# Spatio-temporal beam synthesis and applications to photolithography

Thesis by

Boaz Salik

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, CA

1997

(1 April 1997)

To my parents and brothers

# Acknowledgments

I first wish to thank Prof. Amnon Yariv, my thesis advisor, for being an inspiring teacher and a good friend. Over the years I have worked with him, I have learned more than during all my previous years of school.

I am deeply indebted to Prof. Joseph Rosen, who helped me through my early graduate struggles, both scientific and personal, and has always been open in times of need.

I am grateful to all the scientists with whom I've had the privilege of interacting, including Prof. Aharon Agranat, Dr. Gilad Almogy, George Barbasthatis, Claudine Chen, Prof. Bruno Crosignani, Giang Dao, Jean-Jacques Drolet, Dr. Danny Eliyahu, Doruk Engin, Dr. Jing Feng, Ali Ghaffari, Dr. John Ianelli, Joel Jones, Dr. Anthony Kewitsch, Dr. John Kitching, Roger Koumans, Reginald Lee, Zhiwen Liu, William Marshall, Matthew McAdams, Raanan Miller, Dr. John O'brien, Dr. Sergei Orlov, Dr. Volnei Pedroni, Eva Peral, Rafael Piestun, Dan Provenzano, Dr. Randall Salvatore, Kirill Scheglov, Prof. Axel Scherer, Prof. Ronald Schrimpf, Prof. Mordechai Segev, Dr. Ali Shakouri, Atsushi Sugitatsu, Prof. Miklos Szilagyi,

Maggie Taylor, Xiao-lin Tong, Scott Vanessen, Jian Wu, Yong Xu, Dr. Yuanjian Xu, and Min Zhang.

Other friends who have my sincere gratitude include Lucinda Acosta, Jonnie Burton, Rachel Choppin, Linda Dozsa, Christina Ene, Lin Gilbert, Ben Goldberg, Carmi Goldstein, Dan Jurkowitz, Mehrzad Koohian, Joseph Kruml, Joy Mclees, Jana Rae Mercado, Sam Mordka, Shauna Olen, Rikke Thrane Pagh, Lillian Porter, Grace Rothrock Illiana Salazar, Paula Samazan, Aja Sanzone, Steve Shupper, Tia Smith, Diane Targovnik, Ilana Weiss, Travis Wheeler, and Nathalie Youseffian.

Finally but most importantly, I thank my parents, Hana and Joshua Salik, and my brothers, Erez and Omer Salik, for their unconditional love, support, and food.

# Abstract

This thesis explores techniques for and applications of free-space beam shaping. After reviewing the basic principles of scalar diffraction theory, I discuss and experimentally demonstrate several approaches to two- and three-dimensional transverse beam synthesis; these include analytical solutions of varying complexity as well as methods for computer optimization of beams with arbitrary constraints. Analytical solutions are also presented for the temporal analogy of nondiffracting beams, i.e., nondispersing pulses, and repercussions to time-dependent diffraction theory are discussed.

Next these beam shaping methods are applied to imaging photolithography, addressing ways to improve both resolution and focal depth therein with the use and proper design of phase masks. In this work it is evident that computation time plays a critical role in the applicability of phase masks to photolithography, because phase mask design algorithms tend to scale unfavorably with mask size. I therefore introduce an approximation to the Hopkins equation which reduces the computation time for partially coherent imaging by one to two orders of magnitude. Following this the question of spatial coherence in phase mask-assisted photolithography becomes interesting, and the optimal coherence for such

systems is investigated both theoretically and experimentally. The properties of incoherent imaging are next applied to a slightly different problem--imaging through random media. A new technique for ballistic imaging is presented, discussed theoretically, simulated, and demonstrated experimentally, and its advantages and drawbacks are analyzed. Finally, a theoretical overview of the fundamental limits to space-time beam shaping is presented, several results of which are demonstrated.

Due to the diversity of the subjects discussed, introductions and brief histories are given at the beginning of relevant chapters.

# Table of Contents

# Chapter One

# Scalar diffraction theory

Most of the problems addressed herein deal with scalar diffraction phenomena. Therefore, it is appropriate to begin with an overview of scalar diffraction theory, which I shall supplement with some remarks on time-dependent diffraction.

## 1.1. Preliminaries

Maxwell's equations are, in point form,

$$\begin{aligned} \nabla \cdot (\varepsilon \mathbf{E}) &= \rho \\ \nabla \cdot (\mu \mathbf{H}) &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial (\mu \mathbf{H})}{\partial t} \\ \nabla \times \mathbf{H} &= \mathbf{J}_{free} + \frac{\partial (\varepsilon \mathbf{E})}{\partial t} \end{aligned} \qquad (1.1a\text{-}d)$$

They describe the space-time relationship between the electric (**E**) and magnetic (**H**) field vectors, and are consistent with all observed electromagnetic phenomena. If $\mu$ is scalar and constant in time and position, then taking the curl of (1.1c) yields

$$\nabla \times \nabla \times \mathbf{E} = -\nabla \times \frac{\partial(\mu \mathbf{H})}{\partial t} = -\mu \frac{\partial}{\partial t}(\nabla \times \mathbf{H}) = -\mu \frac{\partial}{\partial t}(\mathbf{J}_{free} + \frac{\partial(\varepsilon \mathbf{E})}{\partial t}) \tag{1.2}$$

Assuming next that $\mathbf{J}_{free} = 0$ and that $\varepsilon$ is a scalar constant, (1.2) becomes

$$\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\mu\varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.3}$$

and, using (1.1a) with $\rho$=0, we arrive at the vector wave equation

$$\mu\varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla^2 \mathbf{E} - \nabla(\nabla \cdot \mathbf{E}) = \nabla^2 \mathbf{E} \tag{1.4}$$

In rectangular coordinates, we see (1.4) is equivalent to a system of three scalar wave equations, one for each component of **E**:

$$\nabla^2 E_i = \frac{1}{c^2} \frac{\partial^2 E_i}{\partial t^2}, \quad i = x, y, z \tag{1.5}$$

Therefore, in linear problems, we can solve the scalar wave equation for each component of **E** independently, and superimpose these components to form the vector solution. Our problem is then reduced to finding a solution of the scalar wave equation

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = 0 \tag{1.6}$$

given initial conditions or boundary conditions on $E(x,y,z,t)$.

### 1.2. Time-harmonic fields

At this point we assume a time dependence of the form

$$E(x,y,z,t) = f(x,y,z)e^{i\omega t} \tag{1.7}$$

i.e., a monochromatic field, which transforms (1.6) into the Helmholtz wave equation,

$$\nabla^2 f = -\frac{\omega^2}{c^2} f = -k^2 f \tag{1.8}$$

We next introduce the polar coordinates

$$\theta = \tan^{-1}\frac{y}{x}$$

$$\phi = \tan^{-1}\frac{\sqrt{x^2+y^2}}{z}, \qquad (1.9a,b,c)$$

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

which imply the inverse transformations

$$x = r\sin\phi\cos\theta$$

$$y = r\sin\phi\sin\theta, \qquad (1.10a,b,c)$$

$$z = r\cos\phi$$

and write (1.8) explicitly in terms of the polar coordinates:

$$\frac{1}{r}\frac{\partial^2}{\partial r^2}(rf) + \frac{1}{r^2\sin\phi}\frac{\partial}{\partial\phi}(\sin\phi\frac{\partial f}{\partial\phi}) + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 f}{\partial\theta^2} = -k^2 f \qquad (1.11)$$

If $f$ has no $\theta$- or $\phi$- dependence, this reduces to

$$\frac{1}{r}\frac{\partial^2(rf)}{\partial r^2} = -k^2 f \qquad (1.12)$$

whose solutions are the spherical waves

$$f(r) = A\frac{e^{ikr}}{r} \tag{1.13}$$

We can use these in Green's Theorem,

$$\iiint_V (g\nabla^2 f - f\nabla^2 g)dv = \iint_S (g\frac{\partial f}{\partial n} - f\frac{\partial g}{\partial n})ds \tag{1.14}$$

by letting $V$ be the volume between a spherical shell $S_1$ of radius $\varepsilon$ centered at $r_0$ and our surface of interest $S_0$ (Figure 1.1), i.e., $S = S_1 + S_0$ (this is necessary because Green's theorem applies only to singularity-free volumes). The left side of (1.14) vanishes if both $g$ and $f$ satisfy (1.8)

$$\iiint_V (g\nabla^2 f - f\nabla^2 g)dv = \iiint_V (gk^2 f - fk^2 g)dv = 0 \tag{1.15}$$

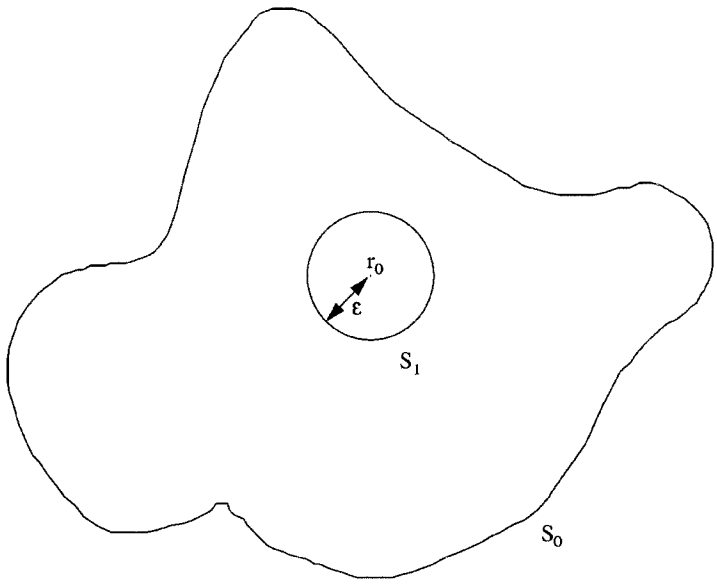



Figure 1.1. The singularity-free integration volume is between our closed surface of interest $S_0$ and a small spherical shell of radius ε around the singularity at $r_0$.

leaving

$$\iint_{S_0}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})ds=-\iint_{S_1}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})ds=4\pi\varepsilon^2(\frac{e^{ik\varepsilon}}{\varepsilon}\frac{\partial u(r_0)}{\partial n}-(\frac{1}{\varepsilon}-ik)\frac{e^{ik\varepsilon}}{\varepsilon}u(r_0)) \tag{1.16}$$

which, letting $\varepsilon \to 0$, yields the Helmholtz-Kirchhoff integral theorem:

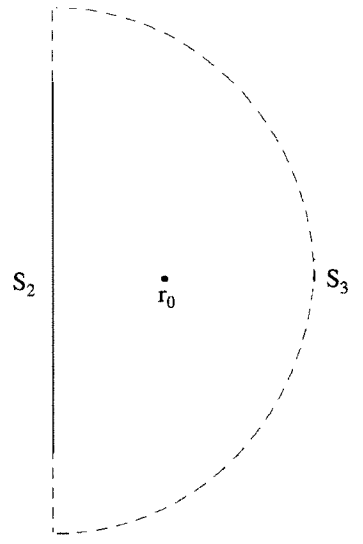



Figure 1.2. The closed surface $S_0$ consists of the infinite plane $S_2$ and the infinite hemispherical surface $S_3$.

$$u(r_0) = \frac{1}{4\pi} \iint_{S_0} \left( \frac{e^{ikr_{0S}}}{r_{0S}} \frac{\partial f}{\partial n} - f \frac{\partial}{\partial n} \frac{e^{ikr_{0S}}}{r_{0S}} \right) ds \tag{1.17}$$

where $r_{0S}$ is the distance from $r_0$ to the surface element under consideration. This specifies $f$ at every point inside a closed surface $S_0$ in terms of its value and derivative on the surface. In the case where we know the field and its derivative on an infinite flat plane $S_2$, we can calculate the field at any point by closing an infinite hemispherical surface $S_3$ around it (Figure 1.2), so that $S_0 = S_2 + S_3$. If $f$ satisfies the radiation condition [1],

$$f \underset{r\to\infty}{\longrightarrow} h(\theta,\phi)\frac{e^{ikr}}{r}, \qquad \frac{1}{f}\frac{\partial f}{\partial r} \underset{r\to\infty}{\longrightarrow} (ik-\frac{1}{r}) \tag{1.18}$$

the integral over $S_3$ disappears and the field everywhere is completely determined by the field and its derivative on the plane. To eliminate the dependence on the field's derivative at the plane, we can adopt the Rayleigh-Sommerfeld Green's function

$$g(r_{0S}) = \frac{e^{ikr_{0S}}}{r_{0S}} - \frac{e^{ik\tilde{r}_{0S}}}{\tilde{r}_{0S}} \tag{1.19}$$

where $\tilde{\mathbf{r}}_{0S}$ is the mirror image of $\mathbf{r}_{0S}$ about the surface $S_2$. Since $r_{0S} = \tilde{r}_{0S}$ and

$$\frac{\partial \mathbf{r}_{0S}}{\partial n} = \cos(\hat{\mathbf{n}}, \hat{\mathbf{r}}_{0S}) = -\cos(\hat{\mathbf{n}}, \hat{\tilde{\mathbf{r}}}_{0S}) = -\frac{\partial \hat{\mathbf{r}}_{0S}}{\partial n}, \tag{1.20}$$

(1.19) vanishes over $S_2$ while

$$\frac{\partial g}{\partial n} = 2(ik - r_{0S}^{-1})\frac{e^{ikr_{0S}}}{r_{0S}}\cos(\hat{\mathbf{n}}, \hat{\mathbf{r}}_{0S}) \tag{1.21}$$

which, substituted in (17), yields the Rayleigh-Sommerfeld integral

$$f(r_{0S}) = \iint_{S_0} i f(r_S)\left(\frac{1}{4\pi r_{0S}} - \frac{1}{\lambda}\right)\frac{e^{ikr_{0S}}}{r_{0S}}\cos(\hat{\mathbf{n}},\hat{\mathbf{r}}_{0S})ds \tag{1.22}$$

### 1.3. Analytic approximations to Rayleigh-Sommerfeld diffraction

There are several approximations we can make to this integral to make it more analytically tractable. First, we might assume our point $r_0$ is many wavelengths from the plane $S_0$:

$$r_{0S} >> \lambda \Rightarrow f(r_0) \cong \iint_{S_0} \frac{-i}{\lambda} f(r_S)\frac{e^{ikr_{0S}}}{r_{0S}}\cos(\hat{\mathbf{n}},\hat{\mathbf{r}}_{0S})ds \tag{1.23}$$

For simplicity, let us assign, without loss of generality, $S_0$ to be the $x$-$y$ plane at z=0. Then

$$\cos(\hat{\mathbf{n}},\hat{\mathbf{r}}_{0S}) = z / r_{0S} \tag{1.24}$$

and our integral becomes

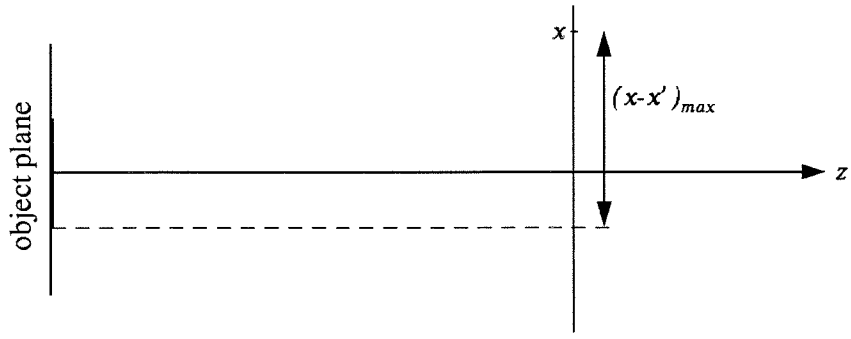



Figure 1.3. Under the Fresnel approximation (shown here in two dimensions), the axial distance ($z$) from the object must be much larger than the maximum transverse distance $(x\text{-}x')_{max}$ from any point on the object.

$$f(r_0) \cong \frac{1}{i\lambda}\iint_{S_0} f(x,y,z=0)\frac{ze^{ikr_{0S}}}{r_{0S}^2}ds$$
$$= \frac{1}{i\lambda}\iint_{S_0} f(x',y')\frac{ze^{ik\sqrt{z^2+(x-x')^2+(y-y')^2}}}{z^2+(x-x')^2+(y-y')^2}dx'dy' \qquad (1.25)$$

where $(x',y') = (x,y)\big|_{z=0}$.

Further simplification ensues if our axial distance from the object is much greater than the lateral displacement from any point on the object (Figure 1.3), i.e.,

$$z^2 >> (x-x')^2 + (y-y')^2 \quad \forall x', y' \ni f(x', y') \neq 0 \tag{1.26}$$

In this case,

$$\frac{z}{z^2 + (x-x')^2 + (y-y')^2} \cong \frac{1}{z} \tag{1.27}$$

and

$$\begin{aligned} e^{ik\sqrt{z^2+(x-x')^2+(y-y')^2}} &= e^{ikz\sqrt{1+\frac{(x-x')^2+(y-y')^2}{z^2}}} \\ &\approx e^{ikz\left[1+\frac{(x-x')^2+(y-y')^2}{2z^2}\right]} = e^{ikz} e^{\frac{ik}{2z}[(x-x')^2+(y-y')^2]} \end{aligned} \tag{1.28}$$

Under this condition, called the Fresnel approximation, (1.25) becomes

$$f(x,y,z) \cong \frac{e^{ikz}}{i\lambda z} \iint_{S_0} f(x', y') e^{\frac{ik}{2z}\left[(x-x')^2+(y-y')^2\right]} dx'dy' \tag{1.29}$$

If our observation point is so far from the object that

$$x >> x', \quad y >> y'$$

i.e., $$xx' >> x'^2, \; yy' >> y'^2 \tag{1.30a,b}$$

while (1.26) is maintained, it resides in the far field or Fraunhofer regime, where

$$f(x,y,z) \cong \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z}(x^2+y^2)} \iint f(x',y') e^{-i\frac{2\pi}{\lambda z}(xx'+yy')} dx'dy' \tag{1.31}$$

i.e., the diffracted field is proportional to a scaled Fourier transform of the object.

### 1.4. Time-dependent diffraction

To solve the wave equation (1.6), we first seek a solution $g$ satisfying

$$\nabla^2 g - \frac{1}{c^2}\frac{\partial^2 g}{\partial t^2} = -4\pi\delta(x-x')\delta(y-y')\delta(z-z')\delta(t-t') = -4\pi\delta(\mathbf{r}-\mathbf{r}')\delta(t-t'). \tag{1.32}$$

Since

$$\delta(\mathbf{r}-\mathbf{r}')\delta(t-t') = \frac{1}{16\pi^4}\iint e^{i(\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-\omega(t-t'))} d\omega d^3k \tag{1.33}$$

and the Fourier transform of $g(\boldsymbol{r}, t, \boldsymbol{r}', t')$, $G(\boldsymbol{k}, w)$, satisfies

$$g(\mathbf{r}-\mathbf{r}', t-t') = \iint e^{i(\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')-\omega(t-t'))} G(\mathbf{k},\omega) d\omega d^3k, \tag{1.34}$$

(1.32) gives us

$$G(k,\omega)=\frac{c^2}{4\pi^3(c^2k^2-\omega^2)} \tag{1.35}$$

whose inverse Fourier transform, assuming outwardly propagating waves, is

$$g(\mathbf{r}-\mathbf{r}',t-t')=\frac{1}{|\mathbf{r}-\mathbf{r}'|}\delta(t'-(t-\frac{|\mathbf{r}'-\mathbf{r}|}{c})). \tag{1.36}$$

Substituting this and (1.32) in Green's theorem (1.14) integrated over time:

$$\int_{\Delta t}\iiint_{V}(g\nabla^2 f-f\nabla^2 g)dv=\int_{\Delta t}\iint_{S}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})ds \tag{1.37}$$

and using the dummy variables $n'$, $s'$, $t'$, $r'$ for integration, we obtain

$$\int_{\Delta t}\iint_{S}(g\frac{\partial f}{\partial n'}-f\frac{\partial g}{\partial n'})ds'dt'=\int_{\Delta t}\iiint_{V}\left[4\pi f(\mathbf{r}',t')\delta(\mathbf{r}'-\mathbf{r})\delta(t'-t)-\frac{1}{c^2}\left(g\frac{\partial^2 f}{\partial t'^2}-f\frac{\partial^2 g}{\partial t'^2}\right)\right]dv'dt'$$

$$=4\pi f(\mathbf{r},t)-\frac{1}{c^2}\iiint_{V}\left[f\frac{\partial g}{\partial t'}-g\frac{\partial f}{\partial t'}\right]_{\Delta t}dv' \tag{1.38}$$

which is valid if we can change the order of the time and volume integrals. If we take the upper boundary of $\Delta t$ large enough that $g$ vanishes there, and call the lower boundary $t_0$, (1.38) can be rewritten

$$4\pi f(\mathbf{r},t) = \int_{\Delta t}\iint_{S}(g\frac{\partial f}{\partial n'} - f\frac{\partial g}{\partial n'})ds'dt' + \frac{1}{c^2}\iiint_{V}(g\frac{\partial f}{\partial n'} - f\frac{\partial g}{\partial n'})\Big|_{t'=t_0} dv' \qquad (1.39)$$

This gives us the value of $f$ at $(r,t)$ in terms of its values and derivatives on the bounding surface $S$ over the time interval $\Delta t$, and within the bounded volume $V$ at the time boundary $t_0$, analogously to the Kircchoff integral (1.17) for time-harmonic fields. The Kircchoff integral is, in fact, a special case of (1.39), from which it can be directly arrived at by assuming $f$ and $\partial f / \partial n'$ vanish at $t_0$ and a time-harmonic field [2]. Still, it is natural to suspect that (1.39) is not yet the most general diffraction integral, because its form is not intrinsically Lorentz-invariant, i.e., it does not treat the time variable symmetrically with the space variables. From a practical viewpoint, (1.39) is inapplicable to radiation from a time-varying surface, due to the change in integration order in (1.38). More general treatments, as well as applications to specific problems, can be found in References 3-6 and in chapter three of this thesis.

## 1.5. Summary

Although it invokes several approximations to Maxwell's equations, scalar diffraction theory explains a wide range of electromagnetic phenomena, most notably in free-space propagation. The paraxial (Fresnel) approximation, in particular, is accurate enough for almost all optical signal processing applications, and simple enough to yield analytic solutions to many problems [4]. With modern laser sources and filtering techniques, illumination systems can be made to approach complete monochromaticity, and even sources with a finite linewidth can usually be treated effectively as superpositions of monochromatic sources [6]. Time-dependent diffraction has therefore lost some of its relevance to classical applications, and its importace will be felt most strongly in modern fields utilizing ultrashort pulses, such as mode locking and high-speed optical communication [7].

**References**

1. J. D. Jackson, *Classical Electrodynamics*, New York: Wiley & Sons (1962), pp. 183-188.

2. Ibid., pp. 280-287.

3. J. Hadamard, *Lectures on Cauchy's Problem*, New York: Dover (1952).

4. J. Goodman, *Introduction to Fourier Optics*, San Francisco: McGraw-Hill (1968), pp. 30-76.

5. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, New York: McGraw-Hill (1953), pp. 843-847.

6. M. Born and E. Wolf, *Principles of Optics*, Oxford: Pergamon (1965).

7. A. Yariv, *Optical Electronics 3ed*, New York: Holt, Rinehart, and Winston (1985).

# Chapter Two

# Beam shaping

## 2.1. Introduction

In this chapter we explore ways to shape electromagnetic beams in two and three dimensions. Because the two-dimensional problem is more involved, I shall begin with the three-dimensional case. In 1987, J. Durnin showed [1] that the three-dimensional field distribution

$$E(x,y,z,t) = e^{i(\beta z-\omega t)} \int_0^{2\pi} A(\theta) e^{i\alpha(x\cos\theta + y\sin\theta)} d\theta, \tag{2.1}$$

where $\alpha^2 + \beta^2 = \omega^2/c^2$ and $A$ is an arbitrary function of $\theta$, is a solution of the three-dimensional scalar wave equation

$$(\nabla^2 + c^{-2}\partial^2/\partial t^2)E = 0. \tag{2.2}$$

Solution (2.1) contains the Bessel beam as a special case, and maintains a constant transverse intensity profile independent of propagation distance (z). In chapter 3 I shall present an alternative way of deriving (2.1) which yields several other interesting solutions. Several variations and generalizations of the Bessel beam have since been proposed [2-5], each with specific relative advantages. A number of unrelated "nondiffracting" beams have also been proposed [6-8]. In 1994, Rosen proposed a general method for optimizing the length of nondiffracting beams given arbitrary constraints on the aperture [9]. This method uses the Rayleigh-Sommerfeld integral [10]

$$u'(z) = u(r = 0, \theta, z) = \frac{1}{i\lambda} \int_0^{2\pi} \int_0^{\infty} f(r', \theta') \frac{e^{ikR}}{R} \cos\phi \cdot r' dr' d\theta' \tag{2.3}$$

to compute the axial field distribution $u'(z)$, where $\lambda$ is the wavelength, $k = 2\pi/\lambda$, $(r', \theta')$ are the aperture coordinates, $f(r', \theta')$ is the aperture field distribution, $R = \sqrt{r'^2 + z^2}$, and $\phi = \tan^{-1}\frac{r'}{z}$. By defining $\rho = r^2$ and $\zeta = z^2$, Rosen showed that we can derive a convolution relation between the axial field $u(\zeta)$ and the circularly averaged aperture distribution

$$t(\rho) = \int_0^{2\pi} f(\rho, \theta) d\theta \tag{2.4}$$

Using a POCS-type optimization algorithm [11,19], we can then impose the desired constraint on our aperture distribution, project it to the axial field, impose the desired axial constraint, and project back to the aperture. By iterating this procedure, we tend to converge to a solution which minimizes (locally) the error in realizing both constraints. This approach has been shown to outperform the various analytical variations of the Bessel beam, given corresponding aperture-plane constraints. By adding a Fourier-transforming lens behind the aperture and using the Fresnel approximation, Rosen and Yariv [12] were able to generalize the POCS optimization scheme to arbitrary axial field distributions.

