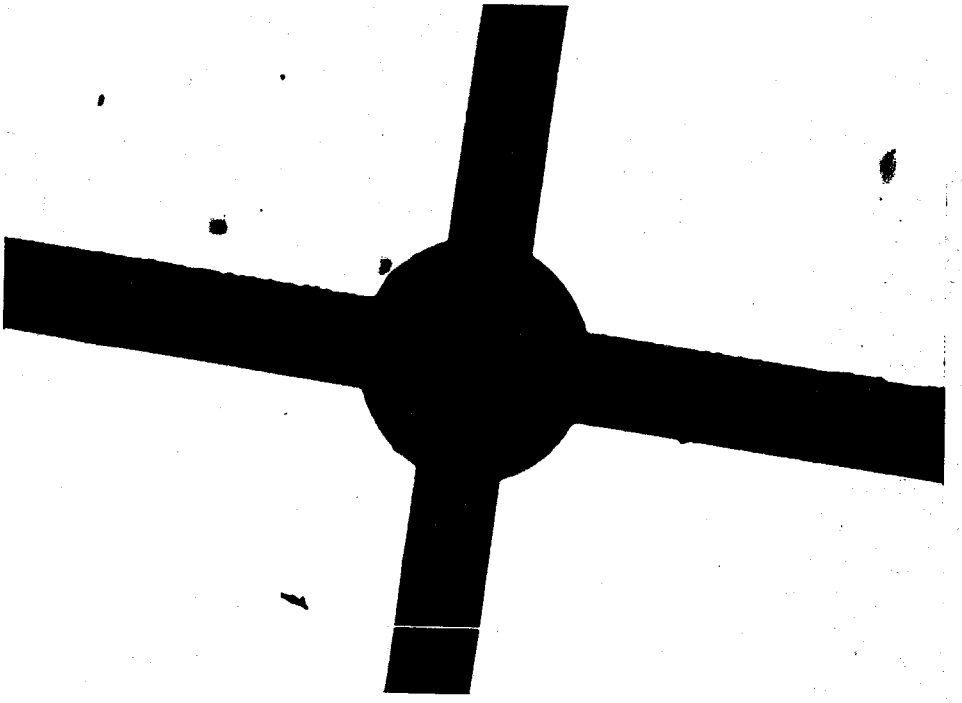


Figure 6.3. Spatial filter used in plane P1 of Figure 6.2.

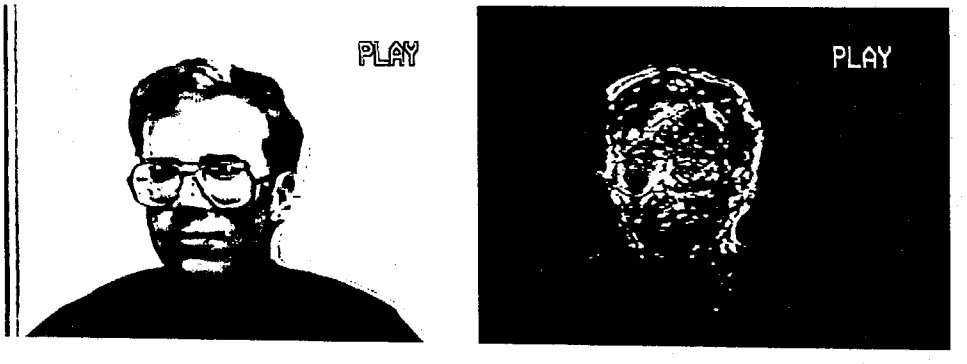


Figure 6.4. Edge enhanced image and original face.

Lens L4 is a Fourier transform lens that produces an image of the edge enhanced input image on a CCD for visual assessment. Lens L3 is also a Fourier transforming lens that produces the response of the first layer at the output plane P4. A linear detector array is placed at P4, where it is used to detect this response.

A beamsplitter placed in front of the array diverts a portion of the light to a CCD camera so that the output of the first layer can be visually monitored. Functionally, the system from the input plane P1 to P4 is an array of image correlators with 1-D shift invariance. To understand this consider the case where a single hologram is recorded in the crystal at a particular angle of the reference beam.

In this case the system is a classic Vander Lugt [13] correlator except that a volume hologram is used and the input has been high-pass filtered. The effect of the volume hologram is to eliminate shift invariance in the horizontal direction in Figure 6.2. This happens because a horizontal shift at plane P1 will change the angle of incidence at plane P3 and cause the hologram to be Bragg mismatched [14,15,16]. Specifically, the light distribution at plane P4 is given by [14] [1]

$$g(x',y') = \iint f(x,y)h(x-x',y-y')\,dx\,dy \cdot \mathrm{sinc}(\alpha x'), \tag{6.1}$$

where $f(x,y)$ and $h(x,y)$ are the input and filter functions, respectively. The input coordinates are (x,y) and the output coordinates are (x',y'). The thickness of the crystal is L, θ is the angle of the reference beam, [2] and

$$\alpha = \frac{\pi L \sin\theta}{\lambda F}. \tag{6.2}$$

We see from Eq. (6.1) that the effect of the thick hologram is to mask off the 2-D correlation pattern except for one vertical strip whose position depends on

---

[1] Please also see the discussion in the appendix of this chapter.

[2] The definition of sinc used in this chapter is

$$\mathrm{sinc}(x) \stackrel{\mathrm{def}}{=} \frac{\sin x}{x}.$$

the the angle of the reference beam. The amount of shift invariance that can be tolerated in the horizontal direction is approximately equal to $1/\alpha$ plus the width of the correlation peak in the horizontal direction. The system retains its shift invariance in the vertical direction. If we change the reference beam angle and record a different hologram at each angle, then we will have at each horizontal location a 1-D strip from a different 2-D correlation function. In the experiment that we will describe, holograms are recorded at 40 separate angles separated by 0.05°, yielding a system that has 40 correlators with 1-D shift invariance.

Figure 6.5. Experiment showing the position of the correlation peak to be proportional to the size of the input face.

The experiment in Figure 6.5 shows the operation of this part of the system. In this case each filter was a recording of the face of the same person at different scales. Figure 6.5 shows the input to the network for 4 different size images, along with the corresponding response at the right-hand side of each picture. As the size of the face increases, the strongest response of the system is at different vertical positions. In the optical setup, the correlation responses shown at the right-hand side of each picture is actually horizontal. This was done by rotating

the CCD camera by 90 degrees. [3]

The second layer performs two tasks. The first task is to take advantage of the vertical shift invariance of the first layer, and the second task is to combine the outputs of the 40 correlators and make the final classification. We will discuss first the shift invariance. Suppose that an image at a particular location at the input produces a strong correlation peak somewhere at the output. If the input is horizontally translated by approximately 0.4 mm then the correlation peak disappears. If the input is translated vertically then the correlation peak moves vertically also. What we really need for shift invariant recognition is a system whose output does not change as the input shifts. To do this we use long detector elements in the vertical direction as shown in Figure 6.2. These long detectors collect the correlation peak and continue to produce a strong output signal as the input image shifts vertically. Unfortunately, we cannot use arbitrarily long detector elements to obtain full shift invariance vertically, because then the detector would simply collect all the diffracted energy from the corresponding filter stored in the hologram. Roughly speaking, all input signals with the same total energy would yield the same response. A shorter detector responds more selectively to the correlation peak, and hence the degree of match between the input and the reference, but it sacrifices shift invariance. Thus there is a basic tradeoff between shift invariance and discrimination capability. In our network we made this compromise by trial and error. By repeating the experiment with a horizontal slit of varying width placed in front of the detector array, the amount of shift invariance in the vertical direction is found to be roughly 3 mm. This is

