Optical Information Storage and Processing

Thesis by

Zhiwen Liu

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2002

(Defended September 27, 2001)

Acknowledgments

First of all, I would like to thank my advisor, Professor Demetri Psaltis. Without his patience, guidance and encouragement, I would not have finished this work. To him, my greatest gratitude.

Lucinda Acosta has provided tremendous help in almost everything during my stay in the group. I owe Yayun Liu so much for her generous assistance in the lab. Iouri Solomatine has helped me a lot in making materials.

Many thanks to Dr. Greg Steckman for teaching me about the optical correlators and other lab skills when I joined the group. Thanks to Dr. Chris Moser for showing me how to use his spectral hole burning setup. Thanks to Dr. Wenhai Liu for all the interesting and productive discussions with him. Thanks to Dr. Ali Adibi for his encouragements and advice. David Zhang gave me great help in the spectral hole burning project as a surf student directly from high school. I would also like to thank George Panotopoulos for upgrading me to "professor" every time I call him "doctor." Thanks to Jose Mumbru for his refreshing jokes. Thanks to Yunping Yang for coinventing a time-multiplexing technique when we were sharing the lab. Thanks to Irena Maravic for some heated philosophic debating. I would like to thank Dr. Greg Billock, Dr. Karsten Buse, Dr. George Barbastathis, Dr. Gan Zhou, Dr. Xin An, Dr. Xu Wang, Dr. Ernest Chuang, Dr. Michael Levene, Dr. Allen Pu, Martin Centurion, Todd Meyrath, Hung-Te Hsieh, Manos Fitrakis, Marc Luennemann, Dimitris Sakellariou and all members of the extended Psaltis group family. I would also like to thank the secretaries of the EE department.

Finally, I would like to thank my family. Their support, understanding and love help me overcome many difficulties in my academic endeavor.

Abstract

Optical information storage and optical information processing are the two themes of this thesis. Chapter two and three discuss the issue of storage while the final two chapters investigate the topic of optical computing.

In the second chapter, we demonstrate a holographic system which is able to record phenomena in nanosecond speed. Laser induced shock wave propagation is recorded by angularly multiplexing pulsed holograms. Five frames can be recorded with frame interval of 12ns and time resolution of 5.9ns. We also demonstrate a system which can record fast events holographically on a CCD camera. Carrier multiplexing is used to store 3 frames in a single CCD frame with frame interval of 12ns. This technique can be extended to record femtosecond events.

Information storage in subwavelength structures is discussed in the third chapter. A 2D simulation tool using the FDTD algorithm is developed and applied to calculate the far field scattering from subwavelength trenches. The simulation agrees with the experimental data very well. Width, depth and angle multiplexing is investigated to encode information in subwavelength features. An eigenfunction approach is adopted to analyze how much information can be stored given the length of the feature. Finally we study the effect of non-linear buffer layer.

We switch gear to holographic correlators in the fourth chapter. We study various properties of the defocused correlator which can control the shift invariance conveniently. An approximate expression of the shift selectivity is derived. We demonstrate a real time correlator with 480 templates. The cross talk of the correlators is also analyzed.

Finally, in the fifth chapter we apply the optical correlator to fingerprint identification and study the performance of the correlation based algorithms. The windowed correlation can improve the rotation and distortion tolerance.

Table of Contents

List of Figures

CHAPTER 1 Introduction

CHAPTER 2 Holographic Recording of Fast Phenomena

1 Introduction

Optics is one of the oldest fields in science, but yet it gives us fascinating new faces. Optics has been found wide applications in physics (laser spectroscopy, quantum optics, BEC, etc.), chemistry (femtosecond chemistry), biology (optical tweezers), and medical practicing (laser eye surgery), to name but a few. The last two decades have witnessed the revolution in the information industry: the boom of optical fiber telecommunications and massive growth of the internet. However, the most widely used information processors are still electronic (RAM, CPU). In this thesis we focus on the topic of information processing using optical techniques.

One of the promising candidates for information storage is holographic memory. It has attracted researchers' interest for decades because of its potential of huge storage capacity (V/λ^3) and fast accessing rate (light speed optically, limited by electronics). In chapter two, we apply holography to monitoring fast phenomena. A reference pulse and a signal pulse instead of two CW beams are used to record a hologram. Different frames are recorded by different pairs of pulses using multiplexing techniques (angle, wavelength, and frequency multiplexing). The resolution is limited by the recording laser pulse width. Individual frames can be read out separately because of the multiplexing selectivity. A nanosecond movie camera is demonstrated, which is used to record the laser induced shock wave propagation. The multiple pairs of recording pulses are generated from a single pulse of a Q-switched Nd:YAG laser using specially designed reference and signal cavities.

Angular multiplexing is employed and five frames with resolution of 5.9 ns are recorded. We also propose and demonstrate recording nanosecond events directly on commercially available low frame rate CCD cameras (30 fps) using carrier multiplexing. The femtosecond version of reference and signal pulse generators are designed. Carrier multiplexing can be extended to record femtosecond movies. Finally we propose another idea to build a femtosecond movie camera using spectral hole burning medium. Some preliminary material characterization of the SHB material (M/#) is included.

In chapter three, we study the diffraction from subwavelength structures. A potential application is to improve the track density of current DVD/CD technologies without using complicated near field scanning. A 2D simulation tool using the FDTD algorithm is developed. We first verify our simulation by comparing the calculated far field scattering pattern with the experimental data. The far field scattering from a trench can be divided into two parts. One is reflection from the wafer surface and the other is diffraction of the light reflected from the bottom of the trench. The phase difference between these two parts can be controlled by the depth of the trench. Information can be encoded by varying the width or depth of the trench and the incident beam polarization. If the scattering is linear, we can stack many trenches together and create a phased array which is equivalent to a mirror. Information can also be encoded by varying the slanting angle of the mirror. In order to analyze how much information can be encoded in a unit length, which is able to be detected in the far field, we employ an eigenfunction approach to investigate the near field to far field transform. Finally we also investigate the effect of adding a nonlinear buffer layer on top of the trench encoded with information. Self-focusing or self-shaping in the nonlinear layer with intensity dependent refractive index or absorption can reduce the effective beam waist.

An approximate expression of the beam waist in the nonlinear medium is derived using a perturbation method. It is found that the waist depends on the effective nonlinear phase shift.

The final two chapters discuss the holographic correlator and its applications. Unlike in a memory device, signal is used to reconstruct reference in a correlator. If the recording material is thin, the correlator is shift invariant in that a shift of the input image simply results in a corresponding shift of the correlation peak in the output plane. A certain degree of shift invariance is required to make the system robust; however, it can limit the number of templates that can be stored. In chapter four, we discuss various properties of the defocused holographic correlator, where shift invariance can be controlled simply by moving the recording material away from the focal plane. We derive the shift selectivity and find out that the selectivity is inversely proportional to the defocusing distance and the image size. Because the side lobes are suppressed in the defocused correlator, cross talk in correlator arrays is reduced. We also demonstrate a real time correlator system with 480 templates limited by shrinkage of the recording material. In chapter five, we apply the optical correlator to fingerprint identification. A false rejection rate of 4.7% and a false classification rate of 4.6% are obtained. We then study the validity of the correlation based algorithms for fingerprint identification. Windowed correlation can improve the distortion and rotation tolerance.

2 Holographic Recording of Fast Phenomena

2.1 Introduction

Monitoring fast phenomena is of interest to science and engineering since it tells us about the dynamics of physical processes [2-1][2-2]. For instance, the pump-probe technique is widely used in non-destructive and repeatable measurements. For 1D imaging (e.g., lifetime measurements), streak cameras with subpicosecond resolution can be used. Recording movies of fast events can be accomplished with a set of sensors [2-3]. The light from the object is gated electronically on to each sensor (intensified CCD) while the image is broadcasted. Limited by weak intensity and silicon circuit speed, about 10 frames can be recorded with frame interval and time resolution of 10 ns.

Since the early days when holography was invented [2-4], people have been studying high speed events using holographic techniques. A well known example is double exposure interferometry [2-5]. In this method, two pulsed holograms are recorded successively with the reference beams in the same direction. Upon reconstruction, the two frames are reconstructed simultaneously and interfere with each other. Multiple frames can be stored and reconstructed separately using multiplexing techniques. In Section 2.2 a system to record nanosecond movies is discussed. We experimentally demonstrate the system by making movies of laser induced shock waves with a temporal resolution of 5.9 ns, limited by the pulse width of the Q-switched Nd:YAG laser used in the experiments. In Section 2.3 we investigate carrier multiplexing on CCD cameras. We record three frames of an air discharge event in a single CCD frame. This technique is useful for low-bandwidth images and

can be extended to femtosecond regime. In Section 2.4, some schemes to record femtosecond movies are investigated. Properties of femtosecond pulsed holograms in $LiNbO_3$ are discussed. Spectral hole burning holography is also investigated as a candidate to record femtosecond movies.

2.2 Nanosecond holographic movie camera [2-13] [2-14]

Previous work has focused on spatial multiplexing where holograms are recorded at different locations of the recording medium. In an early report, a 100 ps laser pulse is used to store five frames [2-6]. Beam splitters are used to generate the reference and signal pulse trains. The time delay between beam splitters determines the frame interval. In another method, a wave front preserving optical delay line (a White cell) and a specially graded beam splitter is used to generate the reference and signal pulse trains [2-7]. The frame interval can be as short as 28.3 ns. Spatial multiplexing by rotating the recording medium has also been reported [2-8], but the speed is either limited by the mechanical scanning or the laser pulse repetition rate. Pulsed holograms have also been angularly multiplexed [2-9] taking advantage of the thickness of the recording medium. In one method, three lasers are used to generate three reference beams with different angles and each laser fires a pulse in a different time [2-10]. The frame interval is about 1μs. A rotating mirror [2-11] or electro-optic switches [2-12] have also been used to generate the reference beams. In these efforts, the speed is limited by electronics or mechanic scanning.

In spatial multiplexing, the number of frames is limited by the recording medium area and the beam diameter, not the dynamic range of the recording material. Furthermore, the direction of signal beam changes from pulse to pulse (except that the recording medium

is rotated) and thus it changes the perspective of the object. It also makes the interferometry between any two of the recorded frames more difficult.

The method we describe uses the angular selectivity of thick holograms to resolve frames that are recorded with adjacent pulses. Two specially designed cavities are used to generate the signal and reference pulse trains. The advantage of our method is that the speed is limited by the pulse width of the laser instead of a scanning mechanism. The duration of the movie (the number of frames) is limited by the dynamic range of the recording material, not its spatial extent.

As shown in Figure 2-1, a sequence of signal and reference pulses are incident on

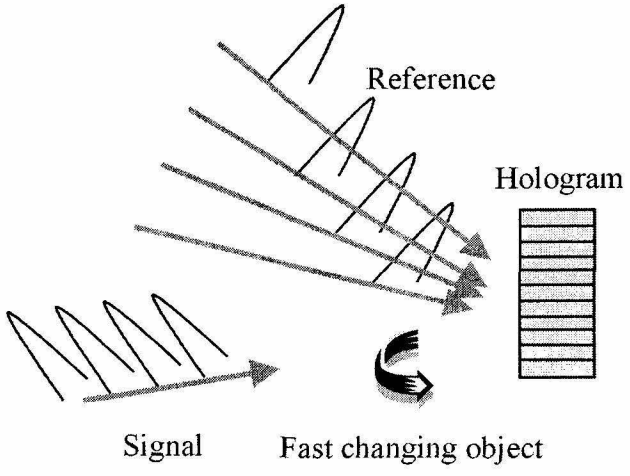


Fig. 2-1. Angularly multiplexing pulsed holograms

the holographic medium during the recording. The signal pulses all travel in the same direction while the reference beam direction changes from pulse to pulse in order to angularly multiplex holograms. After the recording, a CW laser at the same wavelength is used to read out individual frames. Depending on the incidence angle, different frames can be read

out separately due to the angular selectivity of the thick hologram. In the experiments, both the signal and the reference pulse trains are generated by a single pulse from a frequency doubled Q-switched Nd:YAG laser (wavelength 532nm, pulse width 5.9 ns, energy per pulse 300 mJ, beam diameter 9mm).

2.2.1 Nanosecond holographic system

Figure 2-2 shows the cavity that generates the signal pulse train. A polarizing beam

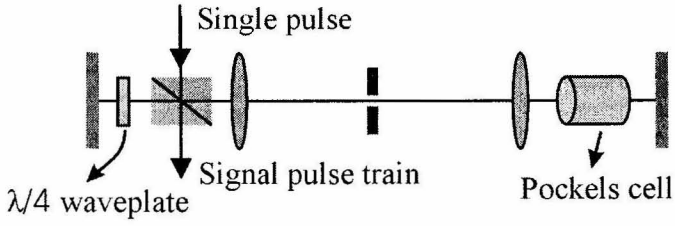


Fig. 2-2. Signal pulse train generation

splitter is used to couple the vertically polarized (perpendicular to the paper) incident pulse into the cavity. The Pockels cell is timed to behave like a temporary λ/4 wave plate to rotate the polarization of the incident pulse to horizontal direction after it first enters the cavity. It is turned off afterwards while the pulse travels back towards the opposite mirror. The pulse is then trapped inside the cavity since the polarizing beam splitter transmits beam with horizontal polarization. A λ/4 wave plate is used to slightly rotate the polarization of the pulse and the induced vertical polarized component is coupled out of the cavity from the polarizing beam splitter. A sequence of signal pulses are then generated and the pulse separation is equal to the round trip time of the cavity. The two lenses inside the cavity form an imaging system that compensates for beam diffraction.

The reference cavity is shown in Figure 2-3. The incident pulse enters the cavity via

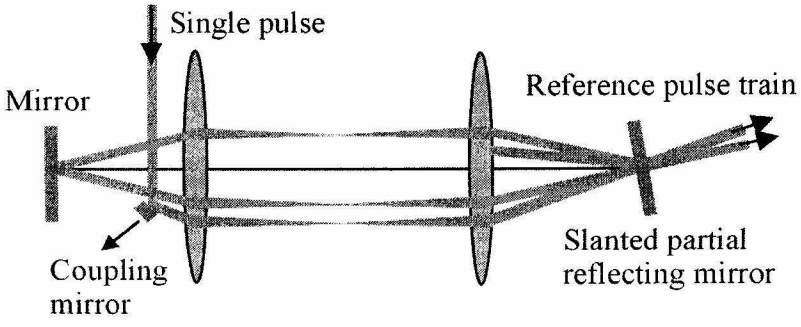


Fig. 2-3. Reference pulse train generation

a small coupling mirror. After the coupling mirror, it travels as if it had originated from the center of the front mirror. The two lenses form an imaging system and the pulse hits the center of the rear mirror. If the rear mirror is parallel to the front one, the reflected pulse would travel backwards symmetrically with respect to the axis. After reflecting by the front mirror, it would then get blocked by the coupling mirror when it tries to retrace the previous path. We break the symmetry of the cavity by slanting the rear mirror slightly. The pulse is then reflected by the rear mirror and travels at a smaller angle towards the axis, just missing the coupling mirror. The pulse hits the center of each cavity mirror at slightly different angle after every round trip. Pulses are coupled out of the cavity by making one of the mirrors partially reflecting.

We generate five signal and reference pulses out of a single pulse from the Nd:YAG Q switched laser using the above method. The pulse separation is about 12 ns which can be changed by tuning the cavity length. Our current signal cavity is quite lossy due to the reflections from the optical components and the spatial filtering that is used to improve the beam profile, and this limits us to only about five signal pulses. Improved cavity design

should alleviate this problem. We recorded five plane wave pulsed holograms. The diffraction efficiency of each frame is shown in Figure 2-4. Both the reference and signal pulse

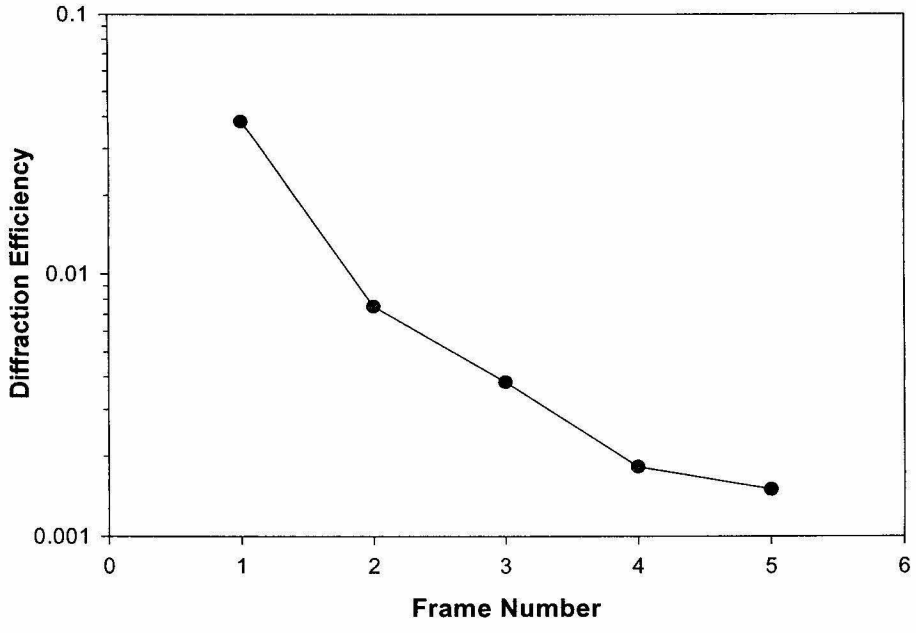


Fig. 2-4. Diffraction efficiency of pulsed holograms

train have a total energy of about 37 mJ. The pulse energy in the reference and signal pulse train decays and the successively recorded holograms get weaker and weaker. The diffraction efficiency of the first hologram is 4% while that of the fifth hologram is about 10^{-3}. Aprilis material [2-15] ULSH500-7A-22 is used as the recording medium. The thickness of the material is 200 μm. The pre-exposure energy is about $2J/cm^2$(white lamp). Figure 2-5 shows the selectivity curve of the first hologram. The measured selectivity is about 0.3 degree. The Aprilis material can yield a diffraction efficiency of approximately 1% for a $1.6mJ/cm^2$ exposure of cw illumination or a single 5.9ns pulse. The diffraction efficiency when M holograms are superimposed goes as $\eta=(M\#/M)^2$ where M/#[2-16] is a

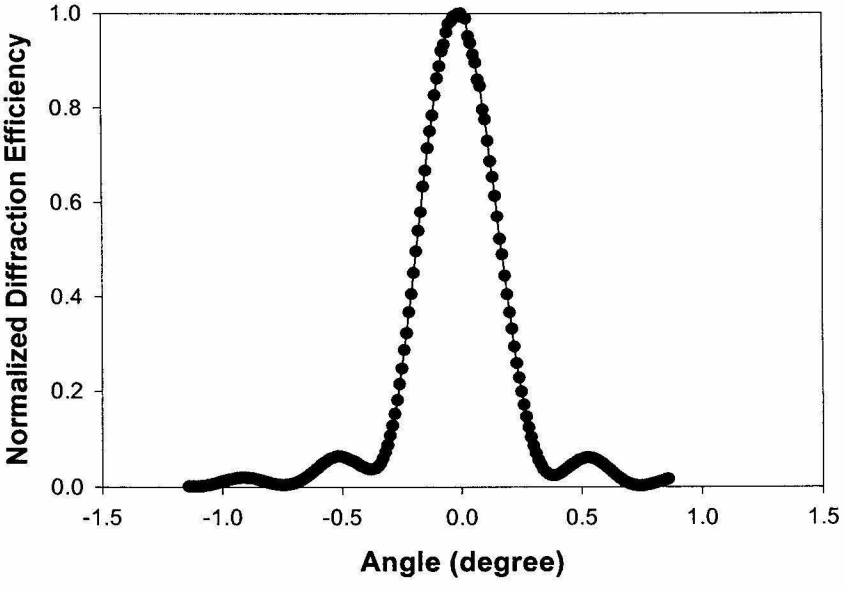


Fig. 2-5. Angular selectivity

material dynamic range parameter. The M/# of the Aprilis material is about 6. Since we typically can obtain high fidelity reconstructions with η of 10^{-4}, movies with several hundreds of frames can be recorded.

We also recorded pulsed image holograms. A mask with random binary pattern was used to modulate the signal beam. Figure 2-6 shows the reconstructed and direct images.

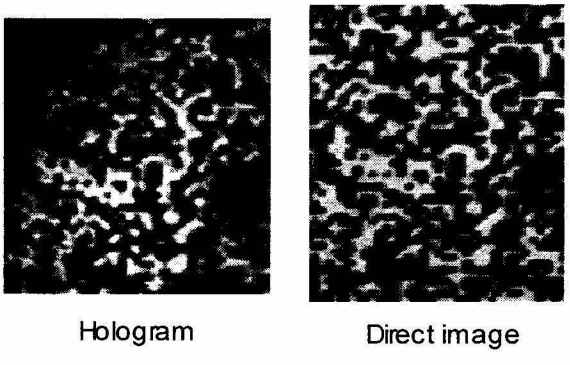


Fig. 2-6. Comparison between the pulsed hologram and direct image

After recording the hologram, the mask becomes yellowish and gets partially burnt by the high power laser pulses. For this reason, a low quality mask was used.

2.2.2 Nanosecond movies of shock waves

We used this movie camera to record optical breakdown events. Our single-pulse-pump-record experimental setup is shown in Figure 2-7. We split the pulse from the laser

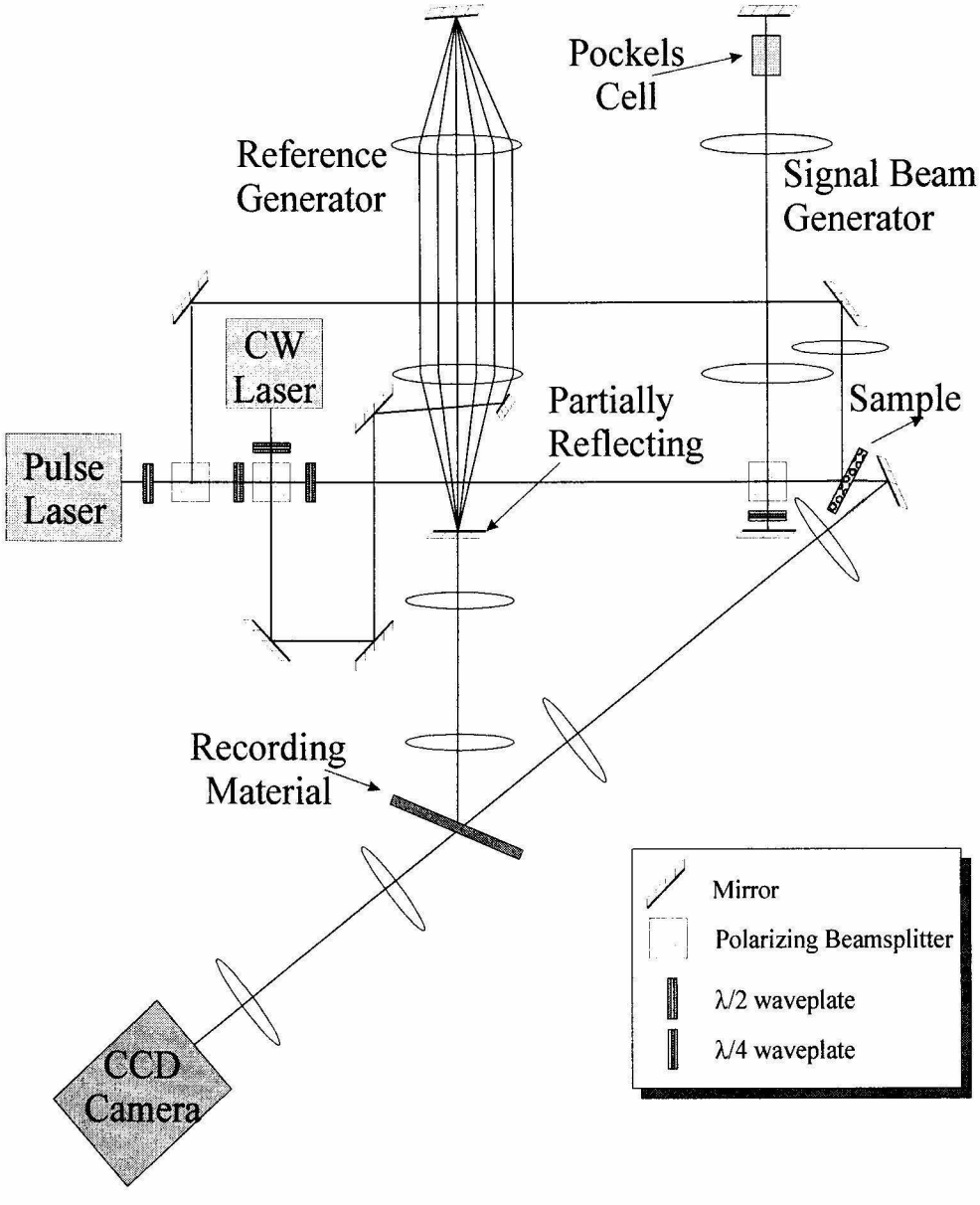


Fig. 2-7. High speed holographic movie camera

and focus it on some sample. This pumping pulse can optically break down the object [2-17]-[2-21]. Figure 2-8 shows the optical breakdown on a PMMA sample. Frame A is

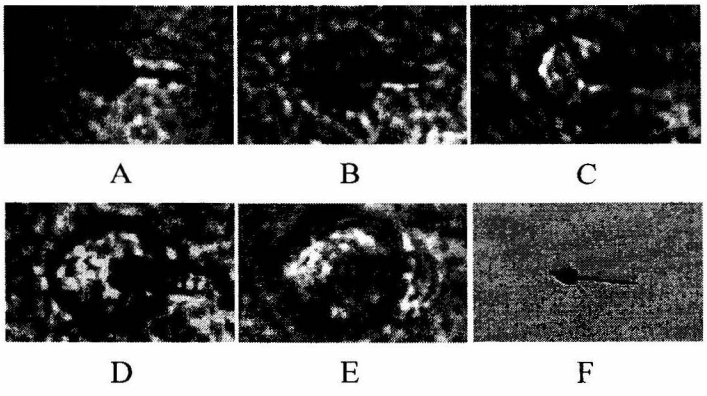


Fig. 2-8. Optical breakdown in PMMA

recorded at about 1ns before the pumping pulse vanishes. A, B, C, D and E are the successively recorded frames and the frame interval is about 12ns. F is the final direct image of the sample after the optical breakdown. The size of the image is 1.74mm × 1.09mm. The intensity of the pumping beam is about $1.6 \times 10^{12} W/cm^2$. Frame A shows the plasma created by the pumping pulse. The tail is likely due to the discharge in the air in front of the sample. In frame B, a shock wave is clearly seen. The average propagating speed of the shock wave between frame A and B is about 10 km/s and that between frame D and E is about 4 km/s. In Figure 2-9 we show the breakdown in air. Similarly, plasma is created in frame A and soon a shock wave forms. The image size is 2.76mm × 1.17mm. The intensity

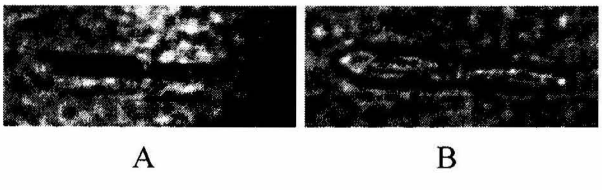
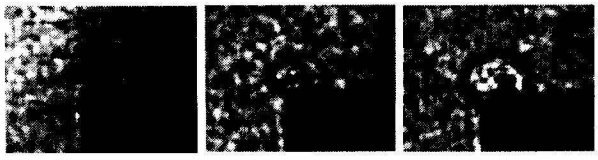


Fig. 2-9. Optical breakdown in air

of the pumping beam is about $5.2 \times 10^{12} W/cm^2$. The air discharge happens in a region near the focal point of the lens and the length of that region is about equal to the depth of focus. A line of spark is visible during the experiment. In Figure 2-10, we focus the pumping pulse



A B C

Fig. 2-10. Optical breakdown in air

near a blade (the dark rectangular shadow). The threshold of optical breakdown is lowered by the presence of the metal blade. The image size is 1.82mm × 1.45mm. The intensity of the pumping beam is about $1.6 \times 10^{12} W/cm^2$. The optical breakdown happens mainly at a small region around the focal point which is close to the metal and produce a more spherical shock wave. We also focused two pumping beams on PMMA to generate two shock waves as shown in Figure 2-11. The image size is 1.48mm × 1.52mm. In the first movie the lower pumping pulse has higher energy as we can see from the plasma size in frame 1a. When the two shock waves meet, the one with higher pressure penetrates as shown in frame 1c. In the second movie the two shock waves roughly have the same pressure, and they balance with each other in the middle.

A unique feature of the holographic movie is that it records the field and thus has both the amplitude and phase information. Phase changes can be detected by interfering two reconstructed frames or interfering the frame with the reference wave. We interfere the second frame of an air discharging movie with the reference (a plane wave). The fringes

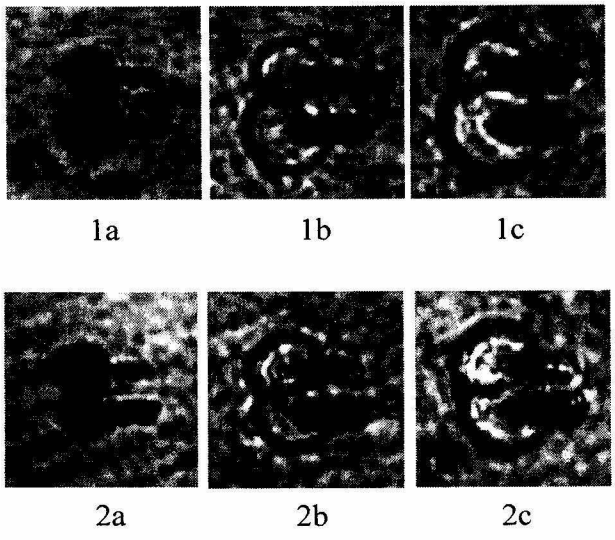


Fig. 2-11. Interaction of double shock waves

are shown in Figure 2-12. Apparently, the refractive index inside the region surrounded by

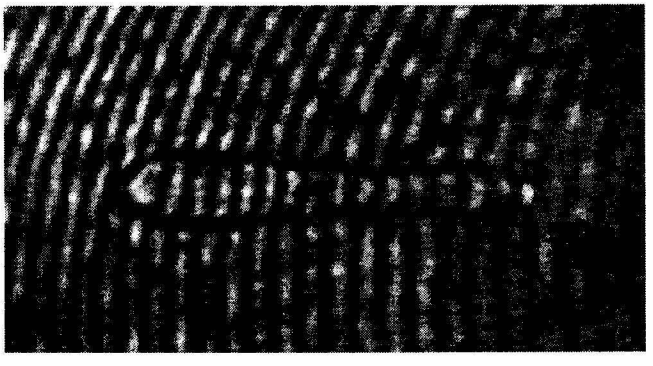


Fig. 2-12. Interferometry between a fast movie frame and its reference wave

the shock front is different from that of the outside and there is an index gradient. In the holographic reconstruction we can focus at different depths since the object field is reconstructed. This is shown in Figure 2-13. In A the plasma created on the PMMA sample is in focus, while in B the shock wave due to the discharge in the air (in front of the sample) comes to focus by changing the position of the CCD camera. The angle between the pump-

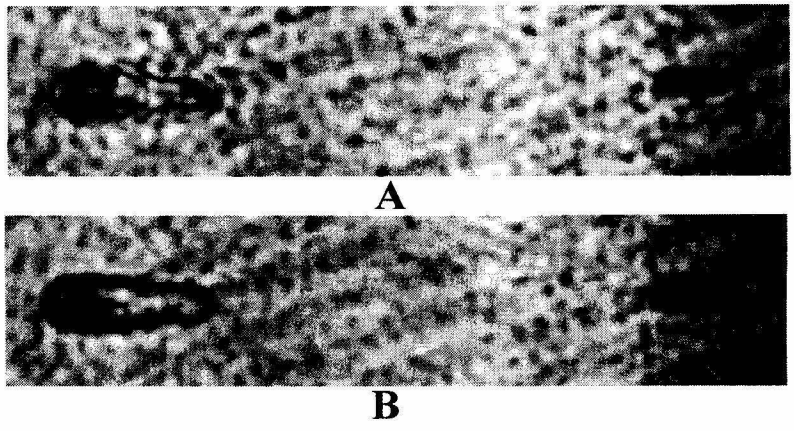


Fig. 2-13. Focusing at different depths

ing beam and the signal beam is about 20 degrees. The pumping pulse is focused at about 1 cm in front of the PMMA sample which is consistent with the measured image depth position difference between A and B.

2.3 Carrier multiplexing [2-26]

The shock wave images in the previous section have low bandwidth. In principle we can also record these holograms in a thin medium. Because of the advancement of CCD technology, there has been a lot of interest in digital holography [2-22]–[2-25] where holograms are recorded in a CCD camera and reconstructed digitally. CCD cameras not only provide a convenient interface to computers but also are an ideal thin recording material, which are sensitive to a very broad spectrum ranging from infrared to ultraviolet and have high responsivity (even hundreds of photons can be detected in one pixel).

Pulsed holograms can be recorded in a CCD camera as shown in Figure 2-14. Successive holograms are recorded with a slightly different reference angle as in angular multiplexing. Because the resolution of the CCD camera is typically about 10 μm, a beam-

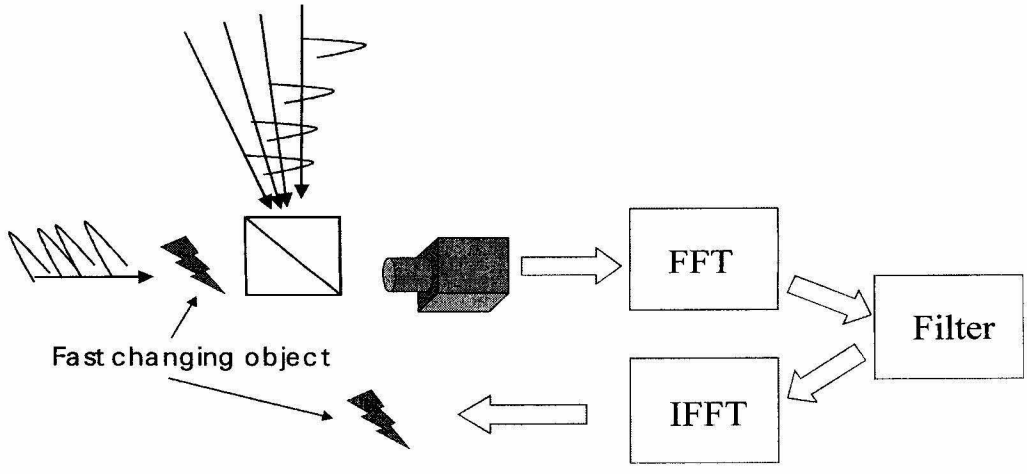


Fig. 2-14. Carrier multiplexing

splitter is used in order to record holograms in Mach-Zehnder interferometer configuration. The CCD camera records the integration of the whole sequence of pulse exposure. All of the holograms are superimposed in one composite CCD frame and each can be independently reconstructed through digital spatial filtering if the image bandwidth of each hologram is sufficiently low. Individual holograms can then be filtered out and reconstructed by first performing a digital Fourier Transform on the composite image, filtering a selected pass-band corresponding to the desired hologram and then performing an inverse Fourier transform on the filtered result. As an example we consider the interference pattern shown in Figure 2-12, which is the hologram of an air discharge event. Figure 2-15a shows the DC filtered Fourier Transform. In general, the recorded pattern on the CCD is proportional to

$$I = |R + S|^2 = |R|^2 + |S|^2 + R^*S + RS^* \qquad (2\text{-}1)$$

where R is the plane wave reference and S is the signal. There are two side bands in Figure 2-15a, which correspond to the last two terms of Equation 2-1. The position of the side band is determined by the interference pattern frequency or the carrier frequency,

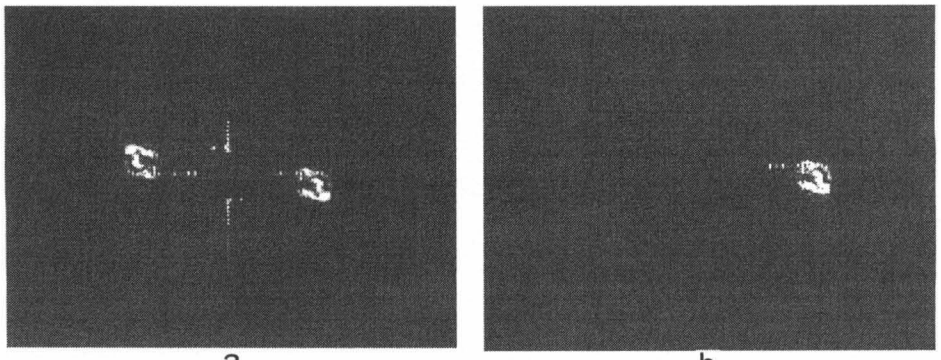


Fig. 2-15. DC filtered Fourier Transform

which can be changed by changing the reference beam incidence angle. We filter out one side band as shown in Figure 2-15b and perform the Inverse Fourier Transform. Either R*S or RS* is obtained. By taking the amplitude, the signal |S| is reconstructed as shown in Figure 2-16. Different holograms can be recorded without overlapping with each other in

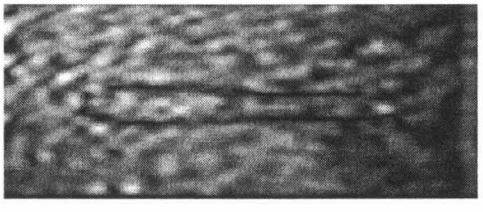


Fig. 2-16. Digitally reconstructed image

the frequency domain if the carrier frequency of each hologram is separated enough. The typical bandwidth of an air discharge pattern is about 0.0024 μm^{-1}. If the CCD pixel size is 10μm × 10μm (maximal bandwidth 0.05μm^{-1}), hundreds of frames can be recorded and reconstructed separately by 2D carrier multiplexing.

Our experimental setup is shown in Figure 2-17. We use the reference cavity in Section 2.2 to generate a sequence of pulses. The generated pulses are relayed by another imag-

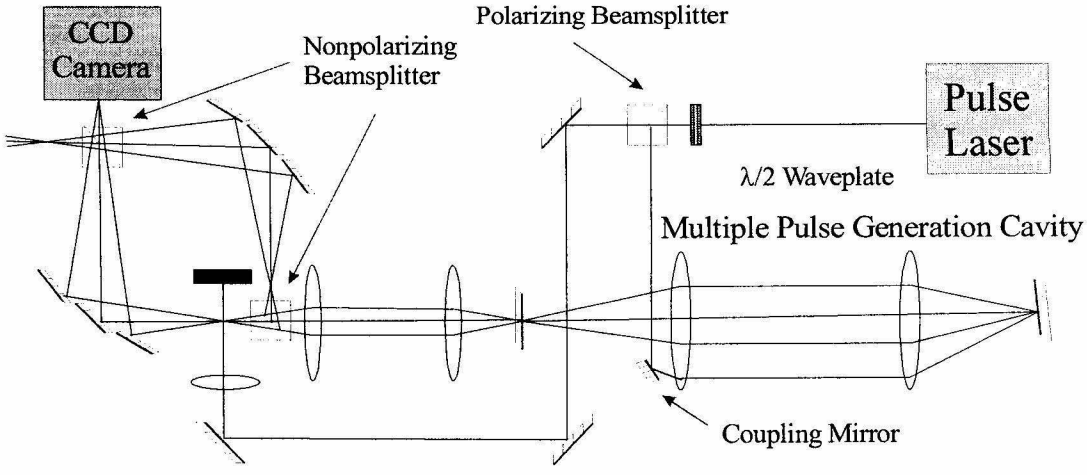
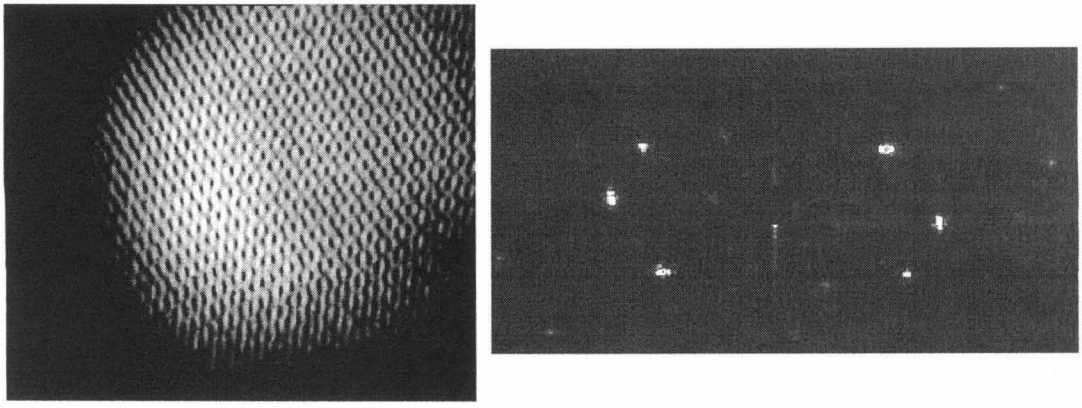


Fig. 2-17. Experimental setup

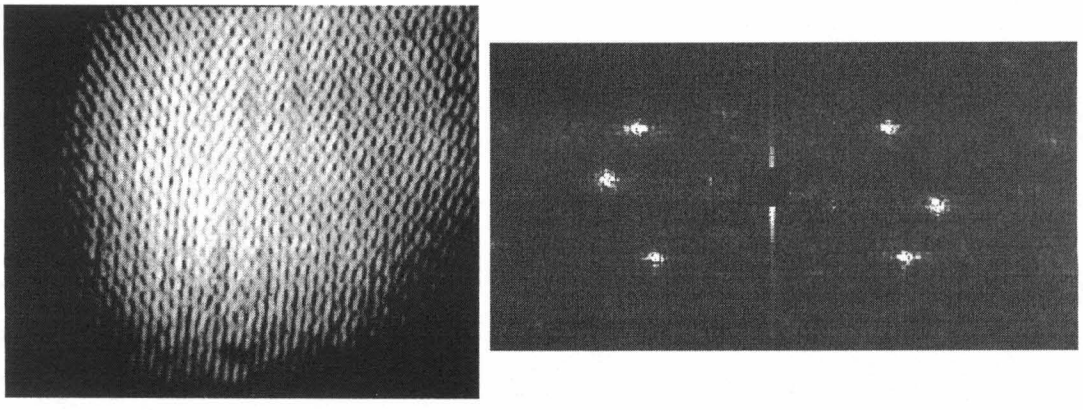
ing system and split into reference and signal by a non-polarizing beamsplitter. We recombine the reference and signal using mirrors and then record holograms on a CCD camera (Pulnix TM-7EX, 30 frames per second, 768 × 494 pixels, pixel size 8.4 × 9.8 μm). A portion of the laser pulse is split off and focus on the signal path to induce the air discharge. Holograms of the optical breakdown in air are then recorded using this setup.