### 2.2. Analytic two-dimensional beams

In two dimensions, the scalar wave equation becomes

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = 0 \tag{2.5}$$

If we require a harmonic $z$-$t$ dependence as in (2.1), i.e.,

$$E(x,z,t) = e^{i(\beta z - \omega t)} f(x), \tag{2.6}$$

we see by substituting (2.6) in (2.5) that the two-dimensional analogy of the Bessel beam is simply a plane wave,

$$f(x) = Ae^{i\sqrt{\omega^2-\beta^2}\,x} + Be^{-i\sqrt{\omega^2-\beta^2}\,x} \tag{2.7}$$

which is indeed nondiffracting but is also uninteresting. In order to obtain an analytical beamlike solution whose intensity remains constant over a finite subset of the propagation axis, Rosen, Salik, and Yariv [13] assumed the following field distribution at the front focal plane of a lens:

$$g(x') = \exp\left[-i2\pi\left(|x'/b|^p - (x'/a)^2\right)\right] \tag{2.8}$$

Assuming Fresnel diffraction, the field at the rear focal plane is

$$u(x,z) = \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}} \int_{-\infty}^{\infty} g(x') \exp\left[-i\frac{k}{f}\left(\frac{zx'^2}{2f} + x'x\right)\right] dx' \tag{2.9}$$

which, substituting (2.8) and using the stationary phase approximation [14], becomes

$$u(0,z) = e^{ikz}[C + D(z/2\lambda f^2 - a^{-2})] = e^{ikz}(C' + D'z), \qquad (2.10)$$

$$-\alpha\Lambda + a^{-2} < z/2\lambda f^2 < \Lambda + a^{-2}$$

where $C$ and $D$ are complex constants, $C' = C - D/a^2$, $D' = D/2\lambda f^2$, $\Lambda = p[4(1+p)]^{(2-p)/p}/(2b^2)$, and $0 < \alpha < 1$. If, on the other hand, we place $g^*(x')$ at the lens's front focal plane, we get the axial field distribution $u^*(0,-z)$, which is linear with the opposite slope of (2.10) over the mirror-image region on the $z$-axis; thus, by placing $[g(x') + g^*(x')]/2 = \cos\left[-i2\pi\left(|x'/b|^p - (x'/a)^2\right)\right]$ before the lens and setting the center of the linear region at zero ($a = \sqrt{2/\Lambda(1-\alpha)}$), we obtain a region of constant intensity centered at $z = 0$. For a given beam width, we have demonstrated [13] a doubling in focal depth (compared to Gaussian beams) using this technique.

### 2.3. Optimized two-dimensional beams

It is natural to try to improve this performance using optimization techniques, as we did in the three-dimensional case; in addition, we expect such an approach to make arbitrary one-dimensional axial distributions realizable. In the Fresnel approximation, the two-dimensional field $u(x,z)$ behind a transparency $g(x')$ is [10]

$$u(x,z) = \frac{e^{ikz}}{\sqrt{i\lambda z}} \int_{-\infty}^{\infty} g(x') \exp\left[\frac{ik}{2z}(x-x')^2\right] dx' \tag{2.11}$$

Next we wish to impose certain conditions on $u(x,z)$ for some subspace $R_0 \subseteq R$, $R=\{(x,z)/z > 0\}$. In general, we can represent these conditions as an error function $E$ to be minimized over $R_0$:

$$E[g(x')] = \iint_{(x,z)\in R_0} e[u(x,z)]dxdz, \tag{2.12}$$

where $e$ is the error density at point $(x,z)$.

For example, if we want the intensity $|u(x,z)|^2$ to come close to some $I_0(x,z)$ for $(x,z) \in R_0$, we might employ a Euclidean distance-square error function

$$E[g(x')] = \iint_{(x,z)\in R_0} [I_0(x,z) - |u(x,z)|^2]^2 dxdz$$

$$= \iint_{(x,z)\in R_0} \left\{ I_0(x,z) - \frac{1}{\lambda z} \left| \int_{-\infty}^{\infty} g(x') \exp\left[\frac{ik}{2z}(x-x')^2\right] dx' \right|^2 \right\}^2 dxdz \tag{2.13}$$

In general, the only way to find the $g(x')$ that minimizes $E$ is to try all possible complex functions--usually an unattractive option. If, however, we write $g(x') = A(x')\exp[i\phi(x_l)]$, where $A$ and $\phi$ are elements of a one-dimensional function space, so that $E[g(x')] = E(A,\phi)$, then assuming that $E$ is smooth in $A$ and $\phi$, we have

$$dE = \nabla_A E \cdot dA + \nabla_\phi E \cdot d\phi \tag{2.14}$$

where $\nabla_A E$ and $\nabla_\phi E$ are energy gradients with respect to amplitude and phase, respectively. Hence we see that sufficient (though not necessary) conditions for $dE = 0$ are that $\nabla_A E = 0$ and $\nabla_\phi E = 0$. Now we can optimize $A$ and $\phi$ independently using any convenient algorithm; although, in general, it may take more than one iteration of each to attain a local minimum for both amplitude and phase, this is still faster in most cases than treating the full ($N$ x $N$)-dimensional problem, which is computationally $N^2$ as complex as the $N$-dimensional problem ($N$ being the length of vectors $A$ and $\phi$ after quantization). A POCS-type algorithm may also be attempted, as done by Rosen for the three-dimensional problem, but we found in general that the algorithm converged very slowly and to poor solutions in the two-dimensional case.

Next we turn to the specific problem of shaping the intensity distribution along the propagation axis, $I(z)$, while maintaining some transverse constraint, $\left|u(x,z=z_0)\right|^2 \in S_x$. Here $S_x$ is the space of allowed transverse intensities.

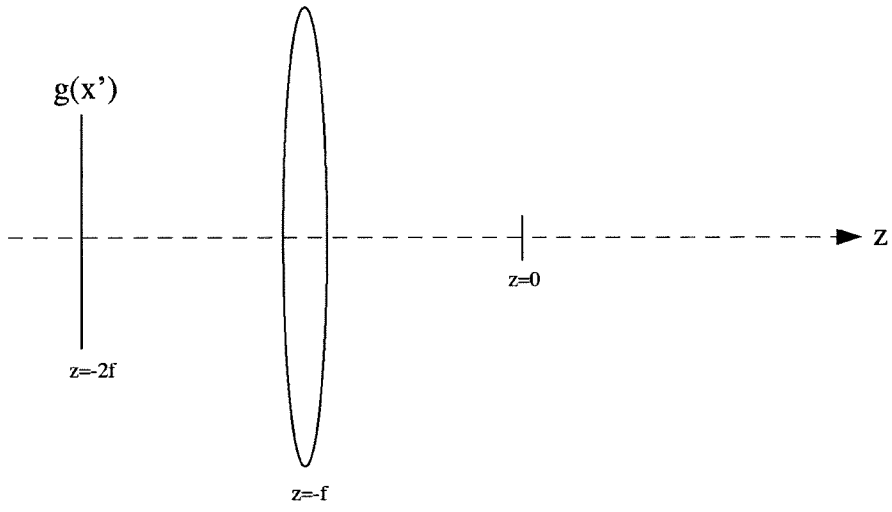



Figure 2.1. Configuration for realizing desired field around $z$=0, with mask at $z$=-2$f$.

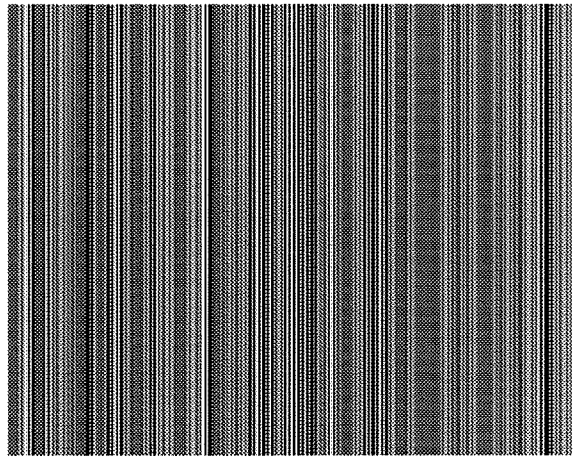


Figure 2.2. Mask used to generate one-dimensional PND beam.

Following Rosen and Yariv [12], we assume for computational convenience (and without loss of generality) that $z_0 = 0$ and there is a lens with focal length $f$ at $z = -f$ (Figure 2.1). If we place our transparency $g(x')$ at $z = -2f$, the one-dimensional Fresnel approximation yields [15]

$$u(x,z) = \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}} \int_{-\infty}^{\infty} g(x') \exp\left[ -\frac{ik(zx'^2 + 2fxx')}{2f^2} \right] dx' \tag{2.15}$$

and there is a convenient Fourier transform relationship between $g(x')$ and $u_t(x) = u(x, z=0)$, easing the computational burden of the transverse constraint. Unfortunately, the axial profile is not as easy to compute. Setting $x = 0$ in (2.15), we have

$$u'(z) = u(x=0,z) = \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}} \int_{-\infty}^{\infty} g(x') \exp\left( -\frac{ikzx'^2}{2f^2} \right) dx' \tag{2.16}$$

Although this can be treated as a Fourier transform in the variable $x'^2$, doing so introduces significant quantization error, which is avoided if we integrate directly. Next we can express $g(x')$ in terms of $u(x, z = 0)$ because of the Fourier relation

$$g(x') = \sqrt{\frac{i}{\lambda f}} e^{-2ikf} \int_{-\infty}^{\infty} u_t(x) e^{ikxx'/f} dx \tag{2.17}$$

Thus we can define a transformation between the $x$- and $z$-axis distributions behind the lens:

$$T[u_t(x)] = u'(z) = \frac{e^{ikz}}{\lambda f}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_t(x)\exp\left[\frac{ik(2fx'-zx'^2)}{2f^2}\right]dxdx' \qquad (2.18)$$

Now we can bypass any mask-plane functions by optimizing $u_t(x)$ subject to the constraint $u_t(x) \in S_x$ to yield $|u'(z)|^2$ as close as possible to our desired $I'(z)$. To optimize $u_t(x)$, we can still decompose it into amplitude and phase, as in (2.14). After arriving at the optimal $u_t(x)$, we simply use (2.17) to obtain our mask, $g(x')$.

In practice, our optical system has a finite aperture, placing limits on the validity of the Fresnel approximation. To quantify this, we add the third term of the binomial expansion in the expression for the longitudinal field distribution behind the lens [10]:

$$u'(z) = u(0,z) = \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}}\int_{-\infty}^{\infty} g(\tilde{x})\exp\left\{-ik\left[\frac{\tilde{x}^2}{2f} - \frac{\tilde{x}^2}{2(z+f)} + \frac{\tilde{x}^4}{8(z+f)^3}\right]\right\}d\tilde{x} \qquad (2.19)$$

where $g(\tilde{x})$ is the field immediately before the lens. There are two points of stationary phase in (2.19): $\tilde{x}_a = 0$ and $\tilde{x}_b = (z+f)\sqrt{-2z/f}$, where the root $\tilde{x}_b$ is

introduced by the fourth-order phase term. In the optical regime ($k >> 1$), in order to ignore the contribution of this root in the integral, we must satisfy the condition $(z+f)\sqrt{-2z/f} > D/2$, where $D$ is the lens diameter. From this condition we conclude that the Fresnel approximation is valid for $z > -f + D/\sqrt{8}$. With the conventional optics that we use (where $f >> D$) our beams do not exist in the region $z < -f + D/\sqrt{8}$, and the Fresnel approximation holds.

In order to realize the complex holograms as real, positive masks, we use a standard off-axis implementation:

$$\hat{g}(x') = 1 + \mathrm{Re}\left[g(x')e^{i2\pi x'\sin\theta/\lambda}\right], \tag{2.20}$$

which is real and positive if $g(x')$ is properly normalized. The exponential shifts the $z$-axis of our desired pattern to a transverse distance of $f\sin\theta$ after the lens. To see this, we substitute $g(x')e^{i2\pi x'\sin\theta/\lambda}$ into (2.15):

$$\begin{aligned} u(x,z) &= \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}} \int_{-\infty}^{\infty} g(x')e^{ikx'\sin\theta} \exp\left[-\frac{ik(zx'^2+2fxx')}{2f^2}\right] dx' \\ &= \frac{e^{ik(z+2f)}}{\sqrt{i\lambda f}} \int_{-\infty}^{\infty} g(x') \exp\left[-\frac{ik(zx'^2+2fx'\hat{x})}{2f^2}\right] dx' \end{aligned} \tag{2.21}$$

where $\hat{x} = x - f\sin\theta$. Now taking the real part adds the same pattern, inverted around $(x,z) = (-f\sin\theta\ ,\ 0)$, and the DC term simply contributes a focused spot at the origin, which does not affect our shifted pattern if $f$ and $\theta$ are large enough. A typical mask is shown in Figure 2.2.

A natural application of the axial beam shaping technique, requiring only one axial and one transverse constraint, is a one-dimensional PND beam. We forced $u_t(x)$ to assume a transverse Gaussian pulse width of $\sigma = 1_{\text{pixel}}$ and a background level < 25% of the peak amplitude (Figure 2.3a). $I'(z)$ was simply set to a constant over $\Delta z = 65_{\text{pixels}}$, with $2f = 128_{\text{pixels}}$, as shown in Figure 2.3b; outside this region there were no constraints on $I'(z)$. We used the error function

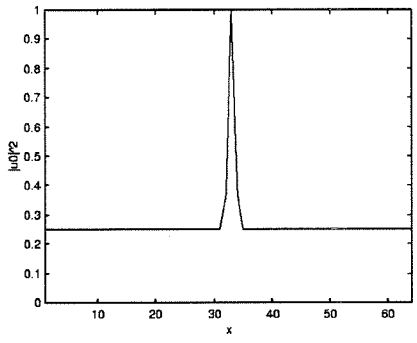
$$E[u_t(z)] = \int_{z\in\Delta z} [|u'(z)| - \sqrt{I'(z)}]^2 dz \tag{2.22}$$

where $u'(z)$ is given by (2.18). This is the Euclidean distance square in field amplitude, and a simple gradient descent [16] was used to alternately optimize $|u_t(x)|$ and Phase$[u_t(x)]$ (the algorithm always covered within eight iterations for vectors of 64 pixels). Figure 3c shows the simulated field amplitude along the $z$-axis (solid curves), compared with a Gaussian beam of the same width (dashed curves), and three transverse cross sections along the beam. Figure 3d shows the two intensities experimentally, over the portion of the $x$-$z$ plane close to the focus ($\Delta z$ = 15 cm, $f$= 30 cm).

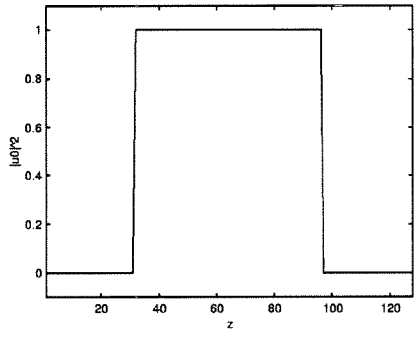
Next we utilize the more general method of considering the entire two-dimensional field distribution to realize a desired axial profile with more than one transverse constraint. Specifically, we wish to produce a sequence of two PND beams separated by a dark region (Figure 2.4a). Using integral (2.15) with a Euclidean error function in amplitude:

$$E[g(x')] = \iint_{(x,z)\in R_0} [\sqrt{I_0(x,z)} - |u(x,z)|]^2 dxdz, \tag{2.23}$$

and again we used a gradient descent to optimize $|g(x')|$ and Phase$[g(x')]$. The simulated and actual beam intensities over the $x$-$z$ plane are shown in Figures 2.4b and 2.4c, respectively, over a distance $\Delta z$ = 8cm. While the analytical solution above doubles the focal depth of a Gaussian beam, this optimization method yields an increase of tenfold in focal depth.

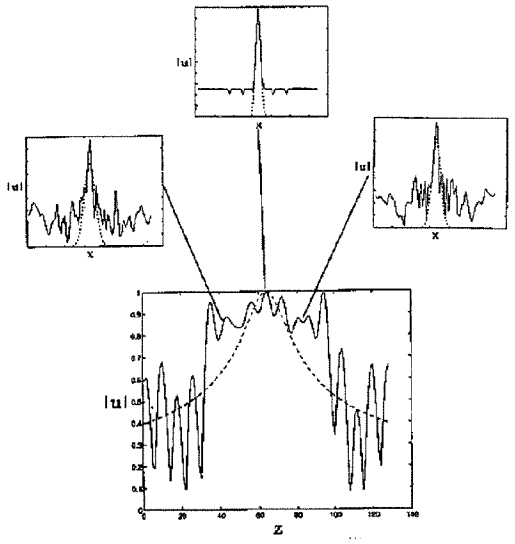



(a)

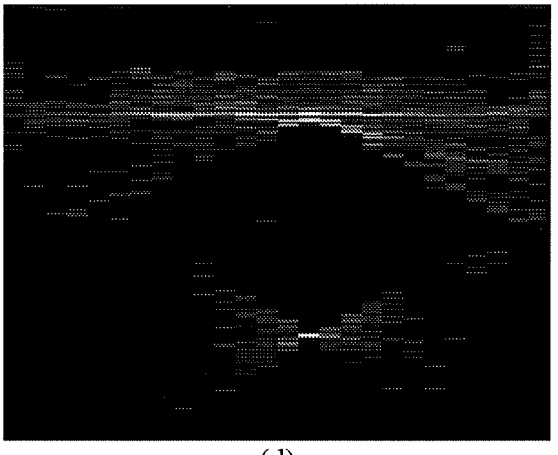



(b)

Figure 2.3. (a) Transverse intensity constraint for one-dimensional PND beam (Gaussian beam with sidelobe intensity < 25% of peak intensity), (b) axial intensity constraint for one-dimensional PND beam (zero indicates no constraint).

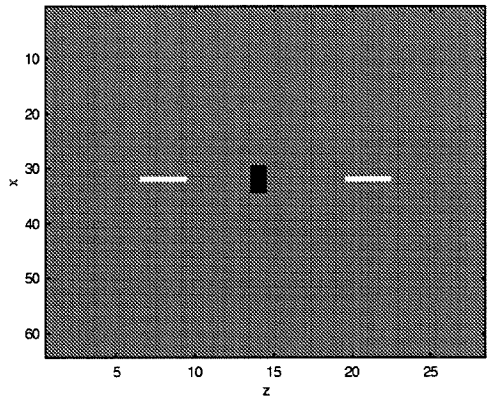



(c)

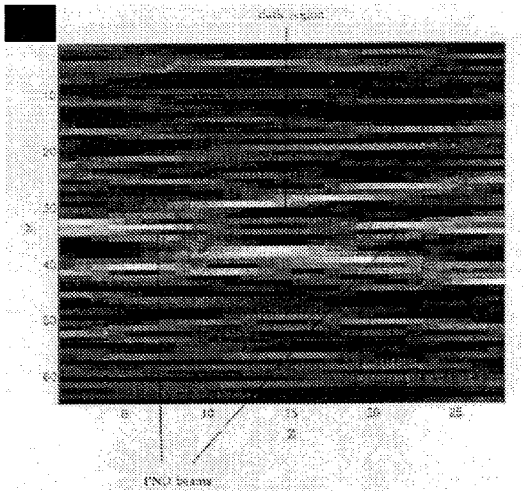


(d)

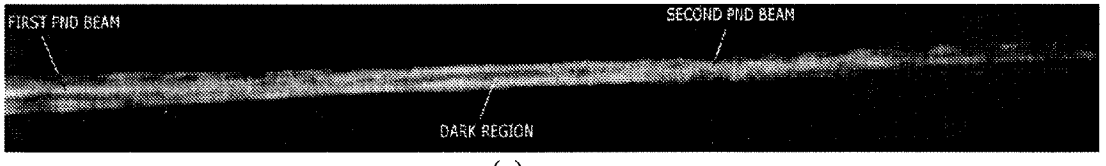
Figure 2.3. (c) Simulated field amplitude of optimized one-dimensional PND beam (solid) on the propagation axis and at three transverse planes along the axis, compared to a Gaussian beam of the same width (dashed). (d) Measured axial field amplitude distribution of optimized one-dimensional PND beam generated by the hologram in Figure 2.2 (top) and Gaussian beam of the same width (bottom).




(a)



(b)




(c)

Figure 2.4. (a) Two-dimensional constraint for realizing two PND beams separated by a dark region, (b) simulated intensity distribution over the $x$-$z$ plane, and (c) Observed intensity distribution over $\Delta z$=8 cm. Lens focal length=30 cm, $\lambda$=633 nm, and mask spatial frequency =206 $cm^{-1}$ (mask width = 1.24 cm).

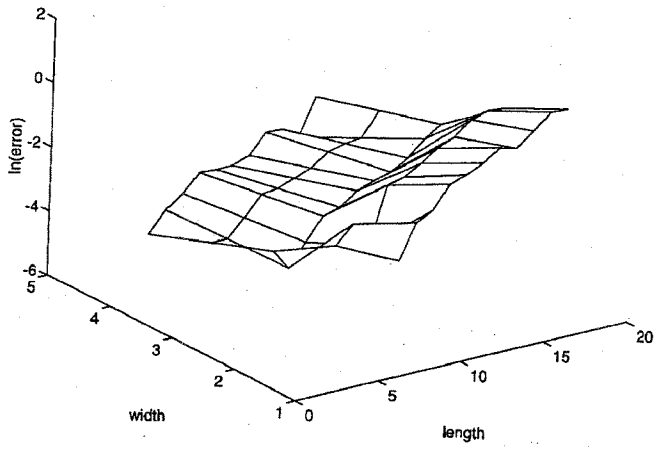



Figure 2.5. Log(total error) of optimized PND beam vs. beam width and beam length. Evidently the error increases with beam length and decreases with beam width. Thus, for a fixed allowable realization error, a larger beam width leads to a longer PND beam length, as in analytical solutions.

## 2.4. Limits of computer-optimized beams

Hafizi and Sprangle [17] have argued that analytical nondiffracting beams are necessarily infinite in extent, and therefore that truncating them reduces the

nondiffracting region to a finite subset of the propagation region. We therefore expect the same scaling laws to apply to optimized PND beams, and tested the dependence of the minimum z-axis error in optimized one-dimensional PND beams on the size of the $z$ constraint, $\Delta z$, and the allowed pulse width $\sigma$. In all finite-power canonical beams, the region of near-constant amplitude, $\Delta z$, increases with the beam width (e.g., for Gaussian beams we have [18] $\Delta z = n\pi\sigma^2/\lambda$). An appropriate analogy here is the region $\Delta z$ over which constant intensity can be attained within a given error. To see this dependence, we optimized PND beams with varying $\Delta z$ and $\sigma$ and plotted these parameters versus the minimum error attained (Figure 2.5). Although there are deviations, the clear trend is an increase in error with larger $\Delta z$ and smaller $\sigma$. Thus, to maintain constant error, any increase in focal depth $\Delta z$ requires an offsetting increase in the beam width $\sigma$. Additionally, it is clear from (2.16) that changing the wavelength $\lambda$ rescales the $z$-axis (i.e., increasing $\lambda$ contracts the $z$-axis) given a constant beam width, meaning the realized focal depth is inversely proportional to wavelength. As expected, optimization allows us to improve depth of focus but does not affect its general dependence on fundamental physical parameters.

**References**

1. J. Durnin, "Exact solutions for nondiffracting beams. I. The scalar theory," JOSA A 4, 651-654 (1987).

2. M. W. Kowarz and G. S. Agarwal, "Bessel-beam representation for partially coherent fields," JOSA A 12, 1324-1330 (1995).

3. R. H. Jordan and D. G. Hall, "Free-space azimuthal paraxial wave equation: the azimuthal Bessel-Gauss beam solution," Opt. Lett. 19, 427-429 (1994).

4. S. R. Mishra, "A vector wave analysis of a Bessel beam," Opt. Comm. 85, 159-161 (1991).

5. F. Gori, G. Guattari, and C. Padovani, "Bessel-Gauss beams," Opt. Comm. 64, 491-495 (1987).

6. J. Rosen, B. Salik, and A. Yariv, "Pseudo-nondiffracting beams generated by radial harmonic functions," JOSA A 12, 2446-2457 (1995).

7. J. H. McLeod, "The axicon: a new type of optical element," JOSA A 44, 592-597 (1954).

8. N. Davidson, A. A. Friesen, and E. Hasman, "Holographic axilens: high resolution and long focal depth," Opt. Lett. 16, 523-525 (1991).

9. J. Rosen, "Synthesis of nondiffracting beams in free space," Opt. Lett. 19, 369-371 (1994).

10. J. Goodman, *Introduction to Fourier Optics*, New York: McGraw-Hill, 1968, Ch. 3-4.

11. H. Stark, ed., *Image Recovery Theory and Application*, New York: Academic, 1987, Ch. 8.

12. J. Rosen and A. Yariv, "Synthesis of an arbitrary axial field profile by computer-generated holograms," Opt. Lett. 19, 843-845 (1994).

13. J. Rosen, B. Salik, and A. Yariv, "Pseudonondiffracting slitlike beam and its analogy to the pseudonondispersing pulse," Opt. Lett. 20, 423-425 (1995).

14. A. Papoulis, *Systems and Transforms with Applications in Optics*, New York: McGraw-Hill, 1968, Ch. 7, p. 222.

15. C. W. McCutchen, "Generalized aperture and the three-dimensional diffraction image," JOSA 54, 240-244 (1964).

16. L. R. Foulds, *Optimization Techniques*, New York: Springer-Verlag, 1981, p. 329-335.

17. B. Hafizi and P. Sprangle, "Diffraction effects in directed radiation beams," JOSA A 8, 705-717 (1991).

18. A. Yariv, *Optical Electronics*, Philadelphia: Saunders, 1991, p. 48-49.

19. B. Salik, J. Rosen, and A. Yariv, JOSA A **12**, 1702 (1995).

# Chapter Three

# Self-focusing pulses in free space

## 3.1. Introduction

Beginning in 1983 with Brittingham's focus wave mode [1], there have been several efforts to find nondispersing/nondiffracting solutions to the free-space scalar wave equation [2,3,4]. In 1987, Durnin introduced the Bessel beam [5], which maintains a constant transverse intensity profile on its propagation axis; this beam has a pulse-like transverse shape (e.g., see [6] for a definition), but is not temporally localized. In this chapter I shall introduce exact, closed-form solutions to the wave equation which are pulse-like both spatially and temporally, and maintain constant transverse shapes as they propagate; furthermore, we will find that the pulse width along its propagation axis narrows as it propagates from the origin, approaching a delta function with infinite distance.