---

[3] Since the LCTV and the CCD are rotated 90 degrees, up and down in the picture become horizontal in the actual system. To avoid confusion, for the rest of this chapter we will use the terms "vertical" and "horizontal" to mean the directions in the actual optical setup. We will use the (admittedly awkward) term "side-to-side" and "up-and-down" to mean the directions for the people sitting in front of the camera. Thus side-to-side motion in front of the camera and in the pictures is really horizontal in the optical setup.

approximately 12 percent of the size of the input image. As we will see later, this choice yields good discrimination capability.

The second layer also puts together all the vertically integrated responses from the first layer and produces the final output. Since the output of the detector array in plane P4 is electronically available the second layer can be implemented either electronically or optically. We have done both with comparable performance. The optical implementation of the second layer is realized by thresholding the output of the detector array and then feeding it to a second LCTV. The inner product between the signal recorded on the LCTV and a weight vector stored in the form of a transparency is then optically formed. This inner product is electronically thresholded to produce the final output. In the current system we describe in this chapter, the operations of the second layer are so simple that it was easier to do them electronically. Specifically, all the weights of the second layer have the same value. In other words, the second layer simply integrates the output of the first layer. The electric signal from each detector is the square of the light amplitude of the total signal incident at each element. The signal from the detector can be thresholded electronically. However, we get the best performance by simply using the square-law non-linearity. In this case, the system becomes similar to a quadratic associative memory [17,18]. Notice that the nonlinearity performed at plane P4 is crucial in this system.

If the outputs of all the correlators from the first layer were somehow coherently added without the inclusion of the nonlinearity, then the overall system would simply be equivalent to a single correlator.

A schematic diagram of the overall system is shown in Figure 6.6. The input images are detected by a standard television camera. The video signal is either stored on a video cassette recorder (VCR) to form a training set or fed directly to the LCTV during real time operation. The 2-layer optical network is the system we described above. A personal computer controls the experiment during the training phase by instructing the VCR to advance the video by one frame, and then pauses the frame so that the training algorithm can be executed. The output

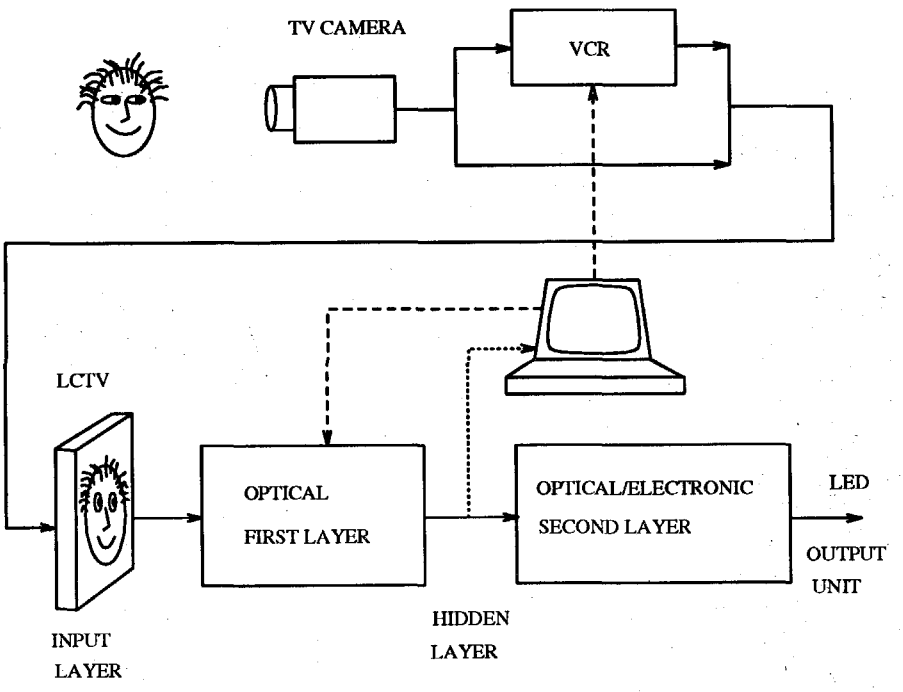


Figure 6.6. Schematic diagram of overall system.

of the hidden layer determines whether the hologram should be modified by the current input image. If a holographic exposure is needed, the computer opens two shutters (one for the signal and one for the reference beam) for a specified time. The hologram is thus recorded. During the execution of the algorithm the computer also controls the angle of the reference beam, so that different hidden units can be trained. After the training is completed, the computer is no longer involved in the operation of the system except to record the output data if desired.

6.3. Training Procedure

The training algorithm that we use is partially motivated by the tiling algorithm [19]. In the tiling algorithm, individual units are trained separately for a fixed number of iterations. Once a unit is trained, the algorithm moves on to a new unit. The new unit is then trained to make up for the deficiencies in the performance obtained with the previous units. In this way networks with multiple layers and many neurons per layer can be built-up and trained. In the standard

tiling algorithm each unit is trained with the perceptron algorithm with the entire training set. In our algorithm each unit is trained by a subset of the training set that consists of similar images. This similarity measure is enforced by training each unit to respond to a contiguous short segment of the training video. In this way, each unit is trained to respond to a specific aspect of the input face. This simplifies the training of individual units and the training procedure results in networks of predictable size.

The flow chart for the algorithm we use is shown in Figure 6.7. We describe more specifically the algorithm. Let $\mathbf{f}^k$ denote the k-th image in the training sequence stored in the VCR, and let w_{ij} denote the weight of the first layer connecting the i-th input pixel to the j-th hidden unit. The training algorithm is as follows:

```
set e = 0                          (e is the number of exposures per hidden unit)
set j = 1                                        (j enumerates the hidden units)
while (''there are more training examples'')
do {                               (go through the training set one frame at a time
)
    h = 0                                (h is the number of hidden units turned on)
    for j' = 1 to j, if Σ_{i'=-I/2}^{I/2} |Σ_i f_i^k w_{i-i',j}|^2 > θ then h = h + 1
                                        (count the number of hidden units that are on)
    if (h < H and Σ_{i'=-I/2}^{I/2} |Σ_i f_i^k w_{i-i',j}|^2 < θ)
                (less than H hidden units are on, and the current unit is off)
        then w_ij = w_ij + f_i^k and e = e + 1                  (make an exposure)
    if (e > E)                          (more than E exposures on current unit)
        then j = j + 1 and e = 0                          (create new hidden unit)
    ''go to next frame''
}
```

The user must select the parameter θ, H, and E before the algorithm begins. The variable j counts the hidden units. We begin training the first unit ($j = 1$) by presenting frames to the system in sequence (incrementing k). The k-th input is added to the weights of the first unit if the response of the first hidden unit is