We generated three pulses with pulse separation of 12ns. Three plane wave pulsed holograms and the DC filtered Fourier Transform spectrum are given in Figure 2-18a and b respectively. Clearly, they are well separated in the frequency domain. We then recorded an air discharge movie. The holograms and the Fourier transform of the composite CCD frame are shown in Figure 2-19a and b respectively. We reconstructed the three holograms and the results are given in Figure 2-20abc. Because the holograms are recorded in the far field, the diffraction of the image needs to be compensated. Consider one of the holograms $R_i^* S_i$. We take the x-y plane as the plane of the CCD camera. The field at z can be calcu-



a b

Fig. 2-18. Plane wave holograms and their DC filtered Fourier Transform



a b

Fig. 2-19. Air discharging movie

lated by convolving the field at z=0 with the Fresnel diffraction kernel, and is given by [2-27]

$$S_i'(x, y, z) = \iint R_i^*(x_1, y_1) S_i(x_1, y_1) e^{j\frac{\pi}{\lambda z}[(x-x_1)^2+(y-y_1)^2]} dx_1 dy_1 \qquad (2\text{-}2)$$
$$= e^{j\frac{\pi}{\lambda z}(x^2+y^2)} \iint R_i^*(x_1, y_1) S_i(x_1, y_1) e^{j\frac{\pi}{\lambda z}(x_1^2+y_1^2)} e^{-j\frac{2\pi}{\lambda z}(xx_1+yy_1)} dx_1 dy_1$$

We only need to multiply the reconstructed field by a chirped phase term and then perform the Fourier Transform. The presence of the reference Ri* merely shifts the position of the

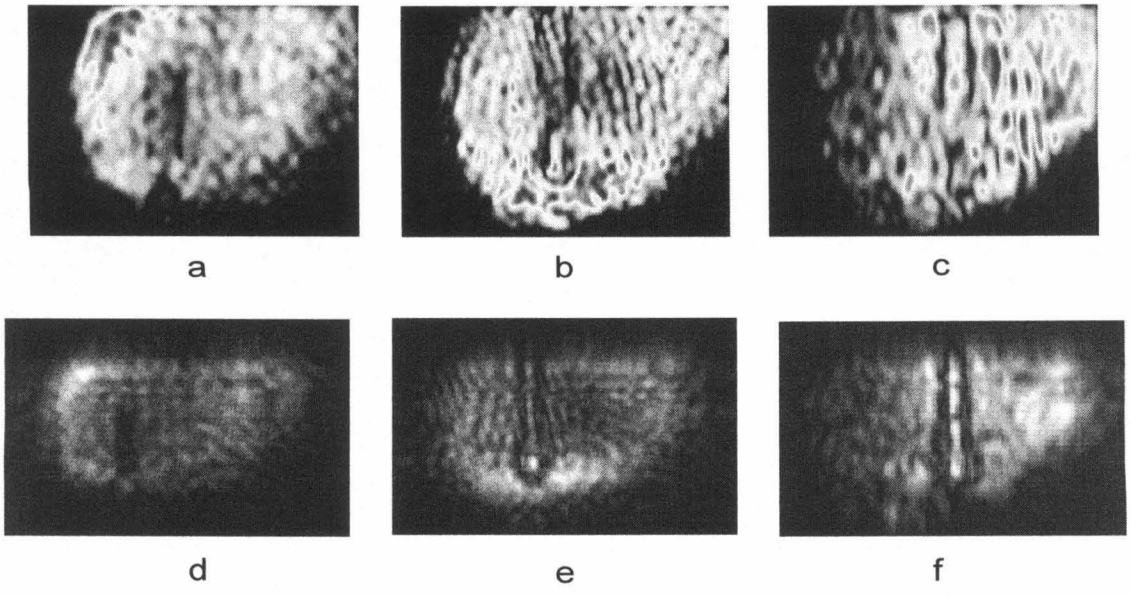


Fig. 2-20. Digital reconstruction with and without diffraction compensation

image slightly from the center. The diffraction can be compensated by using a negative z for $R_i^* S_i$ or a positive z for $R_i S_i^*$ term. The three frames after the diffraction compensation are shown in Figure 2-20def. In the first frame (d) the plasma created by the optical break-down in air is shown, while in the last two frames (e and f) a shock wave is formed and propagates outward.

This method can also be extended to the femtosecond regime if a femtosecond pulsed laser is used. A typical Ti:Sapphire mode locked laser has pulse energy of a few nJ. Let us consider recording a movie using a single femtosecond pulse with 1nJ energy and beam size of $1 cm^2$. We estimate that if the pulse uniformly illuminates the CCD, then each pixel has about 4000 photons. Since we only need several hundred photons per pixel for the recording of a single hologram, we expect that up to ten frames can be recorded with a single 1nJ pulse. An optical amplifier is required for longer movies.

In the nanosecond system we used two cavities to generate the reference and signal pules since the pulse is long (about 1.8 meters spatially). In the femtosecond regime since the pulse is very short (tens of microns spatially) we can simply use delay lines with thickness of a few hundreds of microns. Figure 2-21 shows the reference and signal pulse gen-

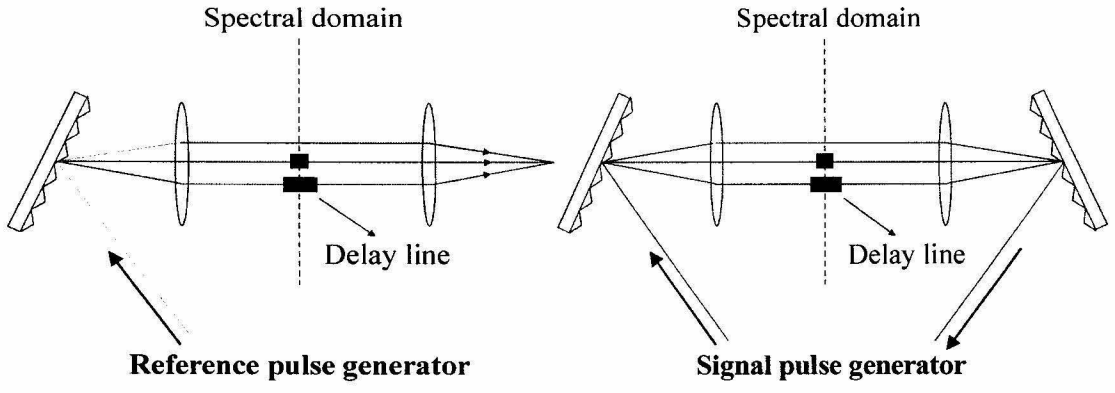


Fig. 2-21. Femtosecond reference and signal pulse generators

erators [2-28][2-29]. An incident femtosecond pulse is first angularly dispersed by a grating. Its different spectral component travels along different direction and is focused on different locations at the back focal plane of the first lens. Hence the focal plane is essentially the spectral domain of the pulse. We divide the spectrum into N parts and delay them by different amount using thin delay lines (e.g., thin glass plates). The two lenses form a 4-F imaging system. A sequence of pulses can then be generated and each pulse is broadened roughly by a factor of N since it has only 1/N of the original spectrum width. The pulse separation is determined by the delay lines. Each pulse is slightly convergent and travels along different angles. This set of pulses can be used as reference pulses. If another grating is put symmetrically at the back focal plane of the second lens (image plane), it can recollimate

the beam and all the pulses travel in the same direction. This set of pulses can then be used as signal pulses.

As an example, we consider a femtosecond pulse given by

$$f(t) = e^{-\frac{t^2}{\tau^2}} e^{-j\omega_0 t} \text{ where } \omega_0 = 2\pi\nu_0 = \frac{2\pi c}{\lambda_0} \tag{2-3}$$

We numerically simulate the generation of 3 and 5 signal pulses shown in Figure 2-22. The

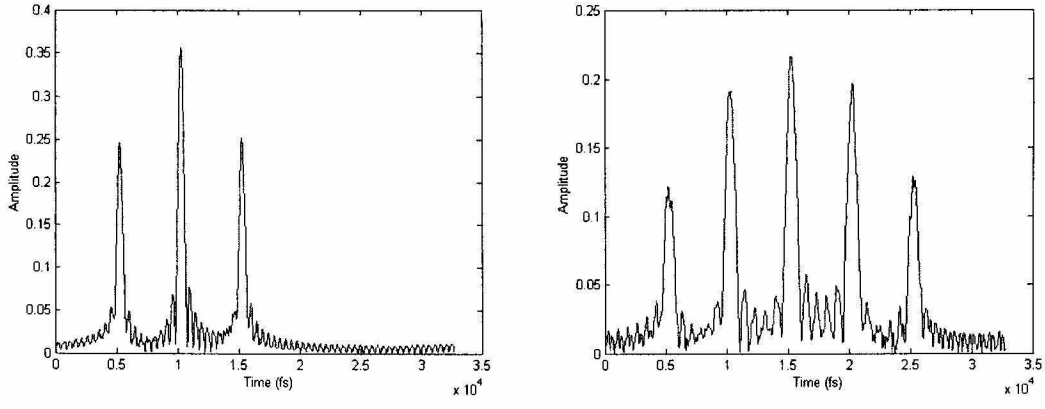


Fig. 2-22. Signal pulse train

central wavelength λ_0 is chosen to be at 750nm and τ is 100 fs. We uniformly divide the spectrum $\left(\nu_0 - \frac{1}{\pi\tau}, \nu_0 + \frac{1}{\pi\tau}\right)$ into 3 or 5 parts with the relative delay between neighboring pulses set to 5ps. The pulse width increases as we generate more signal pulses. There are also small ripples between pulses.

Let us now consider the angle separation of the reference pulses. As shown in Figure 2-23, the grating equation is given by [2-30]

$$K = \frac{2\pi}{\lambda}(\sin\alpha - \sin\beta) \tag{2-4}$$

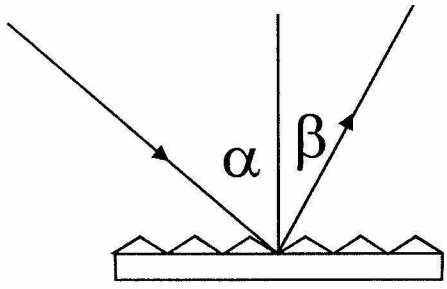


Fig. 2-23. Reflection grating

which can be derived from the momentum conservation in the grating plane. K is the grating vector and is given by

$$K = m\frac{2\pi}{\Lambda} \tag{2-5}$$

where m is the diffraction order and Λ is the period of the grating. For first order diffraction, Equation 2-4 can be rewritten as

$$\frac{\lambda}{\Lambda} = \sin\alpha - \sin\beta \tag{2-6}$$

The angular dispersion is given by

$$D = \left|\frac{d\beta}{d\lambda}\right| = \left|\frac{1}{\Lambda\cos\beta}\right| \tag{2-7}$$

The angular separation between N reference pulses is $\delta\beta = \frac{\Delta\lambda}{N\Lambda\cos\beta}$, where $\Delta\lambda$ is the wavelength bandwidth of the original femtosecond pulse. For a 100 fs pulse the bandwidth $\Delta\lambda$ is about 2–3 nm. $\frac{1}{\Lambda}$ is typically 1200 mm^{-1}. This gives us $\delta\beta \approx 0.03°$ for N=10 and $\beta = 60°$. Bigger angle separation can be achieved if β approaches to 90 degrees.

Carrier multiplexing can also be extended to incoherent imaging system, which may reduce the speckle noise in the recorded frames. In Figure 2-24a, an incoherent image with intensity I(x,y) and two laser pulses are incident on a CCD camera simultaneously.

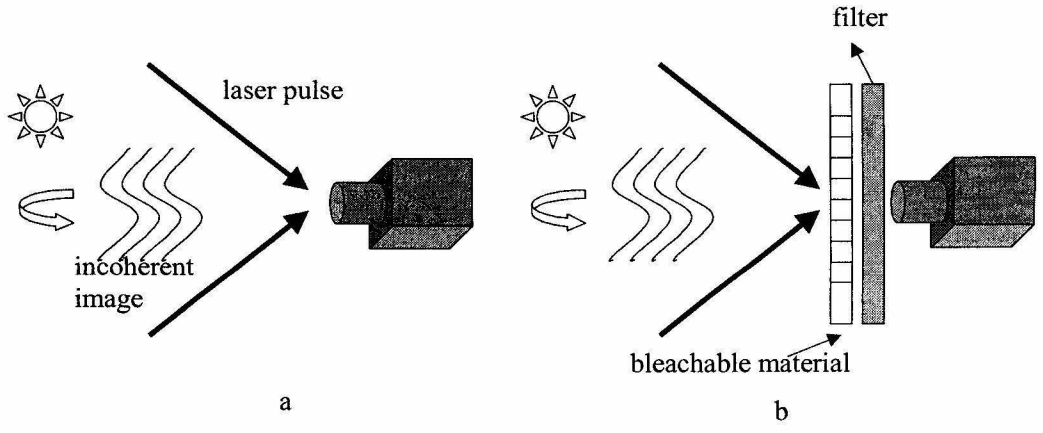


Fig. 2-24. Incoherent carrier multiplexing

The two laser pulses interfere and produce an intensity pattern $I_0\cos(qx)$. If the CCD has some sort of instantaneous intensity detection nonlinearity $D[I_0\cos(qx)+I(x,y)]$, we can get a cross term $I(x,y)\cos(qx)$. We can again record fast events by changing the carrier frequency q with time. A possible implementation is given in b, where we illuminate two pulses on a bleachable material and form an instantaneous absorption grating cos(qx). The incoherent image I(x,y) is then modulated after transmitting the material. A filter is used to filter out the laser pulse. Again fast movies can be recorded by varying the carrier frequency.

2.4 Design of femtosecond holographic movie camera

Recently a single femtosecond pulse recording of an image hologram [2-31] is reported. In order to record multiple frames, the sensitivity of the holographic medium, the generation of the signal and reference beams, the multiplexing methods and the short coherence length of the pulsed laser need to be carefully examined [2-32][2-33]. In this section we will discuss some of these issues.

2.4.1 Angular multiplexing in LiNbO3

The first scenario we will consider is angular multiplexing. We can directly scale the nanosecond system down using femtosecond laser pulses and shorter cavity length (or delaylines). The major uncertainty here is the recording medium. If the image bandwidth is low, CCD cameras can be used as described in the previous section. In the following, we investigate recording femtosecond pulsed holograms in $LiNbO_3$ doped with Ce. The experimental setup is shown in Figure 2-25. The Ti:Sapphire mode locked laser is tuned to

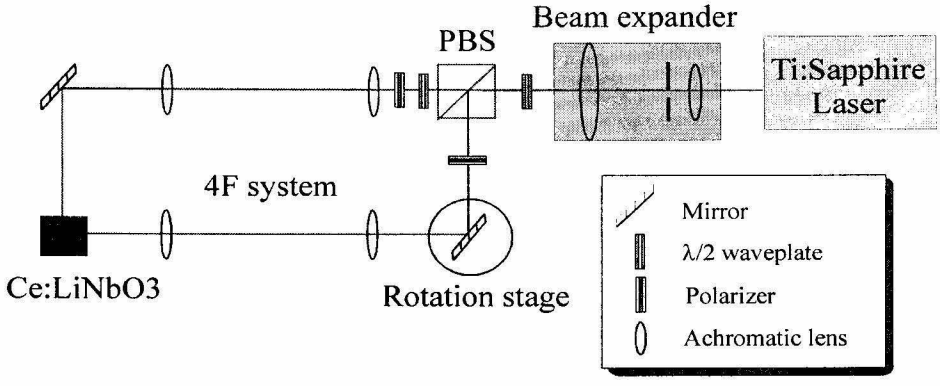


Fig. 2-25. Experimental setup of femtosecond pulsed recording in LiNbO3

740nm. The intensity of the reference and signal beam is equalized and is about 117mW/cm^2. The recording and erasure curves are shown in Figure 2-26. Due to the short coherence length of femtosecond pulses, fanning cannot build up and we get a very clean recording curve which goes all the way to saturation. The grating strength A(t) is defined as the square root of the diffraction efficiency $A(t) = \sqrt{\eta(t)}$.

During recording, the grating strength $A(t) = A_0\left[1 - \exp\left(-\frac{t}{\tau_r}\right)\right]$ and during readout $A(t) = A_0 \exp\left(-\frac{t}{\tau_e}\right)$. The recording time constant τ_r is 4.8326 hours and the erasure time constant τ_e is about 5.1478 hours. We can calculate the M/# and sensitivity S

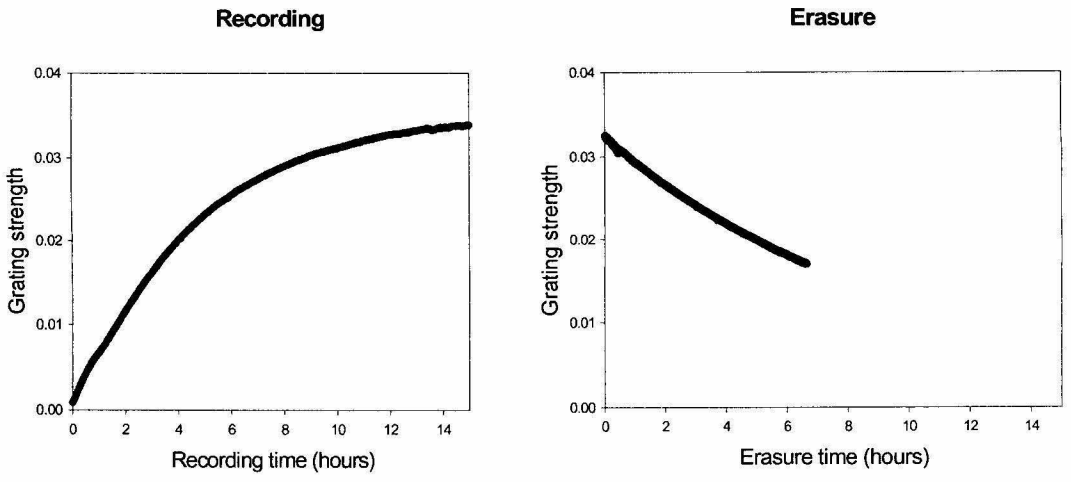


Fig. 2-26. Recording and erasure curves

using $M\# = A_0\frac{\tau_e}{\tau_r}$ and $S = \frac{A_0}{IL\tau_r}$ [2-16][2-34]. τ_r and τ_e are equal unless the recording and erasure dynamics are characterized by a complex time constant, and in that case we should see oscillations in the recording curve. The difference of τ_r and τ_e here is probably due to the short erasing time and error in fitting the erasure curve. We obtain M/# of 0.036 and S of 8.6×10^{-6} which are rather small.

Due to the short coherence length of the laser pulses, the hologram is recorded only in a thin slice of the material along the diagonal direction as shown in Figure 2-27. We assume that the laser pulse is given by $E(t)=f(t)e^{-j\omega t}$. The exposure energy P(x,y) is given as

$$P(x,y) = \int\left|E\left(t-\frac{x}{c}\right)+E\left(t-\frac{y}{c}\right)\right|^2 dt = \int\left|f\left(t-\frac{x}{c}\right)e^{j\frac{\omega}{c}x}+f\left(t-\frac{y}{c}\right)e^{j\frac{\omega}{c}y}\right|^2 dt \qquad (2\text{-}8)$$

$$= \int 2|f(t)|^2 dt + \left[e^{j\frac{\omega}{c}(x-y)}\int f\left(t-\frac{x}{c}\right)f^*\left(t-\frac{y}{c}\right)dt + cc\right]$$

The index grating recorded is given by

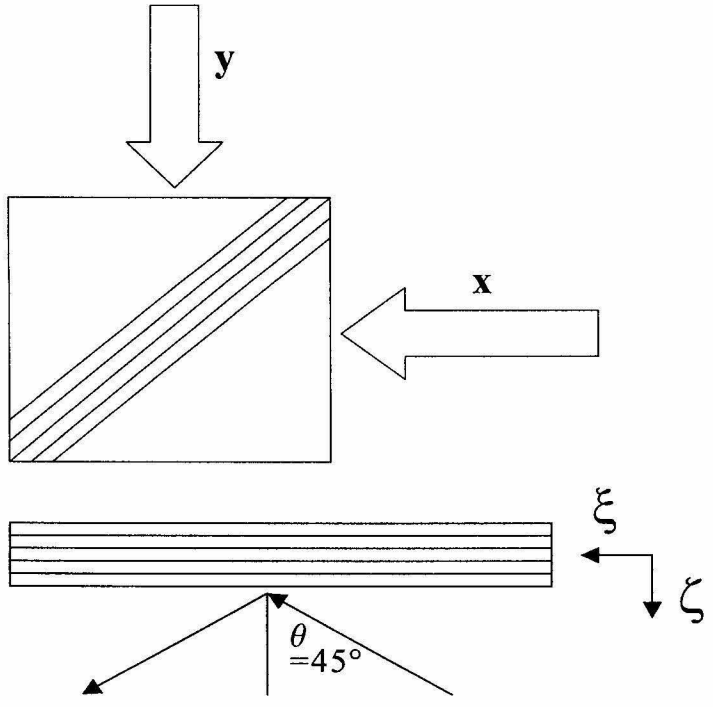


Fig. 2-27. Femtosecond pulsed hologram

$$\Delta n(x, y) \propto e^{j\frac{\omega}{c}(x-y)} \int f\left(t-\frac{x}{c}\right) f^{*}\left(t-\frac{y}{c}\right) dt + cc = e^{j\frac{\omega}{c}(x-y)} AC\left(\frac{x-y}{c}\right) + cc \qquad (2\text{-}9)$$

AC(t) is the autocorrelation function of f(t) and the thickness of the hologram is determined by the coherence length (correlation length) of the laser pulses instead of the thickness of the recording medium. Let's assume a Gaussian pulse.

$$f(t) \propto \exp\left[-\frac{t^2}{\tau^2}\right] \qquad (2\text{-}10)$$

$$AC(t) \propto \exp\left[-\frac{t^2}{2\tau^2}\right]$$

$$\Delta n(x, y) \propto e^{j\frac{\omega}{c}(x-y)} AC\left(\frac{x-y}{c}\right) + cc = e^{j\sqrt{2}\frac{\omega}{c}\zeta} AC\left(\sqrt{2}\frac{\zeta}{c}\right) + cc$$

$$\propto e^{j\sqrt{2}\frac{\omega}{c}\zeta} \exp\left[-\frac{\zeta^2}{c^2\tau^2}\right] + cc$$

where we rotate the (x, y) coordinate system 45 degrees to (ξ, ζ) system as shown in Figure 2-27.

$$\xi = \frac{x+y}{\sqrt{2}}, \zeta = \frac{x-y}{\sqrt{2}} \tag{2-11}$$

Using Born approximation and assuming the hologram extends to infinity along the ξ direction, we can calculate the angular selectivity function.

$$S(\delta\theta) \propto \int \exp\left[-\frac{\zeta^2}{c^2\tau^2}\right] \exp\left[j2\pi \frac{2\sin\theta\delta\theta}{\lambda}\zeta\right] d\zeta \propto \exp\left[-\left(\frac{\delta\theta}{\Delta}\right)^2\right] \tag{2-12}$$

where $\delta\theta$ is the angle change inside the material

$\Delta = \frac{\lambda}{\sqrt{2}\pi c\tau}$, λ&c are wavelength and speed of light in the material

For external angle selectivity, we only need to replace Δ with $n\Delta$ where n is the refractive index of the material.

After recording the hologram, we first use femtosecond pulses to read out the hologram and then tune the Ti:Sapphire laser to cw operation mode (at the same center wavelength) and read out the hologram with a cw beam. Figure 2-28 shows the angular selectivity curves. The selectivity curve of the pulsed readout is broader because different spectral component of the pulse Bragg-matches the hologram at slightly different angle. We measured the autocorrelation of the pulse using an autocorrelator. Figure 2-29 shows the measured intensity correlation $IC(t) = \int |f(t')|^2 |f(t'+t)|^2 dt'$ (equal to $|AC(t)|^2$ for a Gaussian pulse) and the fitted curve. The measured pulse width (τ) is 191 fs. The calculated selectivity is shown in Figure 2-30. They agree with each other very well. The selectivity ($n\Delta$) is about 0.37 degree.

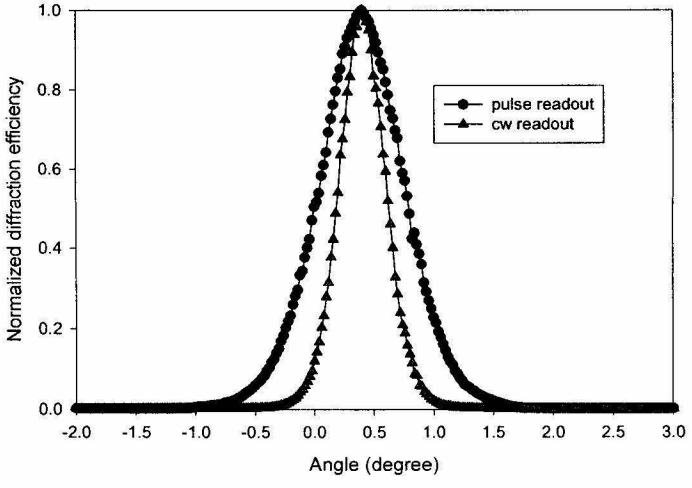


Fig. 2-28. Selectivity curves

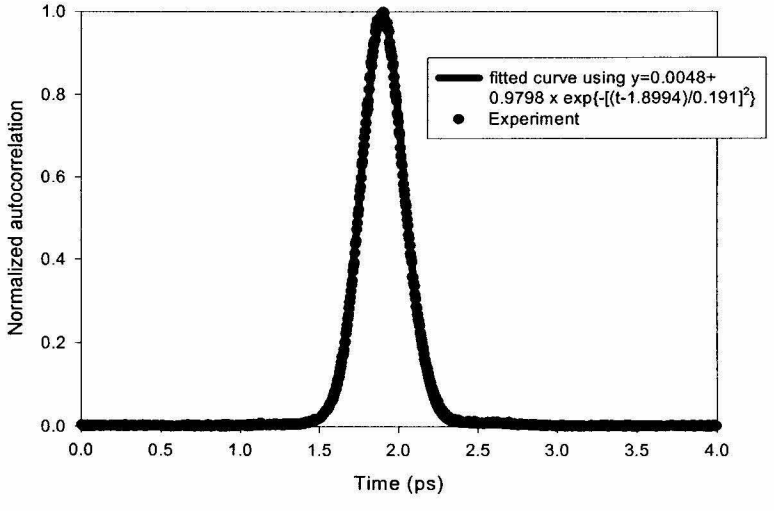


Fig. 2-29. Autocorrelation of the femtosecond pulse

We also measured the selectivity using a He-Ne laser and the result is shown in Figure 2-31. It also gives us a selectivity of about 0.4 degree.

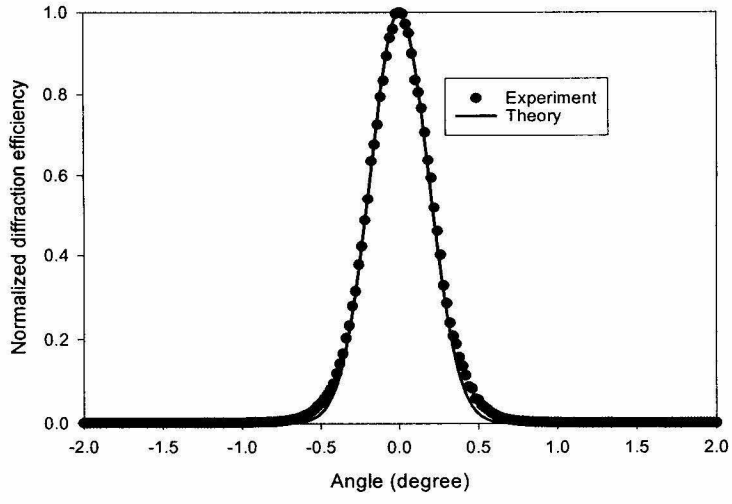


Fig. 2-30. Comparison between theoretical and experimental selectivity

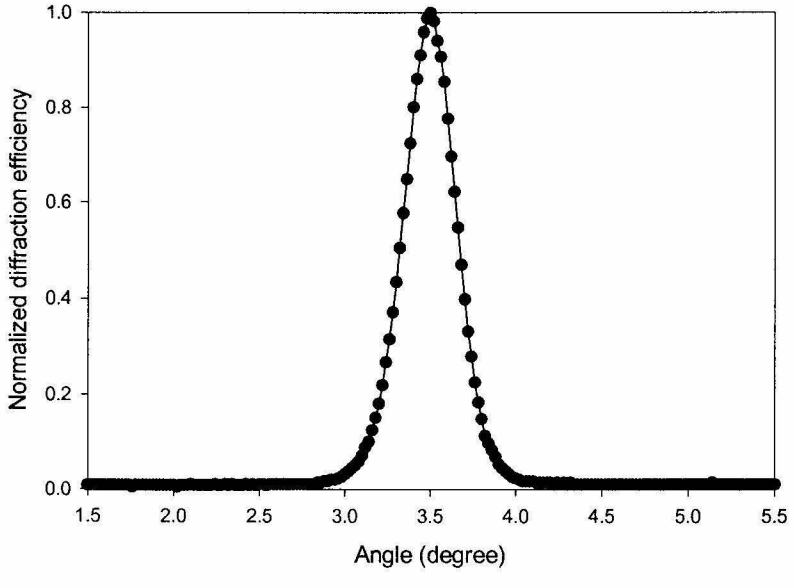


Fig. 2-31. Selectivity measured using He-Ne laser

The experiments in Ce:$LiNbO_3$ show that holograms is recorded only in a thin slice of the material due to the short coherence length of the femtosecond pulse. The angular

selectivity is determined by the laser pulse width. M/# and S are very small. An exposure energy of $1mJ/cm^2$ can only yield a diffraction efficiency of 10^{-6} even if we assume that S=1cm/J is achievable in $LiNbO_3$. It would be very difficult to record a movie using a single femtosecond pulse from the laser.

2.4.2 Wavelength multiplexing

Femtosecond pulse has spectrum width of several nm and it makes wavelength multiplexing possible. Shown in Figure 2-32 is a chirped pulse recording system. A femtosec-

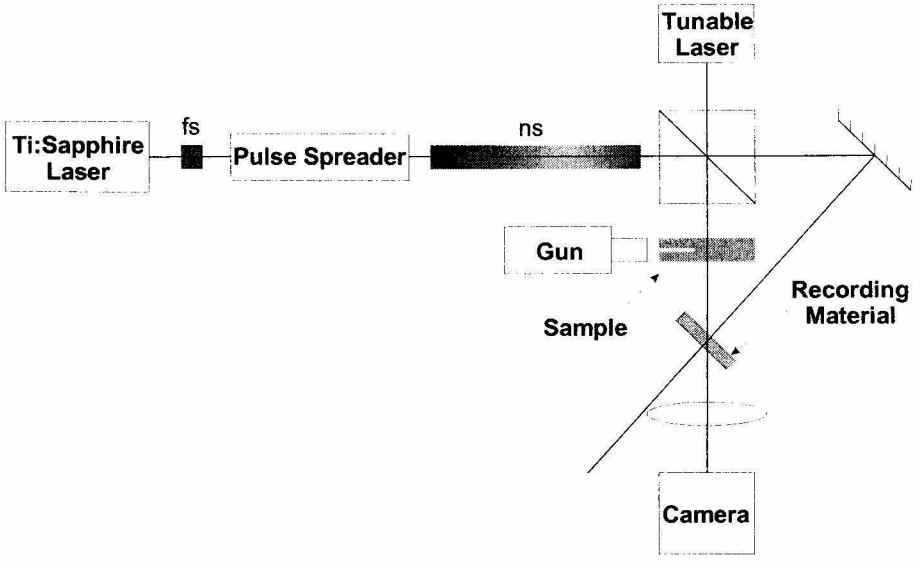


Fig. 2-32. Recording with chirped pulses (courtesy of Dr. Gregory J. Steckman)

ond pulse is chirped and spread to a few nanoseconds. At different time slots holograms are recorded at different wavelength. After the recording, we use a cw beam to read out the holograms. By tuning the wavelength of the cw laser, frames recorded at different times can be played back. A sequence of pulses with different center wavelengths can also be used as described in Section 2.3.

When M holograms are superimposed, it is well known that the diffraction efficiency of each hologram follows the M/# rule $\eta = \left(\frac{M/\#}{M}\right)^2$. M/# is approximately linear in the material thickness M/#=(m/#)L where m/# is the per unit length. The wavelength selectivity of the material is $\delta\lambda \sim \frac{\lambda^2}{L}$. The spectrum of the pulse is $\Delta\lambda$. Since the material cannot resolve wavelength difference smaller than $\delta\lambda$, $M \approx \frac{\Delta\lambda}{\delta\lambda}$. When we readout the pulsed hologram with a cw beam,

$$\eta = \left(\frac{m/\#L}{\frac{\Delta\lambda}{\delta\lambda}}\right)^2 = \left(\frac{m/\#\lambda^2}{\Delta\lambda}\right)^2 = (m/\#L_w)^2 \qquad (2\text{-}13)$$

where $L_w = \frac{\lambda^2}{\Delta\lambda}$ is the coherence length of the pulse. In the case of a pulsed hologram recorded with two femtosecond pulses, this is consistent with the fact that the effective thickness of the hologram is about L_w.

2.4.3 Frequency multiplexing in spectral hole burning medium

The state of the art amplifier system can generate a single femtosecond pulse with energy of mJs. In order to angle or wavelength multiplex holograms with a single femtosecond pulse, the recording medium must have high enough sensitivity. In this section, we will discuss a femtosecond movie system using spectral hole burning medium [2-35][2-36].

Spectral hole burning medium consists of some dopants in a host, for example, polymer matrix or crystal. The dopants absorb incident photons and get excited. Due to the finite lifetime of the excited state, it has homogeneously broadened linewidth which typically ranges from a few MHz to a few GHz in cryogenic temperature (zero-phonon-line). When excited, the dopant can undergo some chemical reactions and arrive in another ground state and thus the material gets bleached. If the material is illuminated with two cw

beams, interference fringes form inside. The material gets bleached in the bright region and an absorption grating is recorded. An interesting feature of spectral hole burning medium is that it also has inhomogeneously broadened linewidth which is typically a few THz. This is because the local environment of the dopant differs from dopant to dopant and it affects the transition frequency as shown in Figure 2-33. We can think of the spectral hole burning

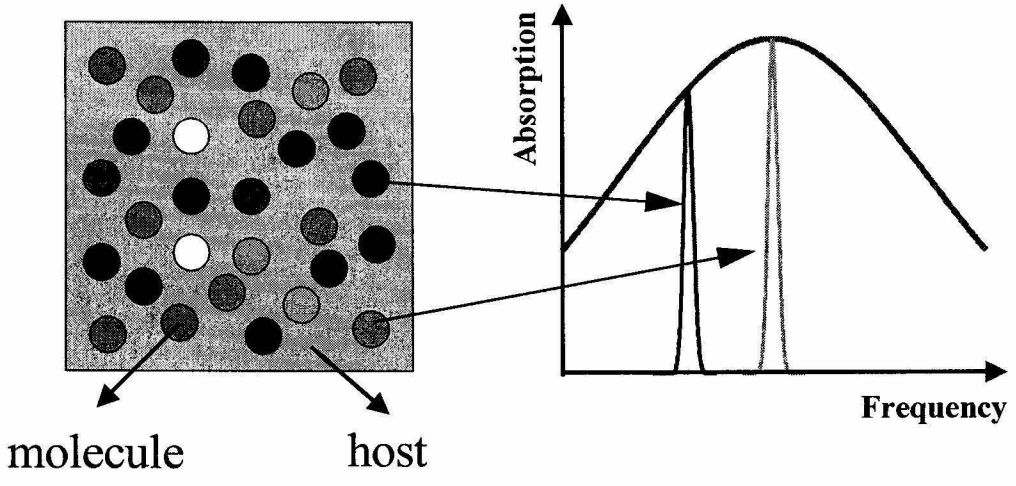


Fig. 2-33. Spectral hole burning medium

medium as a superposition of many different types of media and each type of medium only absorbs photons of certain frequency. We can record holograms in different medium simply by changing the frequency of the laser beams and this is completely different from wavelength multiplexing where we use the Bragg-selectivity of volume holograms. The frequency selectivity of spectral hole burning media is material thickness independent and is determined by the homogeneous linewidth. When a hologram is recorded using femto-second pulses, different spectral component (cw component) records a hologram in differ-ent types of medium. We can record a hologram even if the signal and reference pulses do

not meet inside the SHB medium since the spectral components of the two pulses can still overlap both spatially and temporally. The problem of short coherence length of femtosecond pulses is also solved.

Figure 2-34 shows the femtosecond movie camera system using SHB medium. As

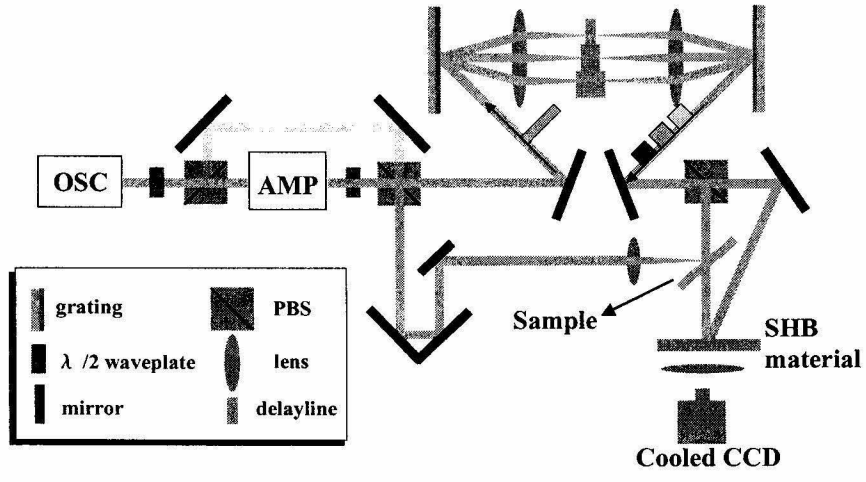


Fig. 2-34. Femtosecond movie camera

described in Section 2.3, multiple pulses can be generated by a single femtosecond pulse from the regenerative amplifier. Since each pulse has a different portion of the original spectrum we can use this sequence of pulses to record femtosecond movie in spectral hole burning medium. After recording, we use the low energy femtosecond pulses from the oscillator to read out the hologram. The pulses from the oscillator go through the same pulse shaping setup and individual frame can be read out by blocking all but one of the delaylines.

In the following, we characterize a specific spectral hole burning medium synthesized in our lab, H_2TBNP in polyvinyl butyral (PVB) with a concentration of 3×10^{-5}mol/l [2-37]. The thickness of the sample is about 400 μm. We used a cw laser diode (λ=790nm)

and measured the M/# in one frequency channel [2-38]–[2-40]. The results can be easily extended to other frequency channels. Figure 2-35 is the experimental setup. The sample is

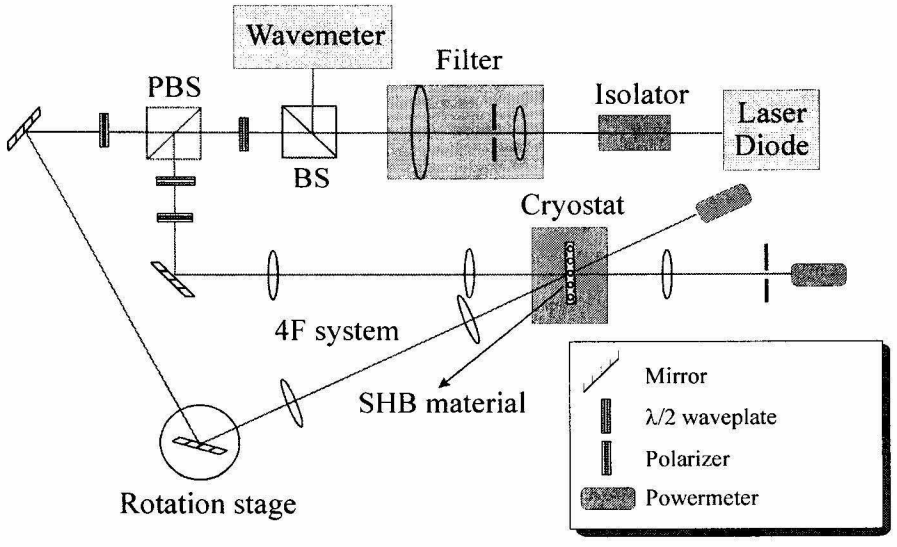


Fig. 2-35. Spectral hole burning holography experimental setup

put in a cryostat immersed with liquid Helium. Figure 2-36 shows the absorption kinetics

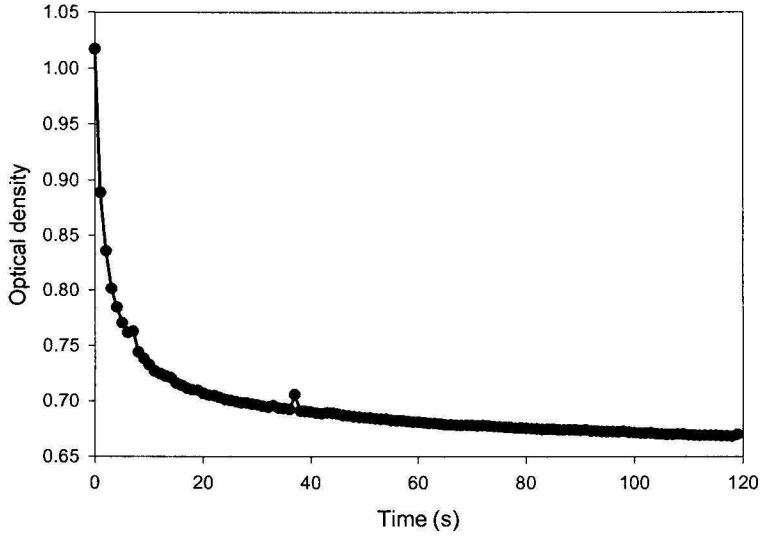
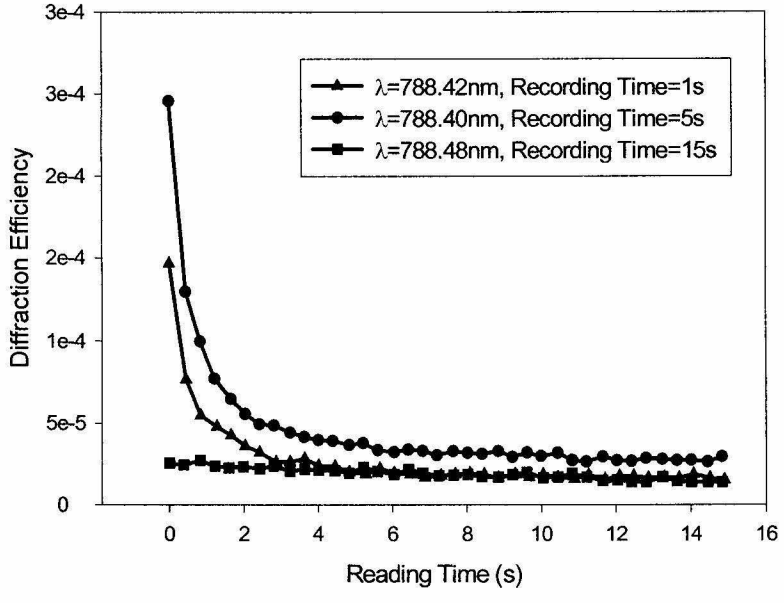


Fig. 2-36. Bleaching kinetics

measured at 2 Kelvin. The bleaching beam intensity is about 10 $\mu W/cm^2$. The intensity transmission coefficient satisfies T=exp(-2αd) where α is the absorption coefficient and d is the thickness of the material. We define the optical density as $OD = \alpha d = -0.5 \ln T$. Single plane wave hologram is recorded at different wavelengths by tuning the temperature of the laser diode with both the reference and the signal intensity of 10 $\mu W/cm^2$. Figure 2-37 shows the reading curves of the single plane wave hologram with the unatten-



p

Fig. 2-37. Reading curves

uated original reference beam. A diffraction efficiency of 10^{-4} can be consistently obtained. The optimal recording time lies between 5-8 seconds. If we record for a longer amount of time the material gets bleached everywhere (intensity modulation index <1). Due to the oscillation of the cryostat we have difficulty to get consistent grating growth curves. Figure 2-38 shows the measured selectivity curve without using the iris filter in front of the

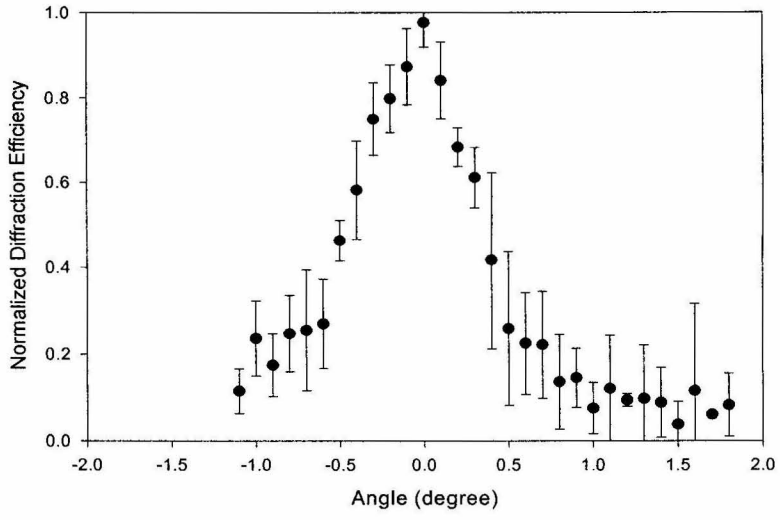


Fig. 2-38. Angular selectivity (without iris filter)

power meter (see Figure 2-35). The error bars stand for the standard deviation. The measured selectivity is about 0.6 degree. However, the data is very noisy due to the scattering from the sample. An iris filter is then used at the back focal plane of the Fourier lens after