### 3.2. Wave equation separability

We begin with the free-space three-dimensional scalar wave equation:

$$\nabla^2 E = \frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} \tag{3.1}$$

Next let us consider functions of the form

$$E(x,y,z,t) = E(x,y,\rho), \tag{3.2}$$

where

$$\rho^2 = z^2 - c^2 t^2 \tag{3.3}$$

With the introduction of the complex angular coordinate

$$\phi = \tan^{-1}\frac{ict}{z} \tag{3.4}$$

it is easy to show that

$$\frac{\partial^2 E}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial E}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 E}{\partial \phi^2} \tag{3.5}$$

Using (3.2) and (3.5), Eq. (1) simplifies to:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}(\rho\frac{\partial E}{\partial\rho}) = -(\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2}) \tag{3.6}$$

We now assume $E$ is separable between $(x,y)$ and $\rho$, i.e.,

$$E(x, y, \rho) = A(x, y)P(\rho) \tag{3.7}$$

so that, dividing (3.6) by $E$, we get

$$\frac{1}{P}\frac{d^2P}{d\rho^2} + \frac{1}{\rho P}\frac{dP}{d\rho} = -\frac{1}{A}(\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2}) \tag{3.8}$$

Since the left side of this equation is independent of $(x,y)$ and the right side is independent of $\rho$, we can equate both sides to some complex constant, say $\alpha^2$:

$$\frac{d^2P}{d\rho^2} + \frac{1}{\rho}\frac{dP}{d\rho} = \alpha^2 P \tag{3.9}$$

$$\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2} = -\alpha^2 A \tag{3.10}$$

Multiplying (3.9) by $\rho^2$, we arrive at the modified Bessel equation

$$\rho^2 \frac{d^2P}{d\rho^2} + \rho \frac{dP}{d\rho} - \alpha^2 \rho^2 P = 0 \tag{3.11}$$

whose solution is

$$P(\rho) = a_1 I_0(\alpha\rho) + a_2 K_0(\alpha\rho) \tag{3.12}$$

where $a_1$ and $a_2$ are arbitrary complex constants and $I_0$, $K_0$ are zero-order modified Bessel functions. Equation (3.10) has several solutions, which we will express as a general sum of plane waves:

$$A(x,y) = \sum_m [c_m e^{\frac{i\alpha(x+my)}{\sqrt{1+m^2}}} + d_m e^{-\frac{i\alpha(x+my)}{\sqrt{1+m^2}}}] \tag{3.13}$$

where $c_m$ and $d_m$ are arbitrary complex constants, and $m$ is real (we can further generalize this solution by summing $E_\alpha = A_\alpha(x,y)P_\alpha(\rho)$ over all possible $\alpha$; see chapter 8 for the repercussions of such generalizations).

If, for example, we assume (analogously to Eq. 3.2) that

$$A(x,y) = A(r) \tag{3.14}$$

where

$$r^2 = x^2 + y^2 \tag{3.15}$$

then (3.10) simplifies to

$$\frac{1}{A}\frac{d^2A}{dr^2} + \frac{1}{rA}\frac{dA}{dr} = -\alpha^2 A \tag{3.16}$$

or

$$r^2\frac{d^2A}{dr^2} + r\frac{dA}{dr} + \alpha^2 r^2 A = 0 \tag{3.17}$$

which has the Bessel-function solutions

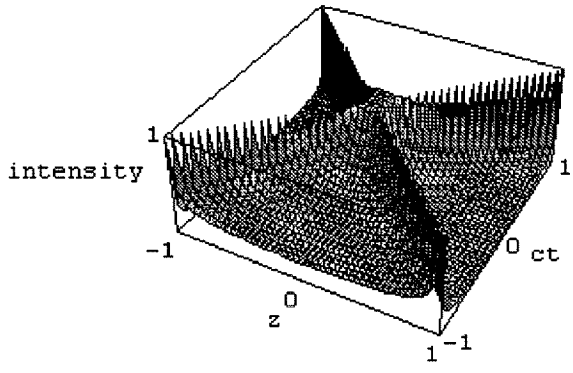
$$A(r) = b_1 J_0(\alpha r) + b_2 Y_0(\alpha r). \tag{3.18}$$

### 3.3. Nondiffracting self-focusing pulses

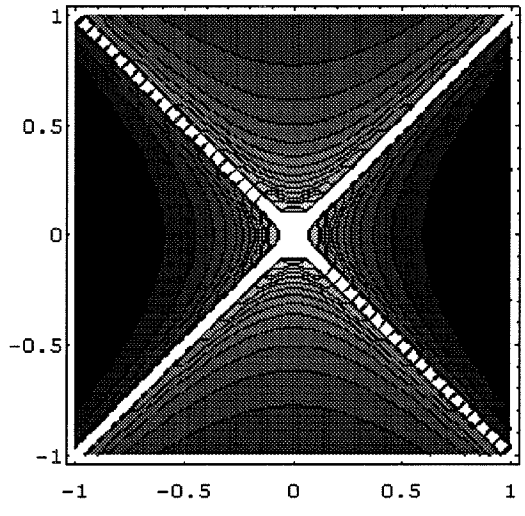
An interesting special case of (3.12) arises when $\alpha$ is real and $a_1 = 0$. We then have

$$E(x, y, z, t) = a_2 A(x, y) K_0(\alpha\sqrt{z^2 - c^2t^2}) \tag{3.19}$$

which has a pulse-like shape in the *z-t* plane. A notable property of solution (3.19) and, in fact, all solutions of the form (3.2), is that the transverse distributions




(a)




(b)

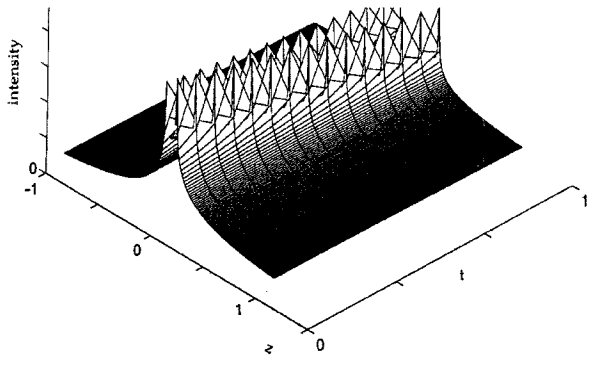
Figure 3.1. (a) Field intensity of solution (15) as a function of $(z,ct)$ with $\alpha=1$. (b) Contour plot of field intensity in $(z,t)$-plane, with c normalized to unity; narrowing contours indicate narrowing pulse width.

$$E(x,y,z,t)\big|_{z=ct} = E(x,y,z,t)\big|_{z=-ct} = A(x,y)P(0) \tag{3.20}$$

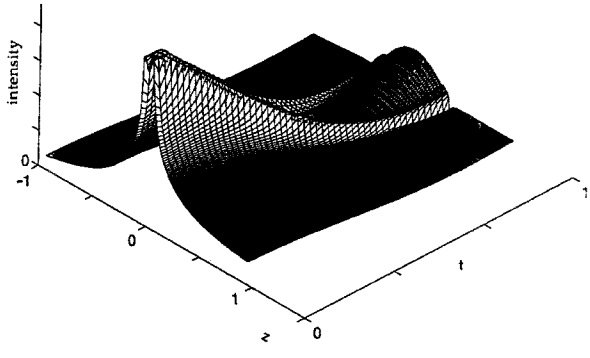
are independent of $z$ and $t$, and are therefore nondiffracting solutions propagating forward and backward on the $z$-axis at velocity $c$; the $z$-$t$ pulse is plotted in Figure 1a, where we see the pulse peak propagating from the origin on the axes $z=ct$ and $z=-ct$. This property is shared by the electromagnetic "splash modes" [7], which can also be written in form (2) and thus used as a basis for expanding the Bessel pulse. If for $A(x,y)$ we use solutions (3.18) with $b_2=0$, our pulses have the transverse distribution $J_0(\alpha r)$ and are therefore localized both spatially and temporally. Furthermore, if we examine the longitudinal pulse width as a function of distance traveled, we find it actually decreases as they propagate from the origin (Fig. 1b: the contours indicate regions of constant intensity; therefore, as the contour lines converge the pulses narrow). This follows from (3.19) if we rewrite it as

$$E(x,y,z,t) = a_2 A(x,y) K_0(\alpha\sqrt{(z+ct)(z-ct)}) \tag{3.21}$$

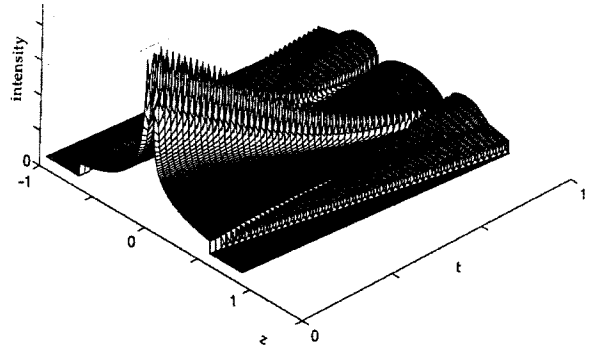
If, for example, we consider the pulse traveling at $z=ct$, the scaling factor $\alpha\sqrt{(z+ct)}$ causes it to narrow as $|z+ct|$ increases, forming a 'self-focusing' pulse (likewise for the pulse at $z=-ct$). Thus the temporal pulse is widest at $(z,t)=(0,0)$ while the spatial pulse maintains a constant width. By using an imaginary $\alpha$, we can




(a)




(b)




(c)

Figure 3.2. Numerical solutions of Eq. (3.9) with (a) unconstrained initial conditions: pulses behave as predicted by Eq. (3.15) and Figure 1; (b) constrained pulse intensity (25% of pulse power is clipped at the origin); and (c) constrained pulse width (pulse is truncated at $z=1/\alpha$).

interchange the $z,t$ and $x,y$ planes; however, these solutions are less useful because $J_0(\alpha\rho)$ diverges for imaginary $\rho$ and we lose the $z$-$t$ pulse shape.

### 3.4. Realization of self-focusing pulses

It should be noted that there are several potential difficulties in realizing such pulses. First, as we would expect from its convergent nature [8], the self-focusing pulse described above has an infinite temporal bandwidth, and therefore can only be approximated in practice (similarly, the $x$-$y$ nondiffracting Bessel beam has an infinite spatial bandwidth). Likewise, both the spatial and temporal pulses must be truncated at some maximal pulse width [5]. Finally, $K_0(\alpha\rho)$ becomes infinite as $\rho \to \infty$, and therefore our pulse must either be normalized by its zero-limit or clipped at some maximum intensity. These practical limitations inevitably alter the pulse properties discussed above; intuitively, we expect spatio-temporal truncation to decrease the pulse's lifetime, and frequency truncation to limit the minimum pulse width [8].

To affirm the above behavior, we calculated numerical solutions of (3.9) using the finite difference method [10, 14]. Figure 3.2a confirms the behavior seen in Figure 3.1 when the Bessel pulse is unconstrained by finite boundary/initial conditions. I simulated the effects of two constraints on the pulses' initial conditions: truncation of maximal pulse width and limiting of maximum pulse intensity. In Figure 3.2b

we see that clipping its peak intensity causes the Bessel pulse to widen into other modes; Figure 3.2c shows how truncating its maximal duration results in escape of the pulse's energy and deterioration of its lifetime. These results are consistent with general theoretical considerations [8,10] and with the qualitative discussion above. It is clear that the effectiveness of devices utilizing these pulses will be limited by their maximum power output and required pulse repetition rate. Practical use will also be complicated by the pulses' bidirectionality, although Einstein causality allows the destruction of one pulse without affecting the other (except the tail, which can be made to contain an arbitrarily small fraction of the pulse energy), since they both propagate at $c$.

### 3.5. Spherical pulses

The complex polar coordinates $(\rho,\phi)$ introduced in (3.3) yielded solutions of the wave equation which were symmetric within the $(x,y)$ and $(z,ict)$ planes, but implicitly distinguished between the two planes. It is natural to attempt to extend this symmetry by seeking solutions which treat all coordinates identically. To this end we introduce the following polar coordinates:

$$\theta = \tan^{-1}\frac{y}{x}$$

$$\phi = \tan^{-1}\frac{\sqrt{x^2+y^2}}{z}$$

$$\varphi = \tan^{-1}\frac{\sqrt{x^2+y^2+z^2}}{ict} \qquad (3.22a\text{-}d)$$

$$r = \sqrt{x^2+y^2+z^2-c^2t^2}$$

so that

$$x = r\sin\varphi\sin\phi\cos\theta$$

$$y = r\sin\varphi\sin\phi\sin\theta \qquad (3.23a\text{-}d)$$

$$z = r\sin\varphi\cos\phi$$

$$ict = r\cos\varphi$$

Note that these coordinates have no immediate geometric meaning, since they can assume complex values for certain values of $(x,y,z,t)$. In a straightforward manner, we can show that the Lorentzian

$$\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} \tag{3.24}$$

becomes

$$\frac{\partial^2}{\partial r^2}+\frac{3}{r}\frac{\partial}{\partial r}=\frac{1}{r^3}\frac{\partial}{\partial r}(r^3\frac{\partial}{\partial r}) \tag{3.25}$$

in the new coordinate system for functions which are independent of the angular coordinates $\theta$ ,$\phi$ ,$\varphi$ . Thus the wave equation becomes

$$\frac{\partial^2 E}{\partial r^2}+\frac{3}{r}\frac{\partial E}{\partial r}=\frac{1}{r^3}\frac{\partial}{\partial r}(r^3\frac{\partial E}{\partial r})=0 \tag{3.26}$$

whose solutions are the functions

$$E(r)=\frac{a}{r^2}. \tag{3.27}$$

Solutions (3.27) evidently are pulses (Figure 3.3a,b), and by introducing the spatial radius

$$\rho^2 = x^2 + y^2 + z^2 \tag{3.28}$$

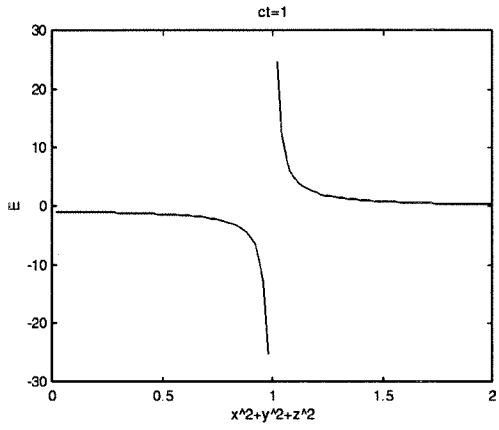
they can be rewritten

$$E(r) = \frac{a}{r^2} = \frac{a}{x^2 + y^2 + z^2 - c^2 t^2} = \frac{a}{(\rho + ct)(\rho - ct)}. \tag{3.29}$$

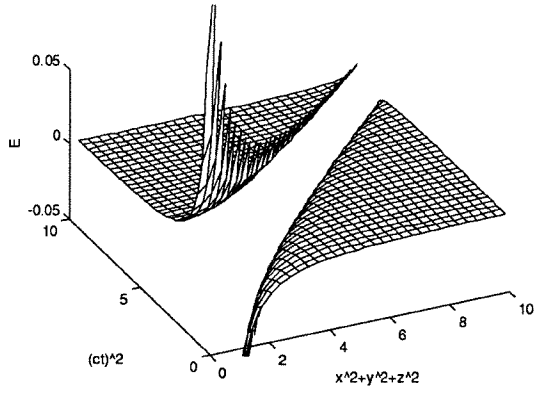
Thus the pulse propagates from the origin at $\rho = ct$, and around this point the solution becomes

$$E(r) \overset{\rho \cong ct}{\rightarrow} \frac{a}{2\rho(\rho - ct)} \cong \frac{a}{2ct(\rho - ct)}. \tag{3.30}$$

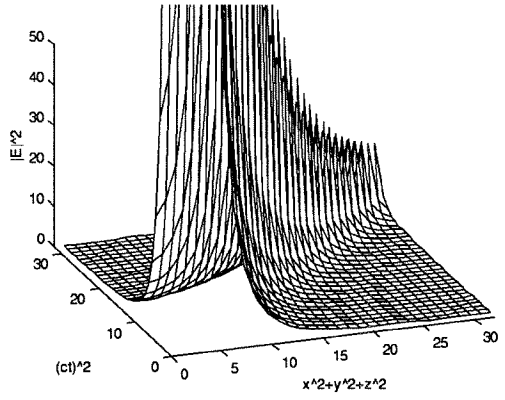
The scaling factor $2\rho = 2ct$ therefore causes the pulse to narrow as $\rho$ and t increase (Figure 3.3c,d), approaching zero width as $\rho = ct \rightarrow \infty$, analogously to the time pulses (3.19). We thus have a spherically symmetric pulse with a time-dependent singularity at radius $ct$, whose width is inversely proportional to its distance from the origin.


ct=1
E
x^2+y^2+z^2
30
20
10
0
-10
-20
-30
0
0.5
1
1.5
2

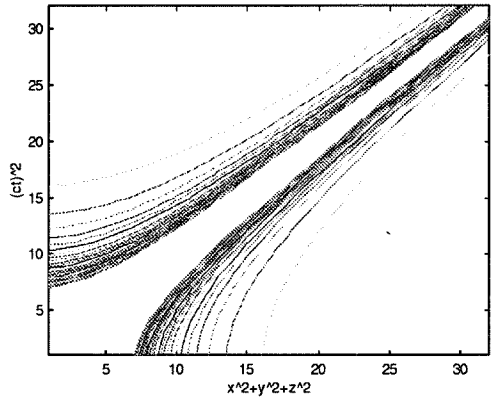
(a)


E
0.05
0
-0.05
(ct)^2
x^2+y^2+z^2
10
5
0
0
2
4
6
8
10

(b)




(c)




(d)

Figure 3.3. (a) Field $E$ of spherical pulse vs. $\rho^2=x^2+y^2+z^2$ given ct=1. Notice the field is odd around the singularity $\rho=1$. (b) Field $E$ of spherical pulse vs. $\rho^2=x^2+y^2+z^2$ and $ct$. The singularity propagates from the origin at speed $c$, i.e., $\rho=ct$. (c) Intensity $|E|^2$ of spherical pulse vs. $\rho^2=x^2+y^2+z^2$ and $ct$. The pulse width (FWHM) decreases as the singularity propagates from the origin. (d) Contour plot of intensity vs. $\rho^2$ and $ct$. Narrowing contours indicate decreasing pulse width with distance from the origin.

### 3.6. Time-dependent diffraction revisited

We next aim to use functions (7) as Green's functions in solving the wave equation. Interesting Green's functions for diffraction theory represent point sources and therefore contain singularities (e.g., time-harmonic spherical waves, or solutions (3.19)). We are particularly interested in using the spherical pulses because their form is inherently covariant, implying the resulting diffraction integral may also be. Because our functions are time-dependent, we must generalize Green's theorem [10],

$$\iiint_{\varsigma}(\not{\nabla}^2 g - g\nabla^2 f)dV = \iint_{\Sigma}\mathbf{N}\cdot(\not{\nabla} g - g\nabla f)d\Sigma = \iint_{\Sigma}(f\frac{\partial g}{\partial n} - g\frac{\partial f}{\partial n})d\Sigma \qquad (3.31)$$

where $\varsigma$ is the volume defined by a closed surface $\Sigma$, to a four-surface $S$ defining a four-volume $V$ [12]:

$$\iint_{S}\iint(\not{\Diamond}^2 g - g\not{\Diamond}^2 f)dS = \iiint_{V}\mathbf{N}\cdot(\not{\Diamond} g - g\not{\Diamond} f)dV \qquad (3.32)$$

Where $\Diamond = \frac{\partial}{\partial x}\hat{\mathbf{x}} + \frac{\partial}{\partial y}\hat{\mathbf{y}} + \frac{\partial}{\partial z}\hat{\mathbf{z}} + \frac{\partial}{\partial q}\hat{\mathbf{q}}$ and $q=ict$, $\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}},\hat{\mathbf{q}}$ being unit vectors along the x, y, z, and t axes, respectively. Here $S$ must be a closed four-surface defining the four-volume $V$. A typical closed four-surface is shown in Figure 3.4; it consists

of a time-dependent closed surface which starts at zero area, ends at zero area, and whose shape varies continuously over time (more complex surfaces can be constructed, but must obey the same criteria that describe surfaces in three dimensions [10,12]). Since

$$\Diamond^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + \frac{\partial^2}{\partial q^2} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \tag{3.33}$$

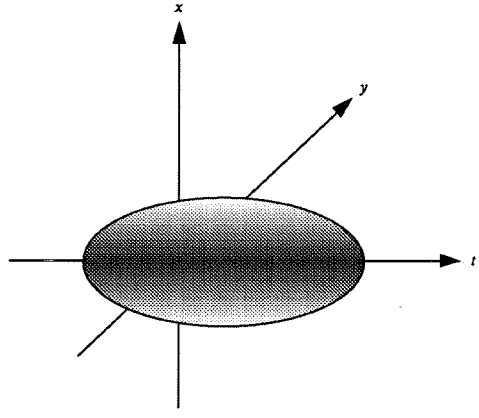



Figure 3.4. Typical closed four-surface projected onto three-dimensional $(x,y,t)$ space. In four dimensions, at each $t$ the four-surface consists of a closed surface over $(x,y,z)$, instead of the closed contours over $(x,y)$ shown above.

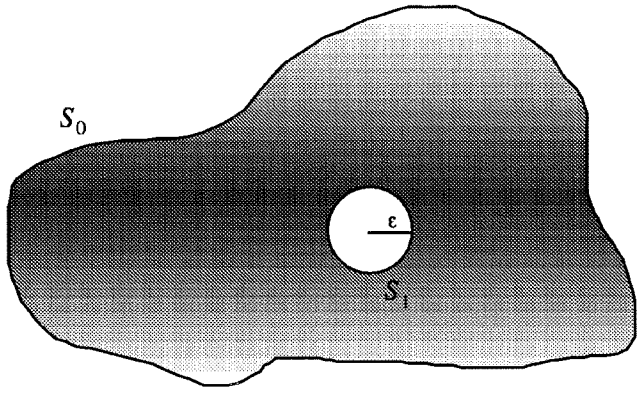



Figure 3.5. Four-volume of integration $V$ is defined by the four-surfaces $S_0$ and $S_1$, which together form the closed four-surface $S$.

the left-hand side of (3.32) disappears, leaving

$$\iiint_S \mathbf{N} \cdot (f\nabla g - g\nabla f)dS = \iiint_S (f\frac{\partial g}{\partial n} - g\frac{\partial f}{\partial n})dS = 0. \tag{3.34}$$

Since Green's theorem only applies to functions continuous on the volume in question (and with continuous first and second derivatives), we introduce, as we did in chapter 1, a small surface $S_1$ around the origin (Figure 3.5), let $S_0$ be our surface of interest, and take our volume of integration between $S_1$ and $S_1$, i.e., $S = S_1 + S_0$. Note that the origin $(x,y,z,t)=(0,0,0,0)$ is the only place where we have to

integrate out the singularity, because the spherical pulse is odd around its singularity everywhere else (Figure 3.3a,b). Thus (3.34) becomes

$$\iiint_{S_0}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})dS=-\iiint_{S_1}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})dV\,. \tag{3.35}$$

Using (3.27), we have at radius $r_{01}$ from our source

$$\begin{aligned} g(r_{01})&=\frac{1}{r_{01}^2},\\ \frac{\partial g}{\partial n}(r_{01})&=-\frac{2}{r_{01}^3}\cos(n,r_{01}) \end{aligned} \tag{3.36}$$

where the cosine is between the radius vector and the outward normal to the 3-surface $S$. We can define $S_1$ as a 4-sphere's shell so that on it $\cos(n,r_{01})=-1$ and

$$\begin{aligned} g(r_{01}=\varepsilon)&=\frac{1}{\varepsilon^2},\\ \frac{\partial g}{\partial n}(r_{01}=\varepsilon)&=\frac{2}{\varepsilon^3}. \end{aligned} \tag{3.37}$$

Thus as $\varepsilon$ gets small,

$$\iiint_{S_1}(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n})dS=2\pi^2\varepsilon^3(\frac{1}{\varepsilon^2}\frac{\partial f(r_{01})}{\partial n}-\frac{2}{\varepsilon^3}f(r_{01}))\overset{\varepsilon\to 0}{\to}-4\pi^2 f(r_{01})$$

(3.38)

and, returning to (3.35),

$$f(r_{01})=\frac{1}{4\pi^2}\iiint_{S_0}\left(g\frac{\partial f}{\partial n}-f\frac{\partial g}{\partial n}\right)dS_0=\frac{1}{4\pi^2}\iiint_{S_0}\left(\frac{\partial f}{\partial n}\cdot\frac{1}{r_{01}^2}-f\frac{\partial}{\partial n}(\frac{1}{r_{01}^2})\right)dS_0 \quad (3.39)$$

which is a generalization of the Helmholtz-Kircchoff integral theorem [13] to include time-varying sources and apertures. Due to the implicitly covariant form (3.22d) of $r_{01}$, we are guaranteed that (3.39) is Lorentz-invariant [15], and simplifies to the familiar (1.39) when we assume an open, time-independent integration surface, from which the familiar results of free-space scalar diffraction can be derived [16].

**References**

1. J. B. Brittingham, J. Appl. Phys. **54**, 1179 (1983).

2. R. W. Ziolkowski, J. Math. Phys. **26**, 861 (1985).

3. J. Lu and J. F. Greenleaf, IEEE Trans. UFFC **39**, 19 (1992).

4. R. W. Hellworth, "Focused one-cycle optical pulses," QELS (1995).

5. J. Durnin, JOSA A **4**, 651 (1987).

6. J. Rosen, B. Salik and A. Yariv, "Pseudo-nondiffracting beam generated by radial harmonic function," to be published in JOSA A.

7. I. M. Besieris, A. M. Shaarawi, and R. W. Ziolkowski, J. Math. Phys. **30**, 1254 (1989).

8. B. Hafizi and P. Sprangle, JOSA A **8**, 705 (1991).

9. G. Indebetow, JOSA A **6**, 150 (1989).

10. C. R. Wylie, *Advanced Engineering Mathematics 3ed*, New York: McGraw-Hill (1966).

11. L. Vicari, Opt. Comm. **70**, 263 (1989).

12. W. Fleming, *Functions of Several Variables 2ed*, New York: Springer-Verlag, 1977.

13. J. Goodman, *Introduction to Fourier Optics*, New York: McGraw-Hill, 1968, Ch. 2.

14. B. Salik and A. Yariv, "Exact solutions for nondiffracting, self-focusing pulses in free space," submitted to J. Math. Phys. (11/96).