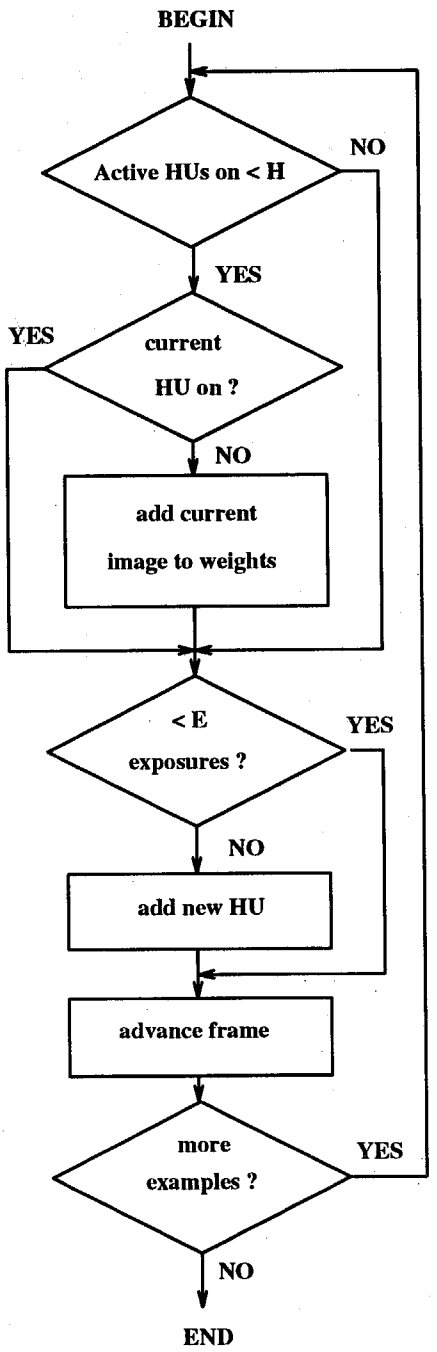


Figure 6.7. Flow chart for the algorithm used to train the network. (HU means "hidden unit").

below a threshold θ. Notice that in the optical system the response of the hidden unit is not simply the inner product between the input and the weight vector, but an integration over I pixels of the center of the correlation function, as we described earlier. If θ is set too high then the units become very highly tuned to respond to the particular images they are trained for. If the threshold is too low then too much cross talk with unfamiliar faces results leading to erroneous classifications. Ideally, θ should be lowered as the training proceeds and hidden units are added, since this weakens all the stored holograms. In the experiment described we used a constant θ. The first unit continues to accumulate training examples in this way until a total of E exposures have been made to it. At that point a new hidden unit is created (j is incremented) by rotating the mirror that controls the angle of the reference beam. We would like to have E large in order to have each unit be responsive to as many training examples as possible. However, since we are only presenting positive examples to the system (i.e., we never subtract anything from the weights but always add to them), if too many examples are accumulated, the weight is simply the average of the subject's faces, which is similar to the average of anybody's face. The unit would then lose its discrimination capability. The first H hidden units are trained in exactly the same manner as the first.

When j exceeds H, the current input frame is added into the weights of the j-th hidden unit only if fewer than H units are above threshold. If H is set to 1, then the training of the early units is identical to the rest. However, this results in a hidden layer response that has only one unit on at a time. We have already commented that we found that this results in poor performance on the training set due to susceptibility to noise and poor generalization.

By requiring that at least H hidden units are on at any one time for the training set, we improve the robustness of the system and improve generalization. If H becomes too large, we would need too many hidden units to enforce this requirement, and the encoding becomes inefficient.

The discussion above describes the basic trends that we predict and exper-

imentally observe as the parameters E, H, and θ are adjusted. The experiment that we will describe in this chapter was carried out with $H = 3$, $E = 6$, and θ was set equal to 3 times above the noise background level. These values were arrived at empirically by running the experiment several times and measuring the generalization performance. The system performance is sensitive to the setting of θ (it should be set relatively low), but not as sensitive to changes in H and E. These settings worked best for all the face recognition experiments we tried. Unfortunately, there is no guarantee that these settings are the best for other problems.

The most attractive feature of this algorithm is that it can be easily implemented with the optical system described in the previous section.

At the same time, it gives remarkably good classification performance, as we will see in the next section. The algorithm requires two basic operations from the optical system: (1) evaluation of the response of the hidden units to an input image, so that the computer can compare it to a threshold, and (2) addition of the current image into the hologram corresponding to the weights of that unit. We have already described how the system evaluates the response of the hidden units. We will discuss here how the weight updates are performed. When a hologram is exposed to light the strength of an individual holographic grating (or connection) w_{ij} is modified according to the following equation [3]:

$$\tau \frac{dw_{ij}}{dt} + w_{ij} = \beta m_{ij}, \tag{6.3}$$

where τ is the time constant of the holographic recording in the photorefractive crystal, β is a constant that depends on the crystal properties, and m_{ij} is the modulation depth of the frequency component (of the illuminating light) corresponding to the grating w_{ij}. For a short light exposure of duration Δt, we can approximate the change in the hologram by

$$\Delta w_{ij} \approx -\frac{\Delta t}{\tau} w_{ij} + \frac{\Delta t}{\tau} \beta m_{ij}. \tag{6.4}$$

In other words, each exposure reinforces each weight in proportion to the strength of the corresponding frequency component of the illuminating light. However,

each exposure also erases all the weights in proportion to their current strength. This is the well known weight decay problem that plagues photorefractive memories [20] and photorefractive neural networks [3]. Several solutions to this problem have been proposed [9,10,21]. We use a simple exposure schedule in our experiment, in which later exposures are linearly shortened to compensate for the decay of the earlier holograms. This results in an approximately uniform final recording. Specifically the m-th exposure, t_m, is set equal to $t_m = 3 - m/240$ seconds. Thus the exposures varied from 3 seconds at the beginning of the exposure sequence to 2 seconds at the end, with a total light intensity equal to 10 mW/cm^2 and a modulation depth approximately 0.1.

The training set for the experiment was a video recording of the face of my colleague, Yong Qiao, moving his head in front of the camera, turning, nodding, tilting his head, smiling, etc. The total number of images in the training set is 5,400 frames. The execution of the algorithm modified the hologram with only 240 of these images. The rest produced an acceptable hidden layer response. Since each hidden unit receives 6 exposures, a total of 40 hidden units were created. The maximum number of hidden units that the system can support is limited by two factors. One is the dynamic range of the photorefractive hologram. In this case a total of 240 holograms are superimposed. If we assume that all these exposures are statistically uncorrelated (i.e., each exposure simply erases all the previously recorded holograms and does not ever reinforce them), then the diffraction efficiency of each hologram would fall by a factor of $(240)^2$ [9] compared to the efficiency with which a single hologram is stored. Since up to 5,000 [22] holograms have been superimposed in lithium niobate crystals, the dynamic range was not a problem in our experiment. The second limitation is the numerical aperture of the optical system to allow all the reference beams to enter the crystal. The system we used in the experiment had the capability to implement in excess of 100 units and it is possible to build systems with more than 1,000 units. Therefore, this particular training set did not stretch the limits of the system's capabilities. The entire training cycle lasted about 40 minutes,

which includes the time for hologram exposure and controlling the system by computer.