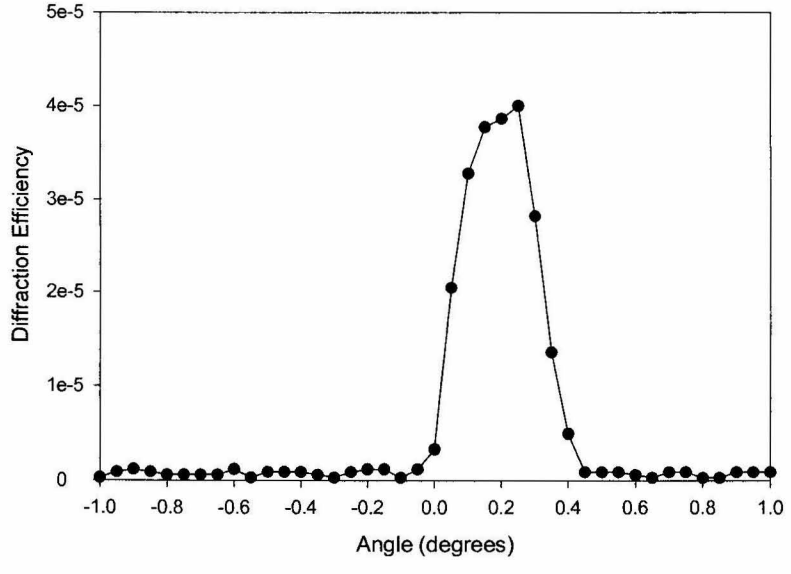


Fig. 2-39. Angular selectivity (with iris filter)

the sample. This serves for two purposes. One is to reduce the scattering and background noise, and the other is to further reduce the selectivity. For thin material, if the reference beam angle is changed slightly the reconstructed signal beam angle is changed accordingly and can get blocked by the iris filter. The filter gives us a combined effect of both angular

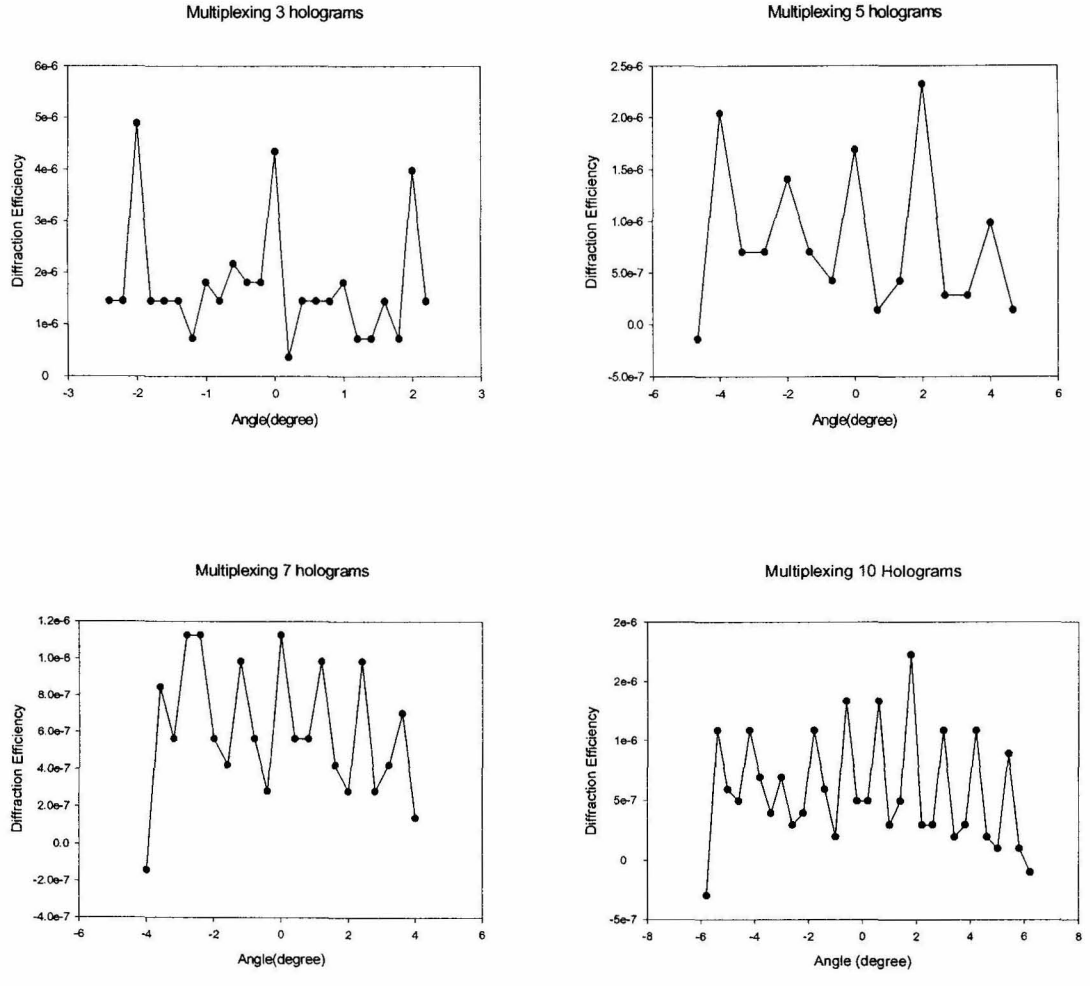


Fig. 2-40. Angularly multiplexing plane wave holograms

and peristrophic selectivity. The selectivity curve with the iris filter is shown in Figure 2-39. A selectivity of about 0.2 degree is obtained, which is consistent with the iris diameter of about 1mm and the Fourier lens focal length of about 24cm. Three, five, seven and ten plane wave holograms are then multiplexed. After the recording, we scan the reference beam angle and measure the dependence of the diffraction efficiency on the reference angle. The obtained comb functions are shown in Figure 2-40. Because the readout process is destructive, a very coarse scanning is used. We use equal time (1s-2s) exposure schedule. Figure 2-41 shows the measured M/# of each case. As we multiplex 3, 5, 7 and

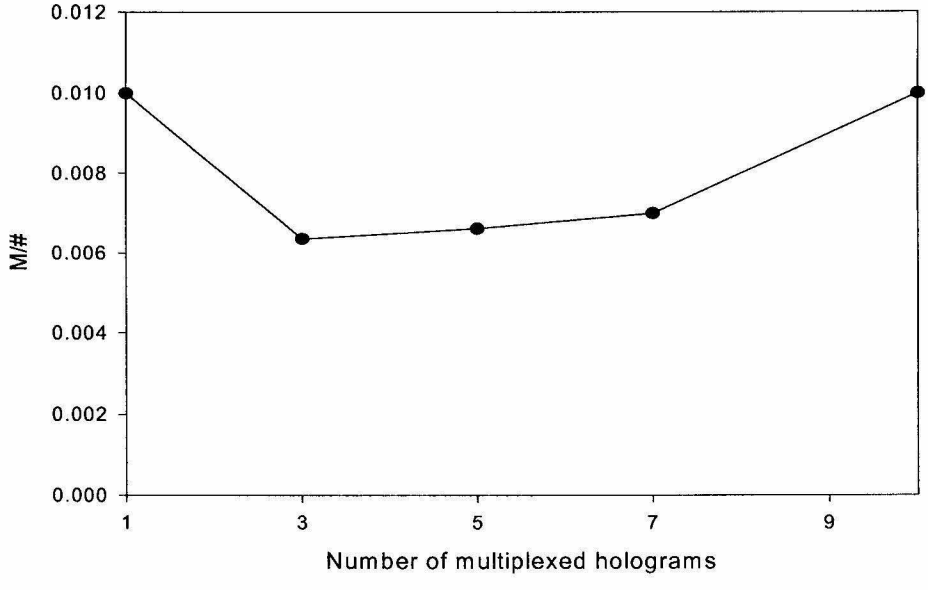


Fig. 2-41. M/# measurement

10 holograms, the M/# has a trend to increase. This is because the total exposure energy increases and we better use the dynamic range of the material. The measured M/# is about 0.01.

From our experiments, about 10μJ/cm^2 of exposure energy can yield a diffraction efficiency of 10^{-4} in one frequency channel. The spectral hole burning medium has high enough sensitivity to record tens of frames using a single pulse with energy of mJs from the laser source if it has similar sensitivity during pulsed recording.

2.5 Conclusion

We demonstrate a nanosecond movie camera system by recording laser induced shock wave evolution. Five frames are recorded and the time resolution is 5.9 ns. Two specially designed cavities are used to generate the reference and signal pulses and angular multiplexing is used to resolve holograms recorded successively. The movies we recorded are among the fastest ever recorded in 2D, comparable with the current state of the art of multi-camera systems (8 frames and 10 ns resolution). Holography provides the added advantages of recording 3-D and phase information. The sensitivity of holographic recording medium is an important issue in order to use the technique in shorter laser pulses or at different wavelengths. We propose and demonstrate carrier multiplexing in commercially available CCD cameras to alleviate the problem if the image bandwidth is low. We also investigate several scenarios to record femtosecond movies and frequency multiplexing in spectral hole burning medium seems to be a good candidate.

References

[2-1] Hyotcherl Ihee, Vladimir A. Lobastov, Udo M. Gomez, Boyd M. Goodson, Ramesh Srinivasan, Chong-Yu Ruan, and Ahmed H. Zewail. "Direct Imaging of Transient Molecular Structures with Ultrafast Diffraction." Science 291: 458-462 (2001)

[2-2] A. J. Rosakis, O. Samudrala, and D. Coker. "Cracks Faster than the Shear Wave Speed." Science 284, 1337-1340 (1999)

[2-3] See Imacon 468, for example, http://www.drs.com/

[2-4] D. Gabor. "A new microscopic principle." Nature 161, 777 (1948)

[2-5] L.O. Heflinger, R.F. Wuerker, and R.E.Brooks. "Holographic interferometry." Journal of Applied Physics 37, 642 (1966)

[2-6] T. Tschudi, C. Yamanaka, T. Sasaki, K. Yoshida, and K. Tanaka. "A study a high-power laser effects in dielectrics using multiframe picosecond holography." Journal of Physics D: Applied Physics Vol.11 177-180 (1978)

[2-7] Michael J. Ehrlich, J. Scott Steckenrider, and James W. Wagner. "System for high-speed time-resolved holography of transient events." Appl. Opt. 31, 5947-5951 (1992)

[2-8] W. Hentschel and W. Lauterborn. "High speed holographic movie camera." Optical Engineering 24, 687-691 (1985)

[2-9] P.J. Van Heerden. "A new optical method of storing and retrieving information." Appl. Opt. 2, 387-392 (1963)

[2-10] Shinichi Suzuki, Yasunori Nozaki, and Hiroshi Kimura. "High-speed holographic microscopy for fast-propagating cracks in transparent materials." Appl. Opt. 36, 7224-7233 (1997)

[2-11] R. G. Racca and J. M. Dewey. "Time smear effects in spatial frequency multiplexed holography." Appl. Opt. 28, 3652-3656 (1989)

[2-12] R. G. Racca and J. M. Dewey. "High-speed time-resolved holographic interferometer using solid-state shutters." Optics and Laser Technology 22, 199-204 (1990)

[2-13] Gregory J. Steckman. Caltech Ph.D. thesis (2001)

[2-14] Zhiwen Liu, Gregory J. Steckman, and Demetri Psaltis. "Holographic recording of fast phenomena." Appl. Phys. Lett. 80, 731 (2002)

[2-15] H.J. Coufal, D. Psaltis, and G.T. Sincerbox (Eds.). Holographic Data Storage, 171-197 (Springer, 2000)

[2-16] Fai H. Mok, Geoffrey W. Burr, and Demetri Psaltis. "System metric for holographic memory systems." Optics Letters 21, 896-898 (1996)

[2-17] Yu. P. Raizer. Laser-induced Discharge Phenomena (Consultants Bureau, New York, 1977)

[2-18] Roger M. Wood. Laser Damage in Optical Materials (Adam Hilger, Bristol and Boston, 1986)

[2-19] Y. R. Shen. The Principles of Nonlinear Optics (John Wiley & Sons, 1984)

[2-20] Hugo Sobral, Mayo Villagran-Muniz, Rafael Navarro-Gonzalez, and Alejandro C. Raga. "Temporal evolution of the shock wave and hot core air in laser induced plasma." Appl. Phys. Lett. 77, 3158-3160 (2000)

[2-21] Y. Tomita, M. Tsubota, K. Nagane, and N. An-naka. "Behavior of laser-induced cavitation bubbles in liquid nitrogen." Journal of Applied Physics Vol. 88 No. 10 5993-6001 (2000)

[2-22] E. Cuche, F. Bevilacqua, and C. Depeursinge. "Digital holography for quantitative phase-contrast imaging." Opt. Lett. 24, 291-293 (1999)

[2-23] G. Indebetouw and P. Klysubun. "Space-time digital holography: A three-dimensional microscopic imaging scheme with an arbitrary degree of spatial coherence." Appl. Phys. Lett. 75, 2017-2019 (1999)

[2-24] S. Schedin, G. Pedrini, H.J. Tiziani, and F.M. Santoyo. "Simultaneous three-dimensional dynamic deformation measurements with pulsed digital holography." Appl. Opt. 38, 7056-7062 (1999)

[2-25] S. Schedin, G. Pedrini, H.J. Tiziani, A.K. Aggarwal, and M.E. Gusev. "Highly sensitive pulsed digital holography for built-in defect analysis with a laser excitation." Appl. Opt. 40, 100-103 (2001)

[2-26] Zhiwen Liu, Martin Centurion, George Panotopoulos, John Hong, and Demetri Psaltis. "Holographic recording of fast events on a CCD camera." Opt. Lett. 27, 22-24 (2002)

[2-27] Goodman. Introduction to Fourier Optics (second edition) (McGraw-Hill, New York, 1996)

[2-28] *Discussion with Professor T. W. Hänsch*

[2-29] A. M. Weiner and Ayman M. Kanan. "Femtosecond Pulse Shaping for Synthesis, Processing, and Time-to-Space Conversion of Ultrafast Optical Waveforms." IEEE Journal of Selected Topics in Quantum Electronics 4, 317-331 (1998).

[2-30] Christopher Palmer and Erwin Loewen. Richardson Grating Laboratory's Diffraction Grating Handbook (fourth edition) (http://www.gratinglab.com)

[2-31] Aleksander Rebane, Mikhail Drobizhev, and Christophe Sigel. "Single femtosecond exposure recording of an image hologram by spectral hole burning in an unstable tautomer of a phthalocyanine derivative." Opt. Lett. 25, 1633-1635 (2000)

[2-32] Kazutaka Oba, Pang-Chen Sun, and Yeshaiahu Fainman. "Nonvolatile photorefractive spectral holography." Opt. Lett. 23, 915-917 (1998)

[2-33] Nils Abramson. "Light-in-flight recording: high-speed holographic motion pictures of ultrafast phenomena." Appl. Opt. 22, 215-232 (1983)

[2-34] Ali Adibi. Caltech Ph.D. thesis (2000)

[2-35] B. Plagemann, F. R. Graf, S. B. Altner, A. Renn, and U. P. Wild. "Exploring the limits of optical storage using persistent spectral hole-burning: holographic recording of 12,000 images." Appl. Phys. B 66, 67-74 (1998)

[2-36] S. Bernet, S. B. Altner, F. R. Graf, E. Maniloff, A. Renn, and U. P. Wild. "Frequency and phase swept holograms in spectral hole-burning materials." Appl. Opt. 34, 22, 4674-4684 (1995)

[2-37] A. V. Turukhin, A .A. Gorokhovsky, C. Moser, I. V. Solomatin, and D. Psaltis. "Spectral hole burning in naphthalocyanines derivatives in the region 800 nm for holographic storage applications." J. Lum. 86, 399-405 (2000)

[2-38] Zhiwen Liu, Wenhai Liu, Christopher Moser, David Zhang, Iouri V. Solomatine, Demetri Psaltis, and Anshel Gorokhovsky. "Derivation and measurement of the M/# in spectral hole burning media." SPIE Annual Meeting, San Diego, CA, July 2001

[2-39] *Zhiwen* Liu, Wenhai Liu, Christopher Moser, David Zhang, Iouri V. Solomatine, Demetri Psaltis, and Anshel Gorokhovsky. "Multiplexing and M# measurement in spectral hole burning medium." CLEO/QELS 2001, Baltimore, MD, May 2001

[2-40] Wenhai Liu. Caltech Ph.D. thesis (2001)

3 Diffraction from Subwavelength Structures

3.1 Introduction

Optical techniques have been widely used to store information. The increasing demand for higher storage density is constantly a driving force of research. Volume holographic memory has been of interest to researchers for decades due to its huge potential in storage capacity and high speed accessing rate. Surface storage, such as DVD/CD technology, has become more and more important recently because of its simplicity. The storage density of these techniques is currently limited by the resolution of the light used to read out the information, i.e., the wavelength of the light source. Researchers have been able to encode information in the nanometer scale, but in order to read out the information complicated near field scanning is required. The main topic of this chapter is how to encode information in subwavelength structures such that the difference of the scattering pattern in the far field is drastic enough to be resolved.

In Section 3.2 we discuss the general properties of the diffraction pattern of subwavelength structures. We compare the experimentally measured far field scattering pattern from some trenches with our FDTD simulation results. We investigate several methods to encode information in Section 3.3 and analyze the general properties of such systems using eigenfunction analysis in Section 3.4. In Section 3.5 we discuss the effects of adding a nonlinear buffer layer on top of the layer encoded with information. Finally, In Section 3.7, we give a brief review of the FDTD algorithm used to calculate the scattering field.

3.2 Comparison of FDTD simulation and IBM data

We developed a 2D simulation tool based on the FDTD algorithm which is briefly reviewed in Section 3.7. In this section, we use the tool to calculate the far field scattering patterns of some subwavelength structures (trenches) and compare them with the experimental data. The fabrication of the trenches and far field scattering measurements were done in IBM [3-1] [3-2]. The trenches were etched in NiP wafer which has a refractive index of 2.07+j2.78 at λ=633nm. They were probed by a focused He-Ne laser beam (λ=633nm, FWHM 5μm). Both the s (the electrical field parallel to the trench axis) and p (the magnetic field parallel to the trench) polarizations were used. For each polarization, two measurements were made with the incident angles (between the incident beam axis and the sample normal) of 5 and 15 degrees. The profiles of the trenches were measured using atomic force microscope (AFM) and we approximate the profiles in our simulation by staircase fittings with a resolution of λ/30. Figure 3-1 to Figure 3-4 show the comparison of the simulation and the measured far field scattering. They agree with each other very well.

A common feature of the scattering patterns is that they have a peak on top of a broad floor. When the probing beam is focused on the trench, it is first reflected by the surface of the wafer. Light is also coupled into the trench and gets reflected by the bottom of the trench. The peak is due to the reflection from the wafer surface while the reflection from the bottom of the trench results in the broad floor because its spatial extension is limited by the width of the trench. If the depth of the trench is close to λ/4, the two reflected beams are out of phase and it produces two sharp dips where the floor meets the peak. We should distinguish these two dips from the dips on the floor which are due to the far field diffraction of the light reflected from the bottom of the trench.

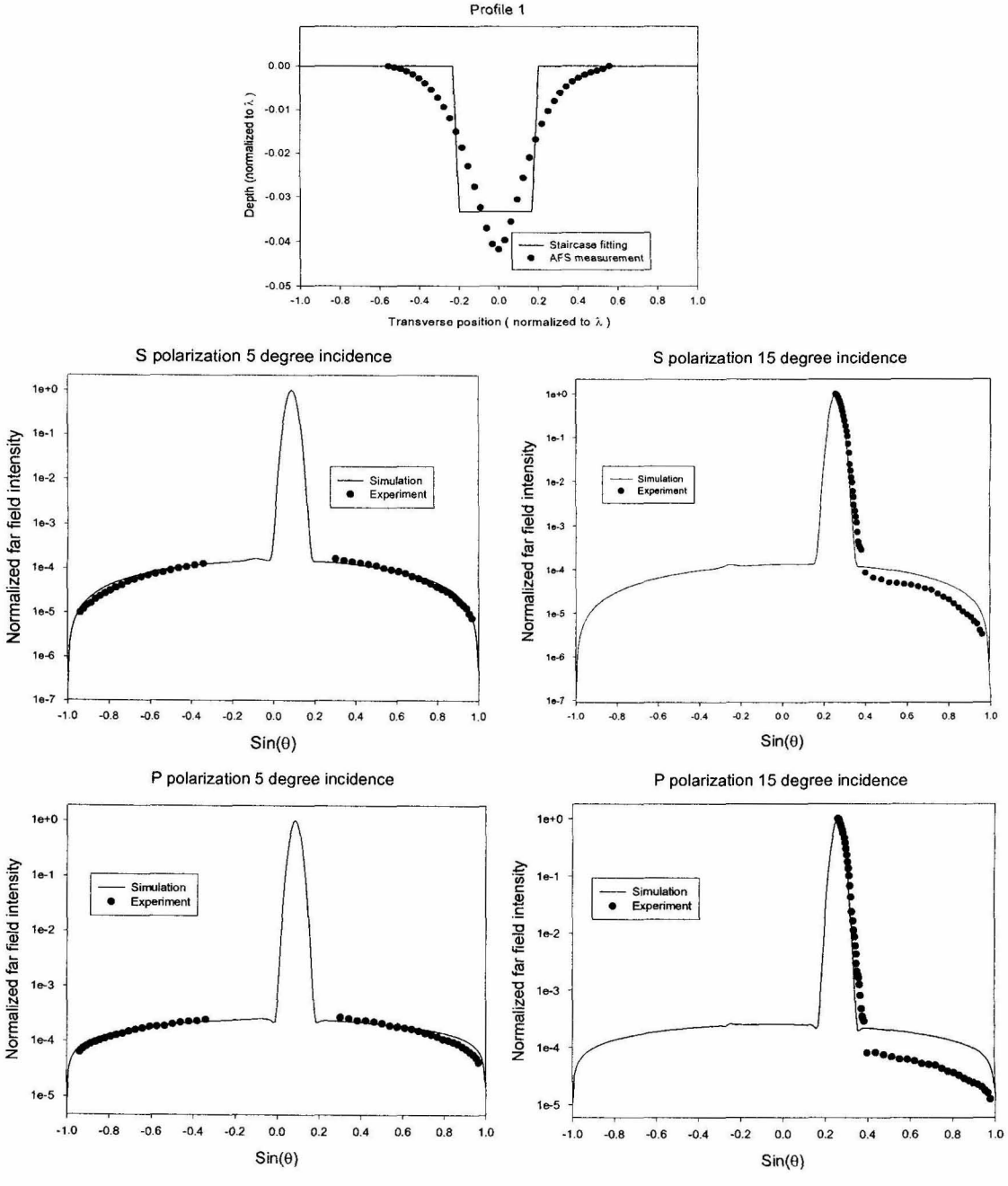


Fig. 3-1. Profile 1 (experimental data from [3-1] [3-2])

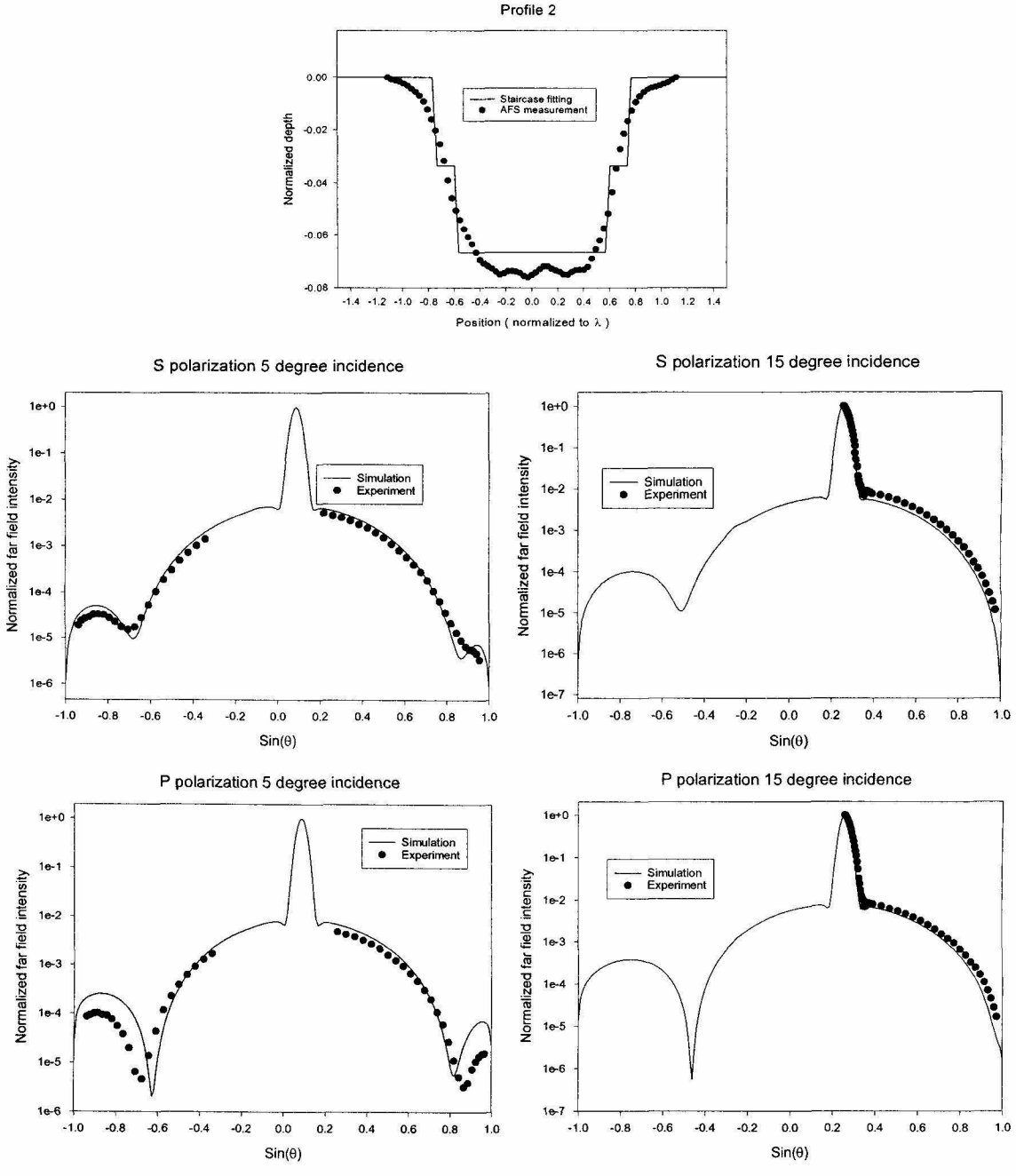


Fig. 3-2. Profile 2 (experimental data from [3-1] [3-2])

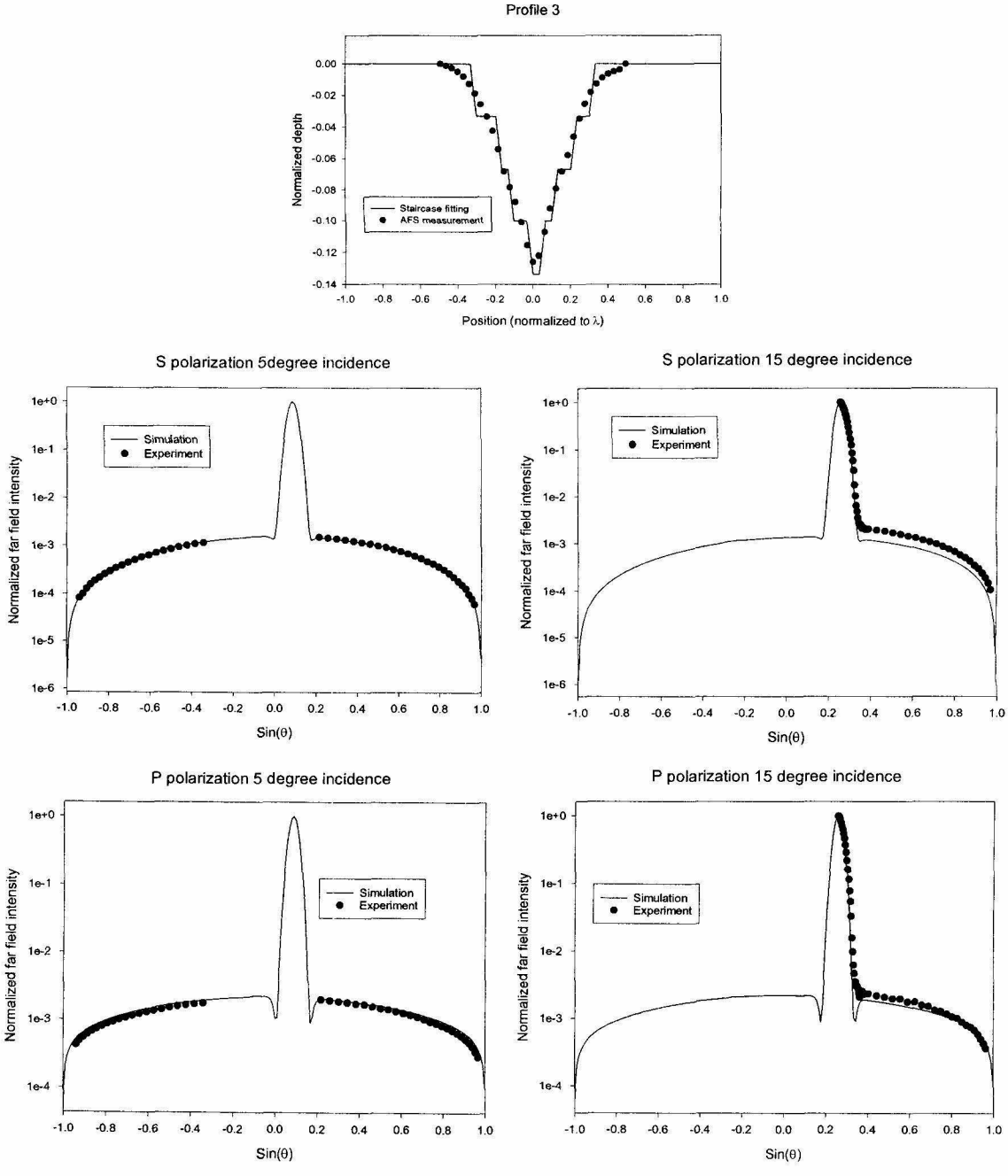


Fig. 3-3. Profile 3 (experimental data from [3-1] [3-2])

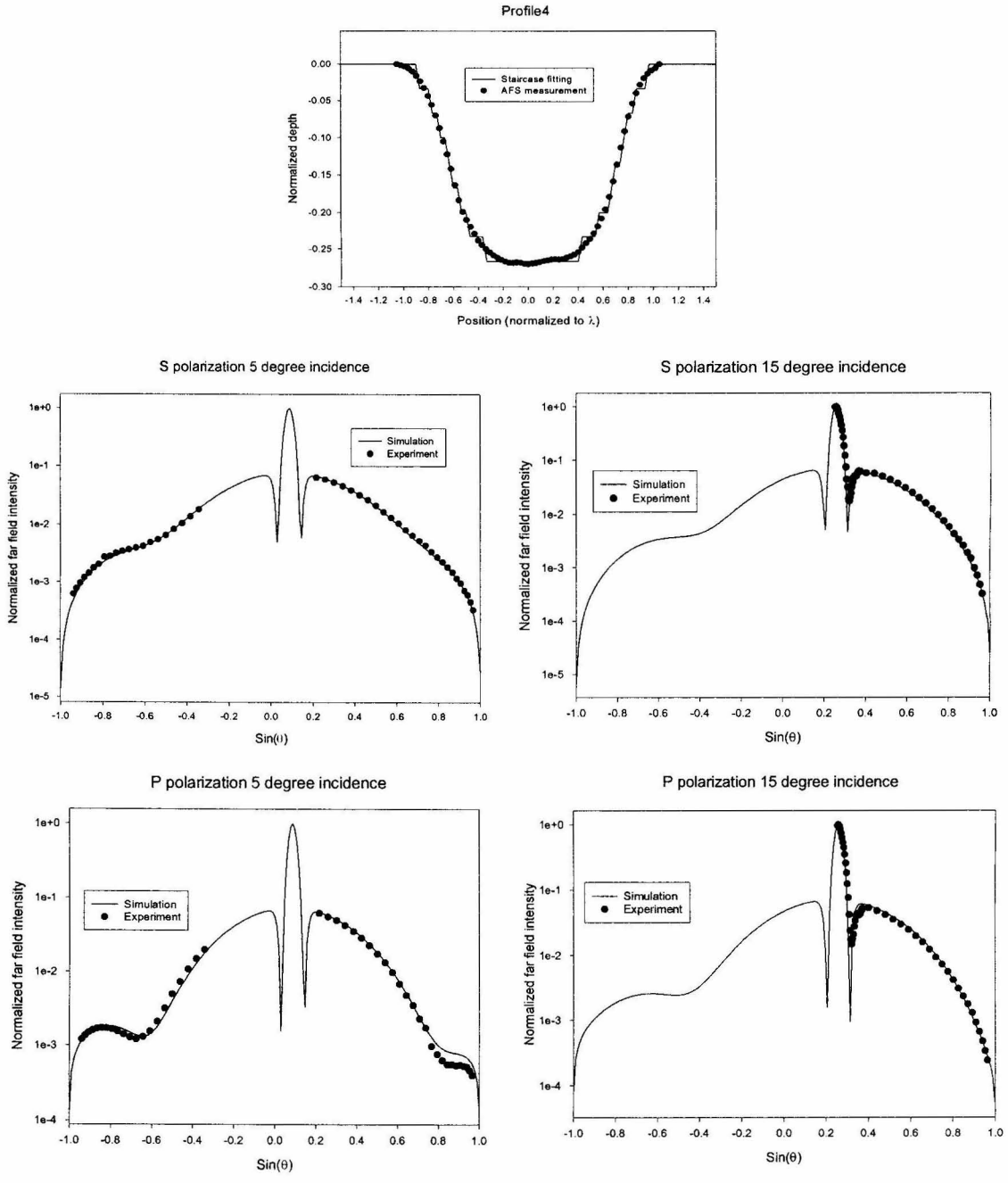


Fig. 3-4. Profile 4 (experimental data from [3-1] [3-2])

In profile 1, the trench is very narrow (0.4λ) and shallow (0.03λ). The two reflections have little phase shift and there is no dip when the floor meets the peak. In profile 2, the depth of the trench is about 0.07λ and again there is no dip when the floor meets the peak. However, the width of the trench is 1.6λ, and the far field diffraction of the light reflected from the bottom of the trench is approximately given by $\mathrm{sinc}\left(W\frac{\sin\theta}{\lambda}\right)$ (Fourier transform of $\mathrm{rect}\left(\frac{x}{W}\right)$), where W is the trench width. The diffraction produces a null (dip) at $\sin\theta \sim \frac{\lambda}{W} \approx 0.63$. In profile 3, the trench width is about 0.4λ and the depth is about 0.14λ. Because the trench is narrow it does not have any diffraction null on the floor, and more importantly, light cannot couple into the trench when illuminated with the s polarized light. Hence, the trench still appears shallow and there is no dip. However, if it is illuminated by the p polarized beam, light can couple into the trench and get reflected. The phase difference between the two reflected beams is significant enough to produce two dips. In Section 3.3, we will discuss the coupling to trench in more detail. In profile 4, the width of trench is 2λ and the depth is λ/4. The two reflections are out of phase and we see two sharp dips when the floor meets the peak. We also see the diffraction dips because the width is larger than the wavelength.

From the above analysis, we can simply model the wafer as a reflective phase modulator. The phase change from the wafer surface to trench is controlled by the depth of the trench if p polarized probing beam is used or the width of the trench is large enough that s beam can also be coupled into it. We use profile 4 as an example. The near field of the scattering wave is approximately given as

$$e^{-\frac{x^2}{w^2}} e^{2jkh\left[\mathrm{rect}\left(\frac{x}{W}\right)\right]} \tag{3-1}$$

where w is the waist of the incident Gaussian beam (assumed to be 4μm), W is the trench width and h is the trench depth. The far field is just the Fourier transform of the near field as given in Equation 3-113 or Equation 3-114. Figure 3-5 gives the results of this

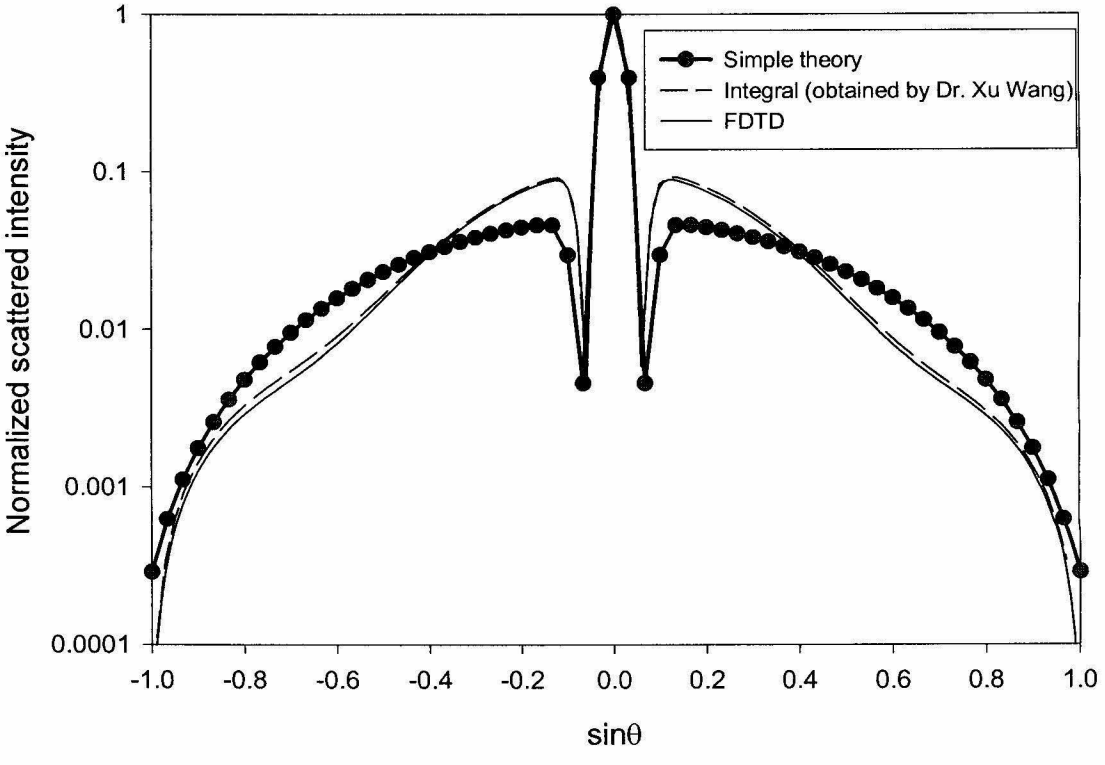


Fig. 3-5. Comparison of simple theory, FDTD and integral method

simple model and the FDTD simulation. The simple theory matches with the FDTD simulation very well and captures the essence of the physics. The result using the integral method obtained by Dr. Xu Wang [3-2] [3-3] is also shown. It agrees with the FDTD simulation result very well.

3.3 Encoding techniques

The trenches discussed in Section 3.2 can be modeled as waveguides as shown in Figure 3-6. For simplicity, the wafer is assumed to be a perfect electric conductor instead

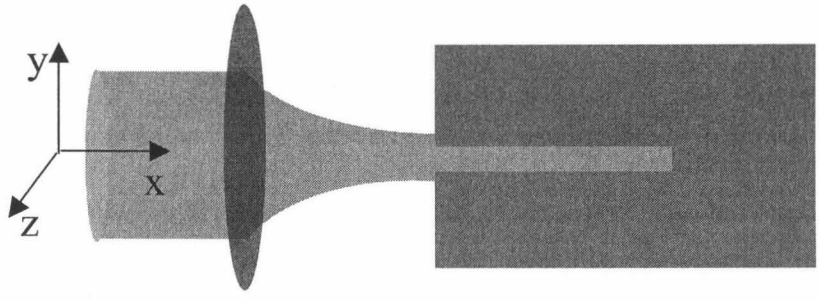


Fig. 3-6. Waveguide model

of NiP. The width of the trench is W. It is well known that there are three types of waveguide modes, TE, TM, and TEM modes [3-4].

1. TE mode

$$\mathbf{E}_m(x, y) = \mathbf{e}_z E_0 \sin\left(\frac{m\pi y}{W}\right) e^{j\beta x} \tag{3-2}$$

$$\mathbf{H}_m(x, y) = \frac{\nabla\times\mathbf{E}_m}{j\omega\mu_0} = \mathbf{e}_x \frac{E_0 m\pi}{j\omega\mu_0 W}\cos\left(\frac{m\pi y}{W}\right) e^{j\beta x} - \mathbf{e}_y \frac{\beta E_0}{\omega\mu_0}\sin\left(\frac{m\pi y}{W}\right) e^{j\beta x}$$

where m=1, 2, ... and $\beta = \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 - \left(\frac{m\pi}{W}\right)^2}$. The cut-off condition is $W \geq \frac{\lambda}{2}$.

2. TM mode

$$\mathbf{H}_m(x, y) = \mathbf{e}_z H_0 \cos\left(\frac{m\pi y}{W}\right) e^{j\beta x} \tag{3-3}$$

$$\mathbf{E}_m(x, y) = \frac{\nabla\times\mathbf{H}_m}{-j\omega\varepsilon_0} = \mathbf{e}_x \frac{H_0 m\pi}{j\omega\varepsilon_0 W}\sin\left(\frac{m\pi y}{W}\right) e^{j\beta x} + \mathbf{e}_y \frac{\beta H_0}{\omega\varepsilon_0}\cos\left(\frac{m\pi y}{W}\right) e^{j\beta x}$$

where m=1, 2, ... and $\beta = \sqrt{\left(\frac{2\pi}{\lambda}\right)^2 - \left(\frac{m\pi}{W}\right)^2}$. The cut-off condition is again $W \geq \frac{\lambda}{2}$.

3. TEM mode (0th TM mode)

$$\mathbf{E}(x, y) = \mathbf{e}_y E_0 e^{jkx} \tag{3-4}$$

$$\mathbf{H}(x, y) = \mathbf{e}_z E_0 \sqrt{\frac{\varepsilon_0}{\mu_0}} e^{jkx}$$

The TEM mode is like a truncated plane wave and it always exists.

Figure 3-7 shows the calculated H (E) field at some instant $t=t_0$ when illuminated with p (s) polarized beam. The trench width is $\lambda/3$. When illuminated with the p polarized beam, TEM mode can be excited in the trench. It propagates along the trench and is reflected back. However if illuminated with s polarized beam, no propagating waveguide modes can be excited and only some evanescent waves are excited.

In the following we discuss various schemes to encode information in subwavelength structures using our FDTD simulation tool. We can encode information by varying the trench width because it changes the ratio between the reflected wave from the wafer surface and that from the bottom of the trench, and the waveguide mode cut-off conditions depend on it. Figure 3-8 shows the far field scattering pattern of trenches with widths of λ and $\lambda/3$ using both the s and p polarizations. The waist of the incident beam is 1.6λ. The depth is fixed to $\lambda/4$. The wafer is assumed to be NiP. The far field scattering is different for different trench width. Because the phase delay of the waveguide mode depends on the depth, we can also encode information by varying the trench depth. Figure 3-9 shows the far fields of trenches with depths of $\lambda/4$ and $\lambda/2$. The width is fixed to λ in this case. The far field scattering is also different for different trench depth.