15. W. G. V. Rosser, *Classical Electromagnetism Via Relativity*, New York: Plenum, 1968.

16. J. D. Jackson, *Classical Electrodynamics*, New York: Wiley & Sons (1962), p. 183-188.

# Chapter Four

# Photolithography and nondiffracting images

## 4.1. Introduction to imaging photolithography

Due to its inherently high throughput, imaging photolithography has become the dominant technology in the semiconductor industry. Photolithographic imaging systems ("steppers") consist of several high-aperture lenses which together are compensated for geometrical and chromatic aberrations at the design wavelength. They are therefore well-modeled as finite-aperture aberration-free imaging systems, as in Figure 4.1, and obey the same general resolution and focal depth relations [1]:

$$\Delta x = k_1 \frac{\lambda f}{D} \propto \frac{\lambda}{NA}, \quad \sigma = k_2 \frac{\Delta x^2}{\lambda} \propto \frac{\lambda}{NA^2} \tag{4.1a,b}$$

where $\Delta x$ is the mask-plane system resolution, $\lambda$ the wavelength, $f$ the focal length, $D$ the aperture, $NA$ the numerical aperture, $\sigma$ the mask-plane depth of focus, and $k_1$ and $k_2$ proportionality constants which depend on the processing method used. In order to increase the device and interconnect density at the wafer, the system resolution $\Delta x$ must be made as fine as possible, which in general means decreasing the wavelength $\lambda$ or increasing the numerical aperture. Increasing numerical aperture, however, dramatically reduces the focal depth, as in (4.1b); when the image's focal depth is exceeded by the photoresist depth or wafer curvature, the photoresist is exposed unreliably and defects may result. Therefore, it is more desirable to improve resolution by decreasing the illumination wavelength, and this strategy has prevailed in the industry for the last ten years. Currently, 248 nm excimer laser illumination is being used to print 0.25-micron features at advanced foundries, while 193 nm illumination and 0.18-micron features are at the intermediate development stages. At such wavelengths, it is challenging to find optical materials with sufficiently low absorption and sufficiently high refractive indices; fused silica is currently the material of choice in 248-nm systems, and is likely to prevail at 193 nm. Below this wavelength, it is unclear whether lenses have a relative advantage over mirrors; furthermore, since there are currently no economical lasers below 193 nm, synchrotron sources may be the only option for high-volume manufacturers. It is therefore clear that any system modifications outside the stepper which improve resolution or focal depth are eagerly pursued by the industry.

$$\Delta x \cong \frac{\lambda f}{D} \cong \frac{\lambda}{NA} \qquad \sigma \cong \frac{\Delta x^2}{\lambda} \cong \frac{\lambda}{NA^2}$$

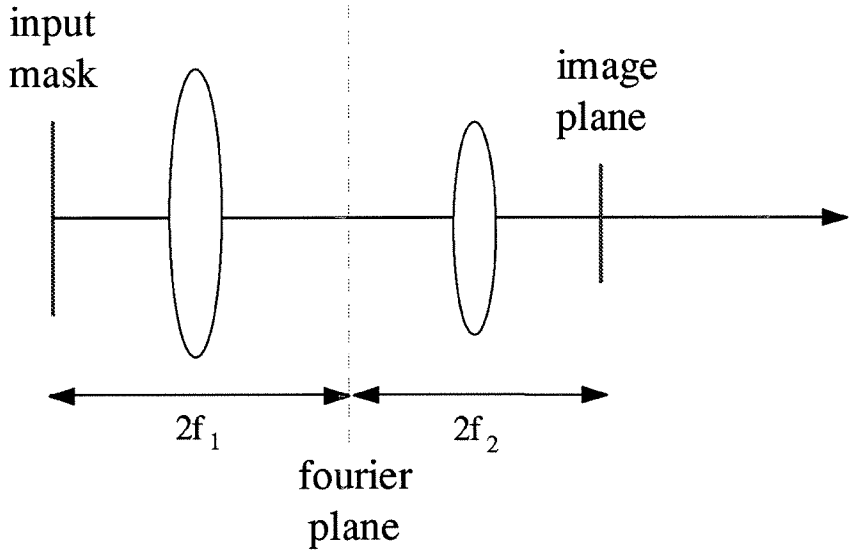
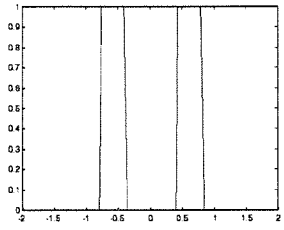


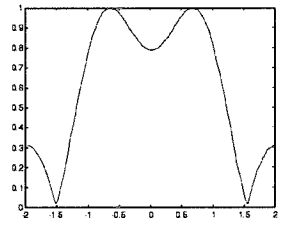

Figure 4.1 Photolithographic imaging systems (steppers) can be modeled as finite-aperture, aberration-free demagnifying 4-f imaging systems at the design wavelength, where most aberrations are compensated for.

**4.2. Phase masks**

Phase masks were introduced to photolithography by Marc Levenson in 1982 [2], and have since been shown to improve both resolution and focal depth therein. Traditional photolithography masks are binary-transmission, meaning they either



(a)



(b)

Figure 4.2. The slits in (a) are not resolved (b) after imaging by a finite aperture whose impulse response is wider than the slit separation.

absorb or transmit the incident illumination at a given pixel (spatial element). Phase masks have the added capability of introducing a phase shift at each pixel. Their performance is based on the destructive interference between adjacent features with opposite phase, after being imaged by a finite aperture. For example, consider the simple two-feature pattern in Figure 4.2a. In the Fresnel

approximation, an aberration-free imaging system multiplies the Fourier transform of the image by the pupil function [3]; thus in one dimension,

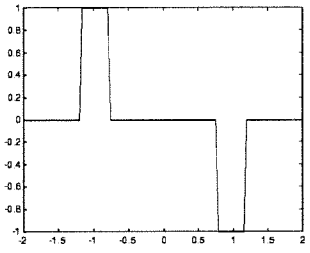
$$F[E_o(x)] = F[E_i(x)]P(x/\lambda f) \tag{4.2}$$

where $E_i(x)$ is the mask pattern (field) to be imaged, $E_o(x)$ is the output field, $P$ is the pupil function (a Rect if no aberrations are present), and $F$ denotes Fourier transforming. In the spatial domain, this equation becomes

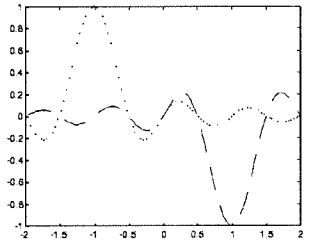
$$E_o(x) = E_i(x) * F^{-1}[P(x/\lambda f)] \tag{4.3}$$

where * denotes convolution. Thus the pattern in Figure 4.2a, composed of two adjacent slits, is unresolved after imaging (Figure 4.2b) because the field between them is too large (and so, therefore, is the intensity). However, when one of the slits assumes a $\pi$ phase shift as in Figure 4.3a, their fields interfere destructively after imaging (Figure 4.3b), and the features remain well-resolved (Figure 4.3c). For partially coherent illumination [4], it is more accurate to use the Hopkins formula

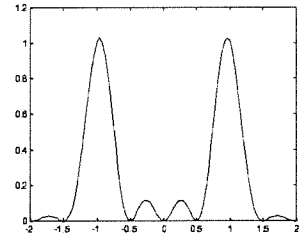
$$\left|E_o(x,z)\right|^2 = \iint E_i(x')E_i{}^*(\tilde{x}')J(x',\tilde{x}')K(x',x,z)K^*(\tilde{x}',x,z)dx'd\tilde{x}' \tag{4.4}$$



(a)



(b)



(c)

Figure 4.3. Fields of two adjacent features (a) interfere destructively (b) after imaging by an aperture whose impulse response is wider than the slit separation, and the features are well resolved when intensity is measured (c).

instead of (2); here $K$ is the impulse response $F[P]$, $J$ is the mutual intensity function, $E_i(x')$ is the field at the mask, and $E_o(x,z)$ *is* the field around the wafer (image plane; see Figure 4.4). The effects of phase masks are then complicated by the fact that the destructive interference between adjacent features is hampered by their less-than-unity correlation. These effects are discussed at length in chapter 6, where it is shown that phase masks always offer some improvement over transmission masks, although the size of the advantage can vary significantly with illumination coherence.

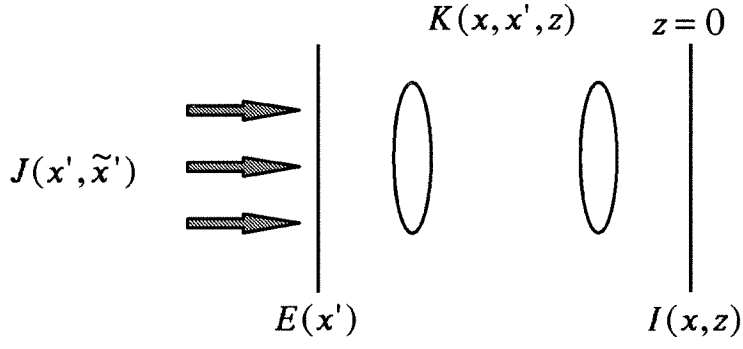



Figure 4.4. Mask field $E$ illuminated by a uniform-intensity, partially coherent source with mutual intensity function $J$, forms image intensity $I$ (given by the Hopkins formula) after propagating through system with impulse response $K$.

One-dimensional images can always be assigned a phase layout in which adjacent features have opposite phase. This, however, is not true for two-dimensional images. Consider, for example, the image in Figure 4.5a. Because it contains three features, each of which is adjacent to the other two, we must assign the same phase to two of the features, as in Figure 4.5b. Thus, while the entire original

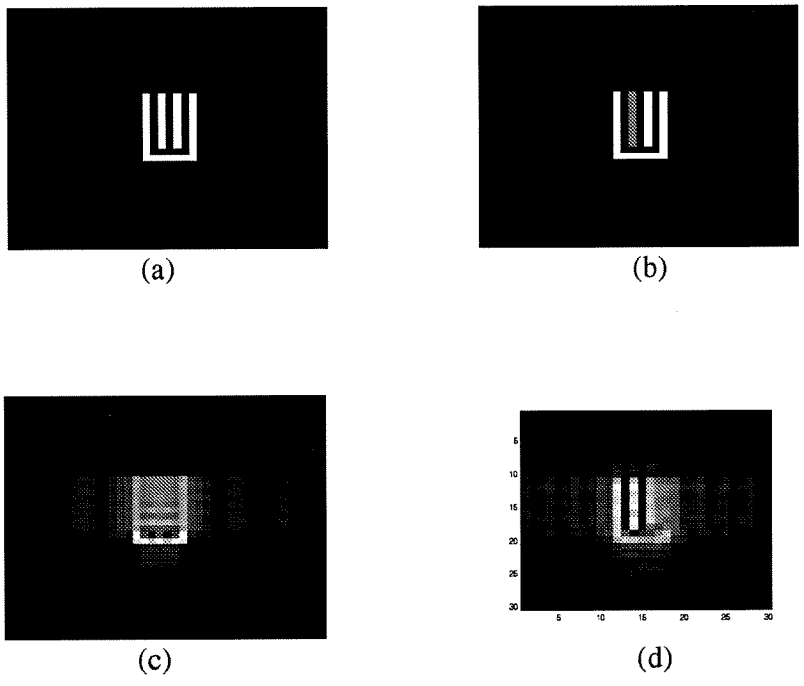



Figure 4.5. Phase conflict: Image in (a) has three adjacent features and therefore suffers phase conflict (b: gray feature is $\pi$-phase shifted). Both the transmission mask and the phase mask patterns are therefore unresolvable after imaging (c and d).

(transmission) mask suffers unacceptable distortion after imaging (Figure 4.5c), the phase mask also becomes unacceptably distorted between the two same-phase features (Figure 4.5d). This effect, intrinsic to images of two (or more) dimensions, is called phase conflict, and is one of the problems this chapter addresses.

Important work in the area of phase mask layout was done by Zakhor's group [5-7]. They have adopted a pixel-by-pixel approach to designing phase masks, which entails finding an optimal amplitude and phase distribution over the mask using an optimization algorithm similar to gradient descent [6]. This approach has several problems. First, their pixel-wise optimization time scales with the number of pixels to the third power (or more), due to the fact that the imaged pattern must be calculated for each mask variation, and these variations must be scanned over all the pixels. Given industrial mask sizes of at least $10^4 \times 10^4 = 10^8$ pixels, the run times become prohibitive. This issue is addressed in their recent paper [7], where they introduce a modified algorithm which saves the scanning over pixels, and therefore scales as the number of pixels squared (for calculating the imaged pattern). Although that is an improvement, it is still not viable for industrial-scale masks. A second problem with this approach is that it requires at least four levels of phase to be effective. This is because it is difficult to distinguish "important" mask areas (where resolution is critical) from "unimportant ones" (where some diffraction is tolerable), and because optimization algorithms such as gradient

descent tend to find local optima, not global ones (due to their faster computation times). This problem is also inherent in the approach of Ref. [7], and makes the technique unappealing due to the difficulty of manufacturing four-phase-level masks.

Another contributor to this area is Watanabe, who takes a multiple-exposure approach to solving phase conflict [8]. Although this approach has been discussed in several places, it is inherently unattractive to the industry because it involves twice the photolithography time as single-exposure techniques, and also introduces alignment and mechanical errors between the two exposures.

Kailath discusses a rule-based approach based on superposition of subsolutions, which scales with the number of pixels squared [9]. However, the improvement introduced is in the computation time for the imaged pattern, and entails an approximation of partially coherent imaging by a fully coherent system. This is accurate only for near-fully coherent imaging due to its reliance on the linearity of field superpositions. Since all lithography currently uses partially-coherent to fully-incoherent illumination, this approach is applicable to only a small fraction of practical cases. Furthermore, the actual optimization employed by Kailath is a Gerschberg-Saxton type algorithm, which is fast but tends to settle on highly suboptimal minima [10], sacrificing phase mask performance for computational

speed. Finally, this approach also requires either four phase levels per mask or two-mask imaging [9], both of which are problematic to manufacture.

Besides phase conflict, the traditional Levenson design also does not guarantee any improvement in focal depth, although it is not inherently incompatible therewith. Given an imaging system, either resolution or focal depth may be the factor limiting feature size, and we would like an approach to phase mask design which can address either or both. It is tempting to adopt the beam shaping methods discussed in chapter 2 to the specific problem at hand.

### 4.3. Phase mask design via beam shaping

We begin by considering the image constraints imposed by the photoresist. A typical photoresist response curve is shown in Figure 4.6. Below some threshold intensity $I_0$ (0.3 in the figure) the incident illumination is not detected; above some intensity $I_1$ (0.7 in the figure) the illumination is detected; and between these two levels the detector behaves unpredictably. It is this intermediate region that we wish to avoid. This allowed intensity variation affords us a degree of freedom in designing the mask--instead of imposing a transmissivity of zero or one (binary transmission), we are allowed in dark regions to transmit an amplitude as high as $\sqrt{I_0}$ and in light regions as low as $\sqrt{I_1}$. Additionally, since the photoresist is insensitive to the field's phase, we may implement an arbitrary phase shift at each

mask pixel and therefore design the phase distribution to maximize our image's focal depth or resolution.

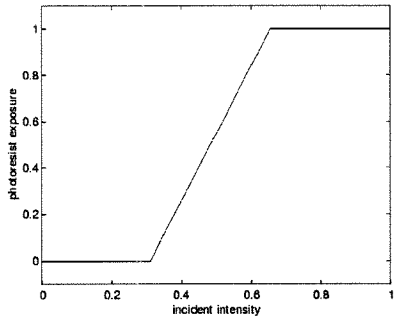



Figure 4.6. Typical photoresist exposure curve; photoresist performs a thresholding. Incident intensities below the threshold interval (.3 - .7 here, in normalized units) leave resist unexposed, incident intensities above the threshold interval leave resist completely exposed, and incident intensities within the threshold interval expose the photoresist unreliably.

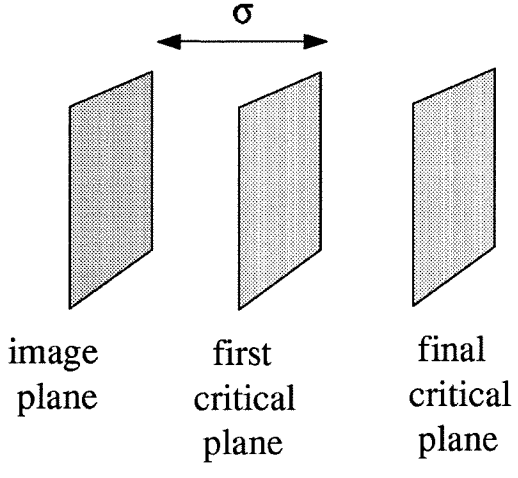



Figure 4.7. Extended focal depth of $k\sigma$ can be divided into subintervals of $\sigma$, over each of which the image must remain focused to insure the overall depth of focus.

We are now in a position to tackle the mask design problem. To extend our original depth of focus $\sigma$ by some multiple, say to $k\sigma$, we must ensure that the image remains focused in the entire volume between the image plane and the final focused plane. Given some minimum feature size $\Delta x$, the distance $\sigma$ over which we know little diffraction will occur is given by (4.1b). Thus, to ensure a focal depth of $k\sigma$, we must check the image at intervals of $\sigma$ (henceforth termed critical planes) and insure that it remains focused at each of these (Figure 4.7). To this end, we define an error measure (e.g., Euclidean distance square--cf. chapter 2) after thresholding that quantifies the deviation of our image from the desired one, calculate it at each critical plane, and sum over all these planes to obtain the total error of our field distribution. Because we are concerned with image coordinates that may stray significantly from the z axis, we use the Rayleigh-Sommerfeld kernel [11]

$$k(x,x',y,y',z) = \frac{ze^{ikR}}{i\lambda R^2}, \qquad R = \sqrt{(x-x')^2 + (y-y')^2 + z^2} \tag{4.5}$$

in the Hopkins equations to calculate the field diffracted from the image plane. The field at the image plane is calculated from the mask-plane field using the kernel in (4.3) generalized to two transverse dimensions, assuming a pupil function

$$P(r/\lambda f) = \mathrm{circ}(r/D) \Rightarrow P(\rho) = \mathrm{circ}(\lambda f \rho / D) \tag{4.6}$$

where $r = \sqrt{x^2 + y^2}$ and $\rho = \sqrt{u^2 + v^2}$ are the space- and frequency-domain radial variables, respectively, meaning

$$k(x,x',y,y',z=0) = \frac{J_1\left(D\sqrt{(x-x')^2 + (y-y')^2}\right)}{D\sqrt{(x-x')^2 + (y-y')^2}} \tag{4.7}$$

Next we independently optimize the mask's amplitude and phase distributions, a process which (as shown in chapter 2) converges to an error minimum for the complex mask distribution. In this case we iteratively perform a pattern search optimization [12, 13] on the amplitude and phase until a common minimum is attained. We optimize until a postthresholding error of zero is reached on all critical planes (if zero error is not attainable, we consider the problem unsolvable by our method, given mask resolution and photoresist thresholding parameters). At this point we have the desired complex mask distribution, which may be physically realized in several different ways [14].

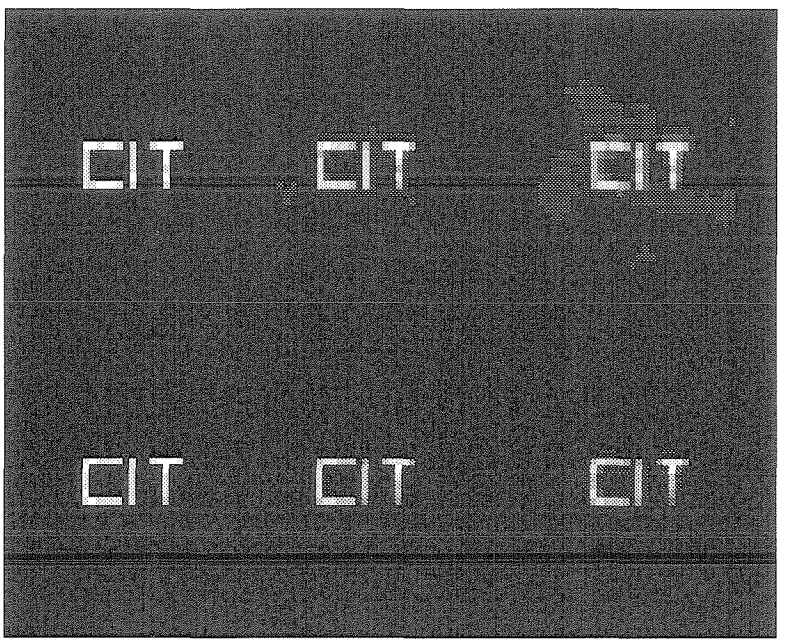



Figure 4.8a. Simulated intensity patterns at three critical planes of optimized amplitude-phase mask (top row) and of binary-amplitude mask (bottom row). Notice the amplitude-phase mask remains well-focused on all critical planes, while the transmission mask defocuses substantially.

Figure 4.8b. Simulated intensity patterns after thresholding of optimized amplitude-phase mask (top row) and of binary-amplitude mask (bottom row). Notice the binary-amplitude mask incurs major defects at two critical planes, while the optimized mask suffers no defects throughout the focal volume.

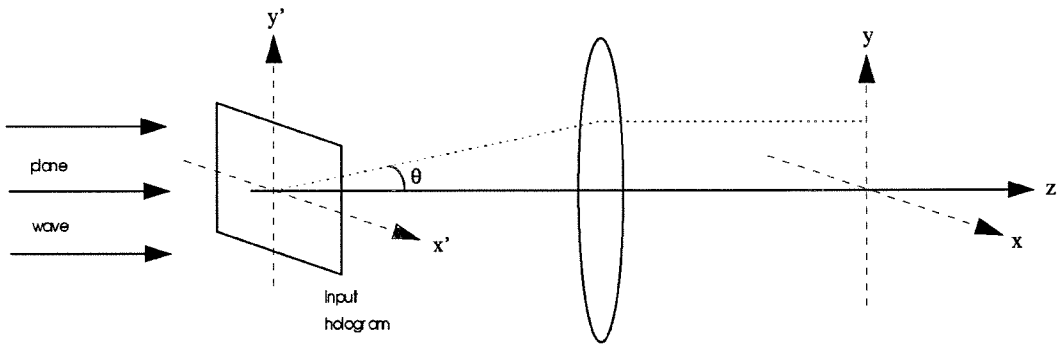
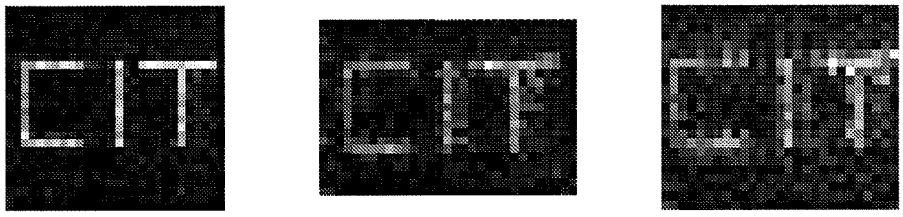
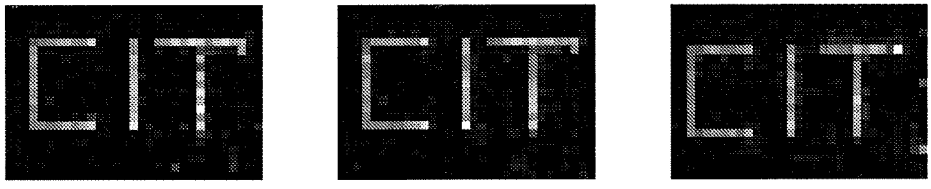



Figure 4.9. Experimental configuration for realizing the mask function $g(x,y)$ using computer-generated holograms.

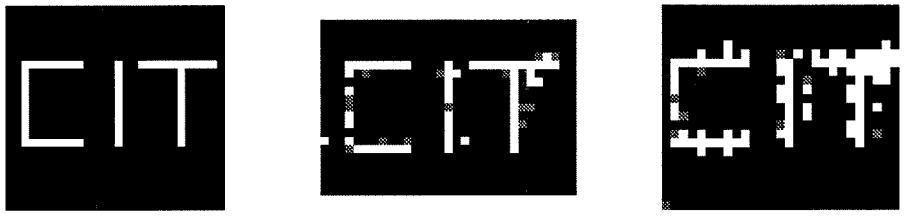


(i) (ii) (iii)



(iv) (v) (vi)

Figure 4.10a. Actual intensity patterns at three critical planes (i)-(iii) of binary-amplitude (transmission) mask and (iv)-(vi) of optimized amplitude-phase mask realized by Fourier holograms. Notice the amplitude-phase mask remains well-focused on all critical planes, while the transmission mask defocuses substantially.



(i) (ii) (iii)

(iv) (v) (vi)

Figure 4.10b. Actual intensity patterns after thresholding at three critical planes (i)-(iii) of binary-amplitude (transmission) mask and (iv)-(vi) of optimized amplitude-phase mask realized by Fourier holograms. The phase mask retains a reliable post-thresholding image throughout the imaging volume, while the transmission mask image is unreliable at two critical planes.

### 4.4. Nondiffracting images

To demonstrate the generality of this technique, we chose the aperiodic, multicornered image shown in Figure 4.8a. The simulated intensity patterns after diffraction for a binary transmission mask and an optimized amplitude-phase mask are shown in Figure 4.8b, and the postthresholding detector response is shown in Figure 4.8c. Although the binary-mask image becomes unacceptably distorted at the final critical plane, the optimized image remains completely undistorted. Next we implemented the amplitude-phase masks as Fourier-transform holograms, as shown in Figure 4.9. Given our desired mask distribution $g(x',y')$, we Fourier transform to $G(x'',y'')$, then take

$$G'(x'',y'') = \min\left(\mathrm{Re}\left\{G(x'',y'')e^{i2\pi u(x''+y'')}\right\}\right) + \mathrm{Re}\left\{G(x'',y'')e^{i2\pi u(x''+y'')}\right\}, \tag{4.8}$$

yielding a real-positive mask. Multiplying $G$ by $e^{i2\pi u(x''+y'')}$ shifts our pattern to a propagation angle $\theta = \lambda u$ in both the $xz$ and $yz$ planes; taking the real part (or, equivalently, adding the complex conjugate) causes the inverted pattern to appear at an angle $-\theta$. Finally, adding the constant simply introduces a focused spot that does not interfere with our pattern if $\theta$ is large enough.

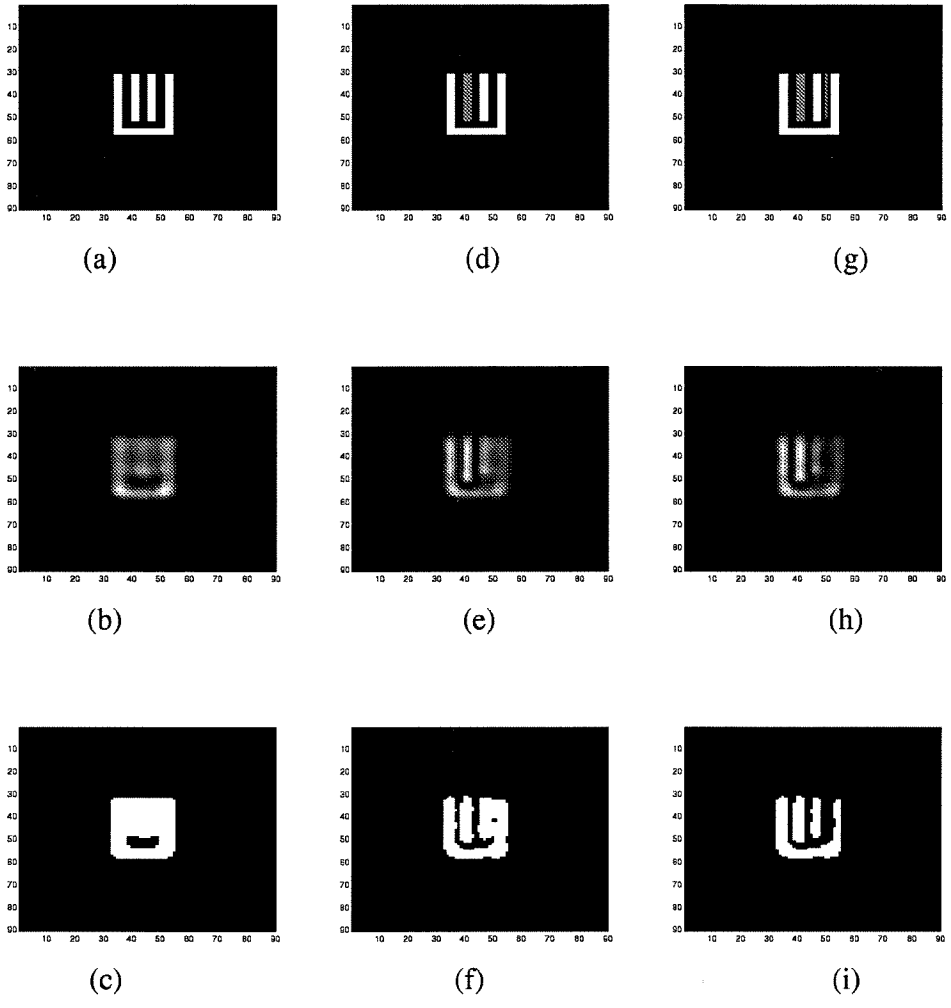



Figure 4.11. Resolution enhancement of optimized phase mask (simulation). Transmission mask in (a) forms pattern in (b) after imaging by a finite aperture and (c) after photoresist thresholding. Levenson-type phase mask in (d) forms image (b) after imaging and (c) after thresholding; notice phase conflict on right side of image. Optimized phase mask in (g) forms image (h) after imaging and (i) after thresholding, solving the phase conflict problem (d, g: gray indicates $\pi$-phase shift areas).