Figure 6.8. Photographs showing part of the training session.

Shown in Figure 6.8 is a composite photograph showing a short sequence of the training session. Each picture in the composite shows the current input frame and on the right, displayed from top to bottom, is the optical response of the hidden units. The first event in the sequence is on the top left in Figure 6.8 and it shows the frame shortly after the hologram is exposed. As time progresses the hidden layer response changes (upper right corner) and gradually dims (lower right corner). Ultimately, there are fewer than 3 units on and the system is triggered to make another exposure (lower right corner). When the region in the picture where the hidden layer normally appears becomes a white ribbon, it means that the crystal is being exposed to light.

6.4. Classification Performance

In this section we describe the performance of the trained network. Once the network is trained it operates in real time, processing 30 frames per second directly from the input TV camera. The outputs from the detector array are simply added together electronically and this sum is then thresholded to produce the final output. The holograms will decay when exposed to light during the testing phase. We can overcome this by either thermally fixing the hologram [23] or by using dynamic copying [10–12]. In this experiment we adopted a simpler route that temporary overcomes this problem. By reducing the readout light intensity to 1/20 of the total writing intensity, we calculate that the holograms will decay after several hours of constant illumination. The holograms were sufficiently strong that the reduction in the readout intensity yielded sufficient signal at the detector. The system was tested with the original training set and a wide variety of test sets, including videos of Yong presented to the system under various conditions, and other people attempting to confuse the system. Shown in Figure 6.9 is the signal at the output of the system before final thresholding. The entire recorded presentation shown in Figure 6.9 lasts for about 10 minutes. The first minute is a portion of the training set. The next 2 minutes is a real-time input of Yong who looks into the TV camera and moves around in a manner similar to the training set. While he does this, he does not have access to any information from the network. The rest of the sequence is the response of the system to two other persons (Allen Pu and myself, Sid). We see that the average response is highest for the training set, and is almost as high for the rest of the time where Yong is the input. The average response for the other two subjects is markedly lower. The variance of the response is higher for Yong, because he was exhibiting a wider range of head perspectives to test the limits of the system. Similar behaviors were observed for all 14 members of our group.

To make the final classification, we need to threshold the signal shown in Figure 6.9. In the actual system, this is done electronically in real time. The

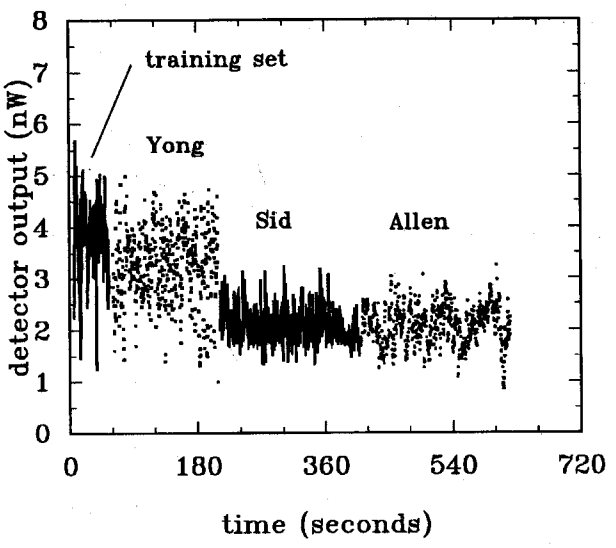


Figure 6.9. System response before thresholding.

optimum threshold was determined from the data shown in Figure 6.9. Shown in Figure 6.10 is a plot of probability of error as a function of the output threshold level. The three curves correspond to the probability of error for Yong, Allen, and myself (Sid) estimated by classifying the data in Figure 6.9 with different thresholds. If we want to minimize the overall probability of error, the optimum threshold level is approximately 2.5 nW, giving a probability of error of about 12%. If we set the threshold slightly above 3 nW, then we almost never make a false recognition while correctly identifying Yong approximately 70% of the time.

We can improve the performance of the system further by using the time domain. If the input face is moving and presents different views to the system, we can eliminate many of the errors by integrating over a time interval longer than the duration of a single frame, and then perform the classification. Specifically, we classify the current frame to be Yong if M out of the N frames give us a positive response. In implementing such an algorithm, we need to select N, M

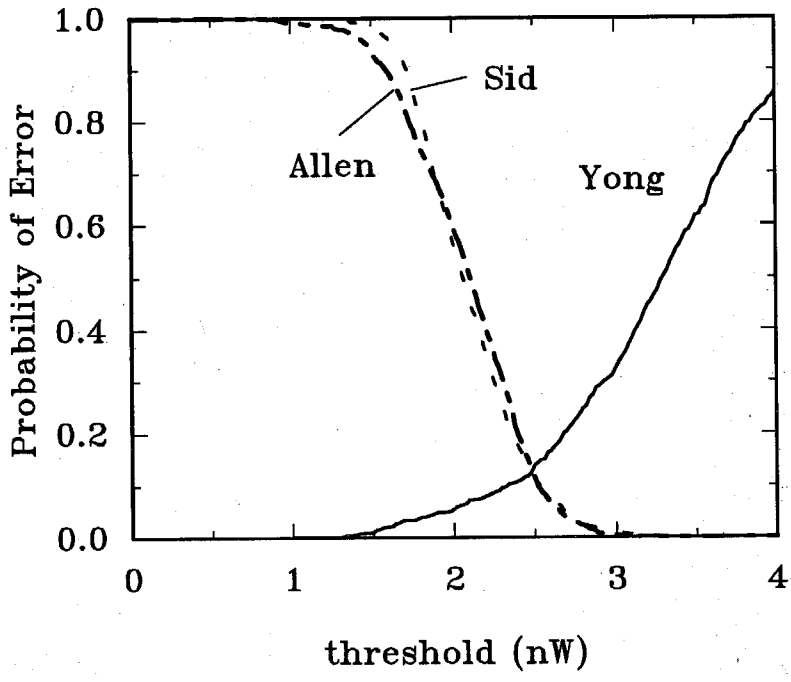


Figure 6.10. Probability of error as function of the output threshold level.