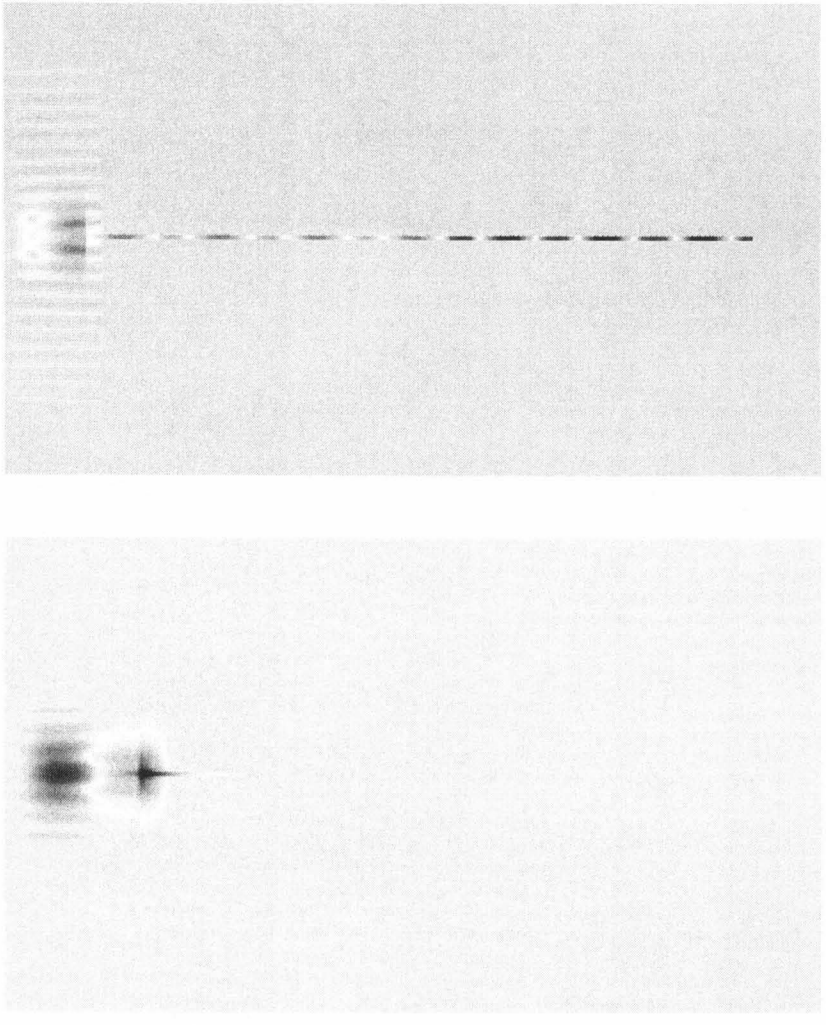


Fig. 3-7. Excitation of waveguide modes with p and s polarization

In the previous discussions, we focus on only one trench. It is interesting to see the effects when multiple trenches are illuminated simultaneously. Consider two trenches closely etched as shown in Figure 3-10. The incident beam has a waist of 6.25λ. We calculate the far field scattering and compare it with the linear superposition of the scattering fields of each individual trench (the reflection from the wafer surface needs to be subtracted because it is taken into account twice). Amazingly they agree very well as shown in

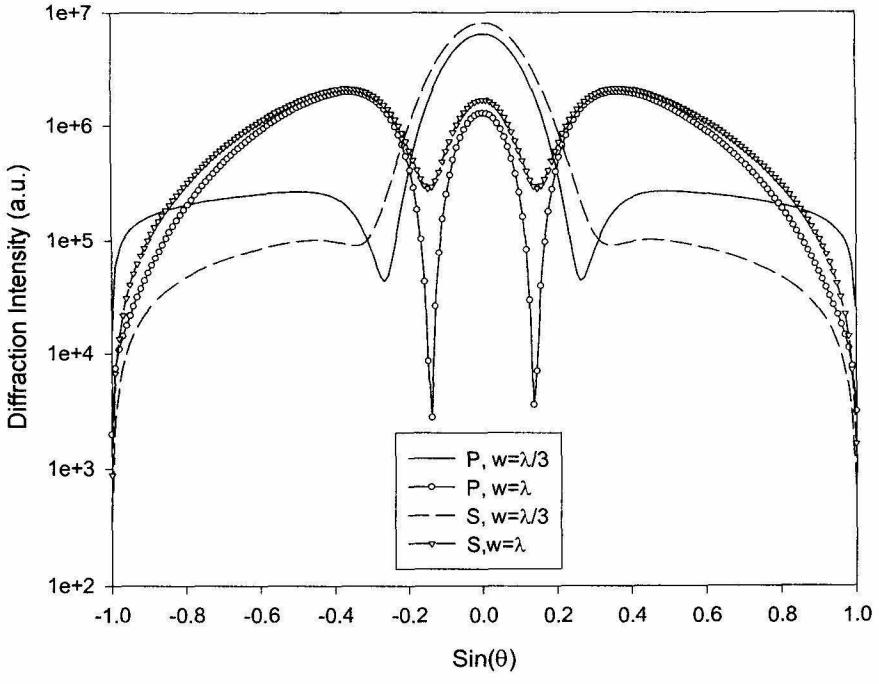


Fig. 3-8. Width multiplexing

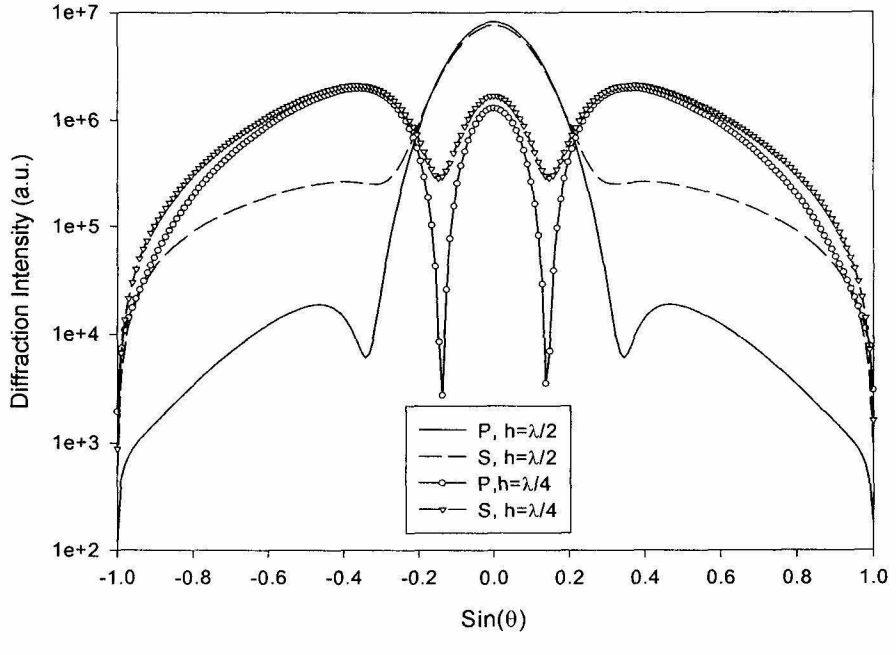


Fig. 3-9. Depth multiplexing

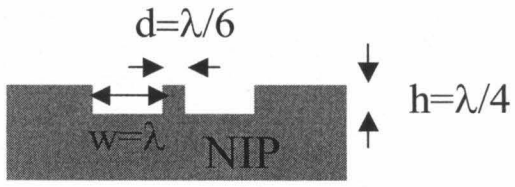


Fig. 3-10. Two close trenches

Figure 3-11. This is because NiP is like PEC and the waveguide modes of the two trenches

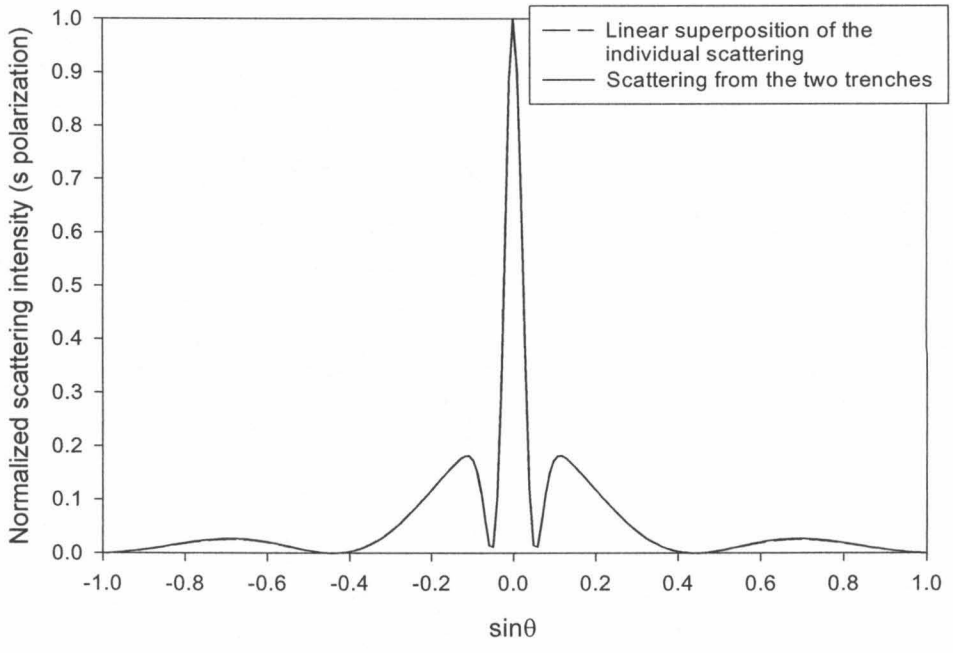


Fig. 3-11. Linear scattering

do not overlap spatially. Thus the presence of one trench does not affect the field in the other trench. This is not the case when surface plasma is excited [3-5] or transparent material is used, for example $LiNbO_3$, because the waves in the two trenches are coupled. We replace NiP with $LiNbO_3$ and repeat the calculation. The results are given in Figure 3-12. The linear superposition does not hold any more in this case.

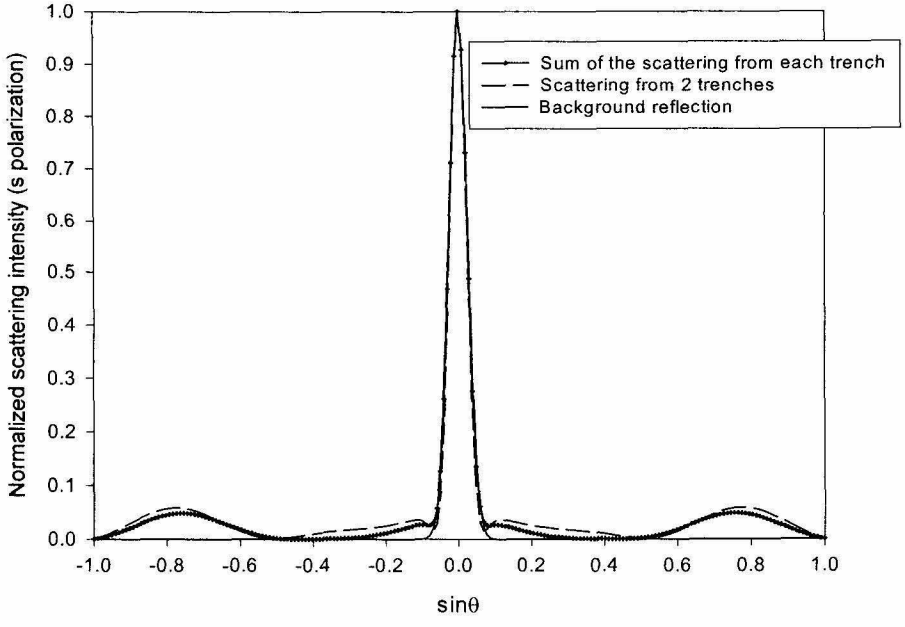


Fig. 3-12. Nonlinear scattering

In most cases, the trench is etched on some metal and if surface plasma is not excited, the scattering may assume to be linear. We can then create a phased array using multiple trenches with linearly chirped depth. The phased array redirect the incident beam toward a direction such that the reflected waveguide modes in all the trenches are in phase. The angle between that direction and the surface normal satisfies the following condition.

$$\sin\theta = \frac{2\delta}{d} \tag{3-5}$$

where δ and d are the depth difference and center-to-center distance of the neighboring trenches respectively. We simulated the far field scattering from 20 PEC trenches which are uniformly distributed within a span of 2 wavelengths. The width of the trench is $1/12\lambda$ and the PEC wall between neighboring trenches has a thickness of $1/60\lambda$. The maximal depth is $1/3\lambda$ and δ is $1/60\lambda$. P polarized focused Gaussian beam with waist w_0 of 0.83λ

($NA=0.6$, $2w_0=1/NA$) is used. The FDTD simulation is given in Figure 3-13. The peak is

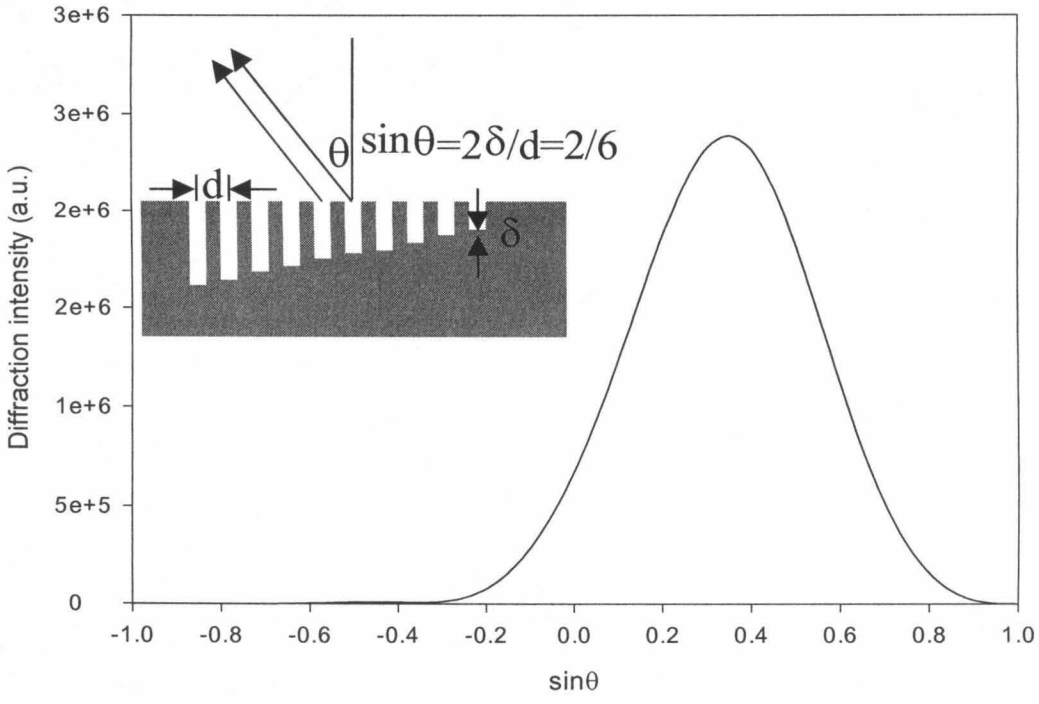


Fig. 3-13. Far field scattering from a phased array

located at $\sin\theta \approx 0.33$ which is consistent with the condition given in Equation 3-5.

The phased array basically is a mirror. We can encode information by varying the angle of the mirror. Figure 3-14 shows the scattering from some mirrors (assumed to be etched in NiP wafer). The slanting angle of the mirror is 15 degrees and the width varies from λ to 10λ. The incident beam has a waist of 1.6λ. When the width is small, it still looks like a trench because part of the beam is reflected by the mirror and the rest is reflected by the wafer surface. As we increase the width of the mirror, it gradually behaves like a mirror. In Figure 3-15, we show the scattering from mirrors with different angles. The mirror width

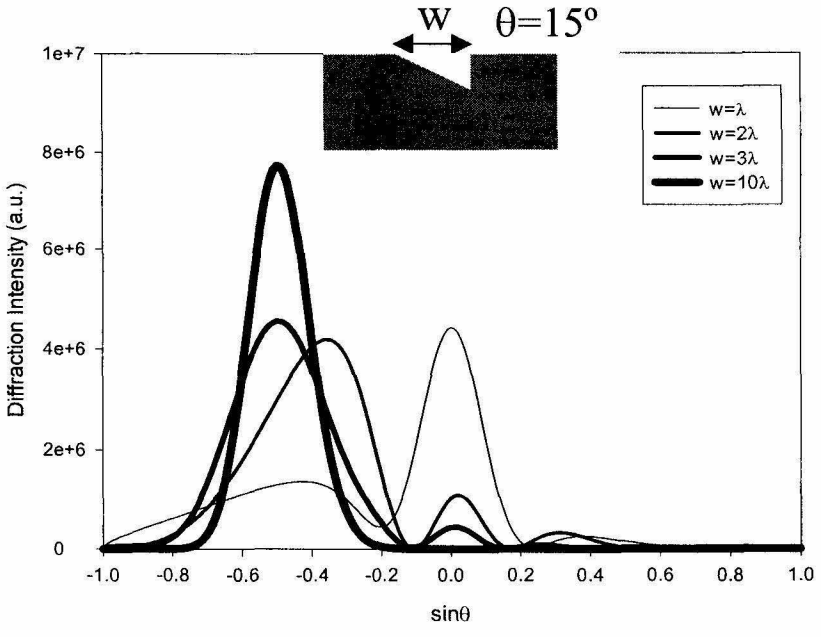


Fig. 3-14. Scattering from mirrors

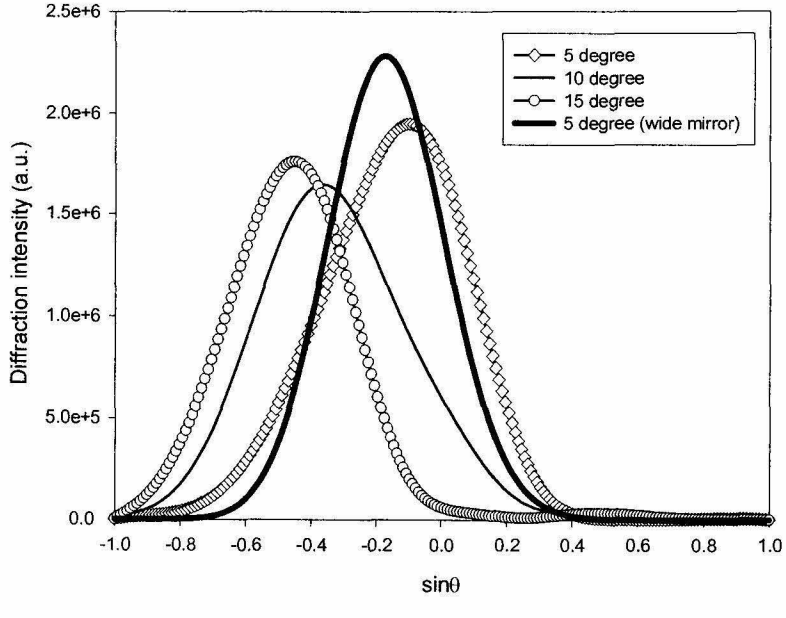


Fig. 3-15. Angle multiplexing

is λ except the wide mirror which has a width of 10λ. The incident beam is p polarized and is assumed to have a waist of 0.83λ.

In summary, we can vary the width or depth of a trench to encode information. A phased array which consists of many trenches is equivalent to a mirror and information can also be encoded by changing the mirror angle.

3.4 Eigenfunction analysis

In this section, we analyze how much information can be stored on a given length in 2D. We start from the near field to far field transformation (see Equation 3-113 and Equation 3-114 in Section 3.7). This result can also be derived from a Fourier optics approach. We assume the scalar field (either E_y of the s beam or H_y of the p beam) distribution at z=0 is f(x, 0) as shown in Figure 3-16. The field at z can be written as

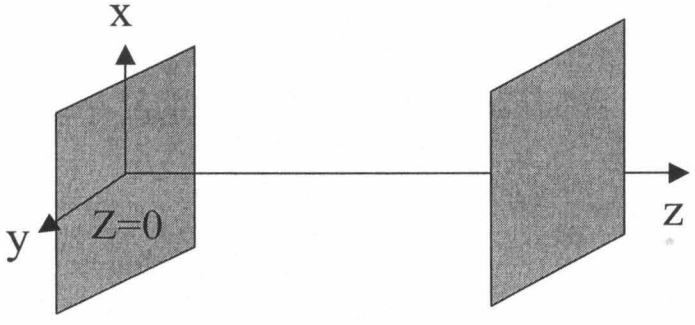


Fig. 3-16. Far field diffraction

$$f(x, z) = \int F(u) e^{j(2\pi ux + 2\pi z\sqrt{1/\lambda^2 - u^2})} du \tag{3-6}$$

where F(u) is the Fourier transform of f(x, 0).

$$F(u) = \int f(x, 0) e^{-j2\pi ux} dx \tag{3-7}$$

Notice that as z approaches to infinity the phase term $\phi(u, z) = 2\pi ux + 2\pi z\sqrt{\frac{1}{\lambda^2} - u^2}$ oscillates a lot except at u_0 where $\frac{\partial}{\partial u}\phi(u_0, z) = 0$.

$$\left.\frac{\partial\phi}{\partial u}\right|_{u_0} = 2\pi x - 2\pi z\frac{u_0}{\sqrt{1/\lambda^2 - u_0^2}} = 0 \tag{3-8}$$

$$\Rightarrow \text{tg}\theta = \frac{x}{z} = \frac{u_0}{\sqrt{1/\lambda^2 - u_0^2}}$$

$$\Rightarrow u_0 = \frac{\sin\theta}{\lambda}$$

The integration in Equation 3-6 has significant contribution only around u_0.

$$f(x, z) \approx F(u_0)|\delta u| \tag{3-9}$$

where δu is the range of u where $\phi(u, z)$ only changes a little bit.

$$\phi(u, z) \cong \phi(u_0, z) + \left.\frac{\partial\phi}{\partial u}\right|_{u_0}\delta u + \frac{1}{2}\left.\frac{\partial^2\phi}{\partial u^2}\right|_{u_0}(\delta u)^2 \tag{3-10}$$

$$\Rightarrow \delta\phi \cong \frac{1}{2}\left.\frac{\partial^2\phi}{\partial u^2}\right|_{u_0}(\delta u)^2 \Rightarrow |\delta u| \approx \sqrt{\left|\frac{2\delta\phi}{\left.\frac{\partial^2\phi}{\partial u^2}\right|_{u_0}}\right|}$$

$$\left.\frac{\partial^2\phi}{\partial u^2}\right|_{u_0} = -\frac{2\pi r\lambda}{\cos^2\theta} \tag{3-11}$$

where $r = \sqrt{x^2 + z^2}$. Plugging Equation 3-11, Equation 3-10 into Equation 3-9, we get

$$f(x, z) \propto \cos\theta F\left(\frac{\sin\theta}{\lambda}\right) = \sqrt{1 - (u\lambda)^2}\,\text{rect}\left(\frac{u\lambda}{2}\right)F(u) \tag{3-12}$$

where $u=\sin\theta/\lambda$ and $\text{rect}(x)=1$ if and only if $|x| \le \frac{1}{2}$.

The far field basically computes the Fourier transform of the near field. There is also a low-pass filtering effect caused by the factor $\sqrt{1 - (u\lambda)^2}\,\text{rect}\left(\frac{u\lambda}{2}\right)$. In the previous

sections we focus on the far field scattering intensity pattern and ignore the phase. Phase information can be detected, for example, by holographic correlators as shown in Figure 3-17. The pattern on the disk is correlated with the stored code templates. However,

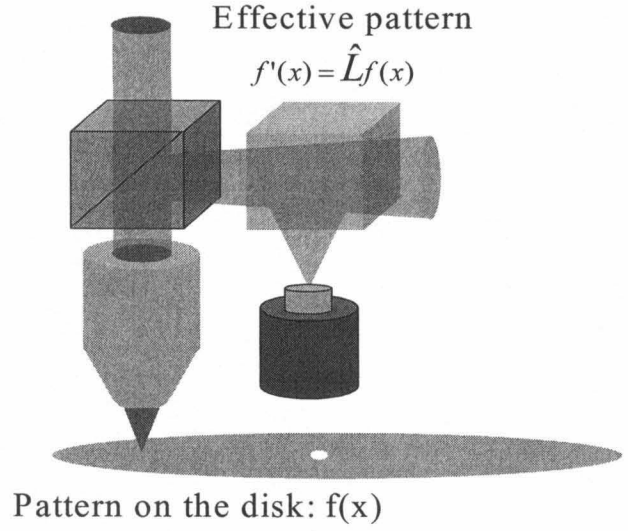


Fig. 3-17. Optical correlation decoder

due to the low-pass filtering effect of the lens, the pattern seen by the correlator is not the near field pattern on the disk.

Consider an ideal case that somehow we are able to measure both the amplitude and the phase of the far field. We can, for example, digitally reconstruct the pattern by performing an inverse Fourier transform. The reconstructed pattern is given by

$$f'(x) = \int \left[F(u)\,\mathrm{rect}\left(\frac{u\lambda}{2}\right)\sqrt{1-(u\lambda)^2} \right] e^{j2\pi ux}\,du = f(x) \otimes t(x) \qquad (3\text{-}13)$$

where $\otimes$ is the convolution operator and we have ignored any constant factor.

$$t(x) = \int\left[\mathrm{rect}\left(\frac{u\lambda}{2}\right)\sqrt{1-(u\lambda)^2}\right]e^{j2\pi ux}du = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\frac{1}{\lambda}\cos^2\theta e^{j2\pi x\sin\theta/\lambda}d\theta \tag{3-14}$$

$$= \frac{1}{2\lambda}\int_{-\pi}^{\pi}\cos^2\theta e^{j2\pi x\sin\theta/\lambda}d\theta = \frac{\pi}{2\lambda}\left[J_0\left(\frac{2\pi}{\lambda}x\right) + J_2\left(\frac{2\pi}{\lambda}x\right)\right]$$

We get the following.

$$f'(x) = \hat{L}f(x) = \int\frac{\pi}{2\lambda}\left[J_0\left(\frac{2\pi}{\lambda}(x+x')\right) + J_2\left(\frac{2\pi}{\lambda}(x+x')\right)\right]f(x')dx' \tag{3-15}$$

In principle, we can also define the transform as

$$f'(x) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left[F\left(\frac{\sin\theta}{\lambda}\right)\cos\theta\right]e^{j2\pi\frac{\sin\theta}{\lambda}x}d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}F\left(\frac{\sin\theta}{\lambda}\right)e^{j2\pi\frac{\sin\theta}{\lambda}x}d\sin\theta \tag{3-16}$$

$$\propto \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}}F(u)e^{j2\pi ux}du$$

If we consider the detection noise, the far field is given by $F\left(\frac{\sin\theta}{\lambda}\right)\cos\theta + N(\theta)$ where $N(\theta)$ is the measurement noise. Equation 3-16 leads to

$$f'(x) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left[F\left(\frac{\sin\theta}{\lambda}\right)\cos\theta + N(\theta)\right]e^{j2\pi\frac{\sin\theta}{\lambda}x}d\theta \tag{3-17}$$

$$= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left[F\left(\frac{\sin\theta}{\lambda}\right) + \frac{N(\theta)}{\cos\theta}\right]e^{j2\pi\frac{\sin\theta}{\lambda}x}d\sin\theta$$

$$\propto \int_{-\frac{1}{\lambda}}^{\frac{1}{\lambda}}[F(u) + \tilde{N}(u)]e^{j2\pi ux}du$$

where $\tilde{N}(u) = \frac{N(a\sin(\lambda u))}{\sqrt{1-(\lambda u)^2}}$ is the equivalent noise density. Notice that $\tilde{N}(u)$ is not uniformly distributed. It approaches to infinity when u is close to $\pm\frac{1}{\lambda}$. For this reason, in the following discussions we use the transform given by Equation 3-15.

It is very easy to check that the operator $\hat{L}$ defined in Equation 3-15 is linear and Hermitian. Thus $\hat{L}$ has real eigenvalues and its eigenfunctions form an orthogonal and complete set. In practical situation, we use a focused laser beam (spot size A) to illuminate the pits. The code pattern f(x) effectively has a finite aperture A and hence we have

$$f'(x) = \hat{L}f(x) = \int_{-\frac{A}{2}}^{\frac{A}{2}} \frac{\pi}{2\lambda}\left[J_0\left(\frac{2\pi}{\lambda}(x+x')\right) + J_2\left(\frac{2\pi}{\lambda}(x+x')\right)\right] f(x')dx' \qquad (3\text{-}18)$$

Here we assume a "hard" boundary, $f(x) = f(x) \times \mathrm{rect}\left(\frac{x}{A}\right)$, to simplify the algebra. It still needs experiments to justify how good this approximation is. In this case, $\hat{L}$ is Hermitian only in a sub-space $\{f \mid f(x) = f(x)\mathrm{rect}(x/A)\}$. However in general the digitally reconstructed pattern $f'(x) = \hat{L}f(x)$ is not limited by the aperture size A. Following the spirit of [3-6], we look for the eigenvalues and eigenfunctions of operator $\hat{L}$ defined in Equation 3-18.

The following algorithm is used to find out the eigenvalues. We start with the delta function $\delta(x)$, which can be represented by a linear combination of eigenfunctions of $\hat{L}$.

$$\delta(x) = \sum_i a_i e_i(x) \qquad (3\text{-}19)$$

where $e_i(x)$ is the eigenfunction of $\hat{L}$ with ith largest eigenvalue E_i. We repeatedly apply the operation $\hat{L}$.

$$\hat{L}^n\delta(x) = \sum_i a_i E_i^n e_i(x) \qquad (3\text{-}20)$$

Then only the eigenfunction with the largest eigenvalue can survive. This is the same as how the laser works. The operation $\hat{L}$ is like the round-trip propagation. Each eigenfunction is like a mode of the laser cavity and the corresponding eigenvalue is the round trip

gain/loss of the mode. Eventually only one mode can survive. Every time after we apply the operation $\hat{L}$, the result g(x) is also normalized such that

$$\int |g(x)|^2 dx = 1 \tag{3-21}$$

The calculated eigenvalue spectrum is shown in Figure 3-18. The eigenvalue is a measure

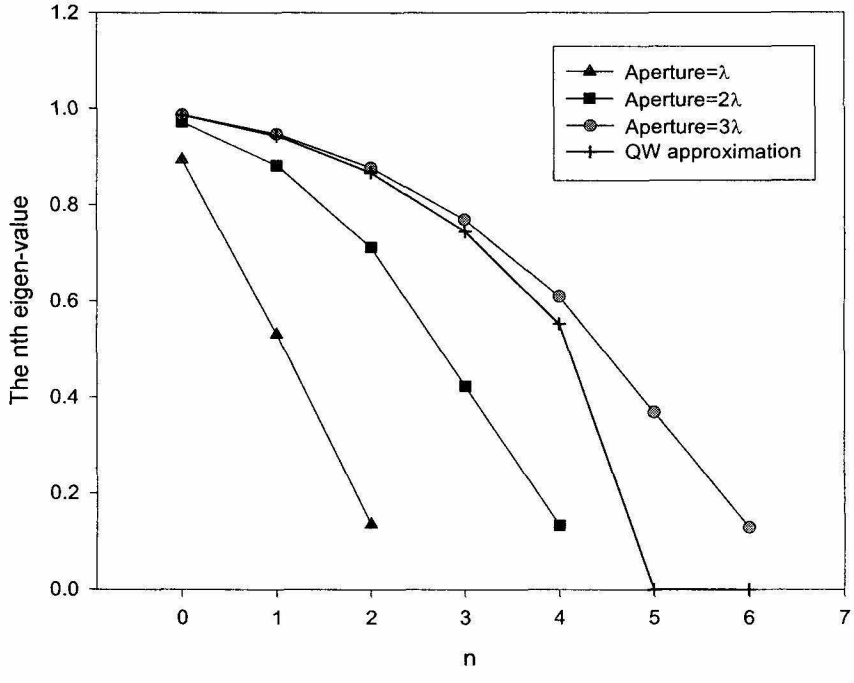


Fig. 3-18. Eigenvalue spectrum

of the noise tolerance. The reconstructed pattern is given by $r_n(x) = \hat{L}e_n(x) + N(x)$ where N(x) is the reconstruction error due to measurements. We get

$$e_n(x) = \frac{r_n(x)}{E_n} - \frac{N(x)}{E_n} \tag{3-22}$$

For small E_n, the reconstruction is very noisy.

If an eigenvalue of 0.1 is tolerable, then we can store 3, 5 and 7 bits with A=λ, 2λ and 3λ respectively. We get the highest density when A=λ. This implies that we should

focus the laser beam as small as possible in order to increase the storage density. Scenarios with multiple bits simultaneously illuminated by a larger beam are not as efficient. The first three eigenfunctions with A=λ are shown in Figure 3-19. The correlation matrix of the

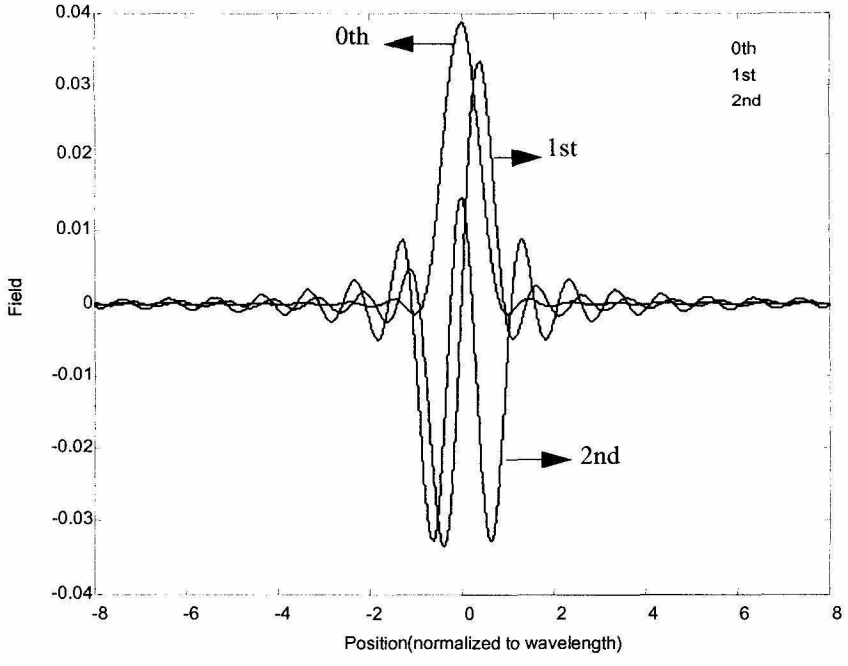


Fig. 3-19. Eigenfunctions

three eigenfunctions is given by

$$C = \begin{bmatrix} \langle e_1|e_1\rangle & \langle e_1|e_2\rangle & \langle e_1|e_3\rangle \\ \langle e_2|e_1\rangle & \langle e_2|e_2\rangle & \langle e_2|e_3\rangle \\ \langle e_3|e_1\rangle & \langle e_3|e_2\rangle & \langle e_3|e_3\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 & -0.1227 \\ 0 & 1 & 0 \\ -0.1227 & 0 & 1 \end{bmatrix} \tag{3-23}$$

$$\text{with } \langle e_i|e_j\rangle = \int e_i^*(x)e_j(x)dx \tag{3-24}$$

The three eigenfunctions are approximately orthogonal to each other.

From Equation 3-13, mathematically if A = ∞ it is easy to check that the eigenvalues and the eigenfunctions are given by the following.

$$E(u) = \sqrt{1-(u\lambda)^2} \qquad (3\text{-}25)$$
$$e(u, x) = e^{\pm j2\pi ux}$$

Notice that physically a plane wave does not have a well defined far field and thus we cannot obtain Equation 3-12. If $|u\lambda|$ is much smaller than 1, then Equation 3-25 becomes

$$E(u) \approx 1 - \frac{1}{2}(u\lambda)^2 = 1 - \frac{(2\pi u)^2}{2m^*} \qquad (3\text{-}26)$$

where $m^* = k^2$. This is like the energy of a hole in bulk material. The wavefunction of the hole is a plane wave. If we limit the spatial extension of the wavefunction (finite A), it is like a hole in a Quantum well. We use the quantization condition of a quantum well [3-7].

$$2\pi u_n A = (n+1)\pi \qquad (3\text{-}27)$$

We get an approximate expression for the eigenvalue E_n.

$$E_n \approx \sqrt{1-\left[\frac{(n+1)\lambda}{2A}\right]^2} \qquad (3\text{-}28)$$

As an example, we calculate the 0th eigenvalue of the case $A=\lambda$. The numerical result is 0.8927 and our approximate result is 0.8660. In Figure 3-18, we show the quantum well approximation of the case $A=3\lambda$ and compare it with the numerical result. They agree very well for the first five eigenvalues.

3.5 Effect of nonlinear buffer layer

From Section 3.4, higher storage density can be achieved by using smaller probing beam. The laser beam can be focused to a spot size of λ/NA, where NA is the numerical aperture of the lens. In order to reduce the spot size, we can either use shorter wavelength lasers (blue or UV lasers for example) or higher numerical aperture lenses (oil immersed objective lenses). We can also use light source with evanescent waves. For example, light

coupled from narrow fiber tip or scattered from sharp needle is used in near field scanning. However, all of these techniques are very complicated and difficult to use. One simpler idea is to bury the trench inside some buffer layer. Because in the buffer layer the wavelength is reduced to λ/n, smaller spot size might be possible. But due to the Snell's law the numerical aperture in the buffer layer is also reduced by a factor of n (refractive index) we still get the same spot size. In this section, we investigate the effects when the buffer layer is nonlinear. In Section 3.5.1, we analyze self focusing in the nonlinear buffer layer using a perturbation method. In Section 3.5.2, we use FDTD and beam propagation method to simulate the beam shaping.

3.5.1 Perturbation method

We consider a nonlinear buffer layer whose refractive index is intensity dependent. The focused incident beam can induce a lensing effect in the otherwise uniform medium and focus to a smaller spot. To analyze this self-focusing problem we start from the wave equation, ignoring absorption.

$$\nabla^2\mathbf{E} + \frac{\omega^2}{c^2}(n_0 + \Delta n)^2\mathbf{E} = 0 \tag{3-29}$$

Here we assume $\nabla\varepsilon \perp \mathbf{E}$ or $\nabla\varepsilon$ small and therefore $\nabla \cdot \mathbf{E} = 0$. We consider a TE wave,

$$\mathbf{E} = \mathbf{e}_y\tilde{E}(x, y, z)e^{jkz} \tag{3-30}$$

where $\tilde{E}$ is the slowly varying amplitude, $k = \frac{\omega}{c}n_0$ and $\Delta n = n_2|\tilde{E}|^2$ is the nonlinear refractive index. Notice that $\nabla^2 = \nabla_t^2 + \frac{\partial^2}{\partial z^2}$ and $\frac{\partial}{\partial z}\tilde{E} \ll jk\tilde{E}$. We get the famous nonlinear Schrödinger equation.

$$2jk\frac{\partial}{\partial z}\tilde{E} = -\nabla_t^2\tilde{E} - 2k^2\frac{n_2}{n_0}|\tilde{E}|^2\tilde{E} \tag{3-31}$$

If n_2=0, it is a diffusion equation and its solution is the well known Gaussian beam [3-7].

$$\tilde{E}(x, y, z) = A_0\frac{w_0}{w(z)}e^{-\frac{r^2}{w^2(z)}}e^{j\left(\frac{kr^2}{2R(z)} - \eta(z)\right)} \tag{3-32}$$

where $w^2(z) = w_0^2\left(1 + \frac{z^2}{z_0^2}\right)$, $R(z) = z\left(1 + \frac{z_0^2}{z^2}\right)$, $\eta(z) = \text{atan}\left(\frac{z}{z_0}\right)$, and $z_0 = \frac{\pi w_0^2 n_0}{\lambda}$.

Let $\tilde{E}(x, y, z) = A(x, y, z)e^{j\phi(x, y, z)}$ where A and ϕ are real. Equation 3-31 can be rewritten as [3-8]

$$k\frac{\partial}{\partial z}A^2 = -\nabla_t \cdot (A^2\nabla_t\phi) \tag{3-33}$$

$$\frac{\partial\phi}{\partial z} + \frac{1}{2k}(\nabla_t\phi)^2 - \frac{k}{2}\left(\frac{\nabla_t^2 A}{k^2 A} + \frac{2n_2A^2}{n_0}\right) = 0 \tag{3-34}$$

Shen solved it by drawing the analogy to the Hamilton-Jacobi equation in classical mechanics [3-8]. Here we give a more straightforward solution based on perturbation method. Consider the 2D case. Without the nonlinear effect, the solution is a 2D Gaussian beam.

$$A = A_0\sqrt{\frac{w_0}{w(z)}}e^{-\frac{x^2}{w^2(z)}} \tag{3-35}$$

$$\phi = \frac{kx^2}{2R(z)} - \eta(z)$$

In the nonlinear case, we still assume this solution. However, w(z), R(z) and $\eta(z)$ are perturbed and satisfy different sets of equations. Plugging Equation 3-35 into Equation 3-33, we obtain

$$\frac{1}{w(z)}\frac{d}{dz}w(z) = \frac{1}{R(z)}. \tag{3-36}$$

Plugging Equation 3-35 into Equation 3-34, we get

$$\frac{1}{2}kx^2\left[\frac{d}{dz}\left(\frac{1}{R(z)}\right)+\frac{1}{R^2(z)}-\frac{4}{k^2w^4(z)}\right]-\frac{kn_2A_0^2}{n_0}\frac{w_0}{w(z)}e^{-\frac{2x^2}{w^2(z)}} \qquad (3\text{-}37)$$

$$-\frac{d}{dz}\eta(z)+\frac{1}{kw^2(z)}=0$$

If we only consider the region $|x| \ll w(z)$, we can expand $e^{-\frac{2x^2}{w^2(z)}} \approx 1-\frac{2x^2}{w^2(z)}$ and get

$$\frac{d}{dz}\left(\frac{1}{R(z)}\right)+\frac{1}{R^2(z)}-\frac{4}{k^2w^4(z)}+\frac{4n_2A_0^2w_0}{n_0w^3(z)}=0 \qquad (3\text{-}38)$$

$$\frac{d}{dz}\eta(z)=\frac{1}{kw^2(z)}-\frac{kn_2A_0^2w_0}{n_0w(z)} \qquad (3\text{-}39)$$

Using Equation 3-36 to eliminate R(z) in Equation 3-38, we obtain

$$\frac{d^2}{dz^2}w(z)=\frac{4}{k^2w^3(z)}-\frac{4n_2A_0^2w_0}{n_0w^2(z)}=-\frac{\partial}{\partial w}V(w) \qquad (3\text{-}40)$$

where $V(w)=\frac{2}{k^2w^2}-\frac{4n_2A_0^2w_0}{n_0w}$ is an effective potential, which is slightly different from Shen's. Equation 3-40 is the same as the Newton's law in mechanics and from the energy conservation law it becomes

$$\frac{1}{2}\left(\frac{dw}{dz}\right)^2+V(w)=E_0 \quad \text{(constant)} \qquad (3\text{-}41)$$

We first consider the linear case, i.e., $n_2=0$. At $z=\infty$, we have $w=\infty$, $V(w)=0$ and $\left.\frac{dw}{dz}\right|_{z\to\infty}=\frac{w_0}{z_0}$. At the waist, we have $w=w_0$ and $\frac{dw}{dz}=0$. Equation 3-41 gives

$$\frac{1}{2}\left(\frac{w_0}{z_0}\right)^2=V(w_0)=\frac{2}{k^2w_0^2} \qquad (3\text{-}42)$$

$$\Rightarrow z_0=\frac{\pi w_0^2 n_0}{\lambda}$$

This is consistent. Now we consider the nonlinear case. At $z=\infty$, we have $w=\infty$. The intensity of the laser beam is so low (approaching to 0) that the medium is essentially linear.

$$\left.\frac{dw}{dz}\right|_{z\to\infty}^{\text{nonlinear}} = \left.\frac{dw}{dz}\right|_{z\to\infty}^{\text{linear}} = \frac{w_{0L}}{z_{0L}} = \frac{2}{kw_{0L}} \tag{3-43}$$

where the subscript L denotes the value when the medium is linear. In the position where the beam has minimal size w_{0NL}, $\frac{dw}{dz} = 0$.

$$\begin{aligned} \frac{1}{2}\left(\frac{2}{kw_{0L}}\right)^2 &= \frac{2}{k^2 w_{0NL}^2} - \frac{4n_2 A_0^2 w_{0NL}}{n_0 w_{0NL}} \\ \Rightarrow \frac{1}{w_{0NL}^2} &= \frac{1}{w_{0L}^2} + \frac{2n_2 A_0^2 k^2}{n_0} \\ \Rightarrow \frac{w_{0L}}{w_{0NL}} &= \sqrt{1 + \frac{2n_2 A_0^2 k^2 w_{0L}^2}{n_0}} = \sqrt{1 + 4\frac{n_2 A_0^2}{n_0} k z_{0L}} = \sqrt{1 + 4\delta\phi_{NL}} \end{aligned} \tag{3-44}$$

where $\delta\phi_{NL} = \frac{n_2 A_0^2}{n_0} k z_{0L} = \frac{2\pi}{\lambda}(n_2 A_0^2) z_{0L}$ is the extra phase shift resulting from the propagation in the nonlinear medium. The effective thickness of the medium is about z_{0L} since the intensity of the beam is high in that region. This shows that in the nonlinear buffer layer we can use the self-focusing effect to achieve a smaller beam waist.