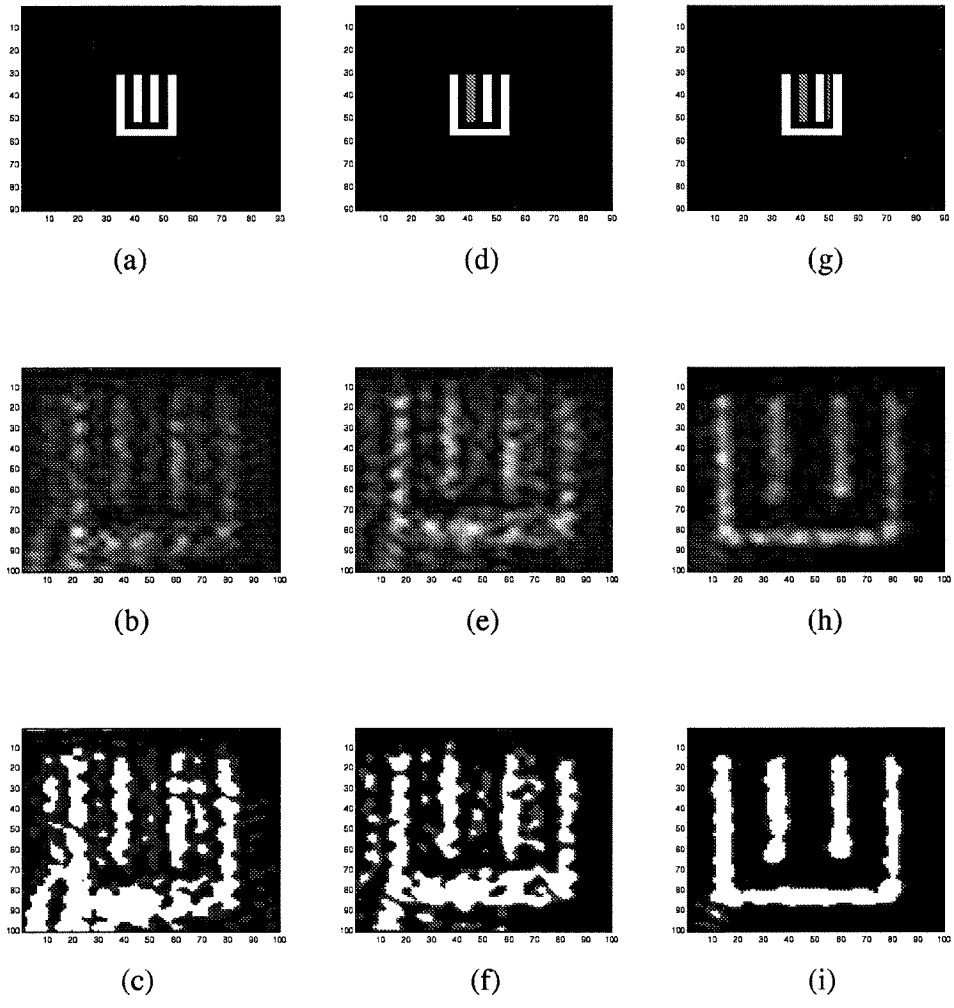



Figure 4.12. Resolution enhancement of optimized phase mask (experiment). Transmission mask in (a) forms pattern in (b) after imaging by a finite aperture and (c) after photoresist thresholding. Levenson-type phase mask in (d) forms image (b) after imaging and (c) after thresholding; notice phase conflict on right side of image. Optimized phase mask in (g) forms image (h) after imaging and (i) after thresholding, solving the phase conflict problem. These results confirm the simulation results in Figure 4.11 (d, g: gray indicates $\pi$-phase shift areas).

Using a CCD to record the transverse images and postprocessing to simulate photoresist thresholding, we obtained the results in Figures 4.10a and 4.10b, respectively. As predicted by our simulations, the optimized mask images remain focused significantly farther than the binary images. In this example, we improved focal depth approximately twofold by using optimized amplitude-phase masks; in practice, the improvement in focal depth depends on the sharpness of the photoresist threshold, attainable mask resolution, and pattern irregularity. Of these, photoresist thresholding is by far the most influential parameter, since a sharper threshold means larger tolerable amplitude fluctuations and more freedom in optimizing the mask.

Our next example is of phase conflict elimination and resolution enhancement. We begin with the simple phase conflicting pattern in Figure 4.5a. In this case, a comparison with Levenson-type phase masks is in order, since they are the standard binary-phase, binary-amplitude tool for resolution enhancement. This comparison will emphasize the need for phase conflict elimination in Levenson-type masks. Figures 4.11a-c and d-f show the mask, image after finite-aperture imaging, and post-thresholding photoresist pattern for transmission and Levenson-type phase masks, respectively. After optimization, the mask-plane field distribution is shown in Figure 4.11g, and the simulated wafer-plane intensity distribution before and after thresholding is shown in Figures 4.11b,c. Clearly

optimization has solved the phase conflict problem and made the features resolvable.

This is confirmed experimentally using the same computer-generated-holographic technique as above to realize the complex mask field. Figures 4.12b-c show the measured image plane intensity and post-thresholding image pattern, respectively, for the transmission mask in Figure 4.12a. The corresponding images are given in Figures 4.12e-f for the Levenson mask in Figure 4.12d and Figures 4.12h-i for the Levenson mask in Figure 4.12g.

Although the optimization method used here compares favorably with other proposed techniques [5-10] in its run times, scaling with the number of pixels square, they are still prohibitively long for industrial-scale masks. Therefore, we implemented a heuristic algorithm which inserts auxiliary phase regions as the one in Figure 4.11a in regions of phase conflict. Due to its rule-based approach, this algorithm's run times scale with the number of mask pixels, which make it viable even for VLSI masks. We tested it on the typical mask layout shown in Figure 4.13a, whose pre- and post-thresholding images are given in Figures 4.13b,c, respectively. Its corresponding Levenson mask (Figure 4.13d) leaves several problem spots before and after thresholding (Figures 4.13e,f, respectively); while the fast-optimized binary phase mask (Figure 4.13g) is completely resolved before (Figure 4.13h) and after (Figure 4.13i) thresholding. Despite its apparent success,

this heuristic approach leads in general to poorer results than the full-mask optimization, and we therefore expect the latter to enable smaller features or larger focal depths. This performance difference can be tempered, without significant run-time increase, by performing a full-mask optimization after the heuristic algorithm has produced the first-order solution. Since the initial layout is now much closer to an error minimum than was the original binary transmission mask, the full optimization converges far more quickly. A detailed exploration of the ideal algorithm combination is beyond the scope of this work.

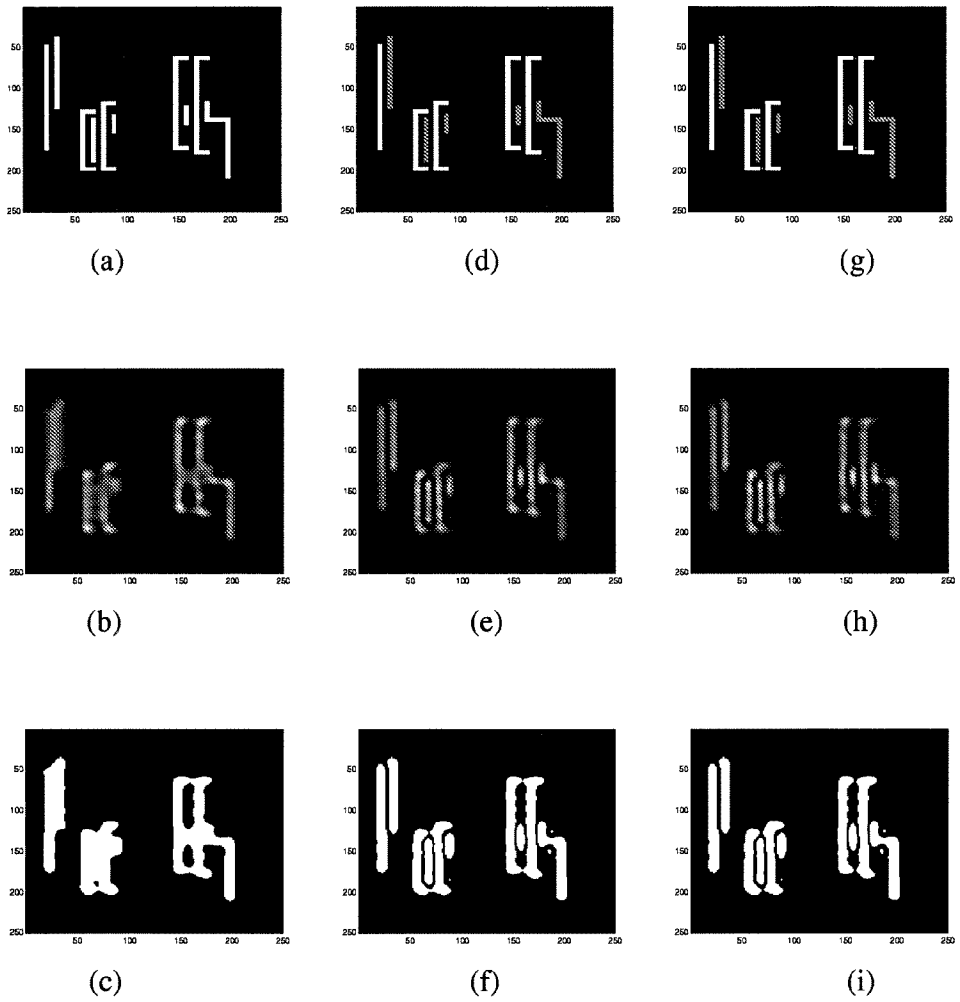



Figure 4.13. Large-scale phase mask optimization. Transmission mask (a) becomes image (b) after imaging by a finite aperture, forming photoresist pattern (c) after imaging. Levenson-type phase mask (d) forms image (e) after imaging and (f) after thresholding; notice phase conflicts at vertical coordinates 70 and 110. Optimized phase mask (g) forms image (h) after imaging and (i) in photoresist, where the phase conflict is eliminated (d, g: gray indicates $\pi$-phase shift areas).

**References**

1. M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, IEEE Trans. Electron. Devices **29**, 1828 (1982).

2. J. W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill (1988), p. 88.

3. J. W. Goodman, *Statistical Optics*. (Wiley & Sons, New York, 1985), p. 308.

4. M. D. Levenson, JJAP 33, 6765 (1994).

5. Y. Liu, A. K. Pfau, and A. Zakhor, SPIE **1674**, 14 (1992).

6. Y. Liu and A. Zakhor, IEEE Trans. Semiconduct. Manuf. **5**, 138 (1992).

7. Y. Liu, A. Zakhor, and M. A. Zuniga, IEEE Trans. Semiconduct. Manuf. **9**, 170 (1996).

8. H. Watanabe, Y. Pati, and R. F. Pease, JJAP I/12B **33**, 6790 (1994).

9. Y. C. Pati and T. Kailath, JOSA A **11**, 2438 (1994).

10. B. Salik, J. Rosen, and A. Yariv, JOSA A **12**, 1702 (1995).

11. J. W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill (1968), p. 45.

12. B. Salik, J. Rosen, and A. Yariv, Opt. Lett. **20**, 1743 (1995).

13. L. R. Foulds, *Optimization Techniques*, New York: Springer-Verlag (1981), p. 329-335.

14. R. A. Ferguson, Proc. Soc. Photo-Opt. Instrum. Eng. 2197, 130 (1994).

# Chapter Five

# Average coherence approximation

## 5.1. Introduction to partially coherent illumination

Monochromatic illumination systems always possess complete spatial coherence [1]. Since all physical illumination sources have a nonzero linewidth, their radiation is more generally described as partially coherent. Thus partially coherent imaging is important in a variety of fields, including astronomy, photolithography, and medicine. Given a coherent impulse response $K$, propagation through a system (Figure 5.1) is described by the Hopkins equation [2]:

$$\left|E_o(x',y')\right|^2 = \iint\iint E(x,y)E^*(\tilde{x},\tilde{y})J(x,y;\tilde{x},\tilde{y})K(x,y;x',y')K^*(\tilde{x},\tilde{y};x',y')dxdyd\tilde{x}d\tilde{y} \tag{5.1}$$

where $E$ is the input field, $J$ is the illumination's mutual intensity function, $(x,y)$ are the system's input coordinates, and $(x',y')$ are the system's output coordinates (see Ref. 10 for an elegant derivation of this equation).

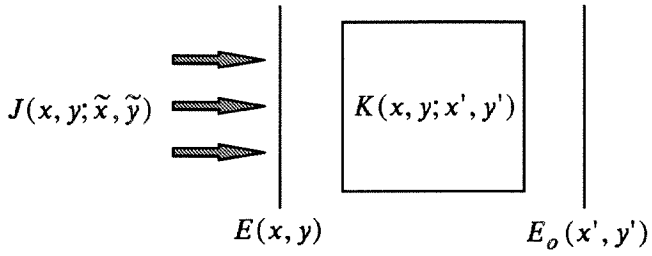



Figure 5.1. Schematic diagram of general optical system described by the Hopkins equation.

Though general, the Hopkins equation is tedious to compute, both analytically and numerically; still, it is used almost universally in analysis and simulation of partially coherent systems [3-5]. There have been several attempts to simplify this expression and derive alternative formulas which are less computationally demanding [6-9], but they have all addressed special cases of the impulse response or mutual intensity functions. In this chapter I introduce an approximation to the Hopkins equation which is valid for a wide range of coherence functions and impulse responses, and considerably reduces the computation time required for determining the output intensity. This approximation is particularly useful when, given $J$ and $K$, the output intensity is desired for varying input fields, e.g., in

iterative routines which optimize the input field to obtain a desired output intensity distribution [3].

### 5.2. One-dimensional approximation

For simplicity, we begin with one transverse dimension, then generalize to two-dimensional problems. The one-dimensional Hopkins integral is, analogously to (5.1),

$$\left|E_o(x')\right|^2 = \iint E(x)E^*(\tilde{x})J(x,\tilde{x})K(x,x')K^*(\tilde{x},x')dxd\tilde{x} \tag{5.2}$$

Our strategy is to decompose the contribution to $E_o(x')$ by each input point $E(x)$ into a coherent component and an incoherent component, then sum the contributions of all points in the $x$ plane to the field intensity at $x'$. The coherent radiation component interferes coherently with the coherent radiation from all other points $x$, while the incoherent component interferes incoherently with all other radiation. Thus we wish to reduce the two-dimensional integral (5.2) to a sum of two one-dimensional integrals, one addressing the coherent radiation component and the other, the incoherent component. While (5.2) treats the interference between every two points at the input plane separately, we shall try to lump the field correlation between a given input plane point $x$ and all other input plane points into an average coherence at $x$. To determine the coherent and

incoherent components of radiation from each point, we must somehow average its coherence $J(x, \hat{x})$ with all other points $\hat{x}$ (Figure 5.2). This average should be weighted by the contribution of the points $\hat{x}$ to the intensity at $x'$, and this contribution is given by $|K(x,x')|^2$. Thus we can define a function $f(x,x')$ which gives the fraction of the intensity at $x$ which interferes coherently with all other coherent field contributions.

$$f(x,x') = \frac{\int |K(x',\hat{x})|^2 \mu(x,\hat{x}) d\hat{x}}{\int |K(x',\tilde{x})|^2 d\tilde{x}}, \qquad (5.3)$$

$$\mu(x,\hat{x}) = \frac{J(x,\hat{x})}{\sqrt{J(x,x)J(\hat{x},\hat{x})}}$$

where $\hat{x}$ and $\tilde{x}$ are dummy variables at the input plane and the functions $K$ and $J$ are normalized to insure $0 \leq f \leq 1$ (this condition is imposed since $f$ is by definition the fraction of incident power which is spatially coherent).

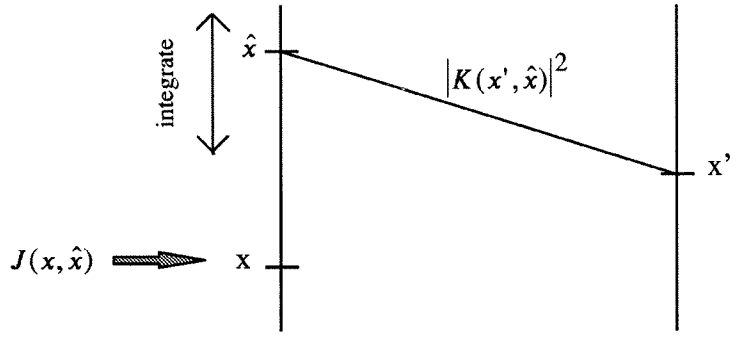



Figure 5.2. Schematic representation of system configuration in the average coherence approximation (x can be an n-dimensional vector, in general).

We are now able to consider the coherent and incoherent contributions to the field at $x'$ independently, using the following relations:

$$|E_o(x')|^2 = |E(x) * K(x,x')|^2 \qquad E(x) \text{ coherent}$$

$$|E_o(x')|^2 = |E(x)|^2 * |K(x,x')|^2 \qquad E(x) \text{ incoherent} \tag{5.4}$$

where $f(x) * g(x,x') \equiv \int f(x)g(x,x')dx$ is a general superposition integral.

In our case, $K(x)$ is assumed to have two components, one coherent and one incoherent, i.e.,

$$\left|K(x,x')\right|^2 = \left|K_c(x,x')\right|^2 + \left|K_i(x,x')\right|^2 = \left|\sqrt{f(x,x')}K(x,x')\right|^2 + \left|\sqrt{1-f(x,x')}K(x,x')\right|^2 \tag{5.5}$$

so that the output intensity is given by

$$\begin{aligned}\left|E_o(x')\right|^2 &= \left|E(x) * K_i(x,x')\right|^2 + \left|E(x)\right|^2 * \left|K_c(x,x')\right|^2 \\ &= \left|E(x) * \sqrt{f(x,x')}K(x,x')\right|^2 + \left|E(x)\right|^2 * \left|\sqrt{(1-f(x,x'))}K(x,x')\right|^2\end{aligned} \tag{5.6}$$

and we have effectively reduced the double integral of the one-dimensional Hopkins equation to a sum of two single integrals. Using the form of $f(x)$ defined in (5.3), (5.6) implies

$$\left|E_o(x')\right|^2 = \left|\int E(x)\sqrt{f(x',x)}K(x',x)dx\right|^2 + \int\left|E(x)\right|^2\left|(1-f(x',x))\right|\left|K(x',x)\right|^2 dx \tag{5.7}$$

### 5.3. Two-dimensional approximation

The two-dimensional generalization is straightforward; (5.3) becomes

$$f(x,y,x',y') = \frac{\iint \left|K(x',y',\hat{x},\hat{y})\right|^2 \mu(x,y,\hat{x},\hat{y}) d\hat{x}d\hat{y}}{\iint \left|K(x',y',\tilde{x},\tilde{y})\right|^2 d\tilde{x}d\tilde{y}} \tag{5.8}$$

while (5.7) is now

$$\left|E_o(x',y')\right|^2 = \left|\iint E(x,y)\sqrt{f(x,y,x',y')}K(x',y',x,y)dxdy\right|^2 + \iint \left|E(x,y)\right|^2 \left|(1-f(x,y,x',y'))\right| K(x',y',x,y)\Big|^2 dxdy. \tag{5.9}$$

Note that although an integration is required to compute $f$, this need only be done once given the system parameters $K$ and $J$, after which computing the output intensity requires only one integration dimension per transverse field dimension. Furthermore, if our system is space-invariant, i.e.,

$$K(x,y,x',y') = K(x-x',y-y') \text{ and } J(x,y,\hat{x},\hat{y}) = J(x-\hat{x},y-\hat{y}) \tag{5.10}$$

then the integrals of (5.3), (5.7), (5.8), and (5.9) become convolutions which are efficiently calculated using fast Fourier transforms (because of the space-invariance, (5.3) and (5.7) become one-dimensional convolutions, and (5.8) and (5.9) become two-dimensional convolutions). Further simplification is possible if we assume specific forms for $K$ and $J$. For example, if we assume one-dimensional imaging by a rectangular aperture of size $D$ then

$$K(x,x') = Sinc(\frac{x-x'}{d}) = \frac{d}{\pi(x-x')}\sin(\frac{\pi(x-x')}{d}) \tag{5.11}$$

where $d = \frac{\lambda f}{D}$ is the system's resolution radius, so that

$$\left|K(x,x')\right|^2 = Sinc^2(\frac{x-x'}{d}) \tag{5.12}$$

and if our illumination source is rectangular [8],

$$J(x,\tilde{x}) = Sinc(\frac{x-\tilde{x}}{a}) = \frac{a}{\pi(x-\tilde{x})}\sin(\frac{\pi(x-\tilde{x})}{a}) \tag{5.13}$$

where $a$ is the coherence radius. Thus we have

$$f(x,x') = \frac{1}{\int Sinc^2(\frac{\tilde{x}-x}{d})d\tilde{x}}\int Sinc^2(\frac{\tilde{x}-x'}{d})Sinc(\frac{\tilde{x}-x}{a})d\tilde{x}. \tag{5.14}$$

Using the substitutions $\hat{x} = \tilde{x} - x'$ and $\bar{x} = x - x'$, this simplifies to

$$f(x,x') = \frac{1}{\int Sinc^2(\frac{\hat{x}}{d})d\hat{x}}\int Sinc^2(\frac{\hat{x}}{d})Sinc(\frac{\hat{x}-\bar{x}}{a})d\hat{x} \tag{5.15}$$

which, by Fourier transforming and assuming $d > 2a$, becomes

$$f(x,x') \propto Sinc^2(\frac{x-x'}{d}). \tag{5.16}$$

This can now be used a priori in (5.7), eliminating the integration due to $f(x,x')$.

### 5.4. Error of the approximation

To evaluate the error of the average coherence approximation, we recall that $J$ is unity for coherent imaging systems and is a delta function for incoherent systems. Therefore (in one transverse dimension), Eq. (5.7) is identical to a Hopkins integral where $J(x, \hat{x})$ is replaced by

$$J'(x,\tilde{x},x') = \sqrt{f(x,x')\,f(\tilde{x},x')} + \sqrt{(1-f(x,x'))(1-f(\tilde{x},x'))}\,\delta(x-\tilde{x}). \tag{5.17}$$

Therefore, the error in field intensity as given by (5.7) is

$$\begin{aligned} e(x') &= \int E(x)E^*(\tilde{x})K(x',x)K^*(x',\tilde{x})J(x,\tilde{x})dxd\tilde{x} - \\ &\int E(x)E^*(\tilde{x})K(x',x)K^*(x',\tilde{x})J'(x,\tilde{x},x')dxd\tilde{x} \\ &= \int E(x)E^*(\tilde{x})K(x',x)K^*(x',\tilde{x})(J(x,\tilde{x})-J'(x,\tilde{x},x'))dxd\tilde{x}. \end{aligned} \tag{5.18}$$

This expression is easier to compute than a direct subtraction of (5.7) from (5.2), and is used in the examples below.

The error will in general depend on the spatial coordinates, input field, system kernel $K$ and coherence function $J$; specifically, the error becomes smaller as $|K(x',x)|$ becomes sharper (its energy is concentrated near some $(x',x)$). If our system is an imaging system, for example, then as its numerical aperture becomes larger, $|K(x,x')|^2 \to \delta(x-x')$ and therefore (assuming $J$ is normalized)

$$f(x,x') = \int \delta(\hat{x}-x')J(\hat{x},x)d\hat{x} = J(x',x) \Rightarrow$$

$$e(x') = \int E(x)E^*(\tilde{x})K(x',x)K^*(x',\tilde{x})J(x,\tilde{x})dxd\tilde{x} -$$

$$\left(\left|\int E(x)\sqrt{J(x',x)}K(x',x)dx\right|^2 + \int |E(x)|^2 |1-J(x',x)||K(x',x)|^2 dx\right)$$

$$= |E(x')|^2 (1-J(x',x')-1+J(x',x'))$$

$$= 0. \tag{5.19}$$

Furthermore, our error goes to zero as we approach full coherence and full incoherence. In the fully coherent case, $J(x,x')=1$, thus $f(x,x')=1$. From Eq. (6) it is clear then that

$$|E_o(x')|^2 = |E(x')*K(x')|^2 \tag{5.20}$$

which is the exact expression for coherent propagation. For the incoherent case,

$$J(x,x') = \delta\,(x - x') \Rightarrow$$

$$f(x,x') = \frac{\left|K(x,x')\right|^2}{J(x,x)\int\left|K(x,x')\right|^2 dx} = 0 \tag{5.21}$$

and, again using Eq. (5.6), we arrive at

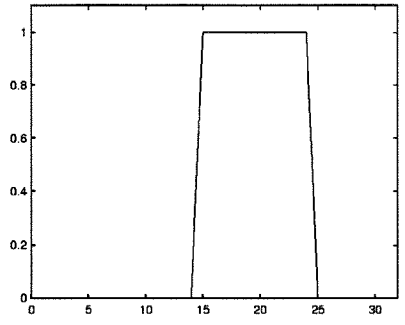
$$\left|E_o(x')\right|^2 = \left|E(x')\right|^2 * \left|K(x')\right|^2 \tag{5.22}$$

which is the exact expression for incoherent illumination. A notable case where the error does not approach zero is when the impulse response $K$ becomes very wide; then our averaging of the coherence actually loses information and the results in general may vary from the Hopkins calculation.