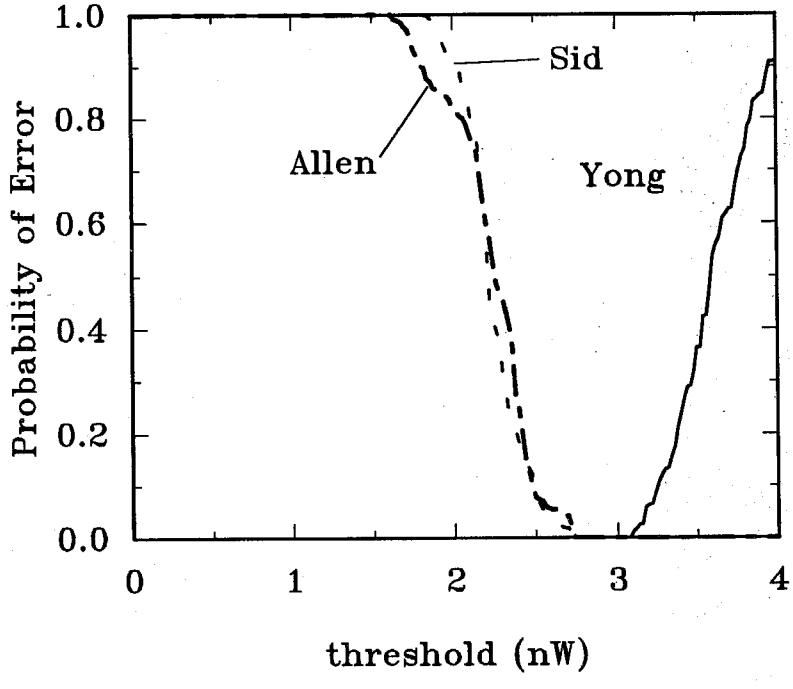


Figure 6.11. Probability of error as function of the output threshold level when the output is observed for 6 seconds to perform the classification.

and the threshold level. Shown in Figure 6.11 is a plot of probability of error on the same three data sets as before as a function of the threshold level for M = 7 and N = 25. Notice that if the threshold level is selected in the range of 2.75 nW to 3 nW, the estimated probability of error is zero. In this example, the decision is made based on observation of the input video for 6 seconds (the computer sampled the output at 4 samples/second). In general, there is a tradeoff between performance and observation time.

The next sequence of experiments we describe were carried out to evaluate the kind of generalization obtained by the network. In this case, the subjects (Yong and others) were allowed to look at the output of the network. Adjustments were made to test the limits of the system. Examples from this series of experiments are displayed in the composite of Figure 6.12. The pictures are arranged in a 4×4 matrix. We assign to each picture a pair of numbers (i, j). The picture at the upper left corner is designated (1,1), and the one at the upper right corner is (1,4). The small black circle within each picture displays the final output of the system after thresholding. If the bright dot appears in the circle, the system makes a positive identification of Yong. Picture (1,1) is an example of Yong being correctly recognized by the system. Picture (1,2) shows Yong illuminated from below and the side, whereas during training the illumination was from above. We can see that the system is sensitive to the direction of illumination because of the edge-enhancement that is performed by the system. As the direction of illumination changes, the edges move around. To obtain invariance to illumination direction, we need to include in the training set examples of different lighting. Picture (1,4) and (2,1) show that key features such as the mouth and the eye are crucial for recognition. However, as picture (2,2) shows, the eyes alone are not enough for a positive identification. Picture (2,3) is meant to display the invariance of the system to up and down motion. It is difficult to assess this from the still photo. However we measured a tolerance to up-and-down shifts of about 5% of the whole scene. As mentioned earlier, the optical system was arranged such that up-and-down shifts of the input image become horizontal shifts on the LCTV.

We did this because we need more tolerance to side-to-side shifts (people move side-to-side much more than up-and-down) and the optical system provides shift invariance in the vertical direction at the LCTV plane. Prior to the training, the tolerance to up-and-down input shifts was 2% of the whole scene. Training more than doubled the tolerance of the up-and-down shift.

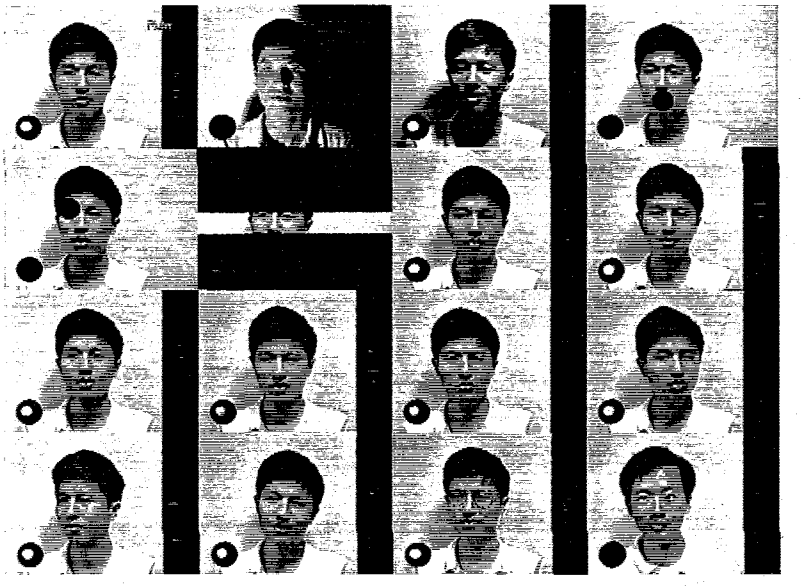


Figure 6.12. Examples demonstrating the generalization capabilities of the system. A bright dot in the circle at the lower-right corner of each photograph indicates that the system classifies the input image as the person it was trained to recognize.

The tolerance of the system to nodding up and down was recorded by measuring the up-and-down motion on the screen of a fixed point on Yong's forehead, as he nodded up and down. According to this measure, the spot on his forehead can move by 1 cm without loss of recognition. From this measurement, and by measuring the dimensions of Yong's head, we obtain a crude estimate of 5 degrees for the maximum tolerable angle of forward head tilt. Picture (3,3) shows an example of the tolerance of the system to side-to-side shifts of the input im-

age. In this direction the optical correlator provides considerable shift invariance. We measured the maximum side-to-side shift to be about 13 percent of the total extent of the input frame. Overall, the system has more than 3 times better tolerance to shifts in the side-to-side than the up-and-down direction. Pictures (3,4) and (4,1) shows the system's ability to tolerate turning of the head, which we measured to be 30 degrees in either direction. The maximum tilt of the head (picture (4,2)) was measured to be 12 degrees in either direction. We did not seriously test the response of the system to scale changes.

6.5. Discussions and Conclusions

The main goal of this experiment was to use a combination of existing optical techniques and algorithmic ideas to build a trainable real time face recognition system that works. This system gave us remarkably good performance and yet it greatly under-utilizes the full capabilities of the optical network. On the other hand, there are many ways we can seek to improve the performance of the system. For instance, to incorporate invariance to scale or illumination, we would need to expand the training set to include all possible *combinations* of scale and illumination conditions of interest, as well as all the invariances that the current system incorporates. For example to accommodate 5 different scales, we would need to expand the size of the training set by roughly a factor of 5. The number of hidden units that are needed with the approach we use usually scales proportionally to the size of the training set. Expanding the size of the optical system from the current 40 hidden units to approximately 1,000 is within reach. It should therefore be possible to expand the variety and range invariances accordingly. In addition, we can seek ways to build in some of the invariances, in addition to the 1-D shift invariance afforded by the Fourier transform holograms. For instance we can have an adaptive optical system that is trained to recognize eyes independently of the identity of the face. This feature detector can then be used to normalize the input for up-and-down position or head rotation.