3.5.2 FDTD and BPM simulations

We simulated the scattering from a trench buried inside a nonlinear buffer layer. The result is shown in Figure 3-20. The nonlinear buffer layer has refractive index of 1.5 and the nonlinear coefficient n_2 of 0.01 or 0.02 (the unit of n_2 is chosen such that an beam intensity of 1 results in a refractive index change of n_2, in this case 0.01 or 0.02). The thickness of the layer is 5λ (λ=633nm). The trench is etched on NiP and has width of λ/3 and depth of λ/4. The incident Gaussian beam is s polarized and has a waist of 1.6λ. In the far field scattering pattern, there are ripples which are possibly due to the multiple reflections between the air/buffer interface and the trench wafer surface. The nonlinear effect lifts the wing of the scattering pattern because of self-focusing. However, we do not see drastic dif-

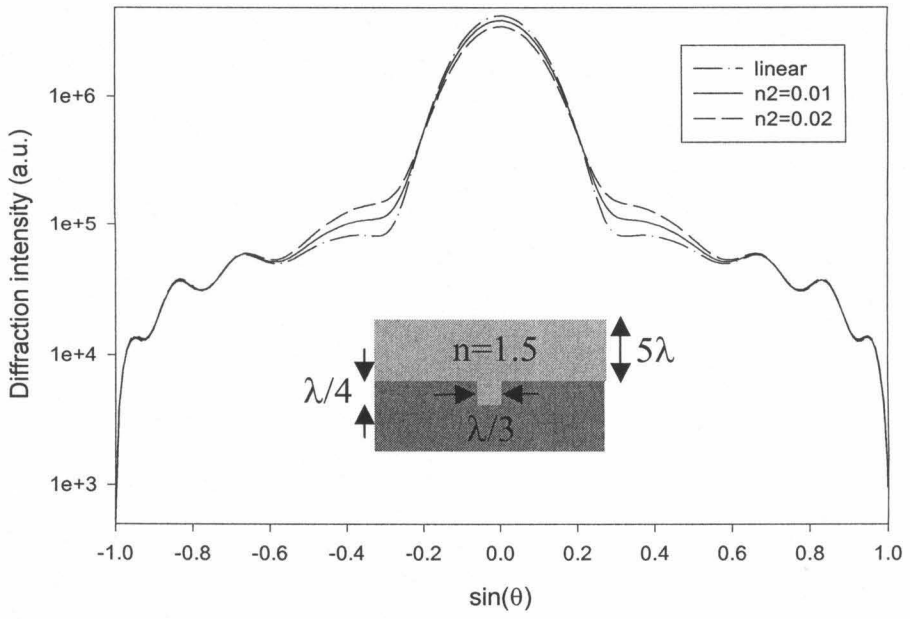


Fig. 3-20. Scattering from a trench in a nonlinear medium

ferences. One possible reason is that the buffer layer is too thin. For thick buffer layer (for example mm to cm thickness), the FDTD algorithm is not efficient any more. In the following we use another method, the beam propagation method (BPM) [3-9][3-10], to analyze the beam shaping in the buffer layer.

The nonlinear Schrödinger equation given in Equation 3-31 can be solved using the following algorithm. As shown in Figure 3-21, we consider a small cell with length Δz. The

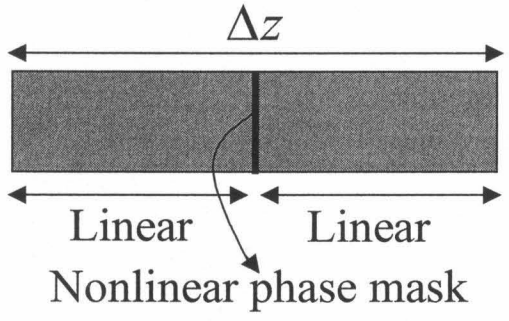


Fig. 3-21. Beam propagation method

distributed nonlinearity is approximated by a lumped phase mask in the center. The field at z=0 propagates a distance of $\Delta z/2$ in a linear medium, and is modulated by the phase mask in the center, and propagates another distance of $\Delta z/2$. This approximation is very good and it can be shown that the error is in the order of $o(\Delta z^3)$. The propagation in a linear medium is just diffraction and we can calculate it using Fourier optics[3-22]. The algorithm consists of the following three steps (Chapter 6 of [3-9]) :

1. Linear propagation from 0 to $\Delta z/2$ (the center):

$$\tilde{E}\left(x, \frac{\Delta z}{2}\right) = IFT\left\{e^{-j\pi\frac{\lambda}{n_0}u^2\frac{\Delta z}{2}} FT[\tilde{E}(x, 0)]\right\} \tag{3-45}$$

where FT is the Fourier transform and IFT is the inverse Fourier transform.

2. Lumped nonlinear phase mask

$$\tilde{E}'\left(x, \frac{\Delta z}{2}\right) = \tilde{E}\left(x, \frac{\Delta z}{2}\right) e^{j\frac{2\pi}{\lambda}n_2\left|\tilde{E}\left(x, \frac{\Delta z}{2}\right)\right|^2\Delta z} \tag{3-46}$$

3. Linear propagation from the center to Δz

$$\tilde{E}(x, \Delta z) = IFT\left\{e^{-j\pi\frac{\lambda}{n_0}u^2\frac{\Delta z}{2}} FT\left[\tilde{E}'\left(x, \frac{\Delta z}{2}\right)\right]\right\} \tag{3-47}$$

The above process is repeated until the desired propagation distance.

Figure 3-22 shows a typical beam waist profile in a nonlinear medium. It is obtained

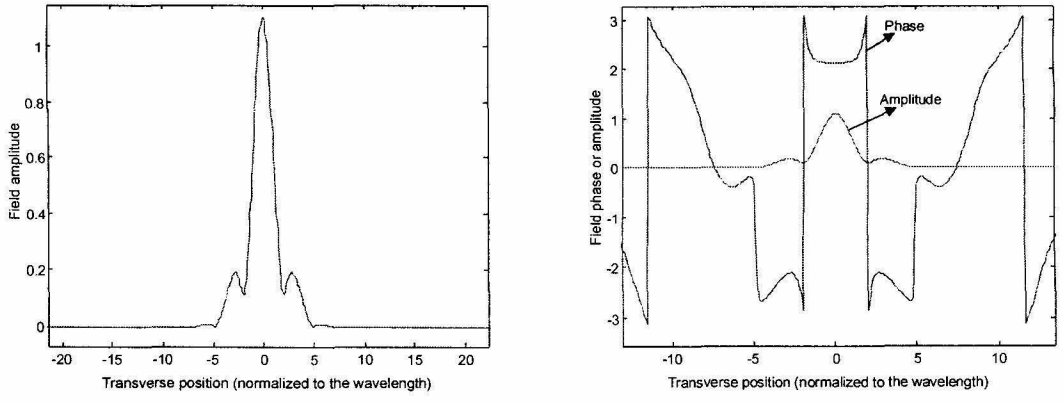


Fig. 3-22. Beam profile in nonlinear medium

using BPM (n_2=0.01, w_{0L}=1.67λ, propagation distance z=491λ, if material is linear the beam waist is at z=500λ). The profile looks like a Gaussian in the center but develops some side lobes on the wing. The center and the side lobe are almost out of phase as shown in the right plot of Figure 3-22. In Figure 3-23, we compare the beam propagation in the linear and nonlinear buffer layer. In the left plot, the amplitudes along the beam axis are shown.

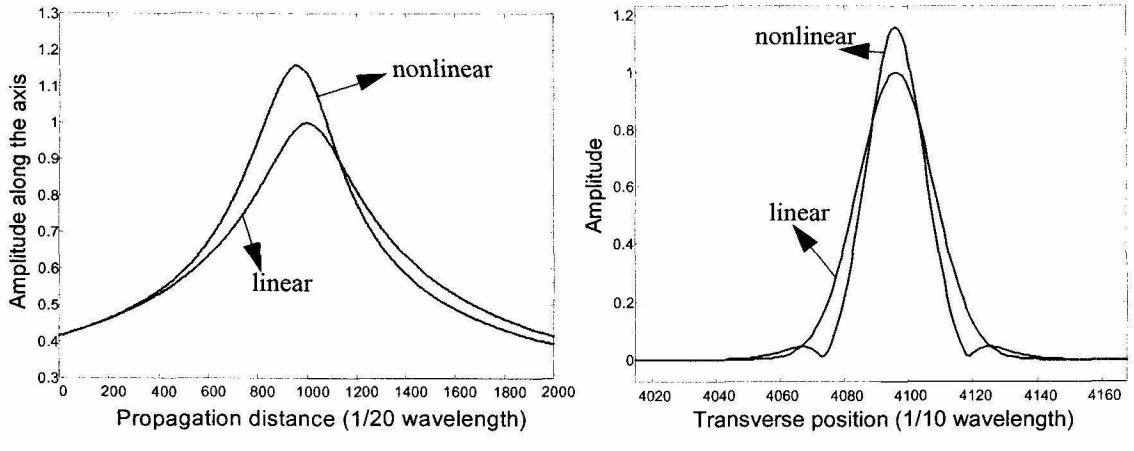


Fig. 3-23. Beam propagation in linear and nonlinear medium

In the nonlinear medium (n_2=0.01), the beam focuses faster and has higher intensity in the center because of the self-focusing. The beam waist profiles are shown in the right plot. Apparently, in the nonlinear medium the waist is smaller.

Because the waist profile typically has side lobes, we redefine the effective waist of the profile as

$$w_{eff} = 2\sqrt{\frac{\int(x - x_c)^2 I(x)dx}{\int I(x)dx}} \tag{3-48}$$

If the beam is Gaussian, the above definition gives w_{eff}=w_0. We calculate the dependence of the effective waist on the propagation distance (from the entrance to the focal plane in the nonlinear buffer layer) in the nonlinear medium and it is shown in Figure 3-24. If the

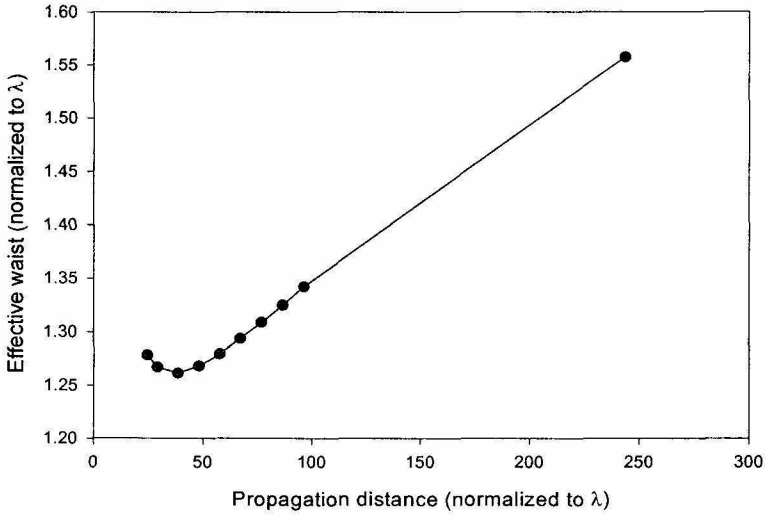


Fig. 3-24. Optimal propagation distance

propagation distance is too long the beam can develop large side lobes and this leads to a

wider effective waist. In the simulation, n_2=0.01 and w_{0L}=1.67λ are used. The optimal thickness is about 50λ and the effective waist of the beam is reduced to about 1.25λ.

The nonlinear coefficient n_2 of most materials is very small and typically high intensity laser beam is required. Recently, it is found that liquid crystals have a large non-linear coefficient [3-11]. In the isotropic phase n_2 is about $10^{-10}cm^2/W$ with a response time of about 100ns and in the nematic phase n_2 is about $10^{-3}cm^2/W$ with a response time of about 1s. However, the change of refractive index is in general rather small comparing with the nonlinear absorption change. The nonlinear absorption effect [3-12] can be very large because we can work in the resonant condition. Consider a material whose nonlinear absorption coefficient is given by

$$\alpha = \frac{\alpha_0}{1 + I/I_{sat}} \tag{3-49}$$

The center of the beam has higher intensity and can bleach the material. After propagating some distance only the center part of the beam is not attenuated and therefore the beam is sharpened. Figure 3-25 shows the dependence of the effective width (normalized to λ) on

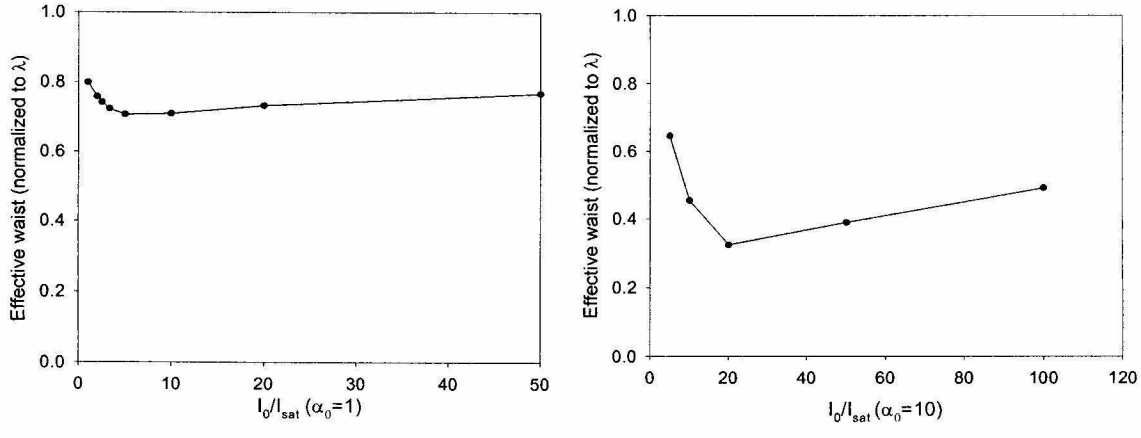


Fig. 3-25. Nonlinear absorption

the intensity at the beam center. If the intensity is too low, the center part of the beam is also attenuated. On the other hand, if the intensity is too high, the peripheral part also bleaches the material. In our simulation, the thickness of the material is chosen to be 3.75λ and the refractive index is 1.5. The incident Gaussian beam has a waist of 0.833λ. Even in such thin layer, the effective waist can be reduced to as small as 0.3λ (α_0=10).

3.6 Conclusion

In this chapter, we developed a simulation tool using the FDTD algorithm to calculate the far field scattering of 2D subwavelength structures. We used this tool to analyze the scattering from subwavelength trenches and the simulation results agree with the experimental data very well. The general feature of the far field scattering from a subwavelength trench consists of a peak on top of a broad floor. The peak is due to the wafer surface reflection while the floor is caused by the reflected waveguide mode (light coupled into the trench). We can encode information by width, depth and angle multiplexing. An eigenfunction approach is used to analyze the near field to far field transform which is related to the trade-off between the storage density and S/N. Finally, we studied the effect of a nonlinear buffer layer. Self-focusing or self-sharpening in the nonlinear buffer layer can achieve a smaller effective beam waist and increase the storage density.

3.7 Appendix: Introduction to the finite difference time domain method (FDTD)

In this section we review the FDTD algorithm which is used to solve the Maxwell's equations. The basic idea is very simple. Given the field distribution **E**(**r**,t) and **H**(**r**,t) everywhere at time t, the Maxwell's equations tell us about the time derivatives of the fields

$\frac{\partial}{\partial t}\mathbf{E}(\mathbf{r}, t)$ and $\frac{\partial}{\partial t}\mathbf{H}(\mathbf{r}, t)$. We can then advance the fields to the next instant t+Δt. Because the Maxwell's equations are directly solved and the boundary conditions are implicitly satisfied, it is very easy to deal with complicated structures. The response in a wide frequency domain can be calculated in one impulse response simulation if the system is linear. The method is also easy to be extended to the nonlinear case by modeling the material parameters field dependent. FDTD can be implemented very efficiently for small structures (in the order of wavelength), for example photonic crystals. However, it takes a lot of memory and CPU time if it is applied to large structures. In the following review we follow [3-13] and [3-14] very closely.

3.7.1 Central difference and Yee cell

Maxwell's equations are differential equations. The first step is to transform them into difference equations. We first consider a scalar function f(t). Its derivative can be approximated using the Taylor expansion as shown in the following.

$$f(t+\Delta t) = f(t) + \left.\frac{df}{dt}\right|_t \Delta t + o(\Delta t^2) \Rightarrow \left.\frac{df}{dt}\right|_t = \frac{f(t+\Delta t) - f(t)}{\Delta t} + o(\Delta t) \qquad (3\text{-}50)$$

In this method, the error in approximating the derivative with the difference is linear with the step Δt. A better way is to use the central difference.

$$f\left(t+\frac{\Delta t}{2}\right) = f(t) + \left.\frac{df}{dt}\right|_t \left(\frac{\Delta t}{2}\right) + \frac{1}{2}\left.\frac{d^2 f}{dt}\right|_t \left(\frac{\Delta t}{2}\right)^2 + o(\Delta t^3) \qquad (3\text{-}51)$$

$$f\left(t-\frac{\Delta t}{2}\right) = f(t) - \left.\frac{df}{dt}\right|_t \left(\frac{\Delta t}{2}\right) + \frac{1}{2}\left.\frac{d^2 f}{dt}\right|_t \left(\frac{\Delta t}{2}\right)^2 - o(\Delta t^3) \qquad (3\text{-}52)$$

Subtract Equation 3-51 from Equation 3-52 and we get

$$\left.\frac{df}{dt}\right|_t = \frac{f\left(t+\frac{\Delta t}{2}\right) - f\left(t-\frac{\Delta t}{2}\right)}{\Delta t} + o(\Delta t^2) \tag{3-53}$$

The error is now quadratic with the step Δt. More generally, consider the following

$$f(t + x_i\Delta t) = f(t) + \frac{df}{dt}x_i\Delta t + \frac{1}{2!}\frac{d^2 f}{dt^2}(x_i\Delta t)^2 + \ldots + \frac{1}{m!}\frac{d^m f}{dt^m}(x_i\Delta t)^m + o(\Delta t^{m+1}) \tag{3-54}$$

where i=1,..., m. We multiply Equation 3-54 with y_i and sum up the m equations.

$$\sum_{i=1}^{m} y_i[f(t + x_i\Delta t) - f(t)] = \frac{df}{dt}\Delta t\sum_{i=1}^{m} x_i y_i + \ldots + \frac{1}{m!}\frac{d^m f}{dt^m}\Delta t^m \sum_{i=1}^{m} x_i^m y_i + o(\Delta t^{m+1}) \tag{3-55}$$

We require x_i and y_i satisfy the following condition

$$\sum_{i=1}^{m} x_i y_i = 1 \text{ and } \sum_{i=1}^{m} x_i^k y_i = 0 \text{ for } k=2,\ldots, m \tag{3-56}$$

The relation can also be written in matrix form

$$\begin{bmatrix} x_1 & x_2 & \ldots & x_m \\ x_1^2 & x_2^2 & \ldots & x_m^2 \\ . & . & \ldots & . \\ x_1^m & x_2^m & \ldots & x_m^m \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_m \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ . \\ 0 \end{bmatrix} \tag{3-57}$$

Equation 3-55 is simplified to

$$\frac{df}{dt} = \frac{\sum_{i=1}^{m} y_i[f(t + x_i\Delta t) - f(t)]}{\Delta t} + o(\Delta t^m) \tag{3-58}$$

The error of the approximation is the mth order of the step Δt. It is easy to find a solution of Equation 3-57 which can be rewritten as

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \\ . & . & \cdots & . \\ x_1^{m-1} & x_2^{m-1} & \cdots & x_m^{m-1} \end{bmatrix} \begin{bmatrix} y_1 x_1 \\ y_2 x_2 \\ . \\ y_m x_m \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ . \\ 0 \end{bmatrix} \tag{3-59}$$

The first matrix is the Vandermonde matrix. It has nonzero determinant if $x_i \neq x_j$ for any $i \neq j$ and hence can be inverted. In the following, the central difference is used due to its simplicity.

Consider the Maxwell's equations

$$\begin{aligned} \nabla \times \mathbf{E} &= -\left(\frac{\partial \mathbf{B}}{\partial t} + \mathbf{J_m}\right) \\ \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J_e} \\ \nabla \cdot \mathbf{D} &= \rho \\ \nabla \cdot \mathbf{B} &= 0 \end{aligned} \tag{3-60}$$

where $\mathbf{J_m}=\sigma_m\mathbf{H}$ is the magnetic current used to model magnetic loss and $\mathbf{J_e}=\sigma_e\mathbf{E}$ is the electrical current. σ_m and σ_e are the magnetic and electric conductivity respectively. We assume $\rho=0$. We can rewrite Equation 3-60 as

$$\begin{aligned} \frac{\partial \mathbf{B}}{\partial t} &= -(\nabla \times \mathbf{E} + \mathbf{J_m}) \\ \frac{\partial \mathbf{D}}{\partial t} &= \nabla \times \mathbf{H} - \mathbf{J_e} \end{aligned} \tag{3-61}$$

Consider a typical equation:

$$\frac{\partial E_x}{\partial t} = \frac{1}{\varepsilon}\left(\frac{\partial}{\partial y}H_z - \frac{\partial}{\partial z}H_y - \sigma_e E_x\right) \tag{3-62}$$

In order to transform it to a difference equation, the space is divided into many cells. The size of a cell is $\Delta x \times \Delta y \times \Delta z$ and typically we choose $\Delta x = \Delta y = \Delta z$. The time step is Δt. Following [3-13], we define $E_x|^n_{i,j,k}$ as a shorthand notation of $E_x(i\Delta x, j\Delta y, k\Delta z, n\Delta t)$. Using central difference, the corresponding derivatives can be written as

$$\left.\frac{\partial E_x}{\partial t}\right|^n_{i,j,k} = \frac{E_x|^{n+\frac{1}{2}}_{i,j,k} - E_x|^{n-\frac{1}{2}}_{i,j,k}}{\Delta t} \tag{3-63}$$

$$\left.\frac{\partial H_z}{\partial y}\right|^n_{i,j,k} = \frac{H_z|^n_{i,j+\frac{1}{2},k} - H_z|^n_{i,j-\frac{1}{2},k}}{\Delta y}$$

$$\left.\frac{\partial H_y}{\partial z}\right|^n_{i,j,k} = \frac{H_y|^n_{i,j,k+\frac{1}{2}} - H_y|^n_{i,j,k-\frac{1}{2}}}{\Delta z}$$

Plug Equation 3-63 into Equation 3-62 and use the following approximation

$$E_x|^n_{i,j,k} = \frac{E_x|^{n+\frac{1}{2}}_{i,j,k} + E_x|^{n-\frac{1}{2}}_{i,j,k}}{2} \tag{3-64}$$

We can solve for $E_x|^{n+\frac{1}{2}}_{i,j,k}$.

$$E_x|^{n+\frac{1}{2}}_{i,j,k} = \left(\frac{1 - \frac{\sigma_e|^n_{i,j,k}\Delta t}{2\varepsilon|^n_{i,j,k}}}{1 + \frac{\sigma_e|^n_{i,j,k}\Delta t}{2\varepsilon|^n_{i,j,k}}}\right) E_x|^{n-\frac{1}{2}}_{i,j,k} \tag{3-65}$$

$$+ \left(\frac{\frac{\Delta t}{\varepsilon|^n_{i,j,k}}}{1 + \frac{\sigma_e|^n_{i,j,k}\Delta t}{2\varepsilon|^n_{i,j,k}}}\right)\left(\frac{H_z|^n_{i,j+\frac{1}{2},k} - H_z|^n_{i,j-\frac{1}{2},k}}{\Delta y} - \frac{H_y|^n_{i,j,k+\frac{1}{2}} - H_y|^n_{i,j,k-\frac{1}{2}}}{\Delta z}\right)$$

Similar equations of E_y, E_z, H_x, H_y, and H_z can be obtained. Notice that we calculate **E** at time step n+1/2 and **H** at time step n. They are interleaved temporally. E_x is calculated in the lattice points (i,j,k). H_y is calculated in the lattice points (i,j,k+1/2). H_z is calculated in the lattice points (i,j+1/2,k). They are also interleaved spatially.

A nice way to visualize the arrangement of the field vectors in space is the Yee cell as shown in Figure 3-26 introduced in 1966 by Yee [3-15], who formulated the original

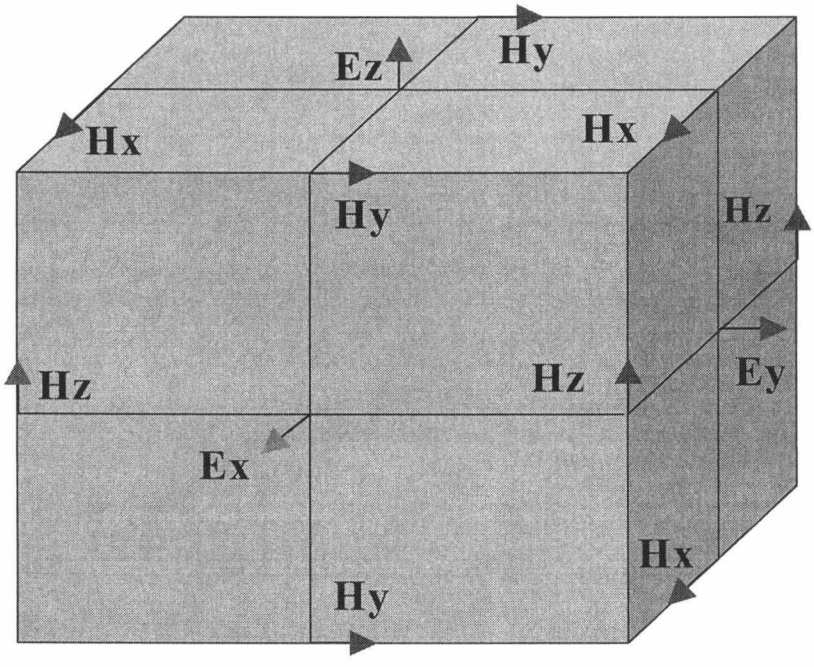


Fig. 3-26. Yee cell. Re-drawn after Fig. 3.1 of [3-13] (Allen Taflove, Computational Electrodynamics, Artech House, Boston, 1995)

FDTD algorithm. In this elegant arrangement, each magnetic field component is surrounded by four electric field components and vice versa which makes the contour integral interpretation of the algorithm very appealing. It is easy to verify that the divergences of **D** and **B** are automatically time independent in a Yee cell. Given the initial fields are divergence free, they remain divergence free all the time [3-13].

In the simulation, we interleave the field components spatially and arrange them in Yee cells. The initial condition must include the magnetic field at time zero $\mathbf{H}(\mathbf{r},0)$ and the electric field at time $t=1/2\Delta t$ $\mathbf{E}(\mathbf{r},1/2\Delta t)$. We first advance the magnetic field to time Δt using the initial conditions, and then advance the electric field to time $3/2\Delta t$ using the just calculated electrical field $\mathbf{H}(\mathbf{r},\Delta t)$ and the initial electric field $\mathbf{E}(\mathbf{r},1/2\Delta t)$. This process repeats itself.

3.7.2 Perfectly matched layer

In computer simulations, we can only simulate field evolution in a finite volume and it results in an artificial boundary which may not exist in reality. In principle, we can make the volume big enough and finish the simulation before any reflection from the artificial boundary contaminates the simulation. However, this is a waist of the computing resources. The solution to this problem is to put some sort of absorbing layer around the structure we are interested in as shown in Figure 3-27. The function of the absorbing layer

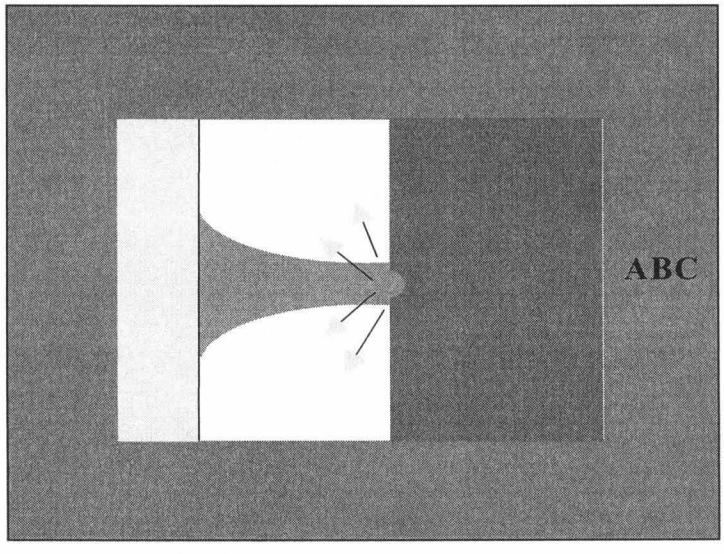


Fig. 3-27. Absorbing boundary

is to absorb any light incident on the layer while not having any reflection. Basically it is a layer with matched impedance. There are many kinds of absorbing boundary conditions (ABC), for example, Mur's ABC [3-16]. In the following we will discuss the perfectly matched layer introduced first by Berenger [3-17] which is the most efficient way to implement the ABC. The original Berenger's splitting field formalism is purely mathematical. We follow a more physical approach given in [3-14] to discuss the PML with uniaxial medium.

First we consider the 1D case, i.e., normal incidence. The material parameters belonging to the structure under study are $\{\mu_1, \varepsilon_1\}$ and those of the surrounding absorbing layer are $\{\mu_2, \varepsilon_2, \sigma_m, \sigma_e\}$. If the impedance is matched at the boundary, the normal incident plane wave enters the absorbing layer without any reflection. This leads to the following condition.

$$\sqrt{\frac{\mu_1}{\varepsilon_1}} = \sqrt{\frac{\mu_2(1-\sigma_m/j\omega\mu_2)}{\varepsilon_2(1-\sigma_e/j\omega\varepsilon_2)}} \text{ or } \mu_2 = \mu_1, \varepsilon_2 = \varepsilon_1 \text{ and } \frac{\sigma_m}{\mu_1} = \frac{\sigma_e}{\varepsilon_1} \tag{3-66}$$

More generally, we require the incident plane wave not get reflected at any incident angle. As shown in Figure 3-28, we assume a uniaxial perfectly matched layer (UPML) and the material parameters are given by

$$\mu_2 = \mu \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix}, \varepsilon_2 = \varepsilon \begin{bmatrix} s_1' & 0 & 0 \\ 0 & s_2' & 0 \\ 0 & 0 & s_3' \end{bmatrix} \tag{3-67}$$

From the apparent requirement of rotation symmetry about the x axis, we have $s_2 = s_3$ and $s_2' = s_3'$. Consider two special cases:

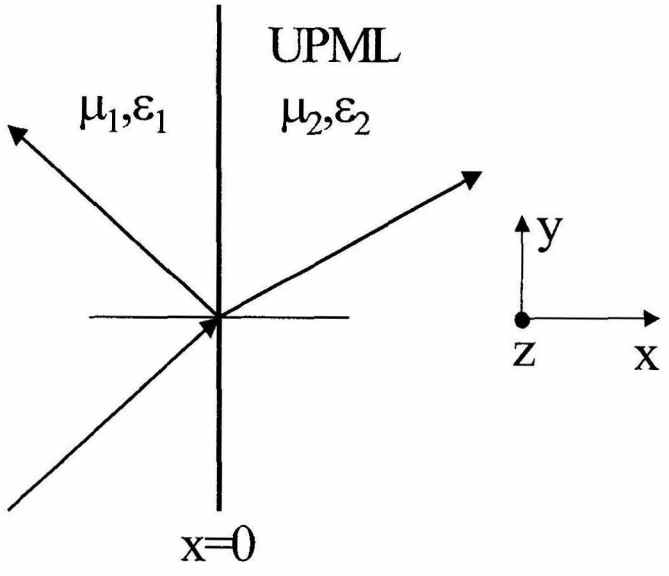


Fig. 3-28. Uniaxial perfectly matched layer

1) normal incidence

It reduces to the 1D case discussed above and therefore

$$\sqrt{\frac{\mu s_2}{\varepsilon s_2'}} = \sqrt{\frac{\mu_1}{\varepsilon_1}} \tag{3-68}$$

It is natural to require that $s_2 = s_2'$, $\mu = \mu_1$, and $\varepsilon = \varepsilon_1$.

2) Incidence angle close to 90 degrees

It again reduces to the 1D case and leads to $s_1 = s_1'$.

Equation 3-67 becomes

$$\mu_2 = \mu_1 \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_2 \end{bmatrix}, \varepsilon_2 = \varepsilon_1 \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_2 \end{bmatrix} \tag{3-69}$$

Consider an s polarized plane wave incident from the material (left side of the interface) to the UPML. In the material, we have

$$\mathbf{E} = \mathbf{e}_z[e^{j(k_x x + k_y y)} + re^{j(-k_x x + k_y y)}] \tag{3-70}$$

$$\begin{aligned}\mathbf{H} &= \mu_1^{-1}\frac{\nabla\times\mathbf{E}}{j\omega} \\ &= \mathbf{e}_x\frac{k_y}{\omega\mu_1}[e^{j(k_x x + k_y y)} + re^{j(-k_x x + k_y y)}] - \mathbf{e}_y\frac{k_x}{\omega\mu_1}[e^{j(k_x x + k_y y)} - re^{j(-k_x x + k_y y)}]\end{aligned} \tag{3-71}$$

In the UPML, we have

$$\mathbf{E} = \mathbf{e}_z te^{j(k_x' x + k_y y)} \tag{3-72}$$

$$\mathbf{H} = \mu_2^{-1}\frac{\nabla\times\mathbf{E}}{j\omega} = \frac{te^{j(k_x' x + k_y y)}}{\omega\mu_1}\left[\mathbf{e}_x\left(\frac{k_y}{s_1}\right) - \mathbf{e}_y\left(\frac{k_x'}{s_2}\right)\right] \tag{3-73}$$

where r and t are the reflection and transmission coefficients respectively. From the continuity of the tangential fields, we get

$$\begin{aligned}1 + r &= t \\ \frac{tk_x'}{s_2 k_x} &= 1 - r\end{aligned} \tag{3-74}$$

Solve for r.

$$r = \frac{t}{2}\left(1 - \frac{k_x'}{s_2 k_x}\right) \tag{3-75}$$

We need to know the dispersion relation in the UPML to find out k_x'. From Maxwell's two curl equations, wave in the UPML satisfies the following equation.

$$\nabla\times(\mu_2^{-1}\nabla\times\mathbf{E}) - \omega^2\varepsilon_2\mathbf{E} = 0 \tag{3-76}$$

Plugging in Equation 3-72, we get the following dispersion relation in the UPML.

$$(k_x')^2 = s_2^2\left(\omega^2\mu_1\varepsilon_1 - \frac{k_y^2}{s_1 s_2}\right) \tag{3-77}$$

If $s_1 s_2=1$, Equation 3-77 leads to

$$k_x' = s_2\sqrt{\omega^2\mu_1\varepsilon_1 - k_y^2} = s_2 k_x \tag{3-78}$$

where the dispersion relation in the material $k_x^2 + k_y^2 = \omega^2\mu_1\varepsilon_1$ is used.

From Equation 3-75, we get r=0 (no reflection) and this is true for any incident angle. In order to attenuate the wave in the UPML along the x direction, we can simply choose a complex s_2. The parameters of the UPML are now given by

$$\mu_2 = \mu_1 S_x,\ \varepsilon_2 = \varepsilon_1 S_x,\ \text{and} S_x = \begin{bmatrix} \frac{1}{s_x} & 0 & 0 \\ 0 & s_x & 0 \\ 0 & 0 & s_x \end{bmatrix} \tag{3-79}$$

To be consistent with the 1D case where the PML is some sort of special conductor, s_x is typically chosen as

$$s_x = \kappa_x - \frac{\sigma_x}{j\omega\varepsilon_1} \tag{3-80}$$

Notice that the PML is matched to the material with any s_x, and this means that UPML's with different s_x's are also matched with each other. UPML's matched along y and z axes are given by

$$\mu_2 = \mu_1 S_y,\ \varepsilon_2 = \varepsilon_1 S_y \text{ and } \mu_2 = \mu_1 S_z,\ \varepsilon_2 = \varepsilon_1 S_z \tag{3-81}$$

$$S_y = \begin{bmatrix} s_y & 0 & 0 \\ 0 & \frac{1}{s_y} & 0 \\ 0 & 0 & s_y \end{bmatrix}, \; S_z = \begin{bmatrix} s_z & 0 & 0 \\ 0 & s_z & 0 \\ 0 & 0 & \frac{1}{s_z} \end{bmatrix} \text{ with } s_y = \kappa_y - \frac{\sigma_y}{j\omega\varepsilon_1} \text{ and } s_z = \kappa_z - \frac{\sigma_z}{j\omega\varepsilon_1} \quad (3\text{-}82)$$

Same results can be derived for the p polarized incident plane wave. In the UPML corner as shown in Figure 3-29, it can be shown that if the parameters are given by $\mu = \mu_1 S_x S_y$

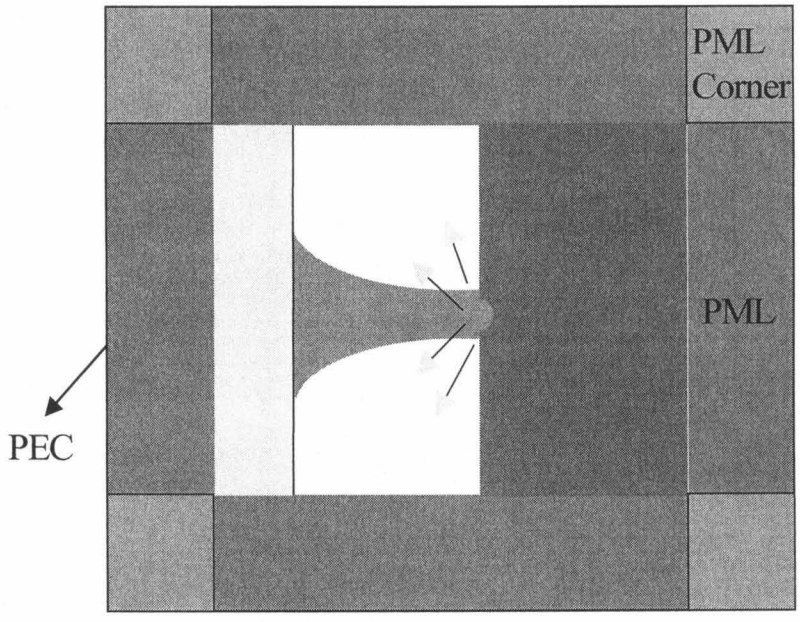


Fig. 3-29. PML corner

and $\varepsilon = \varepsilon_1 S_x S_y$ the UPML corner is matched with both neighboring UPML's whose parameters are given by $\mu_1 S_x$, $\varepsilon_1 S_x$ and $\mu_1 S_y$, $\varepsilon_1 S_y$ respectively. More generally, the UPML parameters are given by

$$\begin{aligned} \mu &= \mu_1 S_x S_y S_z \\ \varepsilon &= \varepsilon_1 S_x S_y S_z \end{aligned} \quad (3\text{-}83)$$

In the following we discuss how to implement FDTD in UPML. Consider the curl of E.

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} = j\omega\mathbf{B} = j\omega\mu_1\begin{bmatrix}\frac{s_y s_z}{s_x} & 0 & 0\\ 0 & \frac{s_x s_z}{s_y} & 0\\ 0 & 0 & \frac{s_x s_y}{s_z}\end{bmatrix}\begin{bmatrix}H_x\\ H_y\\ H_z\end{bmatrix} \tag{3-84}$$

$$= j\omega\mu_1\begin{bmatrix}s_y & 0 & 0\\ 0 & s_z & 0\\ 0 & 0 & s_x\end{bmatrix}\begin{bmatrix}H_x\frac{s_z}{s_x}\\ H_y\frac{s_x}{s_y}\\ H_z\frac{s_y}{s_z}\end{bmatrix} = j\omega\mu_1\begin{bmatrix}s_y & 0 & 0\\ 0 & s_z & 0\\ 0 & 0 & s_x\end{bmatrix}\begin{bmatrix}H_x'\\ H_y'\\ H_z'\end{bmatrix}$$

$$= j\omega\mu_1\begin{bmatrix}\kappa_y & 0 & 0\\ 0 & \kappa_z & 0\\ 0 & 0 & \kappa_x\end{bmatrix}\begin{bmatrix}H_x'\\ H_y'\\ H_z'\end{bmatrix} - \frac{\mu_1}{\varepsilon_1}\begin{bmatrix}\sigma_y & 0 & 0\\ 0 & \sigma_z & 0\\ 0 & 0 & \sigma_x\end{bmatrix}\begin{bmatrix}H_x'\\ H_y'\\ H_z'\end{bmatrix}$$

$$= -\frac{\partial}{\partial t}\mu_1\begin{bmatrix}\kappa_y & 0 & 0\\ 0 & \kappa_z & 0\\ 0 & 0 & \kappa_x\end{bmatrix}\begin{bmatrix}H_x'\\ H_y'\\ H_z'\end{bmatrix} - \frac{\mu_1}{\varepsilon_1}\begin{bmatrix}\sigma_y & 0 & 0\\ 0 & \sigma_z & 0\\ 0 & 0 & \sigma_x\end{bmatrix}\begin{bmatrix}H_x'\\ H_y'\\ H_z'\end{bmatrix}$$

$$\text{where } H_x' = H_x\frac{s_z}{s_x} \tag{3-85}$$

$$H_y' = H_y\frac{s_x}{s_y}$$

$$H_z' = H_z\frac{s_y}{s_z}$$

Consider a typical equation of the intermediate variables,

$$s_x H_x' = s_z H_x \Rightarrow \left(\kappa_x - \frac{\sigma_x}{j\omega\varepsilon_1}\right) H_x' = \left(\kappa_z - \frac{\sigma_z}{j\omega\varepsilon_1}\right) H_x \tag{3-86}$$

$$\Rightarrow \left(j\omega\kappa_x - \frac{\sigma_x}{\varepsilon_1}\right) H_x' = \left(j\omega\kappa_z - \frac{\sigma_z}{\varepsilon_1}\right) H_x$$

$$\Rightarrow \kappa_x \frac{\partial}{\partial t} H_x' + \frac{\sigma_x}{\varepsilon_1} H_x' = \kappa_z \frac{\partial H_x}{\partial t} + \frac{\sigma_z}{\varepsilon_1} H_x$$

We can then use the same methods discussed in Section 3.7.1 to transform Equation 3-84 and Equation 3-86 into difference equations.

3.7.3 Incident source implementation

In this section, we discuss the scattering/total field formalism [3-18] to introduce the incident wave source. The basic idea is illustrated in Figure 3-30. We create an artificial

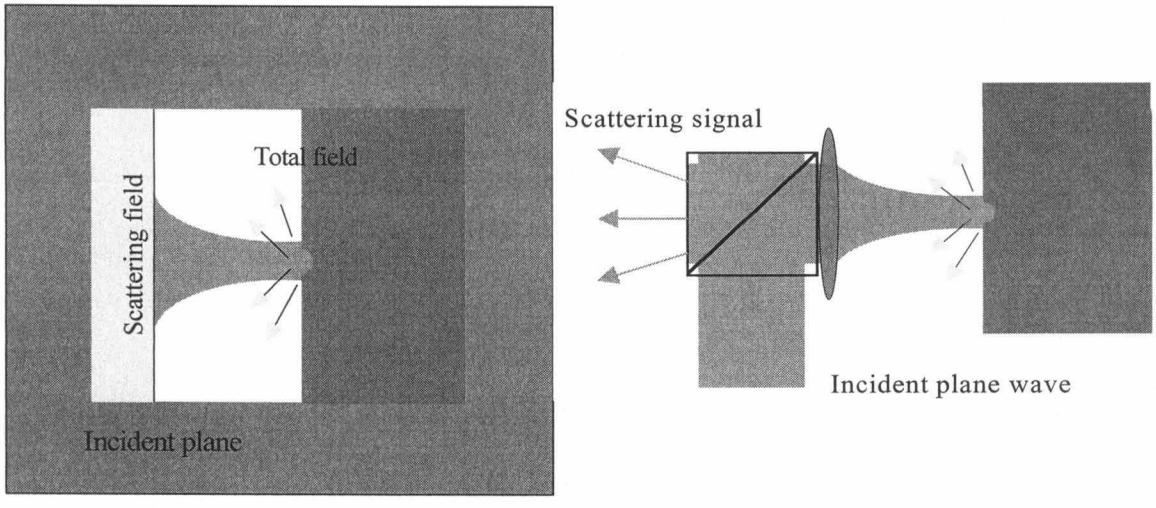


Fig. 3-30. Introducing an incident wave

plane called the incident plane. On the left side of the incident plane it is purely scattering field while on the right side (including the incident plane) it is the total field. This configuration is natural and is a close analog to the experimental setup where we introduce the incident wave and collect the scattering field using a beam splitter. Because in the structure

under study we use the total field, nonlinearity can be introduced without much difficulty. By applying the continuity condition at the incident plane (total field = scattering field + incident field), we can implement any incident wave. In the following, we give an example in 2D.