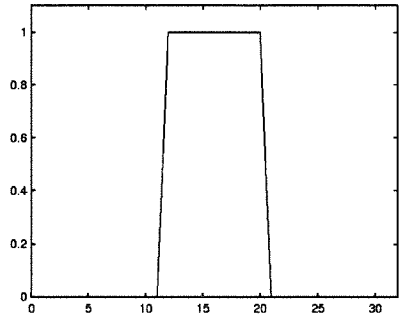
### 5.5. Examples

Empirically, the average coherence approximation (ACA) yields excellent agreement with the exact Hopkins integral for a wide range of input fields and coherence functions. Figure 5.3 compares the ACA and Hopkins integral results [12] for a finite, asymmetric one-dimensional aperture after imaging by a finite

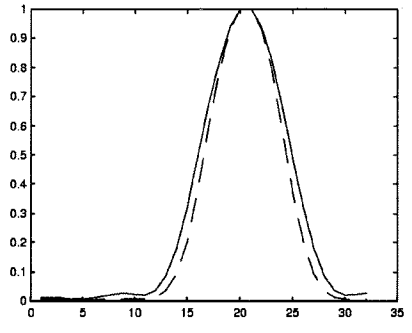
aperture ($J$ is a Sinc with coherence diameter half the image size). We can see that there is agreement to within 5% of maximum intensity throughout the output plane, and the average error/pixel is 0.15%. Figure 5.4 compares the output intensities for a two-dimensional input pattern imaged through a finite aperture. Finally, in Figure 5.5 we plot the ACA's average error per pixel (compared to the Hopkins integral) over the output plane for various widths of the coherence function $J$, assuming Rayleigh diffraction with z=16 pixels. As expected, the error minima lie at full coherence and full incoherence (they are nonzero due to quantization error).




(a)

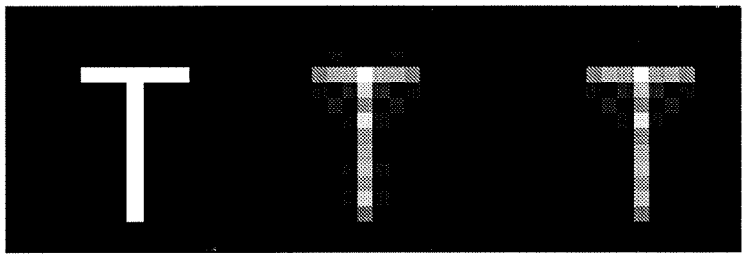



(b)




(c)

Figure 5.3. One-dimensional slit in (a), after imaging through finite aperture (b), yields patterns in (c) using Hopkins integral (solid) and ACA (dashed). Here J(x,x') is a Sinc half as wide as the image.



(a) (b) (c)

Figure 5.4. Pattern in (a) after imaging through a finite aperture becomes (b) under the average coherence approximation and (c) using the full Hopkins equation. Aperture size is half the image bandwidth, and coherence diameter is half the image size; average error/pixel=.95%=.0095.

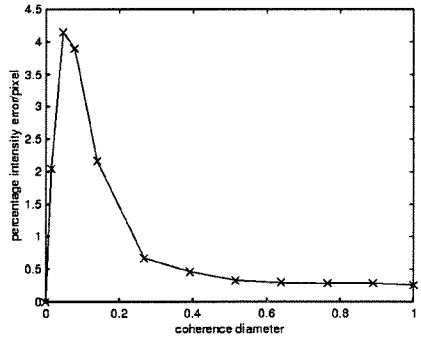



Figure 5.5. ACA error vs. normalized coherence diameter for 1-D finite aperture undergoing Fresnel diffraction.

## 5.6. Computation time

We can estimate the relative computation times of the Hopkins integral and the average coherence approximation by using order-of-magnitude considerations. We assume a grid of $n$ points (i.e., a $j$ x $k$ two-dimensional grid would have $n=jk$). The Hopkins integral, assuming a shift-invariant system, can be computed using a triple correlation for the inner integral, $\int E(\tilde{x})J(x,\tilde{x})K^*(\tilde{x},x')d\tilde{x}$, which requires $6n^2\log n + 2n^2 + 2n^2\log n$ complex operations assuming FFT's with $O(n\ logn)$ are used. The outer integral imposes $n$ iterations of the inner integral plus two complex multiplications, which implies a computational order of $n(8n^2\log n + 2n^2 + 2n^2) = 4n^3(2\log n + 1)$. The average coherence approximation requires two FFT's, a complex multiplication and an inverse FFT to compute $J(x,x')$, plus two integrals, each involving two FFT's, a complex multiplication and an inverse FFT. Thus the total computational order is $2(2n^2\log n) + n^2 + 2n^2\log n + 2(2(2n^2\log n) + n^2 + 2n^2\log n) = 3n^2(5\log n + 1)$.

We can see the ACA scales approximately with $n^2logn$, while the Hopkins integral scales as $n^3logn$, which for a two-dimensional image is a difference of 2 orders of magnitude. Empirically, the ACA has yielded results 10-50 times faster than the Hopkins integral, depending on the size of the input pattern.

## 5.7. Summary

In summary, we have introduced an approximation to the Hopkins integral for propagation of partially coherent fields, which is computationally simpler both analytically and numerically, and yields good agreement with the Hopkins integral for a wide range of input patterns and system transfer functions. We have presented several analytical conditions for the accuracy of this approximation, and numerically analyzed its performance for a range of system parameters. Finally, we have seen that the average coherence approximation can be computed significantly faster than the Hopkins integral for iterative, constant-system applications.

**References**

1. J. W. Goodman, *Introduction to Fourier Optics,* San Francisco: McGraw-Hill (1968), p. 106.

2. H. H. Hopkins, Proc. Roy. Soc. **A208**, 263-277 (1951).

3. Y. Liu, A. K. Pfau, and A. Zakhor, "Systematic design of phase-shifting masks with extended depth of focus and/or shifted focus plane," IEEE Trans. Sem. Man. **6**, 1-21 (1993).

4. D. C. Cole, E. Barouch, U. Hollerbach, and S. A. Orszag, "Derivation and simulation of higher numerical aperture scalar aerial images," JJAP **31**, 4110-4119 (1992).

5. Y. C. Pati and T. Kailath, "Phase-shifting masks for microlithography: automated design and mask requirements," JOSA A **11**, 2438-2452 (1994).

6. J. Perina, *Coherence of Light,* 2nd ed. Dordrecht: Reidel (1985), ch. 4.

7. L. Mandel, JOSA **51**, 1342 (1961).

8. M. Born and E. Wolf, *Principles of Optics*. Oxford: Pergamon (1965), p. 526.

9. W. H. Carter and E. Wolf, JOSA **67**, 785 (1977).

10. J. W. Goodman, *Statistical Optics,* New York: Wiley (1985), p. 286-302.

11. C. R. Wylie, Jr., *Advanced Engineering Mathematics,* 2ed. New York: McGraw-Hill (1966), p. 663-670.

12. B. Salik, J. Rosen, and A. Yariv, JOSA A **13**, 2086 (1996).

# Chapter Six

# Spatial coherence and phase masks

## 6.1. Introduction

Since Marc Levenson's introduction of phase masks to photolithography [1], there have been numerous improvements in their design and fabrication [2-6]. One issue not yet considered is the effect of illumination source spatial coherence on the effectiveness of phase masks in improving resolution and focal depth. As the illumination wavelength has grown shorter (to accommodate the classical resolution limit $\Delta x \propto \frac{\lambda}{NA} = \frac{\text{wavelength}}{\text{numerical aperture}}$), sources have become more spatially coherent [7], and with the use of excimer laser illumination we approach complete spatial coherence. It has generally been assumed that increased spatial coherence improves the performance of phase masks, since it enhances the destructive interference between adjacent (opposite phase) features. Although this is true for periodic patterns, I will show in what follows that more complex images involve a trade-off between the contrast enhancement of coherent alternating

features and the larger effective aperture of incoherent imaging, independently of speckle phenomena. This effect is particularly pronounced in complex 2-D images where phase conflict [7] is a problem.

### 6.2. Minimizing imaging error

Let us begin by defining a criterion for imaging system performance. To quantify image quality, we introduce an error measure [4] that compares the imaged intensity distribution $I(x,y,z)$ to the desired 3-D intensity $I_0(x,y,z)$.

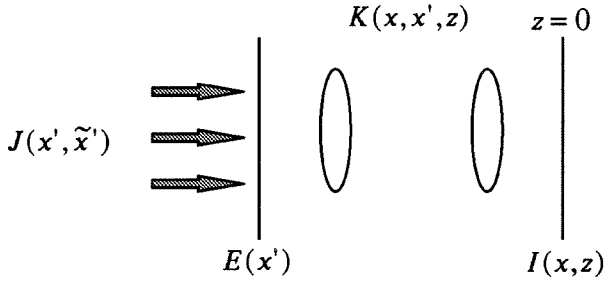



Figure 6.1. Schematic diagram of partially coherent imaging system described by the Hopkins integral.

Shifting to one transverse dimension for notational simplicity, the output intensity due to an input field $E(x')$ (see Figure 6.1) is given by the Hopkins integral

$$I(x,z) = \iint E(x')E^*(\tilde{x}')J(x',\tilde{x}')K(x',x,z)K^*(\tilde{x}',x,z)dx'd\tilde{x}' \tag{6.1}$$

where $J(x',\tilde{x}')$ is the input plane mutual intensity function and $K(x,x')$ is the coherent system impulse response [8]. A Euclidean error over one plane will be, for example,

$$e^2(z) = \int \left| I(x,z) - I_0(x,z) \right|^2 dx \tag{6.2}$$

where $I_0(x,z)$ is the two-dimensional desired intensity distribution, and over a volume it is the sum of single-plane errors

$$e^2 = \int e^2(z)dz = \iint \left| I(x,z) - I_0(x,z) \right|^2 dxdz . \tag{6.3}$$

We can include photoresist thresholding in this error measure by replacing $I(x,z)$ with $T[I(x,z)]$, the photoresist exposure response (Figure 6.2).

If our illumination source is spatially incoherent and uniform along an aperture $d$ then in the far field [9],

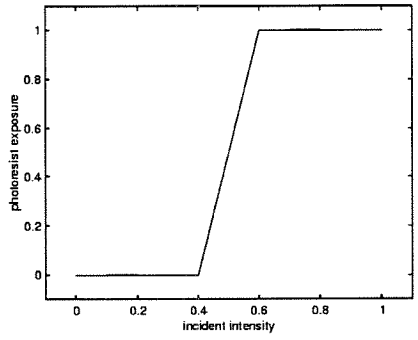



Figure 6.2. Typical photoresist exposure vs. incident intensity (given exposure time). The resist acts as a thresholding agent.

$$J(x',\tilde{x}') = Sinc(\frac{x-\tilde{x}'}{s}) = \frac{s}{\pi(x-\tilde{x}')}\sin\pi(\frac{x-\tilde{x}'}{s}) \tag{6.4}$$

where $s = \lambda L/d$ is the coherence diameter, i.e., the distance within which the fields at two points correlate significantly over time (here $\lambda$ is the wavelength and $L$ the distance from the aperture). We may substitute this into Eq. (6.1), which in turn can be replaced in Eq. (6.3) to yield an explicit expression for $e$ in terms of s:

$$e^2 = \iint dxdz\left|\iint E(x')E^*(\tilde{x}')Sinc(\frac{x'-\tilde{x}'}{s})K(x',x,z)K^*(\tilde{x}',x,z)dx'd\tilde{x}' - I_0(x,z)\right|^2 \tag{6.5}$$

which we differentiate with respect to s and equate to zero to find the values of $s$ that maximize and minimize $e$ (since $e \geq 0$, minimizing $e$ implies minimizing $e^2$). We thus find the following sufficient (but not necessary) alternative conditions for the minimization of $e^2$:

$$\iint E(x')E^*(\tilde{x}')Sinc(\frac{x'-\tilde{x}'}{s})K(x',x,z)K^*(\tilde{x}',x,z)dx'd\tilde{x}' = I_0(x,z), \tag{6.6}$$

and (nonequivalently)

$$0 = \frac{d}{ds}\iint E(x')E^*(\tilde{x}')Sinc(\frac{x'-\tilde{x}'}{s})K(x',x,z)K^*(\tilde{x}',x,z)dx'd\tilde{x}'$$

$$= \iint \frac{E(x')E^*(\tilde{x}')}{\pi(x'-\tilde{x}')}K(x',x,z)K^*(\tilde{x}',x,z)[\sin\pi(\frac{x'-\tilde{x}'}{s}) - \frac{\pi(x'-\tilde{x}')}{s}\cos\pi(\frac{x'-\tilde{x}'}{s})]dx'd\tilde{x}' \tag{6.7}$$

which is satisfied by

$$\tan\pi(\frac{x'-\tilde{x}'}{s}) = \frac{\pi(x'-\tilde{x}')}{s} \Leftarrow s \to \infty. \tag{6.8}$$

Eq. (6.6) simply states that when our output intensity distribution equals the desired intensity $I_0$, the error $e=0$, which must be a minimum since $e \geq 0$. Eq. (6.8) guarantees that a local extremum for the output intensity error exists at the fully coherent limit; however, the nature of this extremum (maximum or minimum)

depends on the input field $E$ and impulse response $K$, since all higher derivatives of the error function also vanish at the coherent limit. Thus, the optimal value for $s$ lies in general somewhere between full coherence and full incoherence.

### 6.3. Experimental verification of coherence-dependent imaging error

Our experiments and simulations [11] suggest that for large images, the most influential factor in determining the optimal coherence $s$ is the smallest resolvable mask feature $\Delta x' = m\Delta x = m\lambda f / D = m\lambda / NA$, where $m$ is the imaging demagnification and $NA$ is the numerical aperture. Initially, we used the experimental setup in Figure 6.3 to test the coherence-dependent resolvability of two narrow slits as the aperture (determined by the width of the Fourier-plane pupil $P$) is varied, i.e., $D_{eff} = \lambda f / P$.

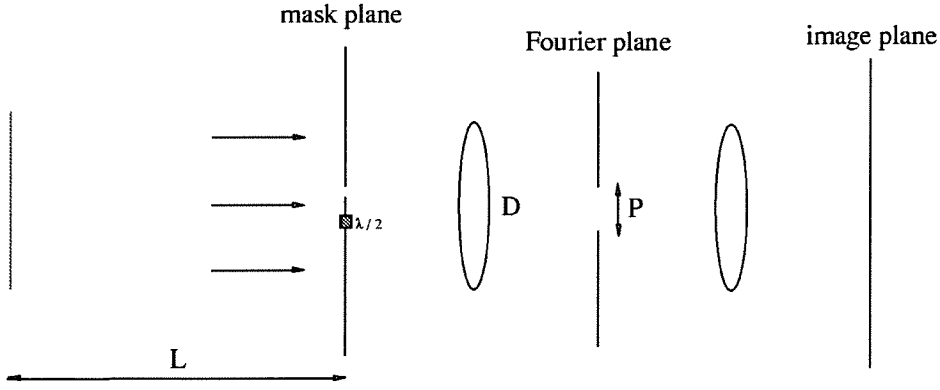



Figure 3. Experimental setup for measuring two-point imaging error as function of illumination spatial coherence.

The results, shown in Figure 6.4, indicate that in general the optimal $s$ ($s_{opt}$) lies between the coherent and incoherent limits, in agreement with the theoretical conclusions drawn from (5). We use here the standard definition of full coherence as a coherence width larger than the mask size (normally tens of millimeters) and full incoherence as a coherence width of zero. Although Figure 6.4 does not extend to the fully coherent case, the normalized error remains saturated at about 4.2 beyond $2s_{opt}$.

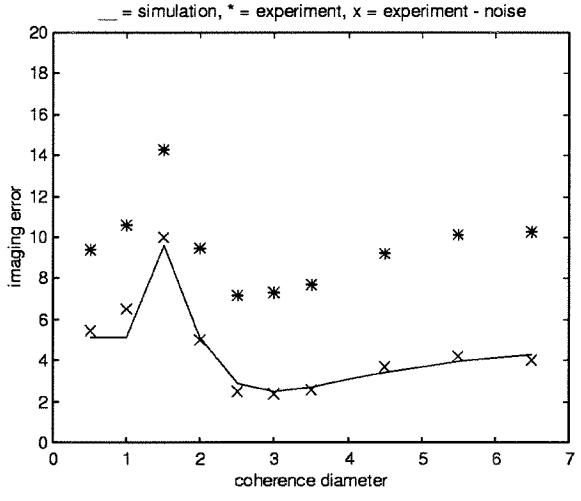



Figure 4. Results of experiment shown in Figure 3 and numerical solutions of (9) (wavelength = 633 nm; system resolution $\Delta x$= 6.3 μm). Speckle and DC noise were measured without image mask and subtracted from *-series to yield x-series. Coherence diameter is in units of $\Delta x$, and error is normalized according to (9). Total error saturates at about 4.2 as we approach full coherence.

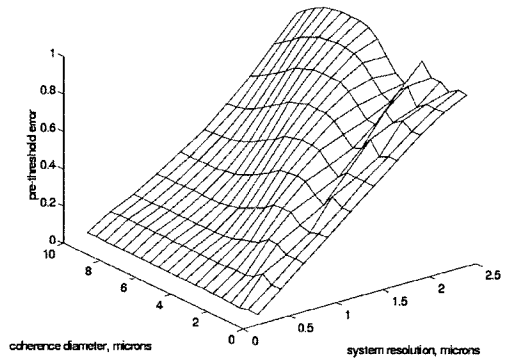
We find that, specifically, $s_{opt}$ varies linearly with system resolution, i.e., inversely with numerical aperture, in both experiments and simulations. This linear variation is shown explicitly in Figure 6.5a, where the two-slit imaging error is calculated as a function of coherence and resolution, and in Figure 6.6a, which shows the same error measure calculated for 20 randomly generated images and averaged. To remove the dependence of error on incident flux, we used the normalized error measure:

$$e^2(z) = \frac{\int \left| I(x,z) - I_0(x,z) \right|^2 dx}{\int \left| I_0(x,z) \right|^2 dx} \tag{6.9}$$

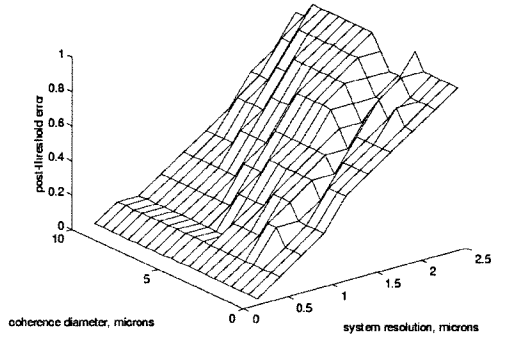
Since the normalization factor is constant in $s$, it does not alter any of the theoretical conclusions drawn in Eqs. (6.4-6.8).

### 6.4. Optimal spatial coherence

In all cases presented here, the optimal coherence diameter is closer to incoherence than to full coherence (assuming a mask size of at least a few millimeters), and increases linearly with system resolution (inversely with numerical aperture). The same variation occurs in a quantized form when we introduce photoresist thresholding in Figures 6.5b and 6.6b.




(a)




(b)

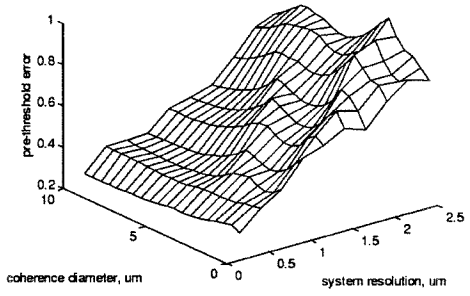
Figure 5. Imaging error plotted vs. illumination coherence and system resolution for two-point imaging: (a) aerial image error, (b) photoresist exposure error. Notice the minimum error varies linearly with the minimum feature size (inversely with the numerical aperture). Simulation wavelength = 248 nm.

Empirically, we find in general that

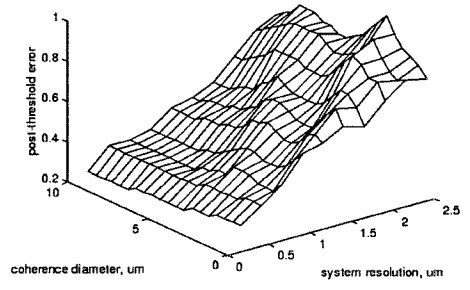
$$s_{opt} \cong 3m\lambda f / D = 3m\lambda / NA \tag{6.10}$$

independently of the mask size (slit separation). Although this is an empirical formula, it is intuitively appealing because a three-resolution-element coherence diameter just allows a feature to destructively interfere with its two (opposite-phase) nearest neighbors while maintaining the larger effective aperture size of incoherent illumination. Thus a smaller coherence diameter would compromise the destructive interference between adjacent pixels, since their fields would add incoherently, while a larger coherence diameter would reduce the width of the transfer function without significantly enhancing inter-pixel interference.

The reciprocal dependence on numerical aperture implies that smaller numerical apertures favor more coherent illumination, which is also intuitive since the relative advantage of phase masks is greatest when the *NA* is small. For typical photolithography masks, $m\lambda f / D \cong 1.5\mu m$ at the mask plane, meaning the optimal coherence diameter for enhancing resolution and focal depth is around 4.5 μm, much closer to the incoherent limit than to coherent illumination. Again, this effect is independent of speckle, which further degrades coherently illuminated images.




(a)




(b)

Figure 6. Imaging error plotted vs. illumination coherence and system resolution, averaged for 20 randomly generated images: (a) aerial image error, (b) photoresist exposure error. Wavelength = 248 nm.

## 6.5. Conclusion

From these results we can draw immediate conclusions for projection photolithography. Most important is that, independently of speckle, coherent illumination is rarely optimal even when using phase masks. The optimal coherence diameter depends explicitly on the imaged pattern, but in general varies inversely with the system's numerical aperture, or linearly with resolution. Although the computation time for numerically solving Eq. (6.5) becomes prohibitive for large images, approximate solution techniques [10] lessen the burden. Finally, since phase masks tend to improve focal depth as well as resolution, we may expect $s_{opt}$ to also increase with focal depth. This is evident from the classical expression $\Delta z \propto \Delta x^2 / \lambda \propto \lambda / NA^2$, where $\Delta z$ is the depth of focus, implying

$$s_{opt} \propto m\sqrt{\lambda \Delta z}. \tag{6.11}$$

Given illumination coherence $s$, this places an upper limit on the allowable focal depth before imaging degradation occurs.

**References**

1. M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, IEEE Trans. Electron. Devices **29**, 1828 (1982).

2. Y. Liu, A. K. Pfau, and A. Zakhor, SPIE **1674**, 14 (1992).

3. Y. C. Pati and T. Kailath, JOSA A **11**, 2438 (1994).

4. B. Salik, J. Rosen, and A. Yariv, Opt. Lett. **20**, 1743 (1995).

5. K. D. Lucas, A. J. Strojwas, and K. K. Low, SPIE **2197**, 489 (1994).

6. B. J. Lin, Circuits & Devices, **March 1993**, 28.

7. M. D. Levenson, Solid State Technology, **February 1995**, 57.

8. H. H. Hopkins, Proc. Roy. Soc. **A208**, 263-277 (1951).

9. M. Born and E. Wolf, *Principles of Optics* (Pergamon, Oxford, 1965), p. 526.

10. B. Salik, J. Rosen, and A. Yariv, "Average coherence approximation for partially coherent optical systems," JOSA A **13**, (1996).

11. B. Salik and A. Yariv, "Effect of spatial coherence on photolithographic phase mask performance," submitted to JJAP (12/96).

# Chapter Seven

# Ballistic imaging through self-interference

In this chapter I will describe a novel method of imaging through random media. This work is related to the preceding material in its utilization of coherence effects to enhance imaging; still, there are several issues unique to this problem, which I will now briefly motivate.

## 7.1. Introduction to imaging through scattering media

In many imaging applications, a scattering medium lies between the object to be imaged and the imaging system (Figure 7.1). This is true for applications such as biological tissue analysis, materials characterization, and remote imaging (in fact it is true, but probably not limiting, in every practical application, including photography and, due to gravitational distortion, even space-based astronomy). As the illumination from the object to be imaged propagates through a random medium, there is in general a fraction thereof which is unscattered, called the

ballistic signal component. This fraction depends on the size, scattering, and absorption of the intervening medium. A prevailing strategy in imaging through random media is separating the unscattered (ballistic) signal from scattered illumination. This has been achieved with a variety of techniques, including time-resolved detection [1], spatial coherence holography [2], and temporal coherence holography [3]. All these techniques require access to the illumination source, through either control of its spatio-temporal behavior or acquisition of a reference pulse. In many applications, e.g., astronomy, space-based imaging, and remote targeting, the illumination source is inaccessible and these methods are therefore inapplicable. In such cases, which often involve turbulent media (e.g., the atmosphere), existing imaging methods such as shift-and-add or speckle interferometry [4] must operate with short exposure times and invoke strong assumptions about the object and propagation medium, e.g., isoplanacity [5]. In this chapter I shall introduce a novel technique for imaging through random and turbulent media which utilizes ballistic photon detection and does not require source manipulation of any kind.

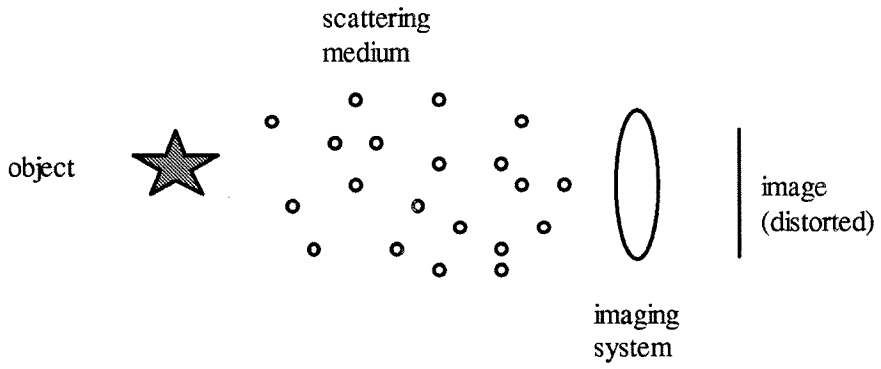



Figure 7.1. Imaging through random media: a scatterer (in general distributed) lays between the object and imaging system.