Appendix

Volume Holographic Correlators

In this appendix, we will analyze more carefully the correlation operation using volume holograms. The approach we use here is basically the Born's approximation method used in Chapter 2. We consider one slice of the volume hologram at a time, and then add the diffracted light from each slice to obtain the total diffracted field. As mentioned in Chapter 2, this method assumes that the incident light is not affected by the holograms.

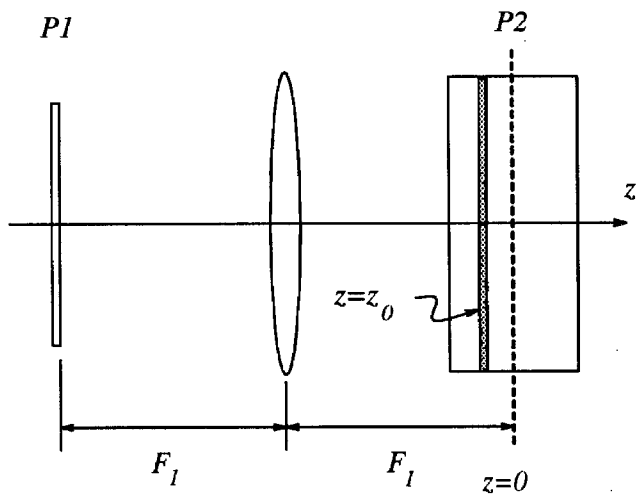


Figure 6.13. The volume holographic correlator.

Consider the Fourier transform system shown in Figure 6.13. The image $f(x,y)$ is placed at the Fourier plane (back-focal plane) $P1$, a distance of one focal length (F_1) away from the Fourier transform lens. The origin of the z axis is taken to be at the front focal plane $P2$, also at a distance F_1 away from the Fourier transform lens. It can be shown that under paraxial approximations, the field distribution at z is proportional to [4]

$$\iint f(\mathbf{x}')e^{j\pi\lambda z u^2}e^{j2\pi\mathbf{u}\cdot\mathbf{x}}\,d\mathbf{x}', \tag{6.5}$$

[4] The implicit time dependence factor is $e^{-j\omega t}$.

where

$$\mathbf{x} = (x, y), \tag{6.6}$$

$$\mathbf{x}' = (x', y'), \tag{6.7}$$

$$\mathbf{u} = (u_x, u_y) = \frac{\mathbf{x}'}{\lambda F_1}, \tag{6.8}$$

and

$$u^2 = u_x^2 + u_y^2. \tag{6.9}$$

Note that here the vectors $\mathbf{x}'$, etc., are 2-D vectors instead of 3-D. In the following, we will work with the spatial frequency variables u_x and u_y. To simplify the notation, we will write $f(\mathbf{u})$ to mean $f(\lambda F_1 \mathbf{u})$, etc.

We first record a hologram using the signal

$$S = \iint f(\mathbf{v}) e^{j2\pi \mathbf{v}\cdot\mathbf{x}} e^{-j\pi\lambda z v^2}\, d\mathbf{v}, \tag{6.10}$$

and the reference beam

$$R = e^{j2\pi \mathbf{u}\cdot\mathbf{x}} e^{-j\pi\lambda z u^2}. \tag{6.11}$$

The hologram is recorded in a volume from $z = z_c - L/2$ to $z = z_c + L/2$. We consider the infinitesimal segment lying between $z = z_0$ and $z = z_0 + \Delta z$. We may treat it as a planar hologram. At $z = z_0$, the hologram recorded is

$$RS^*\big|_{z=z_0} = \iint f^*(\mathbf{v}) e^{j2\pi(\mathbf{u}-\mathbf{v})\cdot\mathbf{x}} e^{-j\pi\lambda z_0(u^2-v^2)}\, d\mathbf{v}. \tag{6.12}$$

If we now apply a second signal

$$S' = \iint g(\mathbf{v}') e^{j2\pi \mathbf{v}'\cdot\mathbf{x}} e^{-j\pi\lambda z v'^2}\, d\mathbf{v}', \tag{6.13}$$

then at $z = z_0$, the hologram is

$$RS^*S'\big|_{z=z_0} = \iint f^*(\mathbf{v}) g(\mathbf{v}') e^{j2\pi(\mathbf{u}-\mathbf{v}+\mathbf{v}')\cdot\mathbf{x}} \cdot e^{-j\pi\lambda z_0(u^2-v^2+v'^2)}\, d\mathbf{v}\, d\mathbf{v}'. \tag{6.14}$$

Let

$$\Delta\mathbf{v} = \mathbf{v}' - \mathbf{v}, \tag{6.15}$$

then assuming $\Delta\mathbf{v}$ is small, we have to first order

$$v'^2 \approx v^2 + 2\mathbf{v}\cdot\Delta\mathbf{v}, \tag{6.16}$$

and

$$u^2 - v^2 + v'^2 \approx u^2 + 2\mathbf{v}\cdot\Delta\mathbf{v}. \tag{6.17}$$

We can then write Eq. (6.14) as [5]

$$RS^*S'\big|_{z=z_0} = \int \left\{ \int g(\mathbf{v}')\ f^*(\mathbf{v}' - \Delta\mathbf{v}) e^{-j\pi\lambda z_0(u^2+2\mathbf{v}\cdot\Delta\mathbf{v})}\ d\mathbf{v}' \right\} \cdot e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}}\ d(\Delta\mathbf{v}). \tag{6.18}$$

If we assume that the widths of f and g are sufficiently small, then we may make the approximation

$$u^2 + 2\mathbf{v}\cdot\Delta\mathbf{v} \approx u^2 + 2\mathbf{v}_0\cdot\Delta\mathbf{v}, \tag{6.19}$$

where $\mathbf{v}_0$ is at the center of the functions f and g. We may then approximate Eq. (6.18) by

$$\begin{aligned} & RS^*S'\big|_{z=z_0} \\ &= \approx \int \left\{ \int g(\mathbf{v}')\ f^*(\mathbf{v}' - \Delta\mathbf{v})\ d\mathbf{v}' \right\} \\ &\qquad \cdot e^{-j\pi\lambda z_0(u^2+2\mathbf{v}_0\cdot\Delta\mathbf{v})}\ e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}}\ d(\Delta\mathbf{v}), \\ &= \int (g\star f)(\Delta\mathbf{v})\ e^{-j\pi\lambda z_0(u^2+2\mathbf{v}_0\cdot\Delta\mathbf{v})}\ e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}}\ d(\Delta\mathbf{v}), \end{aligned} \tag{6.20}$$

where

$$(g\star f)(\Delta\mathbf{v}) = \int g(\mathbf{v}')f^*(\mathbf{v}' - \Delta\mathbf{v})\ d\mathbf{v}' \tag{6.21}$$

---

[5] Actually, the right-hand side of Eq. (6.18) should be the negative of what is written here. However, since this is just an extra constant phase, it is not important and will be omitted for simplicity.

is the cross-correlation function of g and f.