Consider a 2D s polarized ($\mathbf{E} = \mathbf{e}_z E_z$ and $\mathbf{H} = \mathbf{e}_x H_x + \mathbf{e}_y H_y$) Gaussian beam incident on a trench as shown in Figure 3-31.The incident plane is chosen at the plane $i=i_0$.

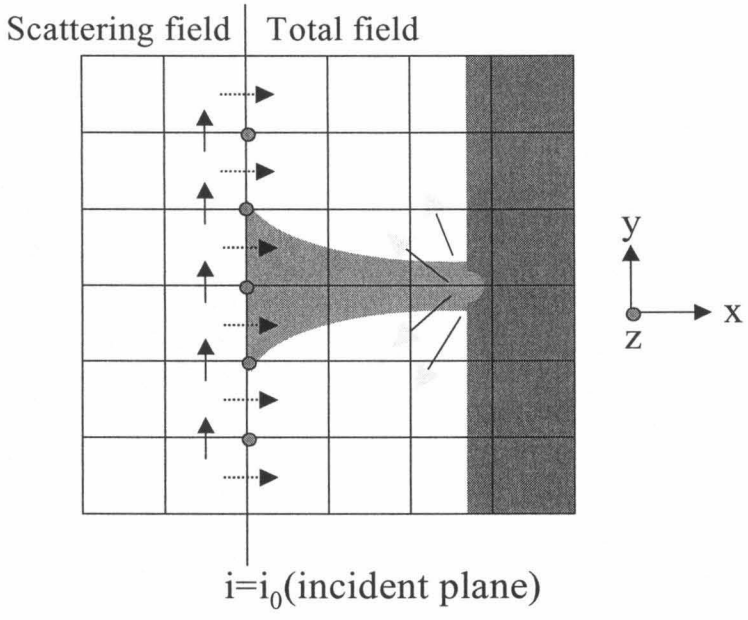


Fig. 3-31. Incident Gaussian beam

We start from the integral form of the Maxwell's equations.

$$\oint \mathbf{H} \cdot \mathbf{dl} = \iint \frac{\partial \mathbf{D}}{\partial t} \cdot \mathbf{dS} \tag{3-87}$$

Apply it to a Yee cell bisected by the incident plane, and we get

$$\left(H_x\Big|^n_{i_0,j-\frac{1}{2}} - H_x\Big|^n_{i_0,j+\frac{1}{2}}\right)\Delta x + \left(H_y\Big|^n_{i_0+\frac{1}{2},j} - H_y\Big|^n_{i_0-\frac{1}{2},j}\right)\Delta y \tag{3-88}$$
$$= \varepsilon\frac{E_z\Big|^{n+\frac{1}{2}}_{i_0,j} - E_z\Big|^{n-\frac{1}{2}}_{i_0,j}}{\Delta t}\Delta x\Delta y$$

Notice that by our convention all the terms except $H_y\Big|^n_{i_0-\frac{1}{2},j}$ in Equation 3-88 represent the total fields. We need to replace it with the total field.

$$\left(H_x\Big|^n_{i_0,j-\frac{1}{2}} - H_x\Big|^n_{i_0,j+\frac{1}{2}}\right)\Delta x + \left[H_y\Big|^n_{i_0+\frac{1}{2},j} - \left(H_y\Big|^n_{i_0-\frac{1}{2},j} + H_{y,\,inc}\Big|^n_{i_0-\frac{1}{2},j}\right)\right]\Delta y \tag{3-89}$$
$$= \varepsilon\frac{E_z\Big|^{n+\frac{1}{2}}_{i_0,j} - E_z\Big|^{n-\frac{1}{2}}_{i_0,j}}{\Delta t}\Delta x\Delta y$$

After some simplification, we get the following.

$$E_z\Big|^{n+\frac{1}{2}}_{i_0,j} = E_z\Big|^{n-\frac{1}{2}}_{i_0,j} \tag{3-90}$$

$$+\frac{\Delta t}{\varepsilon}\left\{\frac{H_x\Big|^n_{i_0,j-\frac{1}{2}} - H_x\Big|^n_{i_0,j+\frac{1}{2}}}{\Delta y} + \frac{\left[H_y\Big|^n_{i_0+\frac{1}{2},j} - \left(H_y\Big|^n_{i_0-\frac{1}{2},j} + H_{y,\,inc}\Big|^n_{i_0-\frac{1}{2},j}\right)\right]}{\Delta x}\right\}$$

where $H_{y,\,inc}\Big|^n_{i_0-\frac{1}{2},j}$ is the y component of the incident magnetic field at the plane $i=i_0-1/2$.

Consider the other integral equation.

$$\oint \mathbf{E}\cdot\mathbf{dl} = -\int\!\!\int \frac{\partial \mathbf{B}}{\partial t}\cdot\mathbf{dS} \tag{3-91}$$

Apply it to a Yee cell bisected by the plane $i=i_0-1/2$.

$$\left(E_z\Big|_{i_0,j}^{n+\frac{1}{2}} - E_z\Big|_{i_0-1,j}^{n+\frac{1}{2}}\right)\Delta z = \mu\left(\frac{H_y\Big|_{i_0-\frac{1}{2},j}^{n+1} - H_y\Big|_{i_0-\frac{1}{2},j}^{n}}{\Delta t}\right)\Delta x\Delta z \tag{3-92}$$

Notice all terms except $E_z\Big|_{i_0,j}^{n+\frac{1}{2}}$ represent the scattering fields. We need to replace it with the scattering field.

$$\left[\left(E_z\Big|_{i_0,j}^{n+\frac{1}{2}} - E_{z,\,inc}\Big|_{i_0,j}^{n+\frac{1}{2}}\right) - E_z\Big|_{i_0-1,j}^{n+\frac{1}{2}}\right]\Delta z = \mu\left(\frac{H_y\Big|_{i_0-\frac{1}{2},j}^{n+1} - H_y\Big|_{i_0-\frac{1}{2},j}^{n}}{\Delta t}\right)\Delta x\Delta z \tag{3-93}$$

We get the following.

$$H_y\Big|_{i_0-\frac{1}{2},j}^{n+1} = H_y\Big|_{i_0-\frac{1}{2},j}^{n} + \frac{\Delta t}{\mu}\frac{\left(E_z\Big|_{i_0,j}^{n+\frac{1}{2}} - E_{z,\,inc}\Big|_{i_0,j}^{n+\frac{1}{2}}\right) - E_z\Big|_{i_0-1,j}^{n+\frac{1}{2}}}{\Delta x} \tag{3-94}$$

where $E_{z,\,inc}\Big|_{i_0,j}^{n+\frac{1}{2}}$ is the incident electrical field at the plane $i=i_0$. In the above discussion we assume that the scattering fields also satisfy the Maxwell's equations which is always true in linear medium (e.g., air). In summary, in order to implement an incident wave, we need to provide information of $E_{z,\,inc}\big|_{i_0,j}$ and $H_{y,\,inc}\big|_{i_0-\frac{1}{2},j}$ or $H_{z,\,inc}\big|_{i_0,j}$ and $E_{y,\,inc}\big|_{i_0-\frac{1}{2},j}$ for the s or p polarized beam. These fields need to be known for all the time, and typically close form expressions are required. Before closing this section, we show an example of p polarized (the magnetic field is along the z direction) 2D Gaussian beam. Figure 3-32 shows the cross-section of the magnetic field at the beam center ($H_z|_{j=j0}$) along the propagation direction at some instant $t=t_0$. The incident plane is at i=25. The beam focuses at i=350 as we can tell from the increasing amplitude of H_z. After the beam waist, the beam

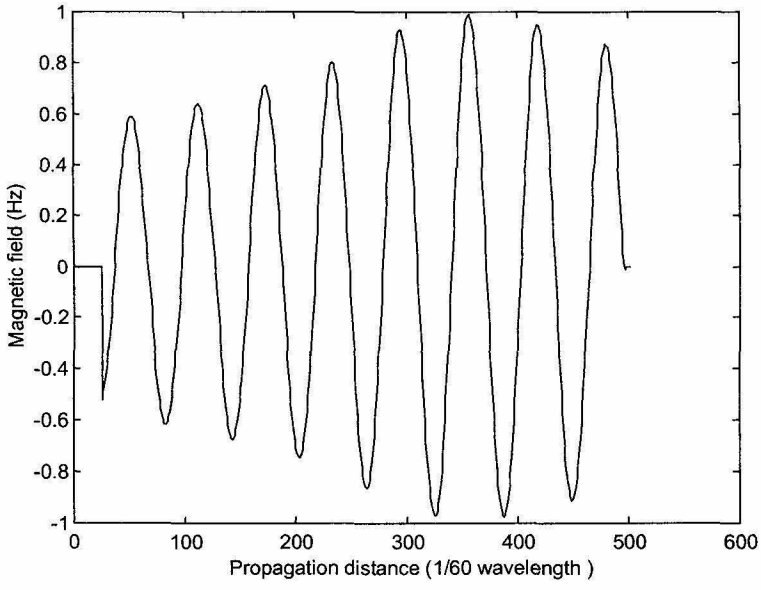


Fig. 3-32. Propagation of a 2D Gaussian beam

diverges and meets the PML. The reflection coefficient of the PML is about 10^{-6} and we do not get any scattering field at the left side of the incident plane.

3.7.4 Material modeling

In this section, we discuss how to implement various material in FDTD algorithm, particularly plasma and nonlinear material.

1. Perfect electrical conductor (PEC)

We only need to set the tangential electric field at the surface and fields inside to be zero.

2. Dielectric material

We just need to specify the parameters (ε_r, μ_r, σ_e) at each lattice point.

3. Plasma

The real metal cannot be modeled as PEC or dielectric material (ε_r is negative). We model it as plasma [3-19] whose permittivity is given by [3-4]

$$\varepsilon = \varepsilon_0\left(1 - \frac{\omega_p^2}{\omega^2}\right) \tag{3-95}$$

Consider the curl of **H**.

$$\nabla\times\mathbf{H} = \frac{\partial\mathbf{D}}{\partial t} + \mathbf{J} = -j\omega\varepsilon\mathbf{E} + \sigma\mathbf{E} = -j\omega\varepsilon_0\mathbf{E} + \sigma\mathbf{E} - \frac{\varepsilon_0\omega_p^2}{j\omega}\mathbf{E} \tag{3-96}$$
$$= \varepsilon_0\frac{\partial\mathbf{E}}{\partial t} + \sigma\mathbf{E} + \mathbf{F}$$

$$\mathbf{F} = -\frac{\varepsilon_0\omega_p^2}{j\omega}\mathbf{E} \Rightarrow -j\omega\mathbf{F} = \varepsilon_0\omega_p^2\mathbf{E} \Rightarrow \frac{\partial\mathbf{F}}{\partial t} = \varepsilon_0\omega_p^2\mathbf{E} \tag{3-97}$$

Now we can use FDTD to solve the new set of Maxwell's equations with an auxiliary variable **F**.

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \tag{3-98}$$
$$\nabla\times\mathbf{H} = \varepsilon_0\frac{\partial\mathbf{E}}{\partial t} + \sigma\mathbf{E} + \mathbf{F}$$
$$\frac{\partial\mathbf{F}}{\partial t} = \varepsilon_0\omega_p^2\mathbf{E}$$

Shown in Figure 3-33 are the reflected plane waves from PEC and NiP (n=1.98+j3.74) mirrors. We monitor the reflected electric field evolution with time at an observing point in the scattering field region (on the left of the incident plane). The incident plane wave is normal to the mirrors. The reflection from PEC is about 100% as expected. The reflection from NiP is about 81% and there is also a phase shift of $\delta\phi = 0.42$ relative to the PEC reflected wave. We can also directly calculate the reflection coefficient.

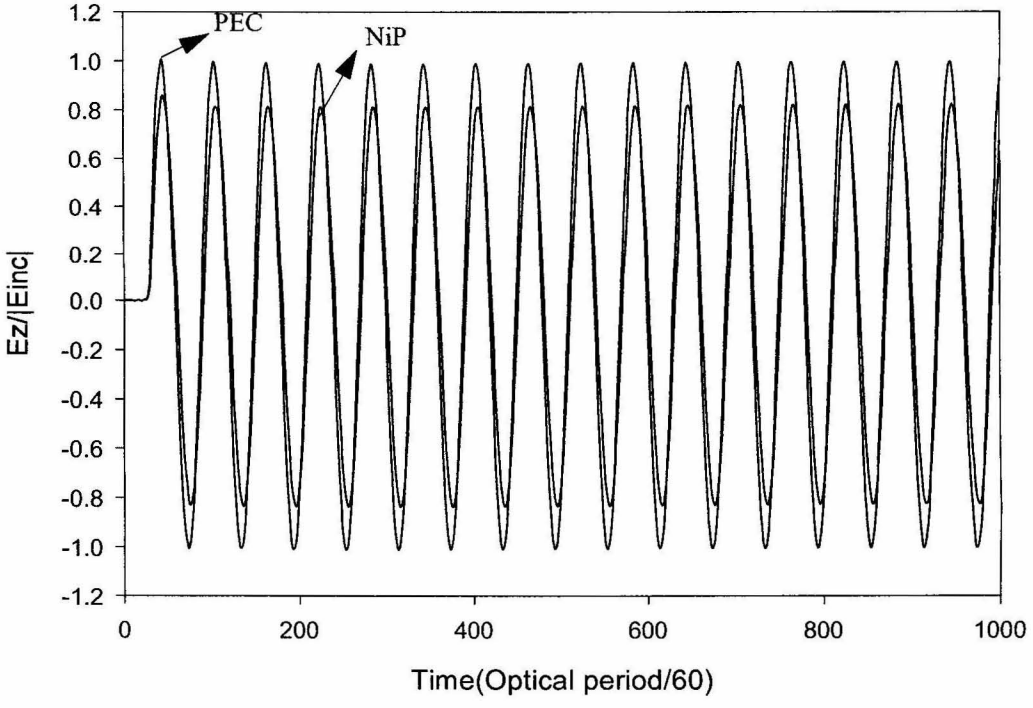


Fig. 3-33. Reflection from PEC and NiP

$r = \frac{1-n}{n+1} = -(0.7394 + j0.3271) = -0.8085e^{j0.4165}$. Notice that the PEC reflected wave has a π phase shift. The results obtained from the FDTD simulation agree very well with it.

4. Kerr nonlinearity

In general, the third order nonlinear polarizability can be written as

$$P_{NL}(t) = \varepsilon_0 \iiint \chi^{(3)}(t-\tau_1, t-\tau_2, t-\tau_3) E(\tau_1) E(\tau_2) E(\tau_3) d\tau_1 d\tau_2 d\tau_3 \qquad (3\text{-}99)$$

It is very difficult to implement a convolution with three variables. We only discuss a simple case here, the Kerr nonlinearity (chapter 9 of [3-13]) (instantaneous nonlinear response).

$$\chi^{(3)}(\tau_1, \tau_2, \tau_3) = \chi^{(3)}_{kerr} \delta(\tau_1)\delta(\tau_2)\delta(\tau_3) \qquad (3\text{-}100)$$

$$P_{NL}(t) = \varepsilon_0 \chi^{(3)}_{kerr} E^3(t) \qquad (3\text{-}101)$$

We only need to replace ε_r with $\varepsilon_r + \chi(t)$ where $\chi(t) = \chi^{(3)}_{kerr} E^2(t)$. In Figure 3-34, a 2D

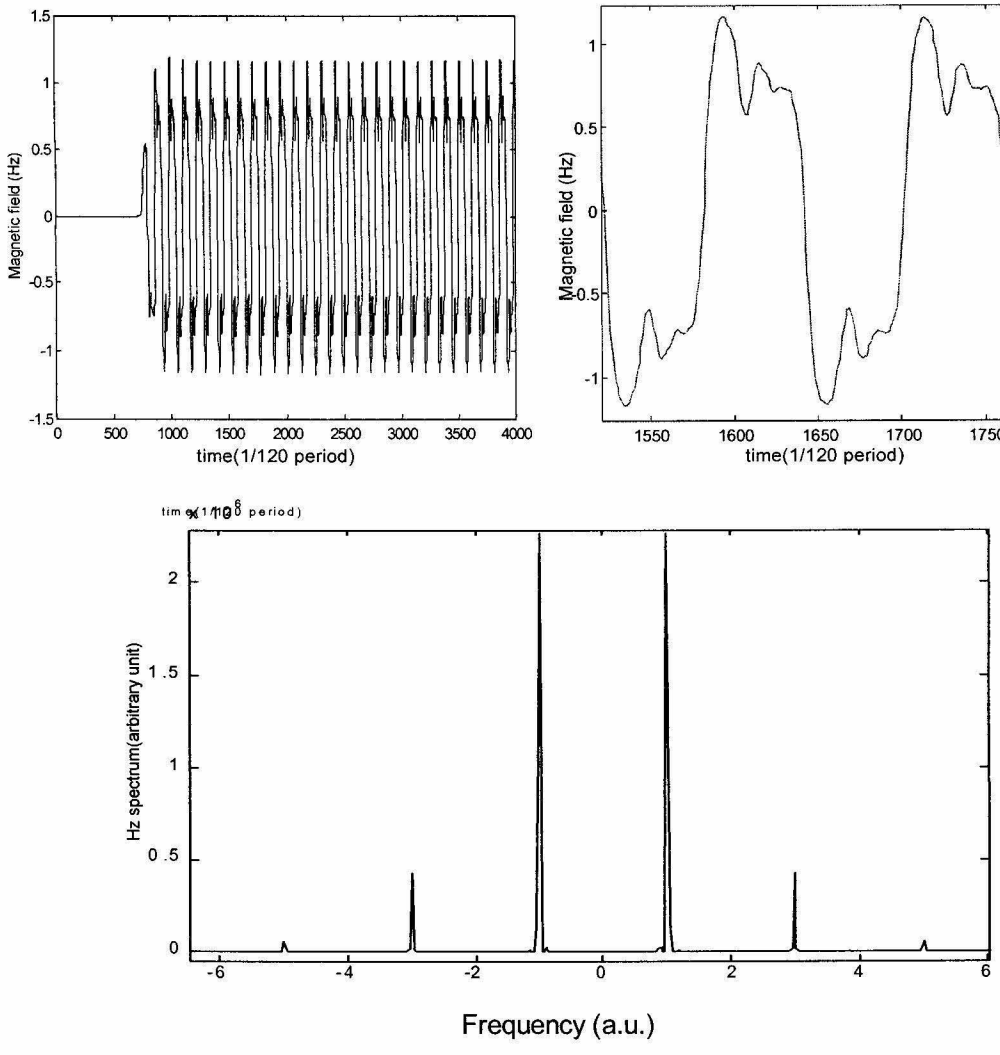


Fig. 3-34. Kerr nonlinearity

p polarized Gaussian beam is propagating in a Kerr medium. We monitor the time evolution of the magnetic field (H_z) at a lattice point. From the Fourier transform of the time evolution, we can see the wave mixing effects. New frequencies are generated in the Kerr medium.

5. Debye model

We can use Kerr medium to model intensity dependent refractive index. However, as a side effect new wavelengths are generated. The Debye model (chapter 4 of [3-9], [3-20], [3-21]) basically eliminates the problem by introducing a proper response time.

$$\tau_R \frac{d}{dt}\chi(t) + \chi(t) = \chi^{(3)}_{kerr} E^2(t) \tag{3-102}$$

If τ_R =0, i.e., instantaneous response, then the Debye model is the same as Kerr. However, by introducing a slow enough response time, the medium cannot respond to the fast optical oscillation and instead only respond to the slowly varying amplitude.

Figure 3-35 shows the time evolution of H_z at a lattice point and its Fourier transform. There is no new frequency generated by wave mixings.

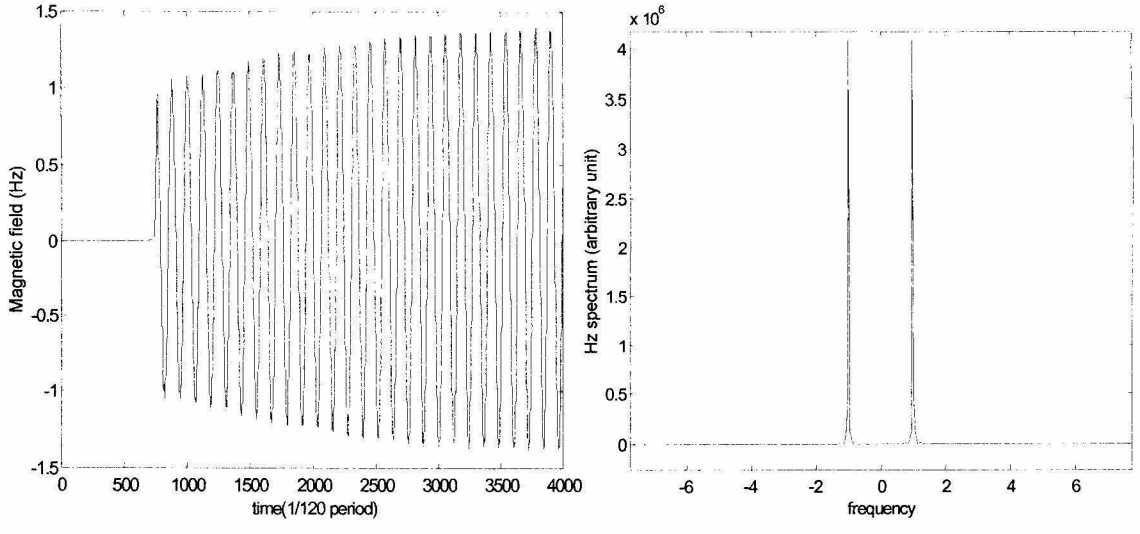


Fig. 3-35. Debye model

3.7.5 Near field to far field transformation

The near field can be calculated by the FDTD algorithm. But sometimes we are more interested in the far field scattering pattern because it is easy to measure experimen-

tally. In this section we discuss how to calculate the far field from the near field information [3-13][3-22].

We consider a 2D s polarized case as shown in Figure 3-36. The scattering field

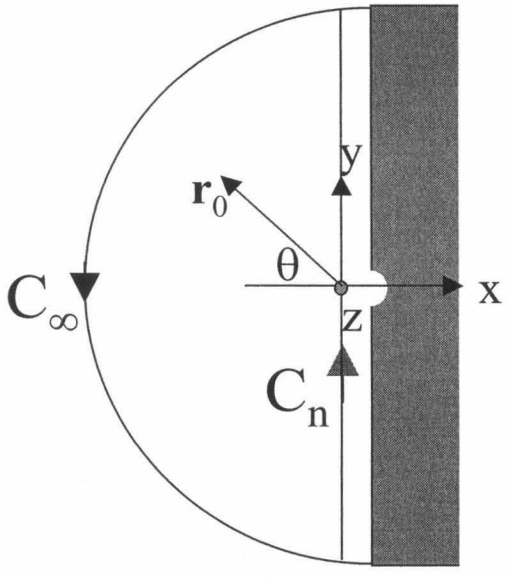


Fig. 3-36. Near field to far field transformation

along C_n (x=0) is already calculated from FDTD. The scattered electrical field satisfies the wave equation in free space.

$$\nabla^2 E_z + k^2 E_z = 0 \tag{3-103}$$

The corresponding Green's function $G(\mathbf{r},\mathbf{r}_0)$ satisfies the following equation.

$$\nabla^2 G + k^2 G = \delta(\mathbf{r} - \mathbf{r_0}) \tag{3-104}$$

We integrate (Equation 3-104 × E_z - Equation 3-103 × G) in the area surrounded by the contour and use the Green's theorem.

$$E_z(\mathbf{r_0}) = \oint_{(C_\infty + C_n)} \left[E_z(\mathbf{r}) \frac{\partial}{\partial n} G - G \frac{\partial E_z}{\partial n} \right] dl \qquad (3\text{-}105)$$

$$= \int_{C_n} \left[E_z(\mathbf{r}) \frac{\partial G}{\partial x} + j\omega\mu_0 H_y(\mathbf{r}) G \right] dy$$

In the above derivation, we assume the integration along C_∞ (the half circle with radius approaching to infinity) is zero for any physical field. n is the outward normal of the contour. Notice that

$$G(\mathbf{r}, \mathbf{r}_0) = \frac{j}{4} H_0^{(1)}(k|\mathbf{r} - \mathbf{r}_0|) \rightarrow \frac{j^{1/2}}{\sqrt{8\pi k}} \frac{e^{jk|\mathbf{r} - \mathbf{r}_0|}}{\sqrt{|\mathbf{r} - \mathbf{r}_0|}} \ (\text{as } k|\mathbf{r} - \mathbf{r}_0| \rightarrow \infty) \qquad (3\text{-}106)$$

In most cases, the near field is significant only in a small region because a focused incident beam is typically used to probe the structure. In the far field $\mathbf{r}_0 \rightarrow \infty$,

$$G(\mathbf{r}, \mathbf{r}_0) \approx \frac{j^{1/2}}{\sqrt{8\pi k r_0}} e^{jkr_0} e^{-jk(\mathbf{r} \cdot \mathbf{r}_0)/r_0} \qquad (3\text{-}107)$$

$$E_z(r_0) = j^{3/2} e^{jkr_0} \sqrt{\frac{k}{8\pi r_0}} \int_{-\infty}^{+\infty} \left[\cos\theta E_z(0, y) + \sqrt{\frac{\mu_0}{\varepsilon_0}} H_y(0, y) \right] e^{-jky\sin\theta} dy \qquad (3\text{-}108)$$

where θ is the angle between $\mathbf{r}_0$ and -x axis. In order to get the far field we only need to calculate the tangential near fields (surface currents). Typically we use the real fields instead of their complex representations in the FDTD algorithm. We can implement an on-the-fly digital Fourier transform to calculate the complex field amplitudes.

$$E_z(0, m\Delta y) = \sum_n E_z\Big|_{0, m\Delta y}^{(n+1/2)\Delta t} e^{j\omega(n+1/2)\Delta t} \qquad (3\text{-}109)$$

$$H_y(0, m\Delta y) = \sum_n \left[\left(H_y\Big|_{-\frac{1}{2}\Delta x, m\Delta y}^{n\Delta t} + H_y\Big|_{\frac{1}{2}\Delta x, m\Delta y}^{n\Delta t} \right) / 2 \right] e^{j\omega n\Delta t}$$

where we use the neighboring values to approximate the magnetic field at the plane x=0.

For the *p* polarized case.

$$H_z(r_0) = j^{3/2} e^{jkr_0} \sqrt{\frac{k}{8\pi r_0}} \int_{-\infty}^{+\infty} \left[\cos\theta H_z(0, y) - \sqrt{\frac{\varepsilon_0}{\mu_0}} E_y(0, y) \right] e^{-jky\sin\theta} dy \qquad (3\text{-}110)$$

Actually, in the 2D case, we can simplify the results by choosing another Green's function (see chapter 3 of [3-22]).

$$G' = \frac{1}{2}[G(\mathbf{r}, \mathbf{r}_0) - G(\mathbf{r}, \bar{\mathbf{r}}_0)] \qquad (3\text{-}111)$$

where $\bar{\mathbf{r}}_0$ is the mirror image of $\mathbf{r}_0$ about the plane x=0. It is easy to check that

$$G'|_{x=0} = 0 \qquad (3\text{-}112)$$

$$\frac{\partial}{\partial x} G' = 2\frac{\partial G}{\partial x}$$

Then we get the following.

$$E_z(r_0) = 2j^{3/2} e^{jkr_0} \sqrt{\frac{k}{8\pi r_0}} \cos\theta \int_{-\infty}^{+\infty} E_z(0, y) e^{-jky\sin\theta} dy \text{ for the s polarization} \qquad (3\text{-}113)$$

$$H_z(r_0) = 2j^{3/2} e^{jkr_0} \sqrt{\frac{k}{8\pi r_0}} \cos\theta \int_{-\infty}^{+\infty} H_z(0, y) e^{-jky\sin\theta} dy \text{ for the p polarization} \qquad (3\text{-}114)$$

These are just the Fourier transforms of the near field multiplied by a Lambertian factor. In the 3D case, we can repeat the same analysis for each component. Please refer to [3-3] and [3-13] for more detailed discussion.

References

[3-1] Xu Wang et al. "Measuring and modeling optical diffraction from subwavelength features." J. OPT. SOC. AM. A 18, 565-572 (2001)

[3-2] Xu Wang. Caltech Ph.D. thesis (2000)

[3-3] David S. Marx. Caltech Ph.D. thesis (1996)

[3-4] Course notes of EE 152 (Engineering Electromagnetics) taught by Dr. William B. Bridges. See also other electromagnetics references, for example, J.D. Jackson. Classical Electrodynamics, 2nd Edition. (John Wiley, New York 1975)

[3-5] T.W. Ebbesen, H.J. Lezec, H.F. Ghaemi, T. Thio, and P.A. Wolff. "Extraordinary optical transmission through sub-wavelength hole arrays." Nature 391, 667-669 (1998)

[3-6] C.W. Barnes. "Object restoration in a diffraction limited imaging system." J. Opt. Soc. Am., 56:575 (1966)

[3-7] Amnon Yariv. Quantum Electronics, Third Edition (John Wiley & Sons, New York, 1989)

[3-8] Y. R. Shen. The Principles of Nonlinear Optics, 303-333 (John Wiley & Sons, 1984)

[3-9] Alan C. Newell and Jerome V. Moloney. Nonlinear Optics (Addison-Wesley, Redwood City 1992)

[3-10] Govind P. Agrawal. Nonlinear Fiber Optics, 2nd edition. 50-54 (Academic Press, San Diego 1995)

[3-11] I.C. Khoo. Liquid Crystals (John Wiley & Sons, New York 1995)

[3-12] Tsuyoshi Tsujioka and Masahiro Irie. "Spot shape on super-resolution optical disks with a photon-mode mask layer." Optical Review 5, 158-162 (1998)

[3-13] Allen Taflove. Computational Electrodynamics (Artech House, Boston, 1995)

[3-14] Allen Taflove (editor). Advances in Computational Electrodynamics, Chapter 5 (Artech House, Boston, 1998)

[3-15] K.S. Yee. "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media." IEEE Transactions on Antennas and Propagation 14, 302-307 (1966)

[3-16] G. Mur. "Absorbing boundary conditions for the finite-difference approximation of the time domain electromagnetic field equations." IEEE Transactions on Electromagnetic Compatibility 23, 377-382 (1981)

[3-17] J.P. Berenger. "A perfectly matched layer for the absorption of electromagnetic waves." J. Computational Physics 114, 185-200 (1994)

[3-18] Korada Umashankar and Allen Taflove. "A novel method to analyze electromagnetic scattering of complex objects." IEEE Transactions on Electromagnetic Compatibility, EMC-24, 397-405 (1982)

[3-19] Steven A. Cummer. "An analysis of new and existing FDTD methods for isotropic cold plasma and a method for improving their accuracy." IEEE Transactions on Antennas and Propagation 45, 392-400 (1997)

[3-20] Richard W. Ziolkowski and Justin B. Judkins. "Applications of the nonlinear finite difference time domain (NL-FDTD) method to pulse propagation in nonlinear media: Self-focusing and linear-nonlinear interfaces." Radio Science 28, 901-911 (1993)

[3-21] Rose M. Joseph and Allen Taflove. "FDTD Maxwell's equations models for nonlinear electrodynamics and optics." IEEE Transactions on Antennas and Propagation 45, 364-374 (1997)

[3-22] J. Goodman. Introduction to Fourier Optics (second edition) (McGraw-Hill, New York, 1996)

4 Defocused Holographic Correlator

In the previous chapters we investigated some topics in optical information storage. From now on, we will discuss optical information processing, or more specifically, holographic correlators and their applications. In this chapter, the defocused holographic correlator is investigated both experimentally and theoretically. An approximate formula of the shift selectivity is derived. The cross talk in correlator arrays is analyzed. A real time correlator system with 480 templates is demonstrated using 100 μm thick Dupont photo-polymer.

4.1 Introduction

Holographic correlator arrays are powerful tools for the pattern recognition problems where correlation is an efficient algorithm because of their massive parallelism and high speed [4-1]-[4-4]. A typical correlator system is shown in Figure 4-1. The input image is first edge-enhanced by filtering its DC component in the Fourier plane. All higher orders [4-5] except the zeroth order diffraction of the spatial light modulator (SLM) are also filtered. The enhanced edge profile of the SLM is then filtered in the image plane using a rectangular filter. These filtering steps are necessary in order to remove the common features of the inputs and improve the contrast of correlation outputs. The filtered image is Fourier transformed by another lens and the recording material is put either in the focal plane of the lens (normal correlator) or slightly defocused (defocused correlator [4-6]). During recording, the direction of the signal beam remains the same while the reference beam angle

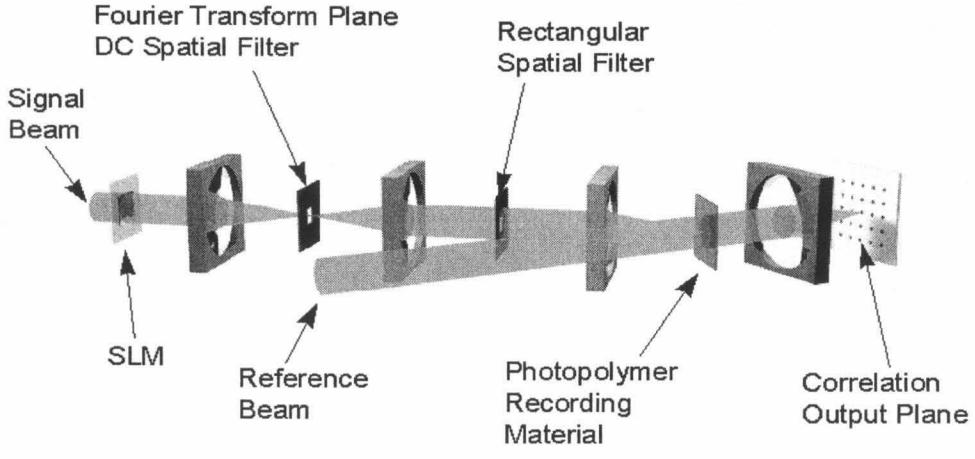


Fig. 4-1. Holographic correlator array (courtesy of Dr. Gregory J. Steckman)

changes to record an array of templates using angular multiplexing [4-7]. Upon readout, an image is displayed on the SLM and it correlates with all the templates simultaneously at the speed of light. If the input is one of the template images, the corresponding reference is reconstructed and focused on the correlation output plane producing a bright correlation peak. A normal correlator is shift invariant if the recording material is thin. A shift of the input image simply results in a corresponding shift of the correlation peak and this makes the system more robust in recognizing the input pattern.

The correlation plane is divided into domains as shown in Figure 4-2.The ith reference beam used to record the ith template focuses at the center of the ith domain and the size of the domain depends on the amount of shift invariance of the system and is chosen such that a correlation peak inside the domain can be attributed to the correlation with the

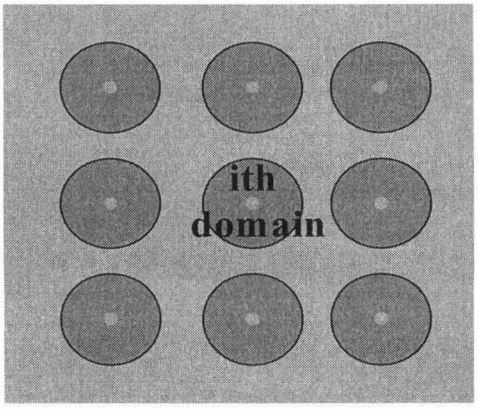


Fig. 4-2. Correlation domains

corresponding template without any ambiguity. Several correlation peaks may appear on the correlation plane. The brightness tells us about the similarity between the input image and the corresponding template. The input is classified to the ith class if the brightest correlation peak lies in the ith domain. There is a trade-off between the number of templates packed and the amount of shift invariance allowed in the system design, because an ambiguity arises when a correlation peak shifts into neighboring domains when the input is shifted. Shift invariance can be reduced by the angular selectivity of volume holograms, but a relatively thick recording medium is required. A technique is proposed in [4-6] to control the shift invariance simply by defocusing the recording material. In this chapter we study the properties of the defocused correlator. In Section 4.3, we simplify the theory and derive an approximate formula of the shift selectivity. In Section 4.4 we investigate the cross talk in correlator arrays. In Section 4.5, we demonstrate a real time correlator system with 480 templates. In Section 4.6, we discuss some interesting properties of the defocused correlator.

4.2 Defocused holographic correlator array

A defocused correlator is shown in Figure 4-3. The recording material is moved

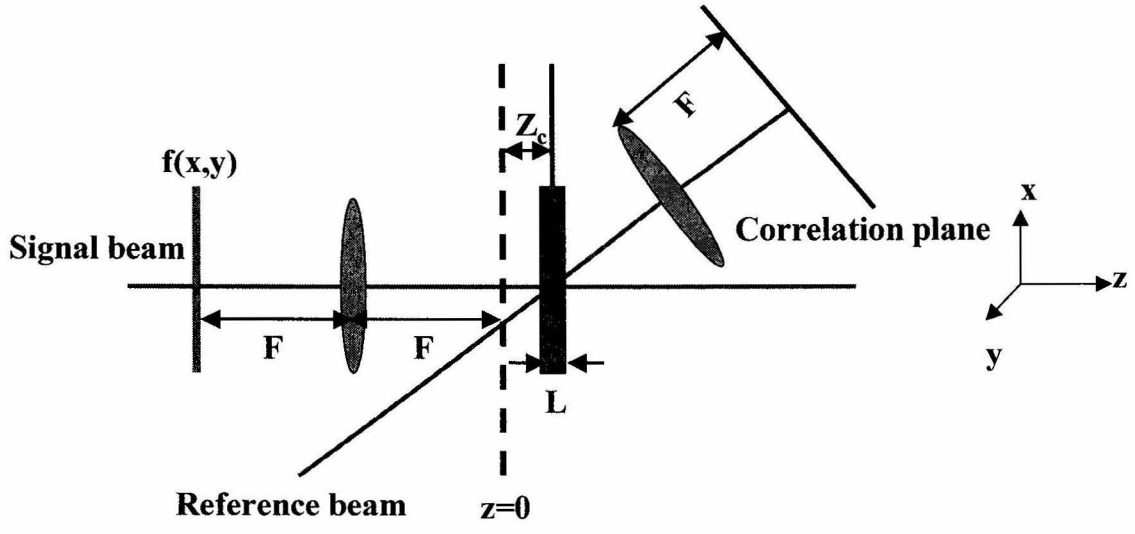


Fig. 4-3. Defocused holographic correlator

away from the focal plane of the Fourier lens and the defocusing distance is z_c. The disturbance at the focal plane of the lens(z=0) is proportional to $G\left(\frac{x}{\lambda F}, \frac{y}{\lambda F}\right)$, where G(u,v) is the Fourier transform of the input image f(x,y) and is given by $G(u, v) = \iint f(x, y)e^{-j2\pi(ux + vy)}dxdy$. The disturbance g(x,y,z) after the focal plane can be obtained from scalar diffraction theory and is given by.

$$g(x, y, z) \propto \left\{ G\left(\frac{x}{\lambda F}, \frac{y}{\lambda F}\right) \otimes \exp\left[j\frac{\pi}{\lambda z}(x^2 + y^2)\right] \right\} e^{jkz} \qquad (4\text{-}1)$$

where $\otimes$ is the convolution operator. By taking the Fourier transform to convert the convolution to a product, we can rewrite g(x,y,z) as

$$g(x, y, z) \propto \iint f(-\lambda Fu, -\lambda Fv)\exp[-j\pi\lambda z(u^2 + v^2)]e^{jkz}e^{j2\pi(ux + vy)}dudv \qquad (4\text{-}2)$$

A hologram is recorded using a plane wave reference $e^{j\mathbf{k}_r \cdot \mathbf{r}}$, and the index grating is given by $\Delta\varepsilon \propto |e^{j\mathbf{k}_r \cdot \mathbf{r}} + g(x, y, z)|^2$. The component that is of interest for correlators is $g^*(x, y, z)e^{j\mathbf{k}_r \cdot \mathbf{r}}$. During readout, an image $f'(x, y)$ is displayed to read out the hologram. The scattering amplitude in the $\mathbf{k_d}$ direction is given by the Born approximation [4-8].

$$
\begin{aligned}
A_d(\mathbf{k_d}) &\propto \langle \mathbf{k_d}|\Delta\varepsilon|g'(x, y, z)\rangle \propto \iiint g^* g' e^{j[(\mathbf{k}_r - \mathbf{k}_d)\cdot \mathbf{r}]}dxdydz \\
&\propto \int dz \iint f^*(-\lambda F u, -\lambda F v)\exp[j\pi\lambda z(u^2 + v^2)]dudv \times \\
&\iint f'(-\lambda F u', -\lambda F v')\exp[-j\pi\lambda z(u'^2 + v'^2)]du'dv' \times \\
&\iint e^{-j2\pi(ux + vy)} e^{j2\pi(u'x + v'y)} e^{j[(\mathbf{k}_r - \mathbf{k}_d)\cdot \mathbf{r}]}dxdy
\end{aligned}
\tag{4-3}
$$

where $g'(x, y, z)$ is the disturbance in the holographic medium generated by $f'(x, y)$.

We assume that the recording material extends to the infinity in the x and y directions. Integrating x and y first, we get the following.

$$
A_d(\mathbf{k_d}) = C\iint f^*(x, y)f'(x + \Delta x, y + \Delta y)e^{j\alpha z_c}\mathrm{sinc}\left(\frac{L\alpha}{2\pi}\right)dxdy \tag{4-4}
$$

where $\mathrm{sinc}(x)$ is defined as $\dfrac{\sin\pi x}{\pi x}$ (4-5)

$$
\Delta x = \lambda F\frac{k_{rx} - k_{dx}}{2\pi}
$$

$$
\Delta y = \lambda F\frac{k_{ry} - k_{dy}}{2\pi}
$$

$$
\alpha = -\frac{\pi}{\lambda F^2}[(\Delta x)^2 + (\Delta y)^2 + 2x\Delta x + 2y\Delta y] + (k_{rz} - k_{dz})
$$

L is the thickness of the material, λ is the wavelength of the laser beam, and C is a constant factor taking care of all the proportional constants. In the following C is omitted.

If $\mathbf{k_d}$=$\mathbf{k_r}$, we have Δx=Δy=0 and this defines the center of a correlation domain. Δx and Δy can be viewed as the position relative to the domain center. In the above derivation we ignored the reflection and refraction at the surface of the material. A more rigorous der-

ivation considering the refraction is given in [4-6]. The Fourier plane appears shifted and z_c needs to be replaced by an effective defocusing distance. The scattering amplitude is given in [4-6]

$$A_d(\mathbf{k_d}) = \iint f^*(x, y) f'(x + \Delta x, y + \Delta y) e^{j\alpha\left[z_c + \left(z_c - \frac{L}{2}\right)(n-1)\right]} \operatorname{sinc}\left(\frac{L\alpha}{2\pi}\right) dx dy \qquad (4\text{-}6)$$

$$\text{where } \Delta x = \frac{\lambda}{n} F \frac{k_{rx} - k_{dx}}{2\pi} \qquad (4\text{-}7)$$

$$\Delta y = \frac{\lambda}{n} F \frac{k_{ry} - k_{dy}}{2\pi}$$

$$\alpha = -\frac{\pi}{\lambda F^2}[(\Delta x)^2 + (\Delta y)^2 + 2x\Delta x + 2y\Delta y] + n(k_{rz} - k_{dz})$$

n is the refractive index of the recording material. λ, $\mathbf{k_d}$ and $\mathbf{k_r}$ are values in the air.

If the recording material is thin ($L \approx 0$) and put in the focal plane (z_c=0), Equation 4-6 is simplified to $A_d(\mathbf{k_d}) = \iint f^*(x, y) f'(x + \Delta x, y + \Delta y) dx dy$, which is a cross correlation and is fully shift invariant. The selectivity of volume holograms (the sinc term in Equation 4-6) and the phase term induced by the defocusing reduce the shift invariance. It is interesting to point out that the selectivity of volume holograms can also be viewed as a special defocusing effect. The volume hologram can be divided into many thin slices and from Equation 4-4 we have $A_d(\mathbf{k_d}, z_c) = \iint f^*(x, y) f'(x + \Delta x, y + \Delta y) e^{j\alpha z_c} dx dy$ for the slice at z_c. Summing up all the contributions we get the following.

$$A_d(\mathbf{k_d}) = \int_{-\frac{L}{2}}^{\frac{L}{2}} A_d(k_d, z_c) dz_c = \iint f^*(x, y) f'(x + \Delta x, y + \Delta y) \int_{-\frac{L}{2}}^{\frac{L}{2}} e^{j\alpha z_c} dz_c dx dy \qquad (4\text{-}8)$$

$$= L \iint f^*(x, y) f'(x + \Delta x, y + \Delta y) \operatorname{sinc}\left(\frac{L\alpha}{2\pi}\right) dx dy$$

where in order to prevent any artificial air/material interface, we use Equation 4-4 instead of Equation 4-6.