## 7.2. Theoretical preliminaries

Let us begin by reviewing the theoretical basis for ballistic imaging. We begin with the inhomogeneous vector three-dimensional wave equation

$$\nabla^2\mathbf{E} + k_0^2\varepsilon\mathbf{E} = -\nabla(\mathbf{E}\cdot\nabla(\ln\varepsilon)) - \frac{2ik_0}{c}\frac{\partial(\varepsilon\mathbf{E})}{\partial t} + \frac{1}{c^2}\frac{\partial^2(\varepsilon\mathbf{E})}{\partial t^2} \tag{7.1}$$

which assumes monochromatic illumination of frequency $\omega = ck_o$ and zero space charge. We now allow the permittivity $\varepsilon$ to vary around some average $<\varepsilon>$:

$$\varepsilon(\mathbf{r},t) = <\varepsilon> (1+\bar{\varepsilon}(\mathbf{r},t)),\ \ k^2 = <\varepsilon> k_0^2 \tag{7.2}$$

which expands (7.1) to

$$\nabla^2\mathbf{E} + k^2\mathbf{E} = -k^2\bar{\varepsilon}\mathbf{E} - \nabla(\mathbf{E}\cdot\nabla\bar{\varepsilon}) - \frac{2ik}{c}\frac{\partial}{\partial t}((1+\bar{\varepsilon})\mathbf{E}) + \frac{<\varepsilon>}{c^2}\frac{\partial^2}{\partial t^2}((1+\bar{\varepsilon})\mathbf{E}) \tag{7.3}$$

To simplify this unwieldy expression, we make some assumptions about the fluctuations in $\varepsilon$. First, we characterize these fluctuations by a characteristic length $l$, characteristic velocity $v$, and characteristic time $\tau = l/v$. Then, if the fluctuation amplitude $\bar{\varepsilon} >> v/c$, we can neglect the last two terms of (7.3) and arrive at the vector stochastic wave equation:

$$\nabla^2\mathbf{E} + k^2\mathbf{E} = -k^2\bar{\varepsilon}\mathbf{E} - \nabla(\mathbf{E}\cdot\nabla\bar{\varepsilon}) \tag{7.4}$$

The rightmost term represents "depolarization," or a coupling between the polarization directions, and gives the equation its vectorial nature (thus, if we explicitly write the equation for each field component, we end up with three coupled equations). If this term is small compared to the first term on the right, the equation decouples into three independent equations, one for each polarization direction, which in rectangular coordinates all assume the form

$$\nabla^2 E_i + k^2(1+\bar{\varepsilon})E_i = 0. \tag{7.5}$$

Next we attempt to use (7.5) to find a diffraction integral for scalar propagation through a random medium, in analogy to free-space scalar diffraction (see chapter 1). We assume, without loss of generality, a Green's function of the form

$$G(\mathbf{r},\mathbf{r}') = G_0(\mathbf{r},\mathbf{r}')G_{\bar{\varepsilon}}(\mathbf{r},\mathbf{r}') \tag{7.6}$$

where $G_0$ is the free-space Green's function of our choice. Normally a Fresnel-type kernel of the form

$$G_0(\mathbf{r},\mathbf{r}') = \frac{1}{i\lambda(z-z')}\exp\left[\frac{ik}{2(z-z')}[(x-x')^2+(y-y')^2]\right] \tag{7.7}$$

is sufficient, due to the large propagation distances involved (here z is the axial coordinate and x,y are the transverse coordinates). Substituting this in the Green's function derivation of diffraction theory (chapter 1), we arrive at the following modified diffraction integral:

$$E(x,y,z=L) = \frac{1}{i\lambda L}\int_{-\infty}^{\infty} e^{\frac{ik}{2L}((x-x')^2+(y-y')^2)} G_{\bar{\varepsilon}}(x,y,L;x',y',0)E_0(x',y')dx'dy' \tag{7.8}$$

where $E_0$ is the object field, $(x',y')$ are the transverse coordinates at the object plane, L is the distance from the object, and $(x,y)$ are the transverse coordinates at the observation plane. A Taylor expansion of $G_{\bar{\varepsilon}}$ in powers of $x$, $y$, $x'$, $y'$, and $L$ produces a DC term corresponding to free-space propagation, and therefore proportional to the amplitude of the ballistic signal.

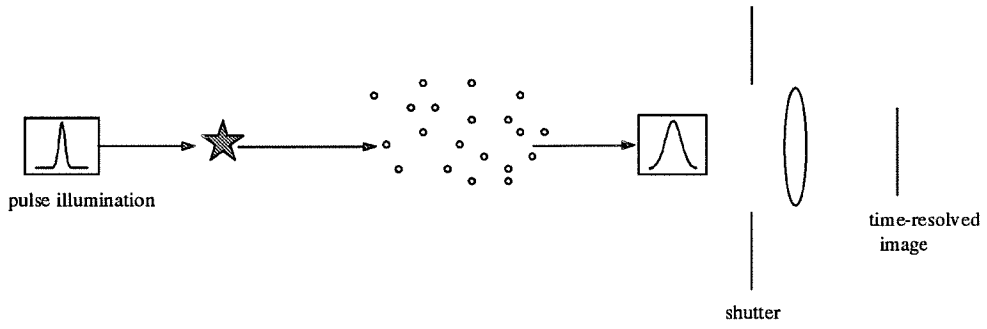



Figure 7.2. Ballistic imaging via pulsed illumination: imaging shutter is synchronized to capture first arriving light.

### 7.3. Methods of ballistic imaging

Let us next review the existing methods for differentiating ballistic from scattered signals. The first such technique was introduced by Duguay and Mattick in 1971 [1], and entails illuminating the object to be resolved with short pulses (Figure 7.2). As the pulsed signal propagates through a scatterer, the distribution of scattering distances will cause a broadening of the pulse, with weakly scattered signal components forming the leading edge of the pulse and highly scattered components forming the trailing edge. By using a shutter synchronized to the illumination pulsing at the imaging aperture, we can eliminate the contribution of scattered photons to the image.

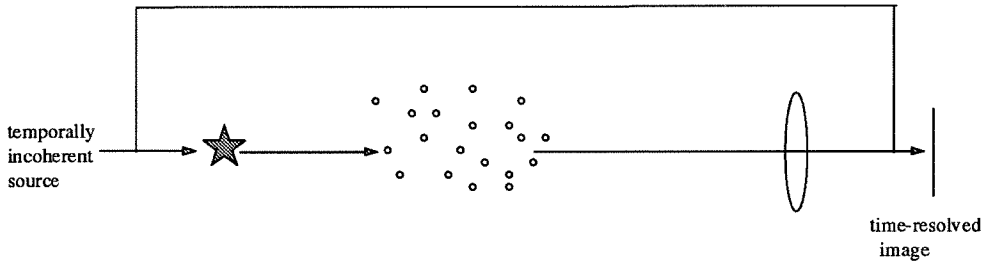



Figure 7.3. Holographic technique for capturing first arriving light: Reference beam is delayed to interfere coherently at image plane with ballistic signal.

A continuous-wave method of isolating ballistic photons was introduced by Abramson in 1978 [3]. This technique requires illumination of the object by a temporally incoherent source (Figure 7.3). When the image-plane field is mixed with a properly delayed

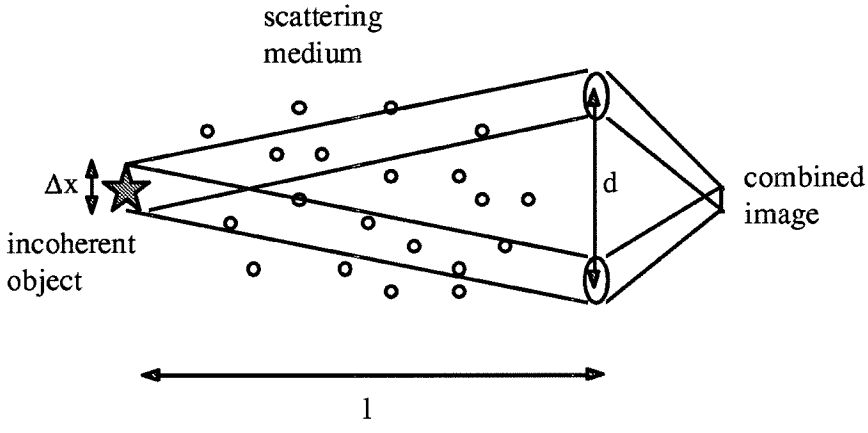



Figure 7.4. Schematic configuration of self-interference imaging. Images propagating through different paths but identical distances in scatterer are aligned and superimposed at the image plane.

reference beam (the delay should equal the delay a ballistic signal would experience in propagating from object plane to image plane), the ballistic signal interferes coherently with the reference beam, forming a grating proportional to the former, while all scattered signal components interfere incoherently due to their inherent time delay, forming a DC background pattern. Thus the ballistic image can be extracted by holographically or electronically retrieving the (high-frequency) grating-superimposed information and discarding all low-frequency components.

These two temporal discrimination methods have direct spatial counterparts [2]. Specifically, time-resolved imaging is analogous to spatially-resolved scanning imaging, where a narrow beam illuminates the object and the imaging aperture accepts only small-angle (low-spatial frequency) signal components, thereby eliminating highly-scattered components (which tend to exit at large angles). Temporal coherence imaging, on the other hand, is analogous to spatial coherence imaging, where the object is illuminated with a spatially incoherent source and the imaged field is mixed with a properly aligned reference beam. Here the interference between image and reference beam is coherent only if an image component is displaced by less than a coherence diameter while propagating through the scatterer. Thus the grating formed at the image plane will almost exclusively contain information from weakly scattered or ballistic components.

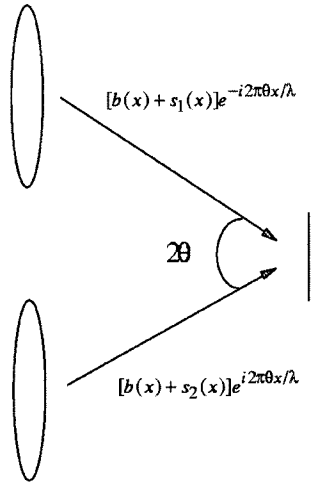



Figure 7.5. Ballistic components b(x) interfere at an angle 2θ to form a grating of frequency θ/λ at the image plane; scattered components $s_i(x)$ interfere incoherently and do not form a grating.

### 7.4. Ballistic imaging via self-interference

Next consider an object illuminated with low-coherence light, as in Figure 7.4. After propagating through a scattering medium, two images of the object will interfere coherently only if their path difference $\Delta d$ through the medium is small compared to the coherence length, i.e., $\Delta d < l_c = ct_c$ (for spatially incoherent objects, the relative displacement must be less than the coherence width). Thus if

we interfere two images of the same object acquired at points equidistant from the object (extended objects are treated below), only the ballistic signals from both images (and the small fraction scattered through the same path lengths to both images) will interfere coherently at the image plane.

When we interfere the images at an angle θ (Figure 7.5), the total intensity at the image ($x$-$y$) plane is

$$\begin{aligned}\left|E(x,y)\right|^2 &= \left|(b(x,y)+s_1(x,y))e^{-i2\pi(\theta_x x+\theta_y y)/\lambda}+(b(x,y)+s_2(x,y))e^{i2\pi(\theta_x x+\theta_y y)/\lambda}\right|^2 \\ &= \left|s_1(x,y)\right|^2+\left|s_2(x,y)\right|^2+4\left|b(x,y)\right|^2\cos^2(2\pi(\theta_x x+\theta_y y)/\lambda)+CT\end{aligned} \tag{1}$$

where λ is the illumination wavelength, $b(x,y)$ is the ballistic (coherent) field component, and $s_i(x,y)$ are the scattered (incoherent) components at the two lenses, whose cross terms ($CT$) disappear over time averages larger than the coherence time. Thus the coherent components $b(x)$ form a grating of spatial frequency θ/λ, on which our ballistic image information is superimposed. To form this grating, the two images must overlap to within the coherence width of the object illumination; this serves as a useful calibration tool, since we can adjust the images' positions until the strongest possible grating forms at the image plane, at which point they overlap exactly. Also, the grating frequency must be large

enough to sample our image without information loss, i.e., it must obey the Nyquist criterion

$$\begin{aligned} u_{grating} &= \theta / \lambda > 2B \\ &\Rightarrow \theta > 2\lambda B \end{aligned} \qquad (2)$$

where $B$ is the image's spatial bandwidth. Two-dimensional systems must obey this relation in each dimension independently.

### 7.5. Object-plane constraints

Note that there is no assumption of isoplanacity in this analysis, i.e., light from different points on the object need not travel identical paths through the scatterer; this distinguishes the method from almost all other atmospheric compensation techniques [4]. Still, we do need to ensure that light from any given point on our object travels the same distance to both imaging systems, to within a coherence length. Assuming the center of the object is equidistant from the centers of the two imaging systems, this requires

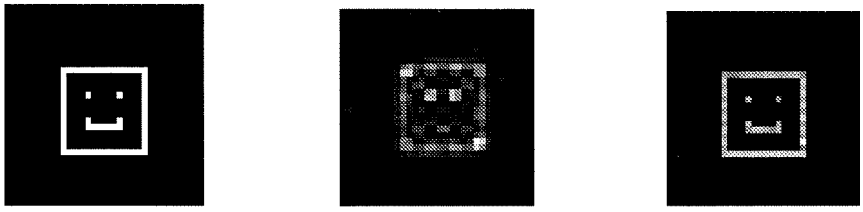
$$d\Delta x/2l < l_c, \qquad (3)$$

where $d$ is the distance between the imaging systems, $\Delta x$ is the object size, $l$ is the object's distance (from the center of the two imaging systems), and $l_c$ is the

illumination coherence length (see Figure 7.4). This requirement is usually far less stringent than isoplanacity, and does not depend on the properties of the scattering medium.

### 7.6. Experimental technique and results

Once the grating is formed, we can extract the ballistic signal in several ways--holographically, by exposing film to the grating and then illuminating it with a coherent plane wave; electronically, by capturing the image with a CCD and frame grabber (assuming the grating period is smaller than the CCD elements) and using discrete Fourier transforms to extract the information superimposed on the grating; or optically, using spatial filtering to isolate the grating information (this requires conversion of our image to a spatially coherent one, which can be done using conoscopic holography [6]). The latter two methods are preferable from a practical viewpoint, since they allow real-time imaging; we demonstrate the second below due to its experimental simplicity.



(a) (b) (c)

Figure 7.6. Simulation of smiley-face (a) to be resolved through a distributed scatterer, (b) imaged without compensation, and (c) imaged with self-interference technique.

Using a distributed scatterer model [7,9] we simulated the performance of the image self-interference technique assuming a ballistic signal fraction $10^{-3}$ of the object intensity (i.e., 0.1% of the light emanating from our object is transmitted ballistically through the scatterer). This model introduces point scatterers throughout the propagation volume, whose distribution produces the desired mean scattering free path. Our simulation results, shown in Figure 7.6, indicate that the self-interfered image (c) has far lower noise than the conventionally acquired image (b); both assume an infinite lens aperture.

Next, to test the degree to which scattered light interferes coherently with the ballistic component, we mixed an incoherent noise beam with a coherently illuminated image, as in Figure 7.7, and attempted to reconstruct the coherent image using self-interference (we extracted the grating electronically; examples of high- and low-coherent component gratings are shown in Figure 7.8a and 7.8c, respectively). The incoherent noise was generated by routing a portion of the coherent illumination beam through a rotating diffuser; thus its center wavelength was unchanged, and separation was necessarily coherence-based. As expected, perfectly aligned images allow the incoherent beam to form a grating with itself, and the degree to which the images are aligned determines the grating strength. In Figure 7.8 we give examples of well-aligned and misaligned images: in Figures 7.8a,b a coherently-illuminated image with incoherent background, and in figures b,c an incoherently-illuminated image. Figure 7.9 shows the variation of image SNR (after self-interference) with radial image displacement, along with a fitted theoretical curve assuming a Gaussian mutual correlation function. The results indicate a coherence width of 850 μm (corresponding to the FWHM of the mutual correlation function), which is the lower bound for detectable scattering distances using this source and diffuser (HeNe laser at 633 nm and 1725 rpm, 14.3 cm diameter diffuser).

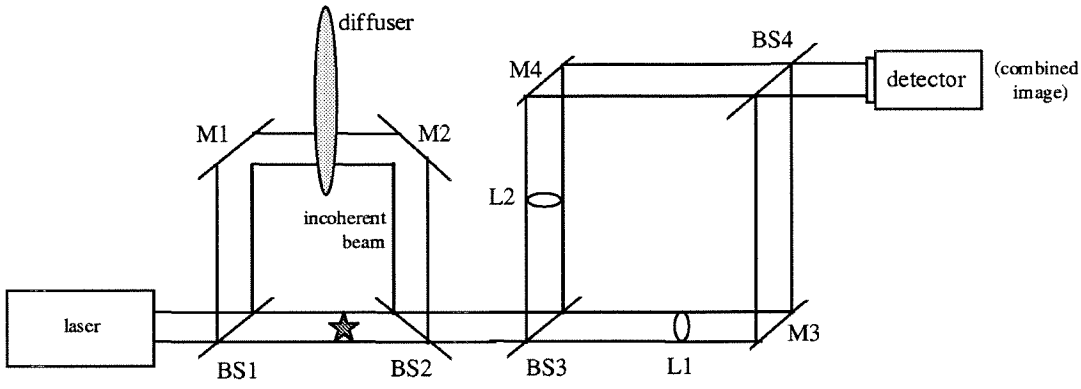



Figure 7.7. Experiment to determine effectiveness of self-interference in eliminating incoherent noise, and to measure coherence length of HeNe laser and rotating diffuser (l=633 nm, diffuser diameter = 14.3 cm, diffuser speed = 1725 RPM).

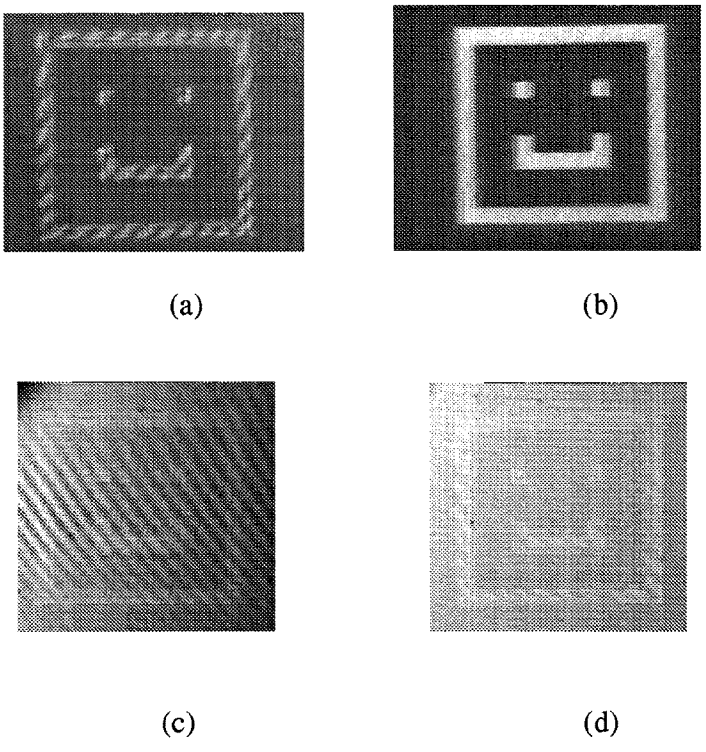



Figure 7.8. Typical self-interference results: incoherently-illuminated image with no background noise (a) aligned and (b) misaligned; and coherently-illuminated image with incoherent background noise (a) aligned and (b) misaligned.

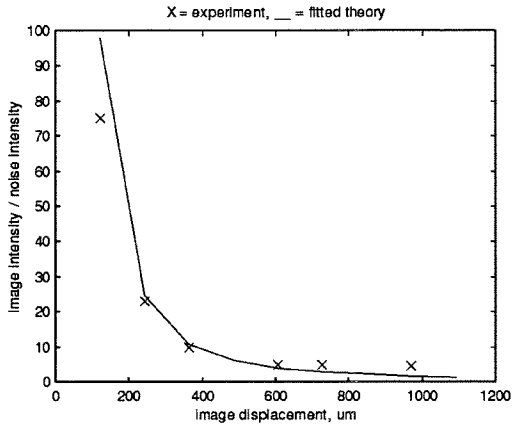
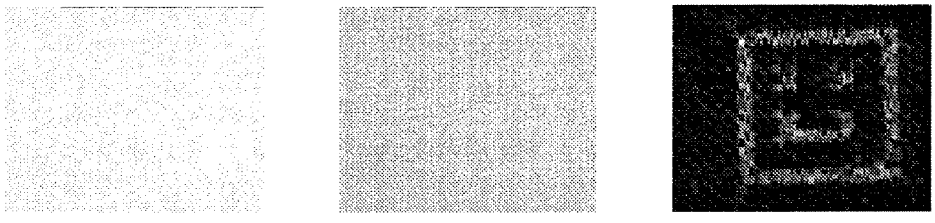



Figure 7.9. Variation of image-plane SNR with misalignment of images. Theoretical (solid) curve is the expected SNR vs. misalignment for a Gaussian mutual-intensity function with an 850 μm coherence diameter.

Finally, we employed the self-interference method to compensate for a distributed scatterer transmitting 2% ballistic illumination (scattering coefficient $\alpha$=.782 $cm^{-1}$; length=5 cm) with an interference angle of .22 radians and wavelength of 633 nm. The grating was recorded on a CCD and extracted digitally; images resulting from no compensation, ordinary spatial filtering to eliminate highly-scattered rays, and self-interference (all normalized by their maximum intensities) are shown in Figure 7.10. These conform to the theory, simulations, and preliminary experiment above, showing a significant restoration of the image via self-interference and a large advantage over spatial filtering alone. The results are somewhat better when

the distributed scatterer is replaced by a thin diffuser as in Ref. 2, but this case is less relevant since self-interference is most useful for long-range imaging. Also, we expect objects reflecting with a polarization preference (unlike the laser-diffuser combination) to enjoy a somewhat better SNR due to the depolarizing effects of scattering [8]. Because of the ballistic component's low intensity, the final image suffers from CCD quantization errors due to limited sensitivity; this problem can be alleviated by using a longer integration time, even in the case of a turbulent medium, since most of the light affected by the turbulence interferes incoherently with the other image.



(a) (b) (c)

Figure 7.10. Result of imaging the smiley face through a 5-cm scatterer with scattering factor $\alpha$=.782 $cm^{-1}$, (a) without any compensation, (b) with spatial filtering to eliminate highly-scattered light, and (c) with image self-interference and electronically extracted ballistic grating.

Compared to reference-beam interferometric techniques, this method produces a much weaker ballistic grating, since the interference intensity is proportional to the ballistic signal intensity, whereas in reference-beam techniques it is proportional to the product of the ballistic and reference beam amplitudes. The resulting degradation in SNR should therefore only be tolerated when an aligned reference beam is not available.

### 7.7. Conclusion

As noted in Reference 2, both temporal and spatial incoherence can be used to differentiate ballistic from scattered illumination. Temporal coherence methods utilize the longer propagation time of scattered light, while spatial coherence methods rely on the spatial displacement thereof. The extent to which these two techniques overlap is unknown, but with self-interference both spatially and temporally incoherent sources may be imaged, and the image quality will increase with their degree of incoherence. Since light scattered less than a coherence length interferes coherently at the image plane, our image resolution varies directly with the illumination coherence length (assuming a large aperture). This, along with its tolerance for large integration times and independence of isoplanacity assumptions, makes the technique useful for many long-range imaging applications.

**References**

1. M. A. Duguay and A. T. Mattick, Appl. Opt. 10, 2162 (1971).

2. E. N. Leith, C. Chen, H. Chen, Y. Chen, J. Lopez, P. C. Sun, and D. Dilworth, Opt. Lett. 16, 1820 (1991).

3. N. Abramson, Opt. Lett. 3, 121 (1978).

4. A. Labeyrie, Astron. & Astrophys. 6, 85 (1970).

5. M. J. Beran and J. Oz-Vogt, "Imaging through turbulence in the atmosphere," *Progress in Optics XXXIII*, ed. E. Wolf, Amsterdam: Elsevier (1994).

6. Gabriel Sirat and Demetri Psaltis, Opt. Lett. 10, 4 (1985).

7. C. Raman, *The Molecular Diffraction of Light*, Calcutta: Calcutta University Press (1922), Chapter 4.

8. D. Eliyahu, M. Rosenbluh, and I. Freund, JOSA A 10, 477 (1993).

9. B. Salik, D. Provenzano, and A. Yariv, "Imaging through random and turbulent media using image self-interference," submitted to Optics Letters (2/97).

# Chapter Eight

# Realizability of arbitrary space-time field distributions

I wish to conclude by addressing the limits to the synthesis of arbitrary desired field distributions over various subsets of four-dimensional space-time.

## 8.1. Introduction

This issue has been treated for many interesting special cases [1-5], and is motivated by the need for "wavefront engineering" in fields from photolithography [6] to super-resolution imaging [7]. Here I will adopt a general approach assuming only scalar free-space propagation. The assumption of scalar fields actually restricts our solution space, and therefore to the extent that a scalar field distribution is realizable, so is at least one vector field distribution with the same intensity.

## 8.2. One spatial dimension

We begin, for concreteness, with a time-harmonic field (this is relevant to most applications), after which we shall generalize the results to arbitrary time dependence:

$$E(x,y,z,t) = f(x,y,z)e^{i\omega t} \tag{8.1}$$

The wave equation then becomes

$$\nabla^2 E = \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2}$$

$$\Rightarrow \nabla^2 f = -\frac{\omega^2}{c^2} f = -k_0^2 f \tag{8.2}$$

where $k_0 = \sqrt{\omega^2 / c^2}$.

When $f$ has no $y$- or $z$-dependence, the Laplacian becomes

$$\nabla^2 f = \frac{d^2 f}{dx^2} \tag{8.3}$$

and the solutions of (8.2) are

$$f(x) = Ae^{-ik_0 x} + Be^{ik_0 x}. \tag{8.4}$$

Given two free parameters, $A$ and $B$, we can specify the field arbitrarily at only two points on the x-axis. More generally, for every frequency $\omega$ (spatial frequency $k_0$) in our solution, we can specify the field arbitrarily at two additional points on the x-axis. Thus a field specified at N points on the x-axis requires in general N/2 frequencies to realize.

### 8.3. Two spatial dimensions

Given $k_0$, there are infinitely many solutions to the two-dimensional analogue of (8.2),

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = -k_0^2 f(x,y). \tag{8.5}$$

We decompose the solutions to separable eigenfunctions:

$$\begin{aligned} f(x,y) &= f_x(x) f_y(y) \\ \Rightarrow \frac{1}{f_x}\frac{\partial^2 f_x}{\partial x^2} &+ \frac{1}{f_y}\frac{\partial^2 f_y}{\partial y^2} = -k_0^2 \end{aligned} \tag{8.6}$$

implying each term on the left must equal a constant, e.g.,

$$\frac{1}{f_x}\frac{\partial^2 f_x}{\partial x^2} = -k_x^2, \frac{1}{f_y}\frac{\partial^2 f_y}{\partial y^2} = -k_y^2,$$
$$k_x^2 + k_y^2 = k_0^2 \tag{8.7}$$

which yield space-harmonic solutions of the form (8.4) in x and y:

$$f(x,y) = (A_x e^{-ik_x x} + B_x e^{ik_x x})(A_y e^{-ik_y y} + B_y e^{ik_y y}) \tag{8.8}$$

Retaining only the forward-traveling waves (this will not affect our results) and labeling $\alpha = k_x/2\pi = 1/\lambda_x$,

$$f_\alpha(x,y) = A_{x\alpha} e^{-ik_x x} A_{y\alpha} e^{-ik_y y} = A_{x\alpha} e^{-i2\pi\alpha x} A_{y\alpha} e^{-iy\sqrt{k_0^2 - (2\pi\alpha)^2}} \tag{8.9}$$

which is a plane wave propagating at an angle $\theta = \tan^{-1}\frac{k_y}{k_x}$ to the x-axis. A general superposition of these functions can be written

$$f(x,y) = \int_{-\infty}^{\infty} f_\alpha(x,y)d\alpha = \int_{-\infty}^{\infty} A_{x\alpha} e^{-i2\pi\alpha x} A_{y\alpha} e^{-i\sqrt{k_0^2 - (2\pi\alpha)^2}\, y} d\alpha\ . \tag{8.10}$$

We wish to realize some $g(x,y)$ defined at a discrete number of points $(x,y)_i$. Since these points are discrete, we can always define a fine enough rectangular mesh of points $x_n$, $y_m$ which comes arbitrarily close to the points $(x,y)_i$ (Figure 8.1). Thus, letting $A_\alpha = A_{x\alpha} A_{y\alpha}$,

$$f(x_k, y_j) = \int_{-\infty}^{\infty} A_\alpha e^{-i2\pi x_n \alpha} e^{-iy_m \sqrt{k_0^2 - (2\pi\alpha)^2}} d\alpha . \tag{8.11}$$

It is straightforward to show that the eigenfunctions (8.8) are mutually orthogonal over the continuous plane $(x,y)$; what we need, however, is a set of linearly independent eigenfunctions over the *discrete* plane $(x_n, y_m)$.