At $z > z_0$, the field $RS^*S'|_{z=z_0}$ becomes (through plane wave propagation)

$$U(z_0) = \int d(\Delta\mathbf{v}) \left\{(g \star f)(\Delta\mathbf{v})\, e^{j2\pi\lambda z_0(\mathbf{u}-\mathbf{v}_0)\cdot\Delta\mathbf{v}}\right\} \cdot \left\{e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}} \cdot e^{-j\pi\lambda z(u^2+2\mathbf{u}\cdot\Delta\mathbf{v})}\right\}. \tag{6.22}$$

The expression above gives the field diffracted from the slice of the volume hologram at $z = z_0$. To find total field at z, we integrate over all the slices:

$$\begin{aligned} U_{total} &= \int_{z_c-L/2}^{z_c+L/2} U(z_0)\, dz_0 \\ &= \int d(\Delta\mathbf{v}) \left\{(g \star f)(\Delta\mathbf{v})\, L e^{j2\pi\lambda z_c(\mathbf{u}-\mathbf{v}_0)\cdot\Delta\mathbf{v}} \cdot \operatorname{sinc}\left[\pi\lambda L(\mathbf{u}-\mathbf{v}_0)\cdot\Delta\mathbf{v}\right]\right\} \cdot \\ &\quad \left\{e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}}\, e^{-j\pi\lambda z(u^2+2\mathbf{u}\cdot\Delta\mathbf{v})}\right\}. \end{aligned} \tag{6.23}$$

The factor

$$e^{j2\pi(\mathbf{u}+\Delta\mathbf{v})\cdot\mathbf{x}}\, e^{-j\pi\lambda z(u^2+2\mathbf{u}\cdot\Delta\mathbf{v})} \tag{6.24}$$

is a plane wave with spatial frequency $\mathbf{u} + \Delta\mathbf{v}$. When the diffracted light is collected by a second Fourier transform lens, the image formed by this lens is [6]

$$\begin{aligned} &(g \star f)\left(\frac{\mathbf{x}}{\lambda F_2}\right) \cdot \left\{L e^{j2\pi z_c(\mathbf{u}-\mathbf{v}_0)\cdot\mathbf{x}/F_2} \operatorname{sinc}\left[\pi\frac{L}{F_2}(\mathbf{u}-\mathbf{v}_0)\cdot\mathbf{x}\right]\right\} \\ &= \left(\frac{1}{\lambda F_1}\right)^2 \iint g(x',y') f\left(x' - \frac{x}{M}, y' - \frac{y}{M}\right) dx'\, dy' \cdot \\ &\quad \left\{L e^{j2\pi z_c(\mathbf{u}-\mathbf{v}_0)\cdot\mathbf{x}/F_2} \operatorname{sinc}\left[\pi\frac{L}{F_2}(\mathbf{u}-\mathbf{v}_0)\cdot\mathbf{x}\right]\right\}, \end{aligned} \tag{6.25}$$

where F_2 is the focal length of the second Fourier transform lens, and $M = F_2/F_1$. In Eq. (6.25), we have translated the coordinate system to the point $\lambda F_2\mathbf{u}$ of the original coordinate system.

---

[6] The second Fourier transform lens does not have to be exactly one focal length away from the hologram, since we are interested only in the intensity distribution.

The diffracted light from volume hologram is therefore the cross-correlation of f and g multiplied by a sinc function. If we assume that $\mathbf{u}$ and $\mathbf{v}_0$ are both parallel to the x axis, and let θ_R (θ_S) be the angles between the z axis and the reference (signal) beam wave vectors, then

$$\mathbf{u} - \mathbf{v}_0 = \frac{\sin\theta_R + \sin\theta_S}{\lambda}\mathbf{e}_x, \tag{6.26}$$

where $\mathbf{e}_x$ is the unit vector in the x direction. The sinc function then becomes

$$\operatorname{sinc}\left\{\pi\frac{L}{\lambda F_2}(\sin\theta_R + \sin\theta_S)x\right\}. \tag{6.27}$$

For the special case where $\theta_S = 0$ (i.e., the signal beam incidents normally on the hologram surface), the expression Eq. (6.27) becomes Eq. (6.1).

As mentioned in Section 6.2, the sinc function in Eq. (6.25) creates a "mask" over the correlation function $g \star f$. This mask is a slit parallel to the y axis. Thus we have shift invariance in the y direction, which is perpendicular to the plane spanned by the reference beam and the signal beam. But we have very little shift invariance in the x direction. Note that from Eq. (6.27), the width of the sinc function is inversely proportional to L, but independent of the position z_c (which contributes only a constant phase factor).

According to the results above, we have shift invariance in the y direction, regardless of whether the hologram is placed at the Fourier transform plane. In reality, of course, the amount of shift invariance in the y direction is finite, and decreases as the hologram is placed further away from the Fourier transform plane. This is true not only for volume holographic correlators, but also for planar holographic correlators. The discrepancy with the predictions of Eq. (6.25) is due to the approximation made in Eq. (6.19). Instead of the exact cross-correlation $g \star f$, we have

$$h(\Delta\mathbf{v}) = \iint d\mathbf{v}\; g(\mathbf{v}) f^*(\mathbf{v} - \Delta\mathbf{v})\, e^{j2\pi\lambda z_c(\mathbf{u}-\mathbf{v}+\Delta\mathbf{v})\cdot\Delta\mathbf{v}} \operatorname{sinc}\{\pi\lambda L(\mathbf{u} - \mathbf{v} + \Delta\mathbf{v})\cdot\Delta\mathbf{v}\}. \tag{6.28}$$

If f and g have the same center $\mathbf{v}_0$, then the peak (or center) of the function $h(\Delta\mathbf{v})$ is at $\Delta\mathbf{v} = 0$, where we have

$$h(0) = \iint g(\mathbf{v}) f^*(\mathbf{v})\, d\mathbf{v}. \tag{6.29}$$

This is the same as the correlation function $g \star f$ at $\Delta\mathbf{v} = 0$. For nonzero $\Delta\mathbf{v}$, the function h starts to deviate from the value of $g \star f$. However, if g and f have strong correlation, then the value of $g \star f$ drops close to zero rapidly as we move away from $\Delta\mathbf{v} = 0$. This is also true for h, so the difference is not significant.

Suppose now that f and g do not have the same center. Let $\mathbf{v}_0$ be the center of f, and let $\mathbf{v}_0 + \Delta\mathbf{v}_0$ be the center of g. If f and g have strong correlation, then since h behaves similar to $g \star f$, the peak of h is at $\Delta\mathbf{v} = \Delta\mathbf{v}_0$. In this case there are two factors that cause the peak value of h to deviate from $g \star f$. The first is

$$O_1 = e^{j2\pi\lambda z_c(\mathbf{u}-\mathbf{v}+\Delta\mathbf{v}_0)\cdot\Delta\mathbf{v}_0}, \tag{6.30}$$

and the second is

$$O_2 = \operatorname{sinc}\{\pi\lambda L(\mathbf{u} - \mathbf{v} + \Delta\mathbf{v}_0) \cdot \Delta\mathbf{v}_0\}. \tag{6.31}$$

As $\Delta\mathbf{v}_0$ increases, the frequency of the O_1 factor increases. At the same time, the O_2 factor drops toward zero. Both factors cause the value of h to be less than $g \star f$.