4.3 Shift selectivity

Assume that a template image $f_1(x,y)$ has been stored holographically in a defocused correlator and it is read out with a new input image $f_2(x,y)$. If the recording material is thin, e.g., a 38μm-thick layer of Dupont polymer, its thickness can be ignored and as derived in Section 4.2 the diffraction intensity distribution in the correlator output plane is given by

$$I_d(\Delta x, \Delta y) = \lim_{L \to 0} |A_d(\mathbf{k_d})|^2 \qquad (4\text{-}9)$$
$$= \left| \iint f_1^*(x, y) f_2(x + \Delta x, y + \Delta y) e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} dx dy \right|^2$$

In the particular case where $f_2(x,y)$ is a shifted version of $f_1(x,y)$, i.e., $f_2(x,y)=f_1(x-\delta x,y-\delta y)$, we can define a selectivity function.

$$S(\delta x, \delta y) = \max\{I_d(\Delta x, \Delta y)\} \qquad (4\text{-}10)$$
$$= \max\left\{ \left| \iint f_1^*(x, y) f_1(x - \delta x + \Delta x, y - \delta y + \Delta y) e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} dx dy \right|^2 \right\}$$

The selectivity function measures the dependence of the correlation peak intensity on the image shift. Figure 4-4 shows a typical measured selectivity function of a random binary pattern. Selectivity is defined either as the FWHM (full width half maximum) or as the first null of the selectivity function. In the following analysis we assume the size of the image is $W_x \times W_y$ and the pixel size is $d_x \times d_y$.

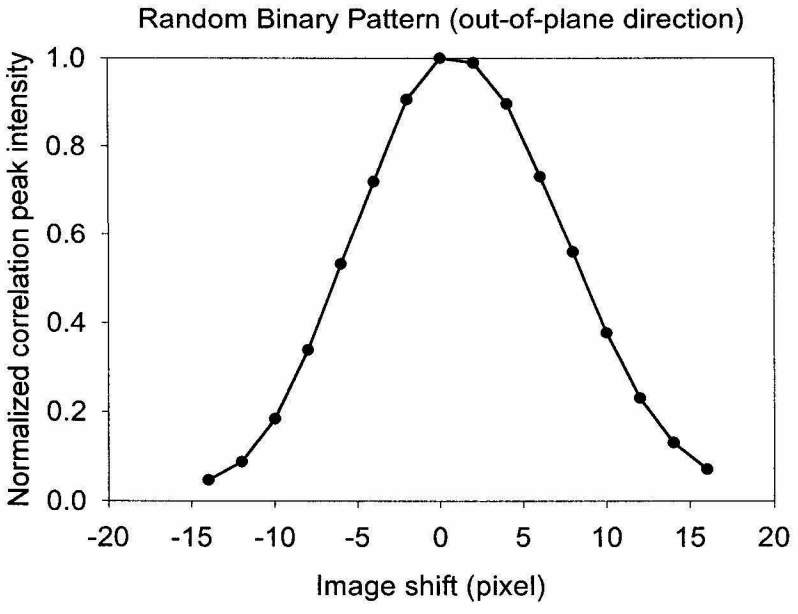


Fig. 4-4. Selectivity function

4.3.1 Gaussian process

Equation 4-9 can be rewritten as

$$I_d(\Delta x,\Delta y) = \frac{1}{4}\int_{-\frac{W_x}{2}}^{\frac{W_x}{2}}\int_{-\frac{W_y}{2}}^{\frac{W_y}{2}} dxdy e^{-j\frac{kz_c}{F^2}(\Delta x x+\Delta y y)}\int_{-(W_x-|x|)}^{(W_x-|x|)}\int_{-(W_y-|y|)}^{(W_y-|y|)} dXdY \qquad (4\text{-}11)$$

$$f_1^*\left(\frac{x+X}{2},\frac{y+Y}{2}\right)f_2\left(\frac{x+X}{2}+\Delta x,\frac{y+Y}{2}+\Delta y\right)f_1\left(\frac{X-x}{2},\frac{Y-y}{2}\right)$$

$$f_2^*\left(\frac{X-x}{2}+\Delta x,\frac{Y-y}{2}+\Delta y\right) = \int_{-\frac{W_x}{2}}^{\frac{W_x}{2}}\int_{-\frac{W_y}{2}}^{\frac{W_y}{2}} dxdy e^{-j\frac{kz_c}{F^2}(\Delta x x+\Delta y y)}(W_x-|x|)(W_y-|y|)$$

$$\frac{1}{(W_x-|x|)(W_y-|y|)}\int_{-(W_x-|x|)}^{(W_x-|x|)}\int_{-(W_y-|y|)}^{(W_y-|y|)} d\frac{X}{2}d\frac{Y}{2}f_1^*\left(\frac{x+X}{2},\frac{y+Y}{2}\right)$$

$$f_2\left(\frac{x+X}{2}+\Delta x,\frac{y+Y}{2}+\Delta y\right)f_1\left(\frac{X-x}{2},\frac{Y-y}{2}\right)f_2^*\left(\frac{X-x}{2}+\Delta x,\frac{Y-y}{2}+\Delta y\right)$$

We assume that $f_1(x_1,y_1)$ and $f_2(x_1,y_1)$ are ergodic stationary joint Gaussian processes.

$$I_d(\Delta x, \Delta y) \cong \int_{-\frac{W_x}{2}}^{\frac{W_x}{2}} \int_{-\frac{W_y}{2}}^{\frac{W_y}{2}} dxdy e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} (W_x - |x|)(W_y - |y|) \tag{4-12}$$

$$E[f_1^*(x_1, y_1) f_2(x_1 + \Delta x, y_1 + \Delta y) f_1(x_1 + x, y_1 + y) f_2^*(x_1 + \Delta x + x, y_1 + \Delta y + y)]$$

where E[ ] is the ensemble average. For Gaussian processes, all statistical properties are completely determined by the first and second order momentum [4-9]. We assume real and edge enhanced images (zero mean).

$$\begin{aligned} &E[\ldots] = \qquad (4\text{-}13)\\ &E[f_1^*(x_1, y_1) f_2(x_1 + \Delta x, y_1 + \Delta y)] E[f_1(x_1 + x, y_1 + y) f_2^*(x_1 + \Delta x + x, y_1 + \Delta y + y)]\\ &+ E[f_1^*(x_1, y_1) f_1(x_1 + x, y_1 + y)] E[f_2(x_1 + \Delta x, y_1 + \Delta y) f_2^*(x_1 + \Delta x + x, y_1 + \Delta y + y)]\\ &+ E[f_1^*(x_1, y_1) f_2^*(x_1 + \Delta x + x, y_1 + \Delta y + y)] E[f_1(x_1 + x, y_1 + y) f_2(x_1 + \Delta x, y_1 + \Delta y)]\\ &= [C_{f_1 f_2}(\Delta x, \Delta y)]^2 + C_{f_1 f_1}(x, y) C_{f_2 f_2}(x, y) + C_{f_1 f_2}(x + \Delta x, y + \Delta y) C_{f_1 f_2}(x - \Delta x, y - \Delta y) \end{aligned}$$

where $C_{fg}(x,y)$ is the cross correlation of f(x,y) and g(x,y).

$$\begin{aligned} I_d(\Delta x, \Delta y) &\cong W_x^2 W_y^2 \operatorname{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right) \operatorname{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) [C_{f_1 f_2}(\Delta x, \Delta y)]^2 \qquad (4\text{-}14)\\ &+ \int_{-\frac{W_x}{2}}^{\frac{W_x}{2}} \int_{-\frac{W_y}{2}}^{\frac{W_y}{2}} dxdy e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} (W_x - |x|)(W_y - |y|)\\ &[C_{f_1 f_1}(x, y) C_{f_2 f_2}(x, y) + C_{f_1 f_2}(x + \Delta x, y + \Delta y) C_{f_1 f_2}(x - \Delta x, y - \Delta y)] \end{aligned}$$

If we consider a special case in which $f_2(x,y)=f_1(x-\delta x, y-\delta y)$, and we assume that the auto-correlation function of $f_1(x,y)$ is a Gaussian,

$$C_{f_1 f_1}(x, y) = Ce^{-\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)} \tag{4-15}$$

then we have the following.

$$I_d(\Delta x, \Delta y) \cong W_x^2 W_y^2 \mathrm{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right) \mathrm{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) C^2 e^{-2\left(\frac{(\delta x-\Delta x)^2}{\sigma_x^2}+\frac{(\delta y-\Delta y)^2}{\sigma_y^2}\right)} \quad (4\text{-}16)$$

$$+ \int_{-\frac{W_x}{2}}^{\frac{W_x}{2}} \int_{-\frac{W_y}{2}}^{\frac{W_y}{2}} dx dy e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} (W_x - |x|)(W_y - |y|)$$

$$\left\{ C^2 e^{-2\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} + C^2 e^{-\left[\frac{(x-\delta x+\Delta x)^2}{\sigma_x^2}+\frac{(y-\delta y+\Delta y)^2}{\sigma_y^2}\right]} e^{-\left[\frac{(x-\delta x-\Delta x)^2}{\sigma_x^2}+\frac{(y-\delta y-\Delta y)^2}{\sigma_y^2}\right]} \right\}$$

$$\cong W_x^2 W_y^2 C^2 \left\{ \mathrm{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right) \mathrm{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) e^{-2\left[\frac{(\delta x-\Delta x)^2}{\sigma_x^2}+\frac{(\delta y-\Delta y)^2}{\sigma_y^2}\right]} \right.$$

$$\left. + \frac{\pi}{2}\frac{\sigma_x \sigma_y}{W_x W_y}\left[1 + e^{-j\frac{kz_c}{F^2}(\Delta x \delta x + \Delta y \delta y)} e^{-2\left(\frac{\Delta x^2}{\sigma_x^2}+\frac{\Delta y^2}{\sigma_y^2}\right)}\right] e^{-\frac{\pi^2}{2}[(\sigma_x^2 z_c^2 \Delta x^2)/(\lambda^2 F^4) + (\sigma_y^2 z_c^2 \Delta y^2)/(\lambda^2 F^4)]} \right\}$$

In the above derivation we assume a sharp autocorrelation peak, i.e., $\sigma_x \ll W_x$ and $\sigma_y \ll W_y$. Since the significant contribution of the integral is from the region $|x| \sim \sigma_x$ and $|y| \sim \sigma_y$, the integration limit is extended to infinity and the approximation $(W_x - |x|) \approx W_x$ and $(W_y - |y|) \approx W_y$ are made. For small z_c, δx, δy the first term in Equation 4-16 dominates.

$$I_d(\Delta x, \Delta y) \cong W_x^2 W_y^2 C^2 \mathrm{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right) \mathrm{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) e^{-2\left[\frac{(\delta x-\Delta x)^2}{\sigma_x^2}+\frac{(\delta y-\Delta y)^2}{\sigma_y^2}\right]} \quad (4\text{-}17)$$

In general if the auto-correlation peak is sharp, we have the following.

$$I_d(\Delta x, \Delta y) \cong W_x^2 W_y^2 \mathrm{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right) \mathrm{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) [C_{f_1 f_1}(\Delta x - \delta x, \Delta y - \delta y)]^2 \quad (4\text{-}18)$$

This shows that the correlation peak is shifted to ($\Delta x=\delta x$, $\Delta y=\delta y$) and the peak intensity drops. The selectivity is $s_x \approx \frac{\lambda F^2}{z_c W_x}$ and $s_y \approx \frac{\lambda F^2}{z_c W_y}$. The system is approximately shift invariant if $|\delta x|<s_x$ and $|\delta y|<s_y$.

4.3.2 Random binary pattern

Intuitively, for small perturbations (z_c, δx, δy small) the system is in general approximately shift invariant (images not necessarily Gaussian processes) and $I_d(\Delta x,\Delta y)$ peaks at $\Delta x=\delta x$ and $\Delta y=\delta y$. Equation 4-10 becomes

$$S(\delta x,\delta y) \approx \left| \iint f_1^*(x,y) f_1(x,y) e^{-j\frac{kz_c}{F^2}(\delta x x+\delta y y)} dxdy \right|^2 \tag{4-19}$$

$$= \frac{1}{4}\iint e^{-j\frac{kz_c}{F^2}(\delta x x+\delta y y)} dxdy \iint \left| f_1\left(\frac{X+x}{2},\frac{Y+y}{2}\right)\right|^2 \left| f_1\left(\frac{X-x}{2},\frac{Y-y}{2}\right)\right|^2 dXdY$$

$$= W_x W_y \iint e^{-j\frac{kz_c}{F^2}(\delta x x+\delta y y)} C(x,y)dxdy$$

where $C(x,y) = \frac{1}{W_x W_y}\iint \left| f_1\left(\frac{X+x}{2},\frac{Y+y}{2}\right)\right|^2 \left| f_1\left(\frac{X-x}{2},\frac{Y-y}{2}\right)\right|^2 d\frac{X}{2}d\frac{Y}{2}$ is the autocorrelation of $|f_1(x,y)|^2$.

We now consider two special cases:

1. Random binary phase modulation or DC filtered random amplitude modulation

In this case, $f_1(x,y) = 1$ or -1, i.e., $|f_1(x,y)|^2=1$, we get

$$C(x,y) = \Delta\left(\frac{x}{W_x}\right)\Delta\left(\frac{y}{W_y}\right) \text{ where } \Delta(x) = \begin{cases} 1-|x| & -1 \le x \le 1 \\ 0 & \text{else} \end{cases}$$

$$S(\delta x,\delta y) = W_x^2 W_y^2 \operatorname{sinc}^2\left(\frac{W_x Z_c \delta x}{\lambda F^2}\right)\operatorname{sinc}^2\left(\frac{W_y Z_c \delta y}{\lambda F^2}\right) \tag{4-20}$$

The shift selectivity along the x-direction is $s_x \approx \frac{\lambda F^2}{z_c W_x}$ and that along the y-direction is $s_y \approx \frac{\lambda F^2}{z_c W_y}$

2. Random amplitude modulation

In this case, $f_1(x,y) = 0$ or 1, $|f_1(x,y)|^2=f_1(x,y)$

$$C(x,y) \approx \frac{1}{4}\left[\Delta\left(\frac{x}{W_x}\right)\Delta\left(\frac{y}{W_y}\right) + \Delta\left(\frac{x}{d_x}\right)\Delta\left(\frac{y}{d_y}\right)\right]$$

$$S(\delta x, \delta y) \approx \frac{1}{4}[W_x^2 W_y^2 \mathrm{sinc}^2(W_x u)\mathrm{sinc}^2(W_y v) + d_x^2 d_y^2 \mathrm{sinc}^2(d_x u)\mathrm{sinc}^2(d_y v)] \quad (4\text{-}21)$$

where $u = \frac{z_c \delta x}{\lambda F^2}$ and $v = \frac{z_c \delta y}{\lambda F^2}$. Because the area of a pixel is much smaller than the area of the image, i.e., $d_x d_y \ll W_x W_y$, we get

$$S(\delta x, \delta y) \approx \frac{1}{4} W_x^2 W_y^2 \mathrm{sinc}^2(W_x u)\mathrm{sinc}^2(W_y v) \quad (4\text{-}22)$$

Again, the shift selectivity along the x-direction is $s_x \approx \frac{\lambda F^2}{z_c W_x}$ and that along the y-direction is $s_y \approx \frac{\lambda F^2}{z_c W_y}$.

In the above discussion, we get the same selectivity with different models. The selectivity is inversely proportional to the defocusing distance and the size of the image. An intuitive way to understand it is that $I_d(\Delta x,\Delta y)$ tends to be zero when the phase term $e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)}$ oscillates fast enough. If $\frac{kz_c}{F^2}\Delta x W_x = 2\pi$ and $\frac{kz_c}{F^2}\Delta y W_y = 2\pi$, the phase term changes from 0 to 2π when we perform the integration across the image and this smooths out the correlation peak, leading to $\delta x \sim \Delta x \sim \frac{\lambda F^2}{z_c W_x}$ and $\delta y \sim \Delta y \sim \frac{\lambda F^2}{z_c W_y}$.

4.3.3 Experiment

In order to verify our analysis, we compare the experimental shift selectivity data in [4-6] with our approximate theory. Figure 4-5 and Figure 4-6 show the dependence of the shift selectivity (normalized to the focal length) on the defocusing distance z_c. In the out-of-plane direction (image shift perpendicular to the incident plane formed by the signal and reference axes), our approximate result matches very well with the experimental data and the rigorous theory (direct numerical integration of Equation 4-6) while in the in-plane direction (image shift in the incident plane) the measured shift selectivity is a little bit smaller. The recording material used in the experiment is a 250μm-thick $LiNbO_3$ crystal.

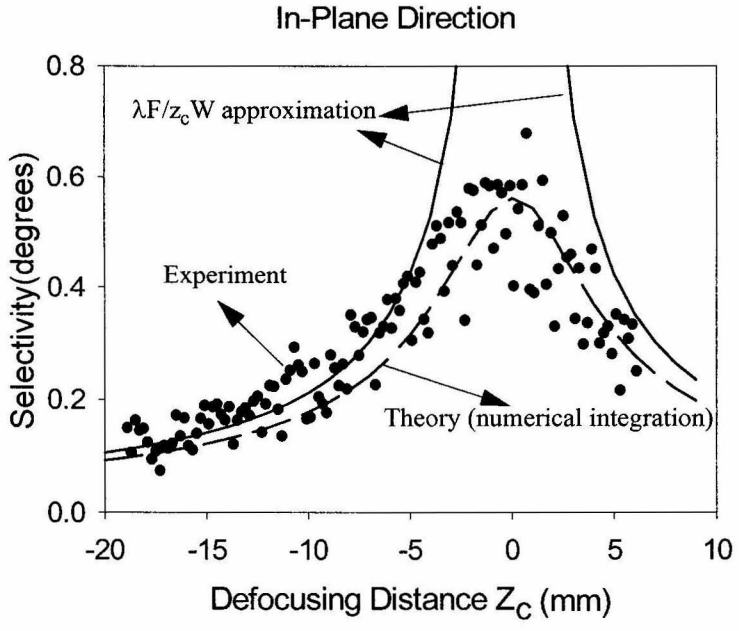


Fig. 4-5. In-plane shift selectivity (experimental and numerical integration data from [4-6])

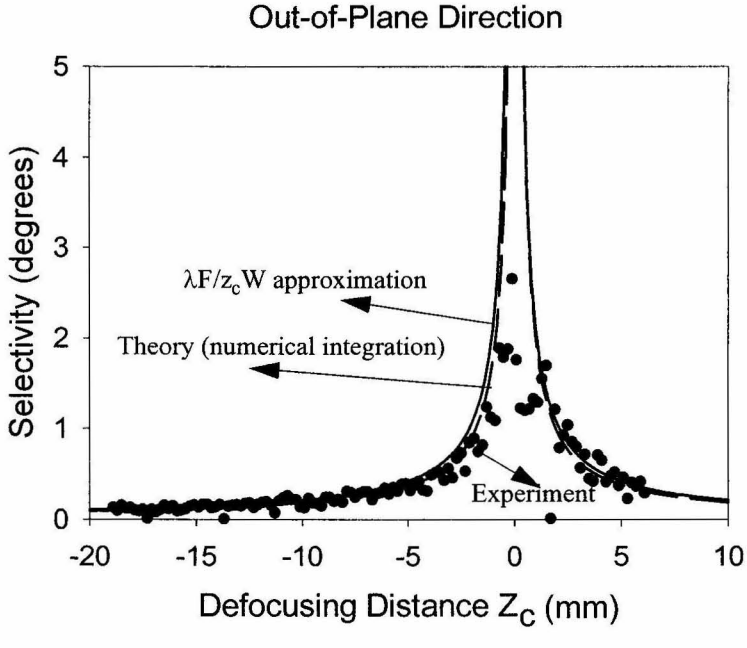


Fig. 4-6. Out-of-plane shift selectivity (experimental and numerical integration data from [4-6])

However, in our analysis the volume hologram selectivity is ignored, which leads to the discrepancy. Since the out-of-plane direction is the Bragg degenerate direction, the volume selectivity effect is weaker and we have better agreement between our approximate theory and the experimental data. The result verifies that the shift selectivity is inversely proportional to the defocusing distance z_c.

We also studied the effect of the image size and pixel size. Our experiment setup is shown in Figure 4-7. The wavelength of the laser is 488nm. The signal travels along the z

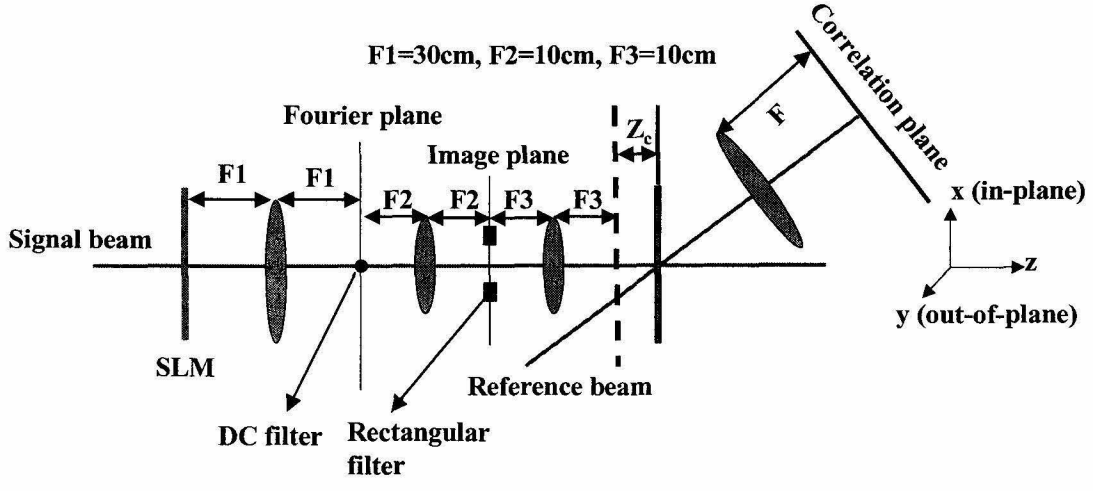


Fig. 4-7. Shift selectivity experiment setup

axis and the reference is in the x-z plane. Both a Koppin SLM (640 × 480, pixel pitch 24μm) and a TVT6000 SLM (480 × 440, pixel pitch 50μm) are used in the experiments. The test image used is a random binary pattern, which takes a value of either 0 or 1 at each super-pixel of n × n SLM pixels. The recording material is 38 μm-thick Dupont polymer. From previous discussion, the shift selectivity on the image plane is

$$s_x \approx \frac{\lambda F_3^2}{z_c W_x} \text{ and } s_y \approx \frac{\lambda F_3^2}{z_c W_y}. \tag{4-23}$$

We want to calculate the corresponding selectivity at the SLM and compare it with experiment. From the imaging condition of a 4-F system, we have

$$\frac{s_x'}{s_x} = \frac{s_y'}{s_y} = \frac{W_x'}{W_x} = \frac{W_y'}{W_y} = \frac{F_1}{F_2} \tag{4-24}$$

where s_x', s_y', W_x', W_y' are the corresponding selectivities and image size at the SLM. Since F_2 is equal to F_3 in our setup, we obtain

$$s_x' \approx \frac{\lambda F_1^2}{z_c W_x'} \text{ and } s_y' \approx \frac{\lambda F_1^2}{z_c W_y'} \tag{4-25}$$

This shows that the recording material can be thought of as being imaged to the back of the Fourier plane and the defocusing distance of the recording material image is still z_c since F_2=F_3.

Figure 4-8 shows two selectivity curves with different image sizes. The TVT6000 SLM is used in this experiment. The defocusing distance z_c is 0.4cm. As shown in the figure, the larger the image, the smaller the shift selectivity. Figure 4-9 and Figure 4-10 show the selectivity curves with different super-pixel sizes in the in-plane and out-of-plane directions. The Koppin SLM is used in these experiments. The defocusing distance z_c is 1cm. Equation 4-20 is plotted (the theoretical curve) and compared with the experimental data. As can be observed, the selectivity is only weakly dependent on the super-pixel size. The experimental results agree very well with the close form approximate results.

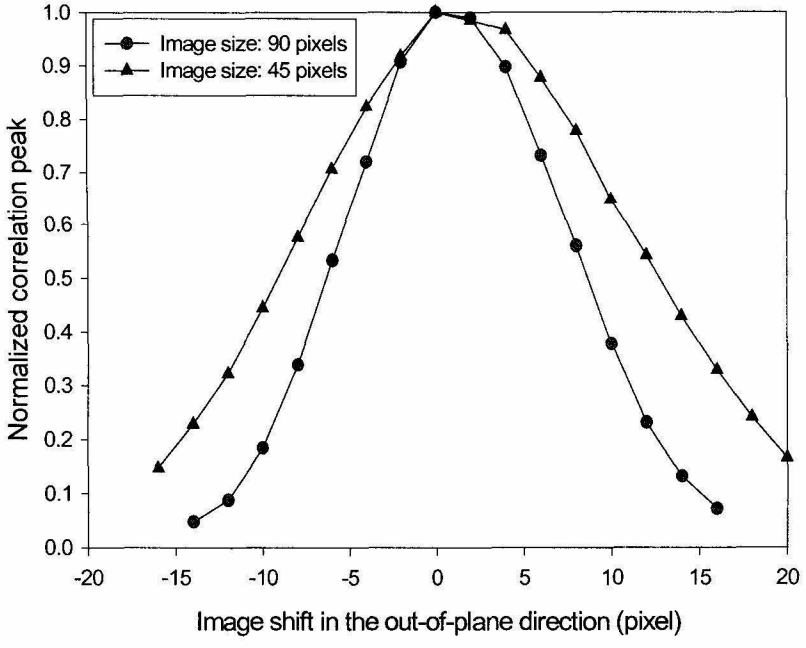


Fig. 4-8. Selectivity with different image size

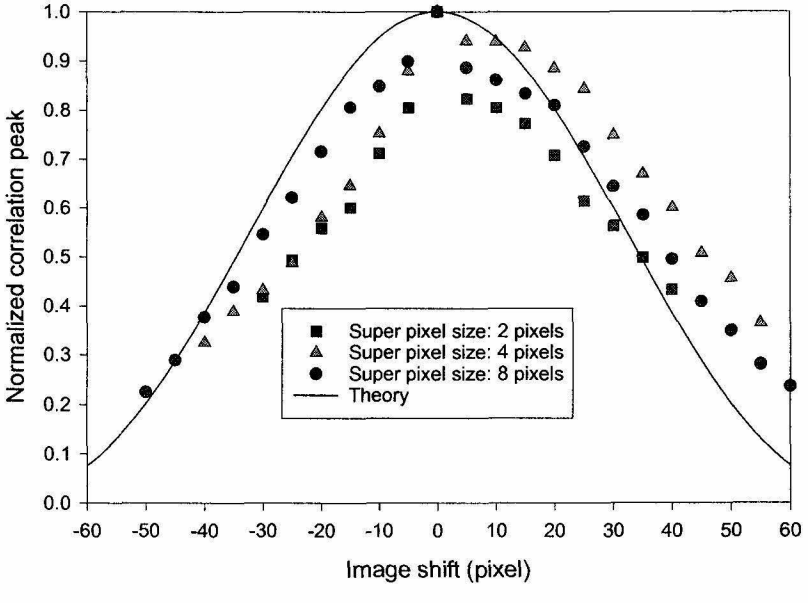


Fig. 4-9. In-plane selectivity curve with different super pixel size

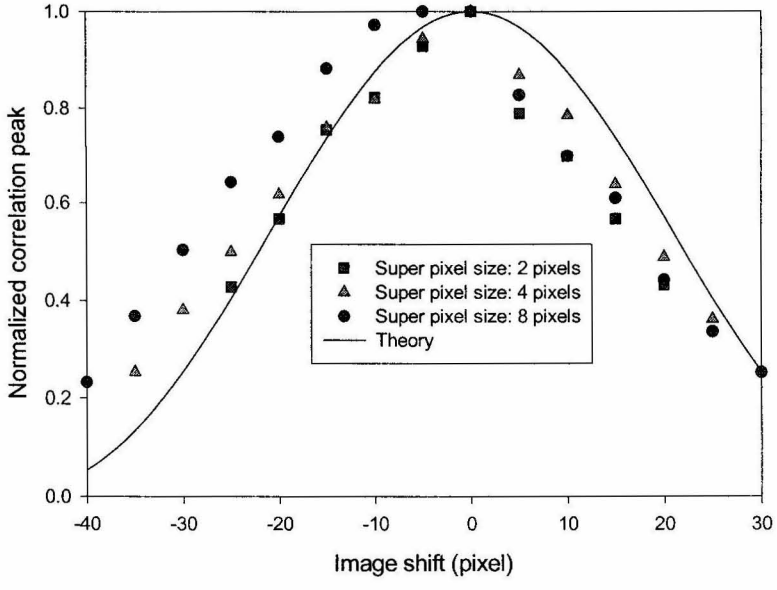


Fig. 4-10. Out-of-plane selectivity curve with different super-pixel size

4.4 Cross talk

In normal correlators, correlations may have side lobes which can extend to neighboring domains and cause cross talk. In a defocused correlator, the side lobes can be suppressed by the defocusing phase term $e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)}$ [4-6], which results in a better S/N.

Following [4-10], to illustrate the idea we calculate the correlation of a test image which consists of four 2mm by 2mm squares as shown in Figure 4-11. The wavelength of the laser is taken as 488nm and the focal length of the Fourier lens is assumed to be 30cm. The autocorrelations with z_c=0 (normal correlator) and z_c=1cm (defocused correlator) are shown in Figure 4-12. Clearly, the side lobes are suppressed in the defocused correlator.

Now suppose we have an array of M correlators, and the templates are $f_i(x,y)$ where i=1,...,M. If the input is $f_1(x,y)$, then there is an autocorrelation peak in the center of the first

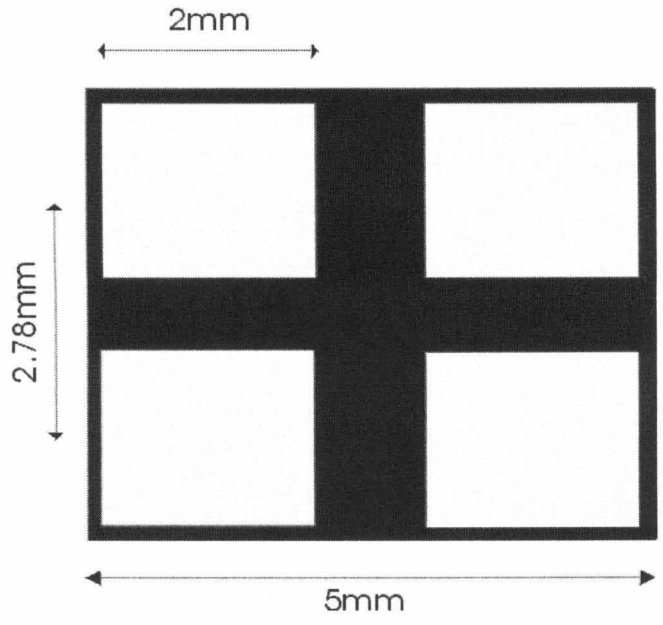


Fig. 4-11. Test image used in cross talk simulations

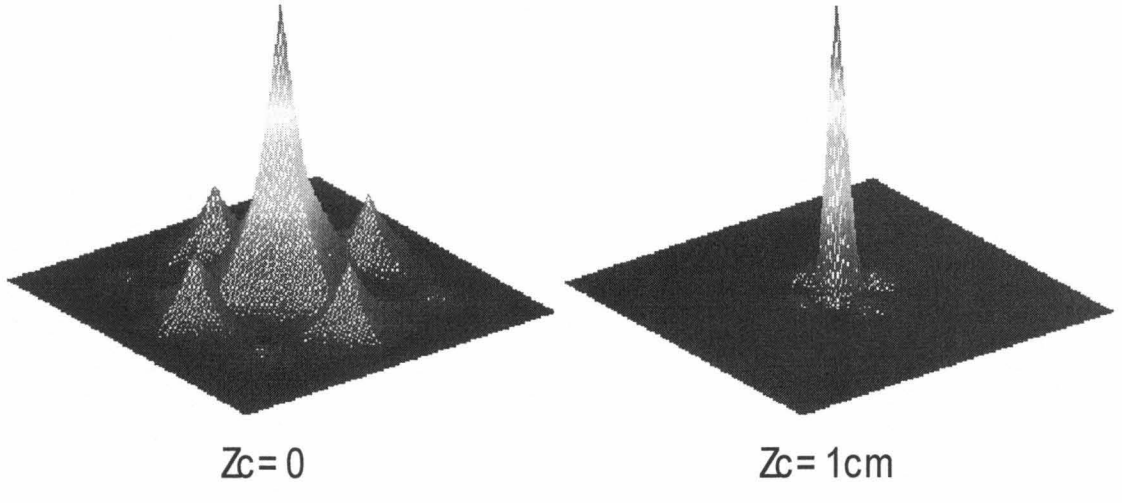


Fig. 4-12. Side lobe suppression in defocused correlators

domain, but there is also cross talk due to the cross correlation between $f_1(x,y)$ and $f_i(x,y)$ where i=2,...,M.

From previous sections,

$$I_d(0, 0) = \left| \sum_{i=1}^{M} A_d^{(i)}(-\Delta x_i, -\Delta y_i) \right|^2 \tag{4-26}$$

where Δx_i and Δy_i are the displacements of the ith domain center from the first domain center and $A_d^{(i)}(\Delta x, \Delta y)$ is given by

$$A_d^{(i)}(\Delta x, \Delta y) = e^{j\phi_i} \iint f_i^*(x, y) f_1(x + \Delta x, y + \Delta y) e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} dxdy$$

$$\phi_i = (k_{rz}^i - k_{dz}^i) n^2 z_c - \frac{\pi z_c (\Delta x_i^2 + \Delta y_i^2)}{\lambda F^2}$$

We can calculate the S/N ratio as

$$S/N = \frac{\left| A_d^{(1)}(0, 0) \right|^2}{\left| \sum_{i=2}^{M} A_d^{(i)}(-\Delta x_i, -\Delta y_i) \right|^2} \tag{4-27}$$

We assume 121(11 by 11) template images are recorded. The S/N in the origin of the center domain is calculated for different values of z_c The correlation domain size is 3.9mm × 3.9mm. Figure 4-13 shows the dependence of the S/N on the defocusing distance

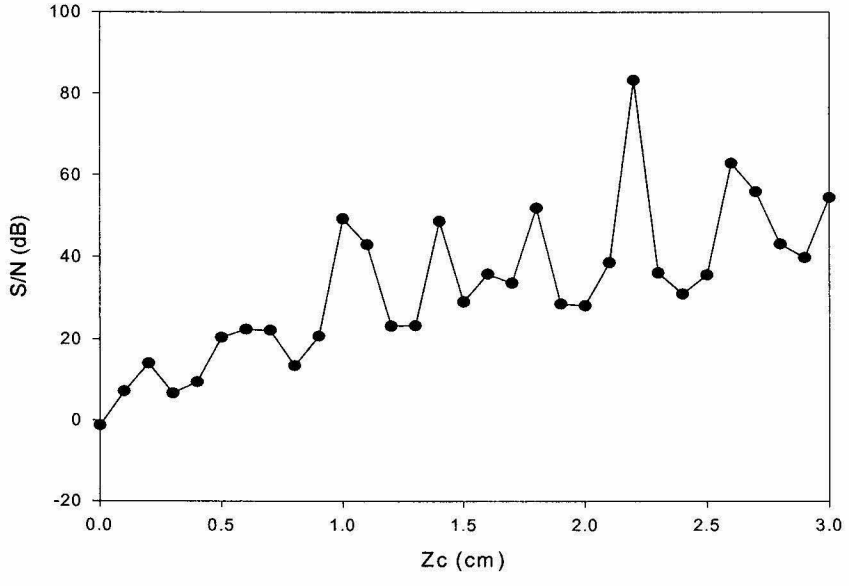


Fig. 4-13. Cross talk in correlators

z_c. As we increase z_c, S/N increases as well. The fluctuation observed in the figure is due to the speckle effect caused by the pseudorandom phase factor ϕ_i.

4.5 Real time correlator system

In this section we describe the real time defocused correlator system [4-11] built in our lab. We recorded 480 templates in 100μm thick DuPont HRF-150 photo-polymer. The number of stored templates is mainly limited by the shrinkage effect [4-12] of the Dupont polymer, not the material dynamic range. One of the stored templates and its correlation output are shown in Figure 4-14.

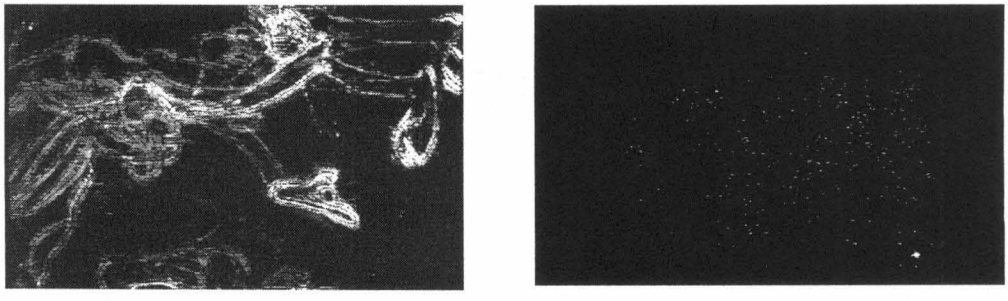


Fig. 4-14. Stored edge-enhanced template and its correlation output

Figure 4-15 shows the measured autocorrelation peak intensity of the 480 stored templates. The correlation peak strength is equalized using the exposure schedule described in [4-13]. The correlation plane is divided into 24 × 20 domains. A raster scan of the maximum correlation is obtained when we sequence through the templates as shown in Figure 4-16. Only 10 out of 480 templates are misclassified. This error rate can be minimized by normalizing the autocorrelation of each template. The template angular selectivity (image shift selectivity normalized to the focal length of the lens) is approximately 0.1 degree and the angular distance between neighboring domains is 0.4 degree. The Bragg

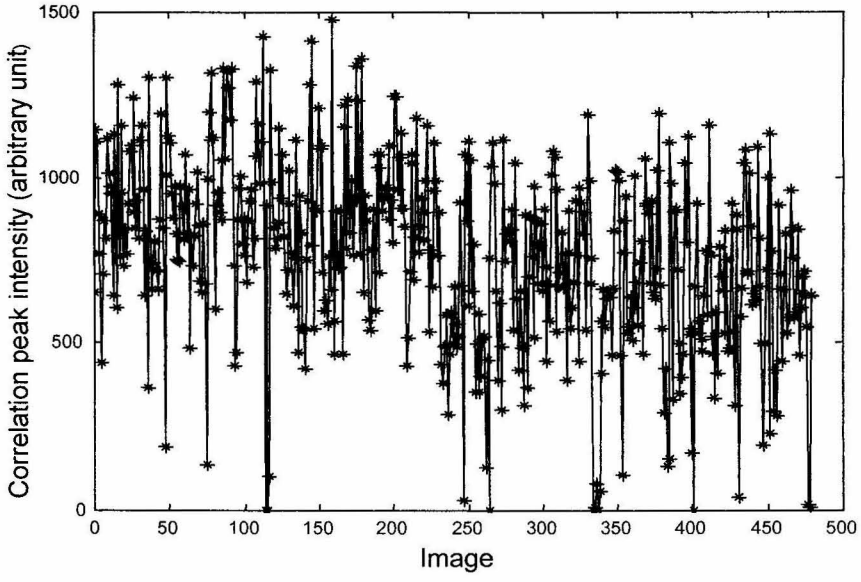


Fig. 4-15. Correlation peak

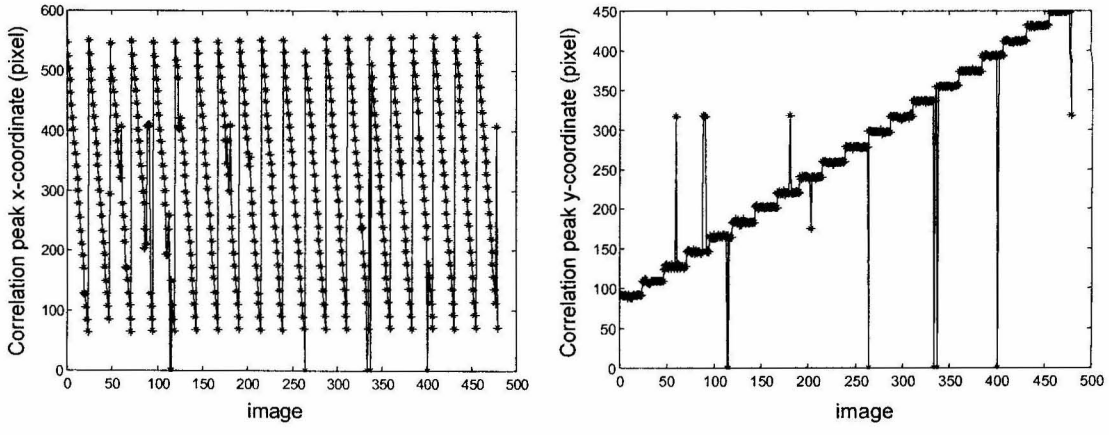


Fig. 4-16. Correlation peak position

selectivity in the in-plane direction is approximately 0.5 degree. Clearly, the defocused correlator system can store more templates than a normal correlator. To implement distortion invariance, synthesized templates may be used. If each template is synthesized with 10 targets, the system then has about 500×10 filters which would be very useful in a content searchable database.

Our system has been used for real time fingerprint and face recognitions. Fingerprint recognition is particularly interesting because it requires a small degree of shift invariance hence many templates can be stored. We will discuss fingerprint identification in more detail in the next chapter. Figure 4-17 and Figure 4-18 show the cross-sections of typical fingerprint and face correlations.