Setting

$$x_n = \frac{nx_0}{N}, \quad y_m = \frac{my_0}{M}, \tag{8.12}$$

$$0 \le n < N, \; 0 \le m < M$$

where $x_0$ and $y_0$ are the maximum $x$ and $y$ distances, respectively, the inner product of $f_{\alpha 1}$ and $f_{\alpha 2}$ is

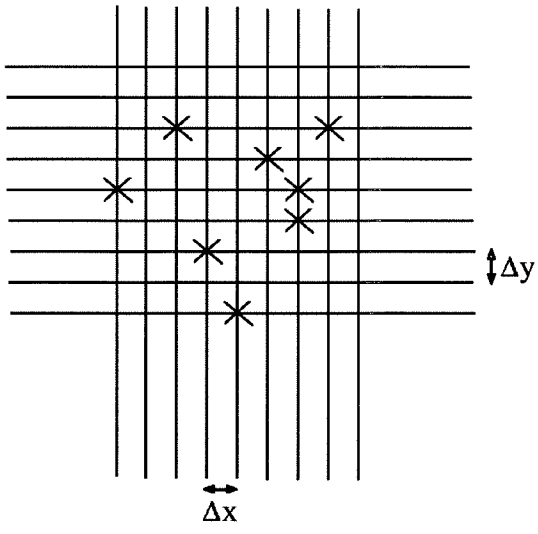



Figure 8.1. An arbitrary discrete set of points can always be covered arbitrarily well by a rectangular mesh.

$$(f_{\alpha_1}, f_{\alpha_2}) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} e^{i2\pi(\alpha_2-\alpha_1)x_n} e^{i2\pi(\beta_2-\beta_1)y_m}$$

$$\sum_{n=0}^{N-1} e^{i2\pi(\alpha_2-\alpha_1)nx_0/N} \sum_{m=0}^{M-1} e^{i2\pi(\beta_2-\beta_1)my_0/M} \tag{8.13}$$

where we assigned $\beta_i = \sqrt{k_0^2 - (2\pi\alpha_i)^2}\,/2\pi = \sqrt{\lambda_0^{-2} - \alpha_i^2}$. We seek a set of *NM* linearly independent eigenfunctions; the first summation will provide *N* orthogonal $\alpha$'s, e.g.,

$$\alpha_j = j/x_0,\ 0 \le j < N\text{-}1 \tag{8.14}$$

which span the space of functions over $\{x_n\}$. With respect to the first summation, any $f_{\alpha_{j+cN}}$ is equivalent to $f_{\alpha_j}$, viz.

$$\sum_{n=0}^{N-1} e^{-i2\pi\left(\frac{j+cN}{x_0}\right)nx_0/N} = \sum_{n=0}^{N-1} e^{-i2\pi\left(\frac{j}{x_0}\right)nx_0/N} e^{-i2\pi cn} = \sum_{n=0}^{N-1} e^{-i2\pi\left(\frac{j}{x_0}\right)nx_0/N} \tag{8.15}$$

so to get a set of *NM* linearly independent eigenfunctions over $\{x_n, y_m\}$ we need for each $f_{\alpha_j}$ a set of *M* functions $\{ f_{\alpha_{j+cN}} \}$ which are linearly independent with respect to $\{y_m\}$. Since

$$\beta_{j+cN} = \sqrt{\lambda_0^{-2} - (\alpha_{j+cN})^2} = \sqrt{\lambda_0^{-2} - \left(\frac{j+cN}{x_0}\right)^2} \tag{8.16}$$

is not guaranteed to be an integer multiple of $y_0^{-1}$, as $\alpha_{j+cN}$ is of $x_0^{-1}$, it is difficult to find a set of M mutually orthogonal functions $\{ f_{\alpha_{j+cN}} \}$ with respect to $\{y_m\}$. We will therefore content ourselves to prove linear independence. To do this, it is sufficient to show that any set of $M$ functions $\{ f_{\alpha_{j+cN}}; c \in C \}$ are linearly independent over $\{y_m\}$ provided

$$\beta_{j+c_1N} - \beta_{j+c_2N} \neq qM / y_0$$
$$\forall q \in Z; \forall c_1, c_2 \in C \tag{8.17}$$

i.e., that no two functions in the set are equivalent with respect to the second summation in (8.13). Suppose we start with a set $\{ f_{\alpha_{j+cN}}; c \in C \}$ which have evenly spaced $\{ \beta_{j+cN} \}$ separated by $1/y_0$, i.e., they are orthogonal with respect to $\{y_m\}$, but are not all solutions of (8.5). An arbitrary function $f_{\alpha_{j+dN}}$ can then be expanded as a sum of these functions over $\{y_m\}$:

$$e^{-i2\pi m y_0 \beta_{j+dN} / M} = \sum_{c \in C} w_c e^{-i2\pi m y_0 \beta_{j+cN} / M},$$
$$0 \leq m < M \tag{8.18}$$

where $w_c$ are the expansion coefficients

$$w_c = \sum_{m=0}^{M-1} e^{-i2\pi m y_0 \beta_{j+dN}/M} e^{i2\pi m y_0 \beta_{j+cN}/M},$$

$$c \in C. \tag{8.19}$$

If now we wish to replace some function $f_{\alpha_{j+c_0N}}$ in our orthogonal set by $f_{\alpha_{j+dN}}$, the new set will still span the space of functions over $\{y_m\}$ if and only if we can express $f_{\alpha_{j+c_0N}}$ over $\{y_m\}$ as a linear superposition of the new functions. But from (8.18) we have

$$e^{-i2\pi m y_0 \beta_{j+c_0N}/M} = \frac{1}{w_0} e^{-i2\pi m y_0 \beta_{j+dN}/M} + \sum_{c \in C-c_0} \frac{w_n}{w_0} e^{-i2\pi m y_0 \beta_{j+cN}/M},$$

$$0 \le m < M \tag{8.20}$$

which is analytic iff $w_0 \neq 0$. Thus the set $\{ f_{\alpha_{j+dN}}, f_{\alpha_{j+cN}}; c \in C - c_0 \}$ spans the space of functions over $\{y_m\}$ iff

$$w_0 = \sum_{m=0}^{M-1} e^{-i2\pi m y_0 \beta_{j+dN}/M} e^{i2\pi m y_0 \beta_{j+c_0N}/M} \neq 0 \tag{8.21}$$

or, equivalently,

$$\beta_{j+dN} - \beta_{j+cN} \neq qM / y_0$$
$$\forall q \in Z; \forall c \in C - c_0 \tag{8.22}$$

which, together with the orthogonality of the original set, is the same as condition (8.17). We can now repeat this procedure by replacing a member of our new basis with another arbitrary function, as in (8.20); we thus show that any set of $M$ functions $\{ f_{\alpha_{j+cN}}; c \in C \}$ are a basis over $\{y_m\}$ iff they satisfy condition (8.17). By choosing $M$ such functions for every $j$, we form a set of $NM$ linearly independent eigenfunctions spanning the set of possible functions over $\{x_n, y_m\}$. In fact, there are (countably) infinitely many such choices we can make for every $j$; recall, furthermore, that we limited ourselves in (8.14) to functions which are orthogonal over $\{x_n\}$, which restricted $\alpha$ to discrete values. Replacing that requirement with linear independence over $\{x_n\}$, as we required over $\{y_m\}$, allows us an uncountably infinite solution space. This means that every discrete field distribution in the $xy$-plane is realizable as a solution of the 2-D scalar wave equation, and in fact a field specified on a rectangular mesh of $NM$ two-dimensional points is physically realizable with $NM$ two-dimensional plane waves.

### 8.4. Three spatial dimensions

In rectangular coordinates, the wave equation (8.2) is

$$\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}=-k_0^2 f(x,y) \tag{8.23}$$

which, decomposing to separable eigenfunctions, becomes

$$f(x,y)= f_x(x) f_y(y) f_z(z)$$

$$\Rightarrow \frac{1}{f_x}\frac{\partial^2 f_x}{\partial x^2}+\frac{1}{f_y}\frac{\partial^2 f_y}{\partial y^2}+\frac{1}{f_z}\frac{\partial^2 f_z}{\partial z^2}=-k_0^2 \tag{8.24}$$

implying

$$\frac{1}{f_x}\frac{\partial^2 f_x}{\partial x^2}=-k_x^2,\ \frac{1}{f_y}\frac{\partial^2 f_y}{\partial y^2}=-k_y^2,\ \frac{1}{f_z}\frac{\partial^2 f_z}{\partial z^2}=-k_z^2$$

$$k_x^2+k_y^2+k_z^2=k_0^2 \tag{8.25a,b}$$

whose solutions are

$$f(x,y,z)=(A_x e^{-ik_x x}+B_x e^{ik_x x})(A_y e^{-ik_y y}+B_y e^{ik_y y})(A_z e^{-ik_z z}+B_z e^{ik_z z}). \tag{8.26}$$

The condition (8.25b) allows us two continuous degrees of freedom, with which we wish to realize a discrete 3-D field distribution. Since we showed above that a discrete 2-D field is realizable given one continuous degree of freedom, we simply

use our extra free parameter (say $k_z$) to span the extra dimension (say $z$). For example, using

$$z_l = \frac{lz_0}{L}, \quad 0 \le l < L \tag{8.27}$$

$z_0$ being the maximum $z$-distance, and setting

$$\gamma_j = k_{zj} / 2\pi = j / z_0, \quad 0 \le j < L \tag{8.28}$$

the inner product over $z$ of two forward-propagating eigenfunctions is

$$(f_{z_j}, f_{z_{j'}}) = \sum_{l=0}^{L-1} e^{i2\pi(\gamma_{j'} - \gamma_j)z_l} = \sum_{l=0}^{L-1} e^{i2\pi \frac{j'-j}{z_0} lz_0 / L} = \delta_{jj'} \tag{8.29}$$

giving us $L$ orthogonal eigenfunctions spanning the set of functions over $\{z_l\}$. For each of these, we can solve the two-dimensional problem of finding $NM$ linearly independent eigenfunctions over $\{x_n, y_m\}$ using the procedure given above, giving us $NML$ eigenfunctions which are linearly independent over $\{x_n, y_m, z_l\}$ and thus form a basis thereon. An arbitrary discrete three-dimensional field distribution defined on a rectangular mesh of $NML$ points is therefore realizable by a superposition of $NML$ three-dimensional plane waves.

## 8.5. Three spatial dimensions + time

To treat this case we must drop the assumption of monochromatic illumination. We therefore revert to the full wave equation

$$\frac{\partial^2 E}{\partial x^2}+\frac{\partial^2 E}{\partial y^2}+\frac{\partial^2 E}{\partial z^2}-\frac{1}{c^2}\frac{\partial^2 E}{dt^2}=0 \tag{8.30}$$

and again assume separability of $E(x,y,z,t)$ to extract the following eigensolutions:

$$\begin{aligned} E(x,y,z,t) = (A_x e^{-ik_x x} + B_x e^{ik_x x})(A_y e^{-ik_y y} + B_y e^{ik_y y})\times \\ (A_z e^{-ik_z z} + B_z e^{ik_z z})(A_t e^{-i\omega t} + B_t e^{i\omega t}) \end{aligned} \tag{8.31}$$

with $k_x^2 + k_y^2 + k_z^2 = \omega^2/c^2 = k_0^2$ as in (8.25b). We now have three free continuous parameters ($\omega$ is not fixed) with which to realize a four-dimensional discrete field distribution. In the same manner that we extended our two-dimensional results to the three-dimensional case, we may use two of the free parameters (say, $k_z$ and $\omega$) to generate sets of eigenfunctions orthogonal over the corresponding lattices (say, $z$ and $t$), then use the last degree of freedom to generate a set of eigenfunctions that spans the remaining two-dimensional lattice, as in the 2-D case discussed above.

## 8.6. Realization of arbitrary fields

In order to obtain the relative weights of the eigenfunctions in our realization, we first write our desired field distribution $u_{nm}$ as a sum of orthogonal functions $g_{nm}^{jk}$ on the grid (we use two dimensions for simplicity):

$$u_{nm} = \sum_{j=0}^{N-1}\sum_{k=0}^{M-1} v_{jk} g_{nm}^{jk} = \sum_{j=0}^{N-1}\sum_{k=0}^{M-1} v_{jk} e^{-i2\pi \frac{j}{x_0}\frac{nx_0}{N}} e^{-i2\pi \frac{k}{y_0}\frac{my_0}{M}} \tag{8.32}$$

Note these functions are not necessarily solutions of the wave equation, since in general

$$(\frac{j}{x_0})^2 + (\frac{k}{y_0})^2 \neq \lambda_0^{-2}. \tag{8.33}$$

However, we know from (8.18) that each of our eigenfunctions can also be written *on the discrete grid* as a sum of the orthogonal functions $g_{nm}^{jk}$:

$$f_{nm}^{c+dN} = e^{-i2\pi(nx_0\alpha_{c+dN}/N + my_0\beta_{c+dN}/M)} = \sum_{j=0}^{N-1}\sum_{k=0}^{M-1} w_{jk}^{cd} e^{-i2\pi\frac{j}{x_0}\frac{nx_0}{N}} e^{-i2\pi\frac{k}{y_0}\frac{my_0}{M}}$$

$$= \sum_{j=0}^{N-1}\sum_{k=0}^{M-1} w_{jk}^{cd} g_{nm}^{jk},$$

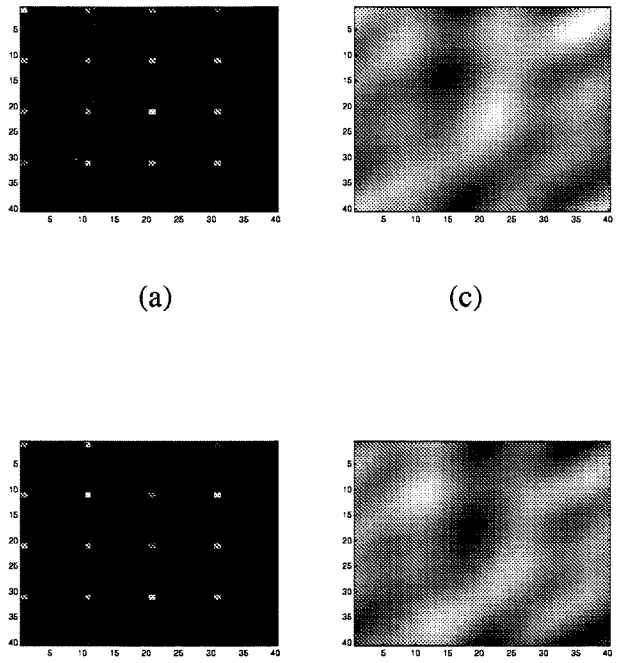
$$0 \le m < M, 0 \le n < N \tag{8.34}$$

Writing (8.34) as a tensor product, $F=GW$ (in the two-dimensional case $F$, $G$, and $W$ are all $N$ x $N$ x $M$ x $M$ tensors), we can express the orthogonal basis $G$ on our grid in terms of $F$, viz. $G = F W^{-1}$, which is nonsingular if both $f^{c+dN}$ and $g^{jk}$ are linearly independent sets of functions. We can also write (8.32) as a tensor product $U=GV$ (in two dimensions $U$ and $V$ are $N$ x $M$), implying

$$U = FW^{-1}V = FV' \tag{8.35}$$

where $V'=W^{-1}V$ is the coefficient matrix for the $F$ representation of $U$.

We used this technique to realize a randomly generated complex discrete field distribution over a grid of 4 x 4 elements. Figure 8.2a,b shows the desired real and imaginary parts of the field at the grid points, while Figure 8.2c,d shows the realized real and imaginary distributions everywhere, which match the desired field at the appropriate points. In so doing, we took advantage of the fact that, using the $g_{nm}^{jk}$ definition in (8.32), $v_{jk}$ is simply the discrete Fourier transform of $u_{nm}$,



(a) (c)

(b) (d)

Figure 8.2. Realization of two-dimensional discrete field distribution. (a) and (b): randomly generated $4_{pixel}$ x $4_{pixel}$ discrete real and imaginary parts of field distribution, respectively; (c) and (d): realized real and imaginary parts of field distribution which match the desired distributions in (a) and (b), respectively, at the appropriate points in 2-D space.

while $w_{jk}^{cd}$ is the DFT of $f_{nm}^{c+dN}$, allowing us to use fast fourier transforms to calculate these tensors.

### 8.7. Converseness with Nyquist sampling

It has been pointed out to me [8] that the theorems proven here are analogous to the Nyquist sampling theorem for continuous signals. Specifically, the Nyquist theorem states [9] that a continuous signal of bandwidth $B$ should be sampled at intervals smaller than $1/2B$ in order to avoid loss of information (in other words, information preservation is guaranteed if the sampling interval is smaller than $1/2B$). The theorem given here for one dimension describes the synthesis of discrete fields by superposition of plane waves, and state in general that a field defined on $N$ points requires $N/2$ plane waves to realize ($N$ plane waves if we disallow backward-propagating solutions). Assuming a discrete spatial interval of $\Delta x$, then, the signal support is $N\Delta x$, meaning our bandwidth is $B=(N/2)/(N\Delta x)=1/2\Delta x$. Thus we have the converse of the Nyquist sampling theorem: Given a discrete field defined on intervals $1/2B$, a continuous field of bandwidth $B$ can always be constructed which equals the discrete field everywhere the latter is defined.

### 8.8. Average intensity on volume elements

For most applications, we are concerned more with the intensity distribution than the field distribution; furthermore, we usually measure not the intensity at

particular points, but integrated over discrete volume elements. Therefore, the question of whether arbitrary local-volume-integrated intensity distributions are synthesizable becomes interesting. We will treat the two-dimensional case here, since it is the simplest case for which arbitrary discrete monochromatic field distributions are synthesizable, and the results are readily generalizable to higher dimensions. We assume throughout a continuous field distribution, which is the case if we have no charges and use a finite number of plane waves to synthesize our field.

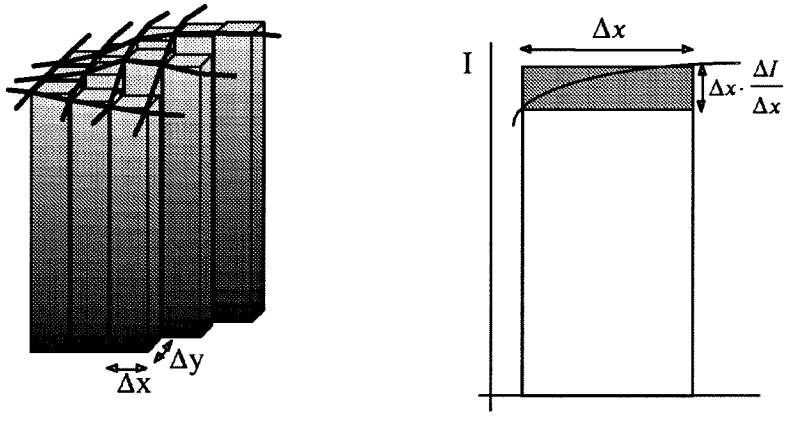
The intensity $I$ at every point is proportional to the square of the field strength; therefore, averaged over a volume element $x \in [\frac{n}{N}x_0, \frac{n+1}{N}x_0)$, $y \in [\frac{m}{M}y_0, \frac{m+1}{M}y_0)$, it is

$$I_{nm}^{avg} = \frac{1}{\Delta x \Delta y} \int_{\frac{n}{N}x_0}^{\frac{n+1}{N}x_0} \int_{\frac{m}{M}y_0}^{\frac{m+1}{M}y_0} |E(x,y)|^2 dy dx \tag{8.36}$$

where $\Delta x = x_0 / N, \Delta y = y_0 / M$ are the pixel dimensions. As $\Delta x, \Delta y \to 0$, this can be approximated by

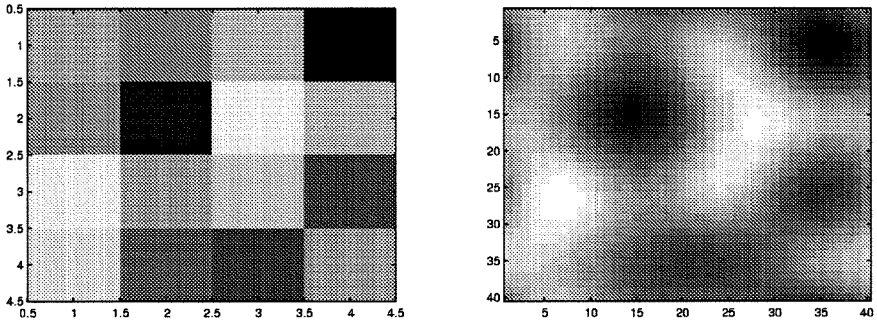
$$\lim_{\Delta x,\Delta y\to 0} I_{nm}^{avg} = \frac{1}{\Delta x\Delta y}\left|E(\frac{n}{N}x_0,\frac{m}{M}y_0)\right|^2 \Delta x\Delta y = \left|E(\frac{n}{N}x_0,\frac{m}{M}y_0)\right|^2 \qquad (8.37)$$

with an error of the order $(\Delta x\Delta y)^2$ at each pixel, implying an integrated intensity error on the $xy$-plane of the order $\Delta x\Delta y$ (Figure 8.3). This order of error can be improved by using higher-order approximations, e.g., the trapezoid rule or Cramer's rule [10].

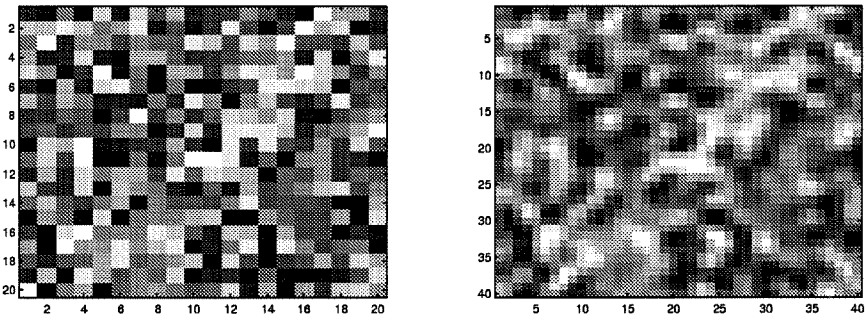



(a) (b)

Figure 8.3. (a) Approximating the intensity distribution by a sequence of rectangles. (b) In one dimension, the integral error over one pixel (rectangle) of width $\Delta x$ is at most $\Delta x \cdot \Delta x\frac{\Delta I}{\Delta x}$, which for continuous $I(x)$ scales as $\Delta x^2$ when $\Delta x$ is small.




(a)




(b)

Figure 8.4. (a) Sparsely specified field (intensity) distribution leads to (b) large intensity fluctuations within area elements, while densely specified intensity distribution (c) leads to smaller deviations within area elements.

Therefore, by making $\Delta x$ and $\Delta y$ smaller in the specification of our discrete field distribution (or, equivalently, making $N$ and $M$ large), we can come arbitrarily close to any desired intensity distribution over the $xy$-plane. This is illustrated in Figure 8.4, where the large $\Delta x$ and $\Delta y$ in (a) lead to the large intensity variations over area elements in (b), while finer $\Delta x$ and $\Delta y$, as specified in (c) lead to smaller fluctuations (d). The amplitudes of the specified discrete field are uniquely determined by (8.37), while the phases are free parameters in this problem. The generalization to three-dimensional and time-dependent intensity distributions is straightforward.

## 8.9. Conclusion

Using monochromatic illumination, it is evidently possible to realize arbitrary discrete two- and three-dimensional field distributions, while one- and four-dimensional distributions require multiple frequencies. These conclusions depend only on the number of free variables resulting from the wave equations, and therefore it does not matter which dimensions we choose to specify; for example, a two-dimensional distribution may be specified on the $xy$-, $yz$-, or $xz$- planes. Furthermore, by dropping the assumption of time harmonic fields, we can include $t$ as one of the specifiable dimensions. It is clear that, using monochromatic plane waves in three dimensions, we can realize arbitrary discrete field distributions in one and two dimensions, since these are special cases of three-dimensional

distributions. Therefore, it is important to remember that the treatment above addresses the realizability of $N$-dimensional discrete fields using plane waves in $N$ dimensions, and thus that the realized fields are constant in the other $(3\text{-}N)$ dimensions.

**References**

1. J. Rosen and A. Yariv, "Synthesis of an arbitrary axial-field profile by computer-generated holograms," Opt. Lett. **19**, 843-845 (1994).

2. B. Salik, J. Rosen, and A. Yariv, "Nondiffracting images under coherent illumination," Opt. Lett. **20**, 1743 (1995).

3. R. Piestun, B. Spektor, and J. Shamir, "Wave-fields in 3 dimensions--analysis and synthesis," JOSA A **13**, 1837-1848 (1996).

4. T. Dresel, M. Beyerlein, and J. Schwider, "Design and fabrication of computer-generated beam-shaping holograms," Appl. Opt. **35**, 4615-4621 (1996).

5. B. Salik, J. Rosen, and A. Yariv, "One-dimensional beam shaping," JOSA A **12**, 1702-1706 (1995).

6. Y. Liu, A. K. Pfau, and A. Zakhor, "Systematic design of phase-shifting masks with extended depth of focus and or shifted focus plane," IEEE Trans. Semic Manuf. **6**, 1-21 (1993).

7. C. J. R. Sheppard, "Leaky annular pupils for improved axial imaging," Optik **99**, 32-24 (1995).

8. J. Obrien, private communication 2/2/97.

9. A. V. Oppenheim and R. W. Schafer, *Discrete-time Signal Processing*, New Jersey: Prentice-Hall (1989).

10. Louis Leithold, *The Calculus with Analytic Geometry*, New York: Harper & Row, 1968, p. 513.