The O_1 factor is 1 when the hologram is at the Fourier plane ($z_c = 0$). However, we still have the O_2 factor. In any case, as $\Delta\mathbf{v}_0$ increases (i.e., as the center of g moves away from the center of f), the function h drops toward zero. From Eq. (6.30), we note that the value of $\Delta\mathbf{v}_0$ where h approaches zero, becomes smaller as L becomes larger. Thus the range of shift invariance decreases as L increases, as we would expect.

Similarly, O_2 oscillates more rapidly for larger values of $|z_c|$ (for the same $\Delta\mathbf{v}_0$). This means that the peak of h decreases more rapidly to zero as the center

of g moves away from the center of f. Thus the range of shift invariance decreases as the hologram moves away from the Fourier plane. [7]

The amount of shift invariance depends primarily on the factors O_1 and O_2, which in turn depend on the thickness L and the position z_c. The second Fourier transform lens just magnifies (the spatial extent) of the correlation function $g \star f$ (or rather its approximation, h), so F_2 does not affect the amount of shift invariance. The focal length of the first Fourier transform lens, (F_1) however, does have an effect on the amount of shift invariance. Recall that in Eq (6.28) above, $g(\Delta\mathbf{v})$ should actually be $g(\lambda F_1 \mathbf{v})$. When F_1 decreases by a factor of m, the spatial extent of $f(\lambda F_1 \Delta\mathbf{v})$ and $g(\lambda F_1 \Delta\mathbf{v})$ (in Eq. (6.28)) are magnified by a factor of m. (The size of the Fourier transforms, on the other hand, shrink by a factor of m.) The amount of shift invariance $\Delta\mathbf{v_0}$, however, is approximately independent of the magnification m (and hence F_1). If we project $\Delta\mathbf{v_0}$ back through the first Fourier transform lens, the corresponding amount of shift is $\lambda F_1 \Delta\mathbf{v_0}$. Since we have not changed the actual f and g functions, and F_1 has decreased by a factor of m, the (relative) amount of shift invariance has decreased by an amount of m. This is also true if we keep F_1 fix and magnify f and g by a factor of m. (In this case, the amount of shift, $\lambda F_1 \Delta\mathbf{v_0}$, does not change, but since f and g are larger, the relative amount of shift invariance has decreased.)

Thus when (the spatial extent of) the Fourier transform is smaller, the amount of shift invariance (in the perpendicular direction) decreases proportionally. In practice, it is desirable to increase the recording speed of the system, which can be done by shrinking the Fourier transform. The tradeoff, as shown above, is that the amount of shift invariance drops.

---

[7] For planar holograms, $O_2 = 1$. Nevertheless, if the the hologram is not placed exactly at the Fourier plane, O_1 still causes the range of shift invariance to decrease.

References

1. K. Wagner and D. Psaltis, "Multilayer Optical Learning Networks," *Appl. Opt*, **26**(23), 5061-5076 (1987).
2. Y. Owechko, G. J. Dunning, E. Marom, and B. H. Soffer, "Holographic Associative Memory with Nonlinearities in the Correlation Domain," *Appl. Opt.*, **26**(10), 1900–1910 (1987).
3. D. Psaltis, D. Brady, and K. Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.*, **27**(9), 1752–1759 (1988).
4. Y. Qiao and D. Psaltis, "Local Learning Algorithm For Optical Neural Networks," *Appl. Opt.*, **31**(17), 3285–3288 (1992).
5. D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in Artificial Neural Networks," *Nature*, **343**(6256), 325–330 (1990).
6. H. Yoshinaga, K. Kitayama, and T. Hara, "All-optical Error-signal Generation for Backpropagation Learning in Optical Multilayer Neural Networks," **Opt. Lett.**, **14**(4), 202–204 (1989).
7. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," Chap. 8 in *Parallel Distributed Processing* Vol. 1, D. E. Rumelhart and J. L. McClelland eds. (MIT Press, Cambridge, 1986).
8. M. A. Neifeld and D. Psaltis, "Optical Implementation of Radial Basis Classifiers," *Appl. Opt.*, **32**(8), 1370–1379 (1993).
9. D. Brady, K. Hsu, and D. Psaltis, "Periodically Refreshed Multiply Exposed Photorefractive Holograms," *Opt. Lett.*, **15**(14), 817–819 (1990).
10. Y. Qiao, D. Psaltis, C. Gu, J. Hong, and P. Yeh, "Phase-locked Sustainment of Photorefractive Holograms using Phase Conjugation," *J. Appl. Phys.* **70**(8), 4646–4648 (1991).
11. H. Sasaki, Y. Fainman, J. E. Ford, Y. Taketomi, and S. H. Lee, "Dynamic Photorefractive Optical Memory," *Opt. Lett.* **16**(23), 1874–1876 (1991).
12. S. Boj, G. Pauliat, and G. Roosen, "Dynamic Holographic Memory Showing Readout, Refreshing, and Updating Capabilities," *Opt. Lett.*, **17**(6), 438–

440 (1992).

13. A. B. Vander Lugt, "Signal Detection by Complex Spatial Filtering," *IEEE Trans. Inform. Theory*, **IT-10**, No. 2, pp. 139–145, 1964.
14. J. Yu, F. Mok, and D. Psaltis, "Capacity of Optical Correlators," *Proceedings of SPIE*, **825**(22), 114–120 (1987).
15. C. Gu, J. Hong, and S. Cambell, "2-D Shift-invariant Volume Holographic Correlator," *Opt. Comm.* **88**(4–6), 309–314 (1992).
16. M. A. Neifeld and D. Psaltis, "Programmable Image Associative Memory using an Optical Disk," submitted to *Appl. Opt.*
17. C. L. Giles and T. Maxwell, "Learning, Invariance, and Generalization in High-order Neural Networks," *Appl. Opt.*, **26**(23), 4972–4978 (1987).
18. D. Psaltis, C. H. Park, and J. Hong, "Higher-order Associative Memories and Their Optical Implementations," *Neural Networks*, **1**(2), 149–163 (1988).
19. M. Mezard and J. P. Nadal, "Learning in Feedforward Layered Networks – the Tiling Algorithm," *J. Phys. A* **22**(12), 2191–2203 (1989).
20. K. Bløtekjaer, "Limitations on Holographic Storage Capacity of Photo -chromic and Photorefractive Media," *Appl. Opt.* **18**, 57–67 (1979).
21. Y. Taketomi, J. E. Ford, H. Sasaki, J. Ma, Y. Fainman, and S. H. Lee, "Incremental Recording for Photorefractive Hologram Multiplexing," *Opt. Lett.* **16**(22), 1774–1776 (1991).
22. F. Mok, "Applications of Holographic Storage in Lithium Niobate," (presented at OSA 1992 Annual Meeting) *OSA 1992 Annual Meeting Technical Digest*, Vol. 23, **WE1**, 102, Sept 1992.
23. D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple Storage and Erasure of Fixed Holograms in Fe-Doped $LiNbO_3$," *Appl. Phys. Lett.*, **26**(4), 182–184 (1975).
24. H.-Y. Li, Y. Qiao, and D. Psaltis, "Optical Network for Real Time Face Recognition," to appear in *Appl. Opt.* (1993).
25. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill Books, New York, 1968).