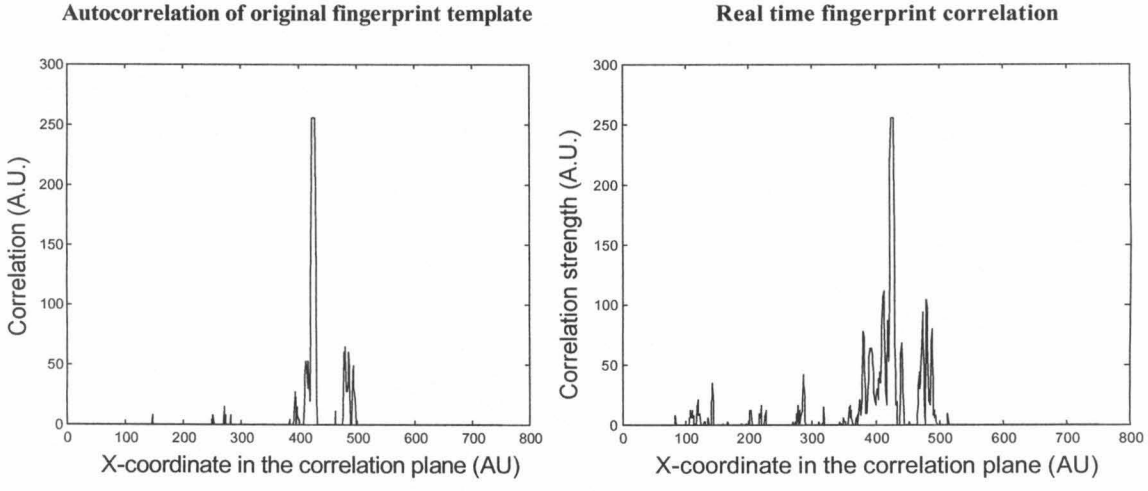


Fig. 4-17. Fingerprint correlation

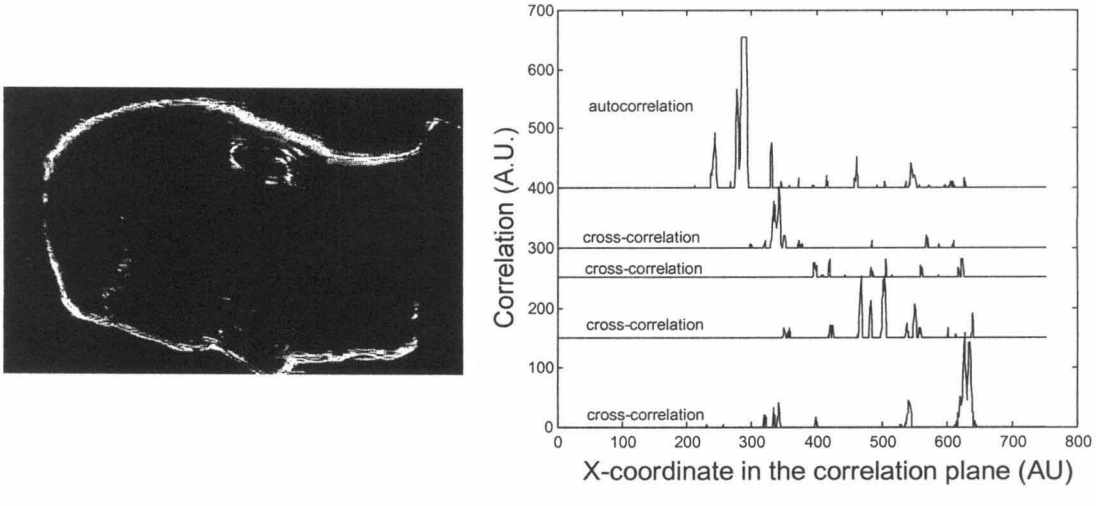


Fig. 4-18. Face recognition

4.6 Discussion

4.6.1 Incoherent correlator

In defocused correlators, the shift invariance is reduced by the fast oscillating phase term $e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)}$ in Equation 4-9. If the laser beam is modulated by a moving diffuser during readout, the input image can be represented by $f_2(x, y)e^{j\psi(x, y, t)}$ where $f_2(x, y)$ is the image displayed on the SLM and $\psi(x, y, t)$ is the pseudo-random phase modulation induced by the moving diffuser. If a thin recording material is used, the diffraction intensity is given by

$$
\begin{aligned}
I_d(\Delta x, \Delta y) &= |A_d(\mathbf{k_d})|^2 \qquad (4\text{-}28)\\
&= \left\langle \left| \iint f_1^*(x, y) f_2(x - \Delta x, y - \Delta y) e^{j\psi(x-\Delta x, y-\Delta y, t)} e^{-j\frac{kz_c}{F^2}(\Delta x x + \Delta y y)} dx dy \right|^2 \right\rangle \\
&= \iint |f_1(x, y)|^2 |f_2(x - \Delta x, y - \Delta y)|^2 dx dy
\end{aligned}
$$

where <...> is the average over time. The averaging is necessary because of the slower response of the detector (CCD camera). The system becomes an incoherent correlator which is shift invariant again.

4.6.2 Information security

Usually, the difference between holographic correlator and memory is whether the signal is used to reconstruct reference (correlator) or the reference is used to reconstruct the signal (memory). However, the defocused holographic correlator is not a good memory device because holograms are not recorded with enough angular separation between the reference beams. Multiple reconstructed images would appear on the detector and there

would be severe cross talk if it is used as a memory device trying to steal the template information.

As shown in Figure 4-19, even if the correlation peak disappears the total diffracted

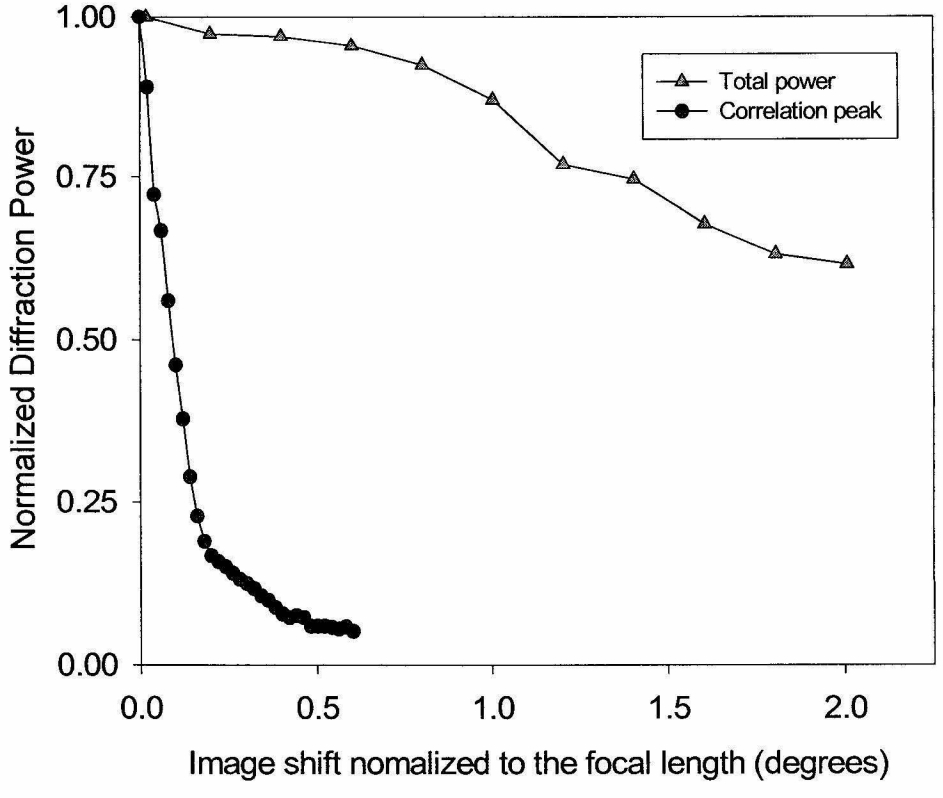


Fig. 4-19. Total power and correlation peak power

power is still significant which is randomly distributed in the correlation plane and builds up a noise floor. Since the reconstructed reference is focused, correlators are more noise tolerant than memory devices.

4.6.3 Spatial frequency domain representation

$I_d(\Delta x, \Delta y)$ can also be written in the spatial frequency domain.

$$I_d(\Delta x, \Delta y) = \left| \iint f_1^*(x, y) f_2(x + \Delta x, y + \Delta y) e^{-j2\pi \frac{z_c}{\lambda F^2}(\Delta x x + \Delta y y)} dx dy \right|^2 \tag{4-29}$$

$$= \left| FT[f_1^*(x, y) f_2(x + \Delta x, y + \Delta y)] \right|_{u=\Delta u,\, v=\Delta v}^2$$

$$= \left| FT[f_1^*(x, y)] \otimes FT[f_2(x + \Delta x, y + \Delta y)] \right|^2$$

$$= \left| \iint F_1^*(-u, -v) F_2(\Delta u - u, \Delta v - v) e^{j2\pi[\Delta x(\Delta u - u) + \Delta y(\Delta v - v)]} du dv \right|^2$$

$$= \left| \iint F_1^*(u, v) F_2(\Delta u + u, \Delta v + v) e^{j2\pi[\Delta x(\Delta u + u) + \Delta y(\Delta v + v)]} du dv \right|^2$$

$$= \left| \iint F_1^*(u, v) F_2(u + \Delta u, v + \Delta v) e^{j2\pi \frac{\lambda F^2}{z_c}(\Delta u u + \Delta v v)} du dv \right|^2$$

where $\Delta u = \frac{z_c \Delta x}{\lambda F^2}$, $\Delta v = \frac{z_c \Delta y}{\lambda F^2}$, FT[...] is the Fourier transform operation and $F_1(u,v)$, $F_2(u,v)$ are the Fourier transform of $f_1(x,y)$ and $f_2(x,y)$ respectively.

The formula of $I_d(\Delta x, \Delta y)$ is highly symmetric in the spatial and the spatial frequency domains. If $z_c \ll F$, from the representation in the spatial domain we get the following.

$$I_d(\Delta x, \Delta y) = \left| \iint f_1^*(x, y) f_2(x + \Delta x, y + \Delta y) e^{-j2\pi \frac{z_c}{\lambda F^2}(\Delta x x + \Delta y y)} dx dy \right|^2 \tag{4-30}$$

$$\cong \left| \iint f_1^*(x, y) f_2(x + \Delta x, y + \Delta y) dx dy \right|^2$$

which is a spatial domain correlation. If $z_c \gg F$, using the representation in the spatial frequency domain we get the following.

$$I_d(\Delta x, \Delta y) = \left| \iint F_1^*(u, v) F_2(u + \Delta u, v + \Delta v) e^{j2\pi \frac{\lambda F^2}{z_c}(\Delta u u + \Delta v v)} du dv \right|^2 \tag{4-31}$$

$$\cong \left| \iint F_1^*(u, v) F_2(u + \Delta u, v + \Delta v) du dv \right|^2$$

which is a spatial frequency domain correlation. This means that the defocused correlator is a mixture of spatial and frequency domain correlators. Intuitively, if we put the recording

material too far away from the focal plane, then due to Fraunhofer diffraction the template image, rather than its Fourier transform, is recorded in the hologram. The defocused correlator continuously transforms itself from a spatial domain correlator to a frequency domain correlator as we increase z_c. The spatial correlator is shift invariant while the frequency domain correlator is not. The shift invariance is controlled by the amount of "doping" of the frequency domain correlator.

We can also use the formula to analyze the position of the correlation peak. Assume $f_2(x,y)=f_1(x-\delta x, y-\delta y)$. First we consider the case $z_c \ll F$

$$I_d(\Delta x, \Delta y) = \left| \iint F_1^*(u, v) F_1[u + \Delta u, v + \Delta v] \times e^{j2\pi \frac{\lambda F^2}{z_c}[(\Delta u - \delta u)u + (\Delta v - \delta v)v]} du dv \right|^2 \quad (4\text{-}32)$$

where $\Delta u = \frac{z_c \Delta x}{\lambda F^2}$, $\delta u = \frac{z_c \delta x}{\lambda F^2}$, $\Delta v = \frac{z_c \Delta y}{\lambda F^2}$ and $\delta v = \frac{z_c \delta y}{\lambda F^2}$

If $\Delta u=\delta u$ and $\Delta v=\delta v$, or $\Delta x=\delta x$ and $\Delta y=\delta y$, the rapid oscillating phase vanishes and $I_d(\Delta x, \Delta y)$ has its maximal value. This verifies the shift invariance assumption we made in Section 4.3.2.

We consider the other case $z_c \gg F$. Using the spatial domain representation we get

$$I_d(\Delta x, \Delta y) = \left| \iint f_1^*(x, y) f_1(x + \Delta x - \delta x, y + \Delta y - \delta y) e^{-j2\pi \frac{z_c}{\lambda F^2}(\Delta x x + \Delta y y)} dx dy \right|^2 \quad (4\text{-}33)$$

The phase term is fast oscillating unless $\Delta x=\Delta y=0$. This means the peak locates at the center of the correlation output domain.

Sometimes, due to this mixture of the spatial and frequency domain correlator we can see twin peaks with one always at the center (frequency domain correlation peak) and

the other moving when the input is shifted (spatial correlation peak). Figure 4-20 shows the measured twin correlation peaks.

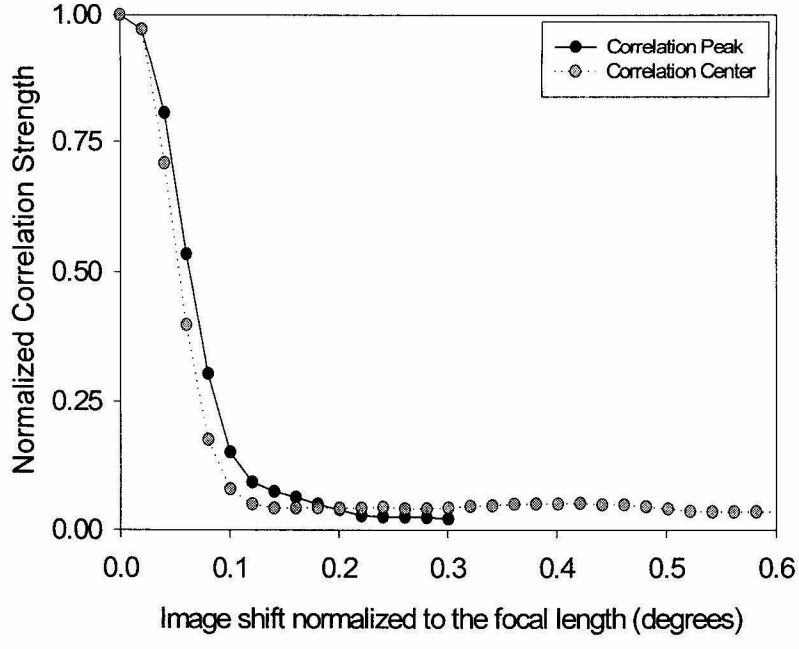


Fig. 4-20. Twin peaks

We now apply the frequency domain formalism to a variation of the defocused correlator, in which we consider that the SLM is not in the front focal plane of the Fourier lens as shown in Figure 4-21. We can assume an effective image $f'(x, y)$ in the focal plane which produces the same field in the recording material as the image f(x,y) displayed on the SLM with a distance of L from the focal plane.

$$f(x, y) = f'(x, y) \otimes e^{j\frac{\pi}{\lambda L}(x^2 + y^2)} \tag{4-34}$$

or in the frequency domain,

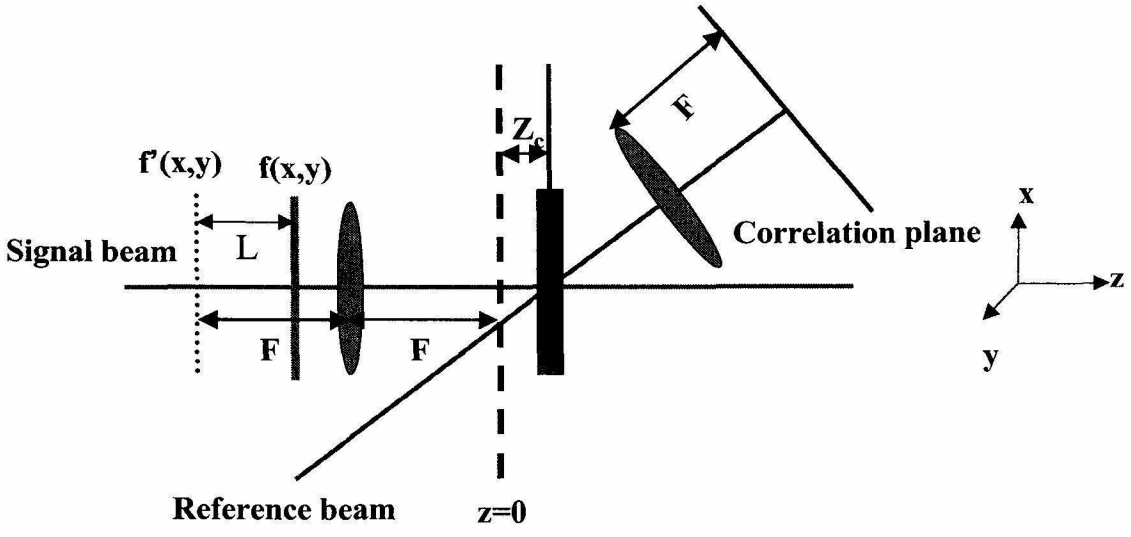


Fig. 4-21. Effect of SLM displacement

$$F(u, v) = F'(u, v)e^{-j\pi\lambda L(u^2 + v^2)} \tag{4-35}$$

where F(u,v) and $F'(u, v)$ are Fourier transform of f(x,y) and $f'(x, y)$.

$$I_d(\Delta x, \Delta y) = \left| \iint F_1^*(u, v)e^{j\pi\lambda L(u^2 + v^2)} F_2(u + \Delta u, v + \Delta v) \right. \tag{4-36}$$
$$\left. \times e^{-j\pi\lambda L[(u + \Delta u)^2 + (v + \Delta v)^2]} e^{j2\pi\frac{\lambda F^2}{z_c}(\Delta u u + \Delta v v)} du dv \right|^2$$
$$= \left| \iint F_1^*(u, v)F_2(u + \Delta u, v + \Delta v)e^{j2\pi\frac{\lambda F^2}{z_c}\left(1 - \frac{Lz_c}{F^2}\right)(\Delta u u + \Delta v v)} du dv \right|^2$$

If we choose L=F (SLM placed directly in front of the lens) and z_c<<F,

$$I_d(\Delta x, \Delta y) \cong \left| \iint F_1^*(u, v)F_2(u + \Delta u, v + \Delta v)e^{j2\pi\frac{\lambda F^2}{z_c}(\Delta u u + \Delta v v)} du dv \right|^2 \tag{4-37}$$

which is the same as if the SLM was in the front focal plane.

4.7 Conclusion

In this chapter, we investigate the properties of the defocused holographic correlator which can conveniently control the shift invariance by tuning the defocusing distance.

A simple expression of the shift selectivity is derived ($s_x \approx \frac{\lambda F^2}{z_c W_x}$ and $s_y \approx \frac{\lambda F^2}{z_c W_y}$), which agrees very well with the experimental results. The cross talk in the correlator array is analyzed and we show that the defocused correlator can improve the S/N. We also demonstrate a real time correlator system with 480 templates.

References

[4-1] Demetri Psaltis and Fai H. Mok. "Holographic Memories". Scientific American 273, 70-76 (1995)

[4-2] J. Goodman. Introduction to Fourier Optics (second edition) (McGraw-Hill, New York, 1996)

[4-3] Michael Levene. Caltech Ph.D. thesis (1998)

[4-4] Gregory J. Steckman. Caltech Ph.D. thesis (2001)

[4-5] F.T.S. Yu, Guowen Lu, Mingzhe Lu, and Dazun Zhao. "Application of position encoding to a complex joint transform correlator." Applied Optics 34, 1386-1388 (1995)

[4-6] Michael Levene, Greg J. Steckman, and Demetri Psaltis, "A method to control the shift invariance of optical correlators." Applied Optics 38, 394 (1999)

[4-7] P.J. Van Heerden. "A new optical method of storing and retrieving information." Applied Optics 2 (4) 387-392 (1963)

[4-8] G.W. Burr. Caltech Ph.D.thesis (1996)

[4-9] J. Goodman. Statistical Optics (John Wiley & Sons , New York, 1985)

[4-10] Claire Gu, Hui Fu, and Jenn-Ren Lien. "Correlation patterns and cross-talk noise in voulme holographic optical correlators." J. Opt. Soc. Am. A 12, 861 (1995)

[4-11] Zhiwen Liu, Xin An, Gregory J. Steckman, Michael Levene, Gan Zhou, and Demetri Psaltis. "Defocused Holographic Correlator Array." Technical digest of optics in computing 94 (1999)

[4-12] J.T. Galo and C.M. Verber. "Model for the effects of material shrinkage on volume holograms." Applied Optics 33, 6797-6804 (1994)

[4-13] Allen Pu, Kevin Curtis, and Demetri Psaltis. "Exposure schedule for multiplexing holograms in photopolymer films." Opt. Eng. 35 , 2824-2829 (1996)

5 Correlation Based Fingerprint Identification

In the previous chapter we investigated the defocused holographic correlators which can pack many templates with controlled shift invariance. In this chapter we focus more on the correlation algorithm itself instead of the holographic system design. In Section 5.2 we discuss the performance of the optical correlation of fingerprints. In Section 5.3 we introduce the windowed correlation to improve the distortion tolerance. In Section 5.4, we give detailed simulation results and discuss the performance of various strategies.

5.1 Introduction

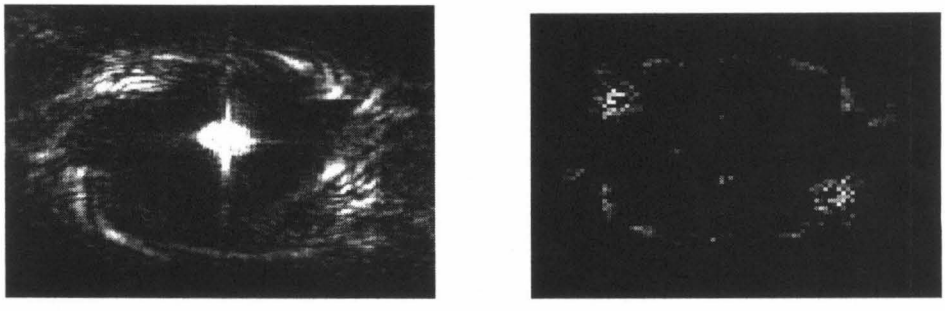
Fingerprint recognition has been studied extensively for many years. Minutia based algorithms [5-1] are very efficient and robust and they are now widely used in fingerprint verification and identification [5-2]. False acceptance rates (FAR) as small as 0.01% can be obtained with false rejection rate (FRR) of 1.4% [5-3]. However, if the database comprises millions of fingerprints the inherent serial algorithm nature may limit the speed in identification applications. The huge amount of memory needed also increases the cost.

The advantage of the correlation method [5-4] [5-5] is that it can be optically implemented. Holographic correlator arrays have massive parallelism, high speed and huge storage capacity. In the previous chapter we demonstrate that at each location of a 100μm thick Dupont polymer about 500 templates can be stored. Thus one holographic disk of a CD size can store approximately a million templates.

The major problem with the correlation method is that it is too sensitive to distortion and rotation. In the following sections, we will investigate the correlation method in detail and try to improve its performance.

5.2 Optical correlation

The Fourier transform of the correlation of two fingerprints is a product of their spectrum. Before discussing the optical correlation, we first take a look at the spectrum of fingerprint. The DC filtered optical and digital Fourier transforms of a fingerprint are shown in Figure 5-1. The spectrum looks like a ring due to the circular shape of ridges and



Optical Spectrum Digital Spectrum

Fig. 5-1. Spectrum of fingerprints

valleys. This is a typical spectrum of fingerprint. Because the spectral information concentrates around the carrier frequency (υ_0) [5-5], the fingerprint correlation are sensitive to rotation (spectrum rotates $\theta\upsilon_0$ with a rotation angle θ) and distortion (different local rotations).

The real time holographic correlator discussed in the previous chapter was used to optically compute the correlation functions of fingerprints. We collected 50 samples of one

fingerprint and one of the sample images was recorded as the template. After the recording, all 50 samples were displayed on the SLM sequentially and the corresponding correlation output was captured by a CCD camera. The value of each correlation peak was then found. The optically computed correlation peak intensity is shown in Figure 5-2. The correlation

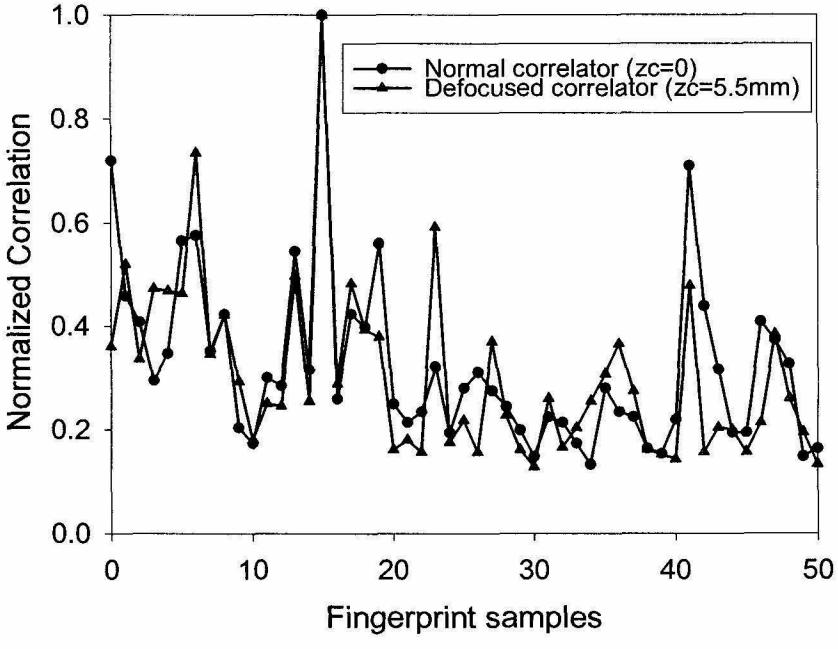


Fig. 5-2. Correlations using normal and defocused correlators

functions are computed using both the normal and the defocused correlators. The performance of both correlators is very close although the shift invariance is reduced in the defocused correlator.

In order to estimate the performance that we can get on a database with multiple classes, we collected a database of 1000 samples from 20 different fingerprints (classes). Each class has 50 samples. Five samples of each class were recorded as templates, and the rest of the 45 samples were used as the test set. After the recording, samples in the test set

were displayed sequentially on the SLM and correlated with all the recorded templates. First the correlation peak in each template was found, and then the peak values of the correlations with the five templates of each class were summed up. The input was classified to the class with the maximal sum if it was above the threshold. Otherwise, the input is rejected. Figure 5-3 shows the performance of our optical fingerprint identification system.

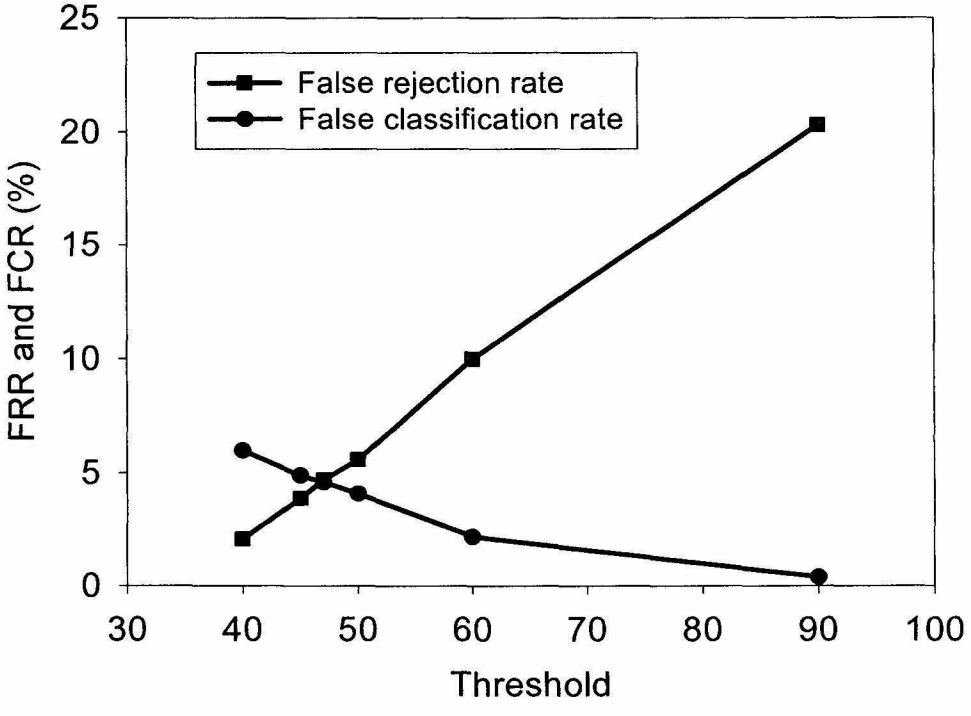


Fig. 5-3. Optical fingerprint identification

If a sample belonging to one of the 20 classes is rejected because its correlation peaks are below the threshold, a false rejection error occurs. On the other hand, if a sample belonging to class A is classified to class B, a false classification error occurs. At the optimal threshold, our system yields an error rate of 4.6% (false classification rate = false rejection rate).

The performance we obtained with our small database is inferior to that obtained with minutia based algorithms. In order to isolate the reasons for this we digitally implemented the correlation algorithm to eliminate possible errors due to the optical realization. Figure 5-4 shows the comparison between the digital and optical correlations. A sample is

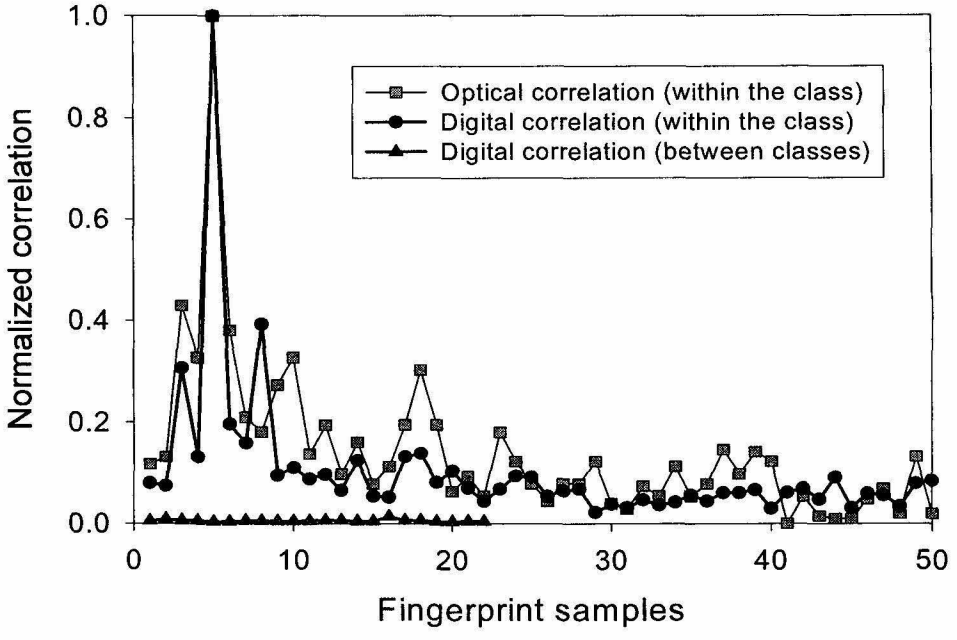


Fig. 5-4. Digital and optical correlations of fingerprints

optically and digitally correlated with all the 50 samples of the same class (within the class) and is digitally correlated with samples from other classes (between classes). The digital correlation is squared to be consistent with the optical correlation (CCD camera detects intensity of light). The optical and digital correlation in general agree with each other except for samples whose optical correlation is close to the noise level and it can lead to a false classification or rejection. The digital correlation has better performance for these

samples. From now on we focus on the digital simulations to investigate the correlation algorithm itself independent of the holographic correlator system implementation.

5.3 Windowed correlation

Distortion and rotation can be thought of as non-uniform shifts of different parts of an image. The basic idea of the windowed correlation [5-6] is to break the image into pieces or small windows and to correlate each piece sequentially with the template. Each piece of the image may produce a correlation peak at different locations in the correlation plane. The windowed correlation response is the sum of all the correlation peaks. We expect improved performance since if the whole image is used instead as in normal correlation, the peaks due to different windows (pieces) may not add up constructively.

We first compared the ability of normal versus windowed correlation to deal with rotation using 24 fingerprint samples (each of a different class). We computed the correlation of each sample and its rotated version. In windowed correlation we divided the rotated image uniformly into nine windows with the width and height of each window being one third of the original size. We correlated the partial image within each window with the entire unrotated image to get the corresponding partial correlation peak value. All nine partial correlation peak values were then detected, squared and summed up. The average dependence on the rotation angle is shown in Figure 5-5. for both normal and windowed correlations. The error bars are the standard deviations of the normalized correlation response. The windowed correlation improves the rotation invariance by a factor of two at the 50% correlation level. In Figure 5-6 we show that the windowed correlation (9 uniform windows) also improves the discrimination for the more practical case of correlating mul-

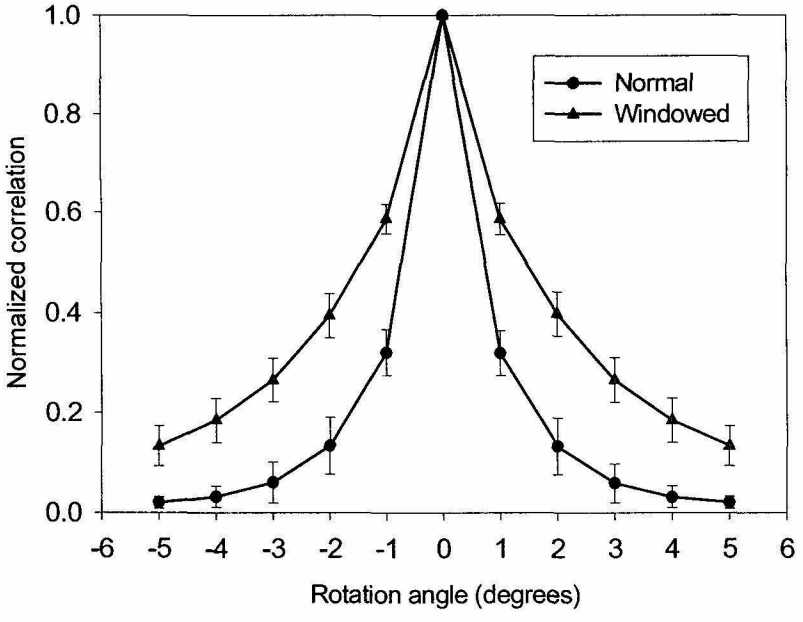


Fig. 5-5. Dependence of correlation on rotation

tiple instances of fingerprints that belong to the same class or the templates of other classes. We calculated the Mahalanobis distance between the within-class and between-class correlation, defined as

$$M = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{5-1}$$

where μ_1 and σ_1 (μ_2 and σ_2) are the mean and standard deviation of within-class (between-class) correlation. The Mahalanobis distance of normal correlation is 1.28 while that of the windowed correlation is 2.81. In this special case, no error would be made using either method. However, we expect that the increased margin provided by the windowed correlation is crucial for improved performance in a large database.

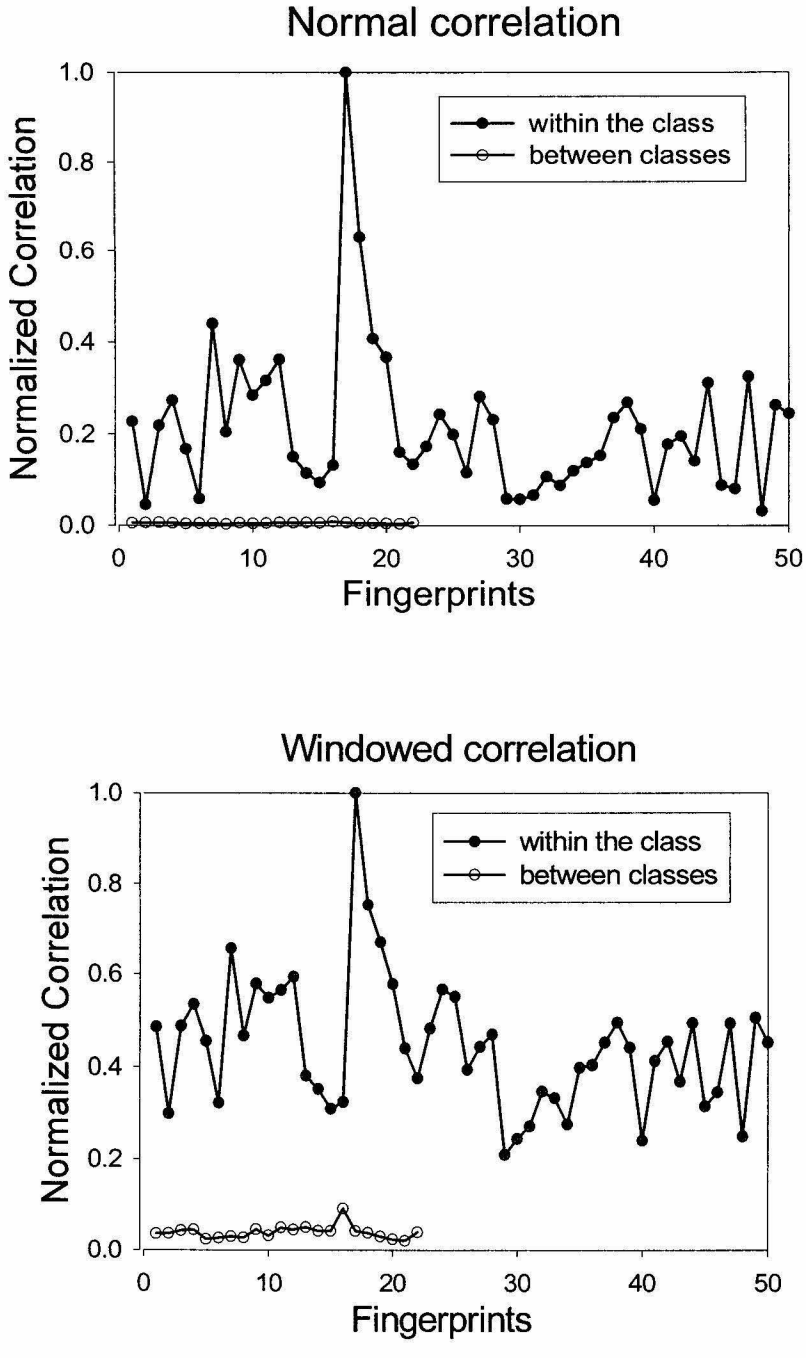


Fig. 5-6. Comparison of the normal and windowed correlations

5.4 Digital correlation

In this section, we study the performance of correlation algorithm with various strategies. We collected a new fingerprint database that consists of 24 classes with 50 samples per class, i.e. 1200 fingerprints total.

First, we investigate the normal correlations with 1, 2, 3, 4 and 5 templates per class (denoted as N1, N2, N3, N4 and N5). The templates are normalized so that the autocorrelation of any template is 1. The input fingerprint is classified according to the peak values of its correlations with all the templates. In the case of multiple templates per class, the response is the sum of the peak values of its correlations with all templates of the class. The input fingerprint is classified to the class which has maximal correlation response.

We then broke the input fingerprint into four (W4) or nine (W9) windows uniformly to perform the windowed correlation as shown in Figure 5-7 a & b. We also adapted our

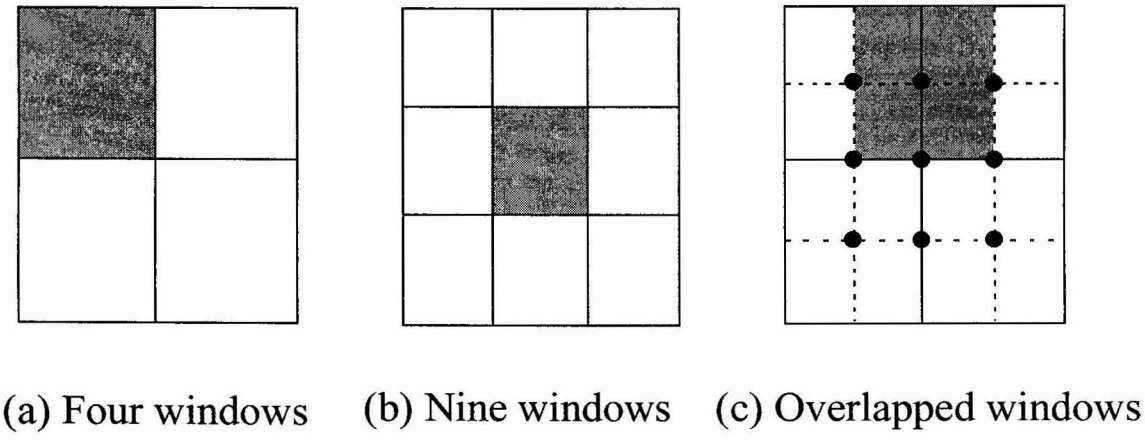


Fig. 5-7. Window sizes and positions. (The gray region represents a typical window)

algorithm and used nine overlapped windows (OW) whose width and height are half of the original size. In this scenario the window was centered at the nine dots shown in

Figure 5-7c and the partial fingerprint image within the window was correlated with the templates at each time. As in the earlier discussion the peaks of the correlations between windows of the input and a template are detected and their values were squared and summed up. In windowed correlation there are cross correlations between different parts (subfields) of the fingerprint. To reduce this noise we use the normal correlation peak position as a reference point. We only search around that reference point when computing the correlation of a window of the input and the templates since the cross correlation between different subfields is further away from the reference point. We used this subfield cross correlation suppression (SCCS) strategy in the windowed correlation (nine uniform windows). The correlation-peak-searching-window in the correlation plane has the same size as that of the input window.

Finally, synthetic filters [5-7]-[5-10] or composite templates are studied. A simple filter is the direct sum of the available templates. Because 4 templates per class (see Table 5-1) can result into very low FCR (false classification rate), we sum up the 4 templates to make a composite image (C4). But these 4 templates may have displacements relative to each other. We align these templates according to their cross correlation peak position and normalize their auto-correlation to 1 before the summation (NC4). We also apply the overlapped windowed correlation using the normalized composite template (NC4OW). The simulation results are shown in Table 5-1. In our simulation we always

Table 5-1. FCR of correlation based fingerprint identification

Single template					Multiple templates				Composite template		
N1	W4	W9	OW	SCCS	N2	N3	N4	N5	C4	NC4	NC4OW
4.7%	3.7%	2.7%	1.3%	1.0%	0.67%	0.17%	0%	0.083%	5.9%	1.2%	1.1%

classify the input fingerprint to some class and thus FRR is 0. We use the FCR as the metric of performance.

5.5 Conclusion

The performance of the correlation based fingerprint identification is summarized in Figure 5-8.The correlation algorithm and its variations can be very efficient (FCR ~ 0

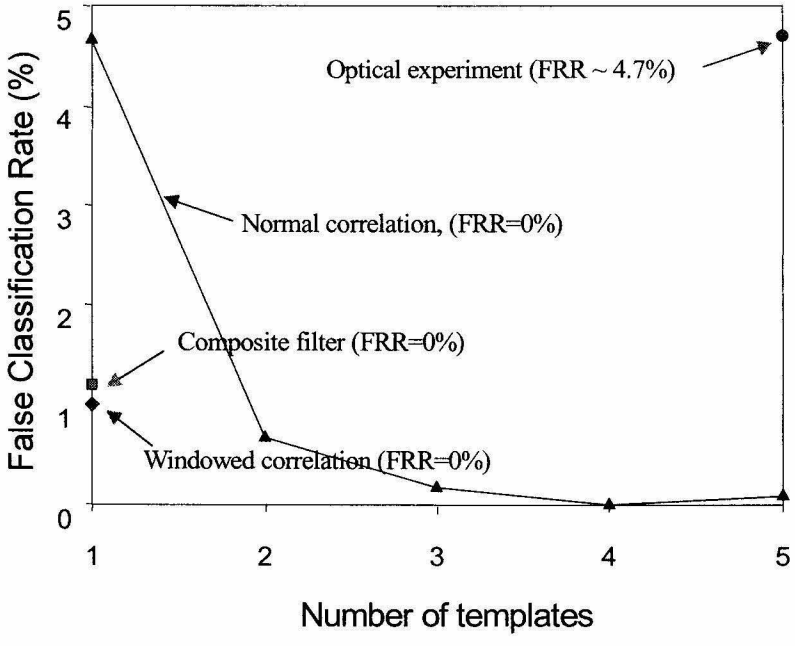


Fig. 5-8. Performance summary

using four templates per class and FCR ~1% using windowed correlation). In the case of searching a large database where the cost of memory and the speed are very important, we can turn to optics for help and combine the minutia based algorithm with the optical correlator to take advantage of the low cost of holographic memory and massive parallelism and high speed of optics.

References

[5-1] Anil Jain, Ruud Bolle and Sharath Pankanti (Eds.). Biometrics-Personal Identification in Networked Society (Kluwer Academic Publishers, Boston, 1999)

[5-2] See, for example, http://www.digitalpersona.com/

[5-3] DigitalPersona white paper "Guide to fingerprint identification" (http://www.speak-it.com/documents/fingerprintidentification.pdf, http://www.digitalpersona.com)

[5-4] Milos Klima *et. al.* "Mode of 2D Optical Correlator for Fingerprint Identification." IEEE AES Systems Magazine, 3-9 (1997)

[5-5] V. Soifer, V. Kotlyar, S. Khonina, and R. Skidanov. "Optical-digital methods of fingerprint identification." Optics and Lasers in Engineering 29, 351-359 (1998)

[5-6] CNS 248 class projects. Dmitri Linde, "A new approach to fingerprint verification." Tom Olick and Stefan Schlamp, "Fingerprint verification."

[5-7] David Casasent. "Unified synthetic discriminant function computational formulation." Appl. Opt. 23, 1620-1627 (1984)

[5-8] B.V.K. Vijaya Kumar. "Tutorial survey of composite filter designs for optical correlators." Appl. Opt. 31, 4773-4801 (1992)

[5-9] A. Mahalanobis, B.V.K. Vijaya Kumar, and S.R.F. Sims. "Distance-classifier correlation filters for multiclass target recognition." Appl. Opt. 35, 3127-3133 (1996)

[5-10] B.V.K. Vijaya Kumar and Daniel W. Carlson. "Optimal trade-off synthetic discriminant function filters for arbitary devices." Opt. lett. 19, 1556-1558 (1994)