OPTICAL NEURAL NETWORKS

USING

VOLUME HOLOGRAMS

Thesis by

Claire Xiang-Guang Gu

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1990

(Submitted September 19, 1989)

ACKNOWLEDGEMENTS

I am indebted to my advisor, Professor Demetri Psaltis. Since I joined this group, he has shown me the way of doing research and led me to the accomplishment of this thesis.

It is a pleasure to acknowledge previous and present members of the optical information processing group. Thanks to Dr. Hyuk Lee, Dr. John Hong, Dr. Jeffrey Yu, David Brady, Dr. Ken Hsu, Dr. Eung Gi Paek, Dr. Kelvin Wagner, Dr. Fai Mok, Dr. Robert Snapp, Cheol-Hoon Park, Nabeel Riza, Scott Hudson and Alan Yamamura for many useful discussions.

I am also grateful to Mrs. Helen Carrier and Ms. Su McKinley for their help and kindness.

Finally I would like to thank Peiwei Gu and Manqi Tan, my parents. For many years, they have paid enormous attention to my education and given me great encouragement. I would also like to thank Yanong Zhu, my husband. With his love and help, my stay at Caltech was more enjoyable.

ABSTRACT

The optical implementation of neural networks utilizing volume holograms is investigated. The intrinsic degeneracy effect that limits the number of independent interconnections are identified and analyzed by applying the **K**-space analysis. Basic relationships between the number of neurons, the number of interconnections and the size of the optical system that is used to implement the neural network are derived. Systematic methods for selecting the positions of the neurons to achieve the maximum number of independent interconnections are described. Experiments of global and local connectivities accomplished by using fractal sampling grids for eliminating the degenerate interconnections are presented. The degrees of freedom of the volume hologram and of the planar hologram are compared.

Table of Contents

1. INTRODUCTION

Motivated by understanding perception, pattern recognition, learning, and other functions of the brain, neural biologists detect electrical signals sent by individual neurons, trace the propagation of these signals, study the bio-chemical basis for generating these signals and try to develop a picture of how neurons operate collectively to realize complicated functions. They have accomplished a great amount of knowledge about nervous systems. But the variety and complication of neural systems require much more effort before they can be understood in sufficient detail.

A neural network consists of a large number of neurons connected by synapses. Nerve cells (neurons) have different shape and size. But the basic function of all neurons is the same. Each neuron sends out pulses to other neurons through the synapses (or interconnections), and processes the pulses received from many other neurons. Typically each neuron is connected with thousands of other neurons.

Mathematical models are created to describe neural networks. Based upon the basic structure and the observations of experimental neural biologists, theoretic analysis are conducted to model the functioning of a large number of interconnected neurons. Different models are created to describe the neural network in different ways. These models help to understand the functioning of the human brain and provide new algorithms to build new hardware with computational power imitating the neural system.

Electrical and optical systems are built to implement neural network models. The signal transmission in an electrical circuit is similar to that in a biological

neural network. By designing a complicated circuit, neural network models can be implemented using many processors connected by wires. The area required for these wires, however, limits the number of connections and makes the implementation of dense interconnections difficult. Optics provides another way to implement neural network models. Its large capacity and intrinsic parallelism have great advantages over electronics. The optical neural network, neural network architecture combined with the advantages of optics, brings new hope for computers with higher intelligence.

1.1 Neural Network Models

1.1.1 Neural Network Architecture

All neural network models follow the same basic architecture, simulating that of the biological neural system. A neural network consists of many identical elements (neurons) linked by interconnections (synapses). The differences between different models lie in the way neurons are connected and the way interconnections are formed. Neurons are usually depicted as points and interconnections as lines. Fig.1.1.1 shows the basic architecture.

Each neuron receives signals from and sends signals to many other neurons. The received signals are magnified or attenuated by the interconnections and summed by the neuron. Suppose neuron i receives a total input signal x_i^{in} from c different neurons, and the output signal sent by neuron j is x_j^{out}. The relationship between the received and transmitted signals is

$$x_i^{in} = \sum_{j=1}^{c} w_{ij} x_j^{out}, \tag{1.1}$$

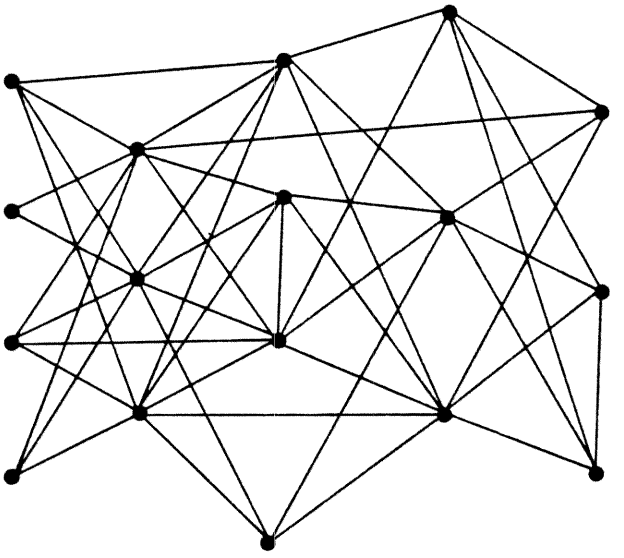


Fig.1.1.1 Neural network architecture.

where w_{ij} is the interconnection weight between neurons i and j. The total received signal is then processed by neuron i to send out an output signal

$$x_i^{out} = g(x_i^{in}). \tag{1.2}$$

The commonly used function, $g(x)$, has a lower bound and a upper bound with continuous values in between, for example, a sigmoid function, as in Fig.1.1.2. When the input is very small, the output is the minimum value. When the input is some moderate value, the output is something in between. When the input is very high, the output reaches its maximum value. An extreme case of the function is obtained when the output jumps from its minimum to its maximum at certain threshold input value, also shown in Fig.1.1.2. In this case, the output has only two possibilities, low, if the input is below the threshold value and high, otherwise. The processing performed by the neuron

in this case is a threshold-logic function. This step function is called a hard thresholding function. In analogy, a continuous function $g(x)$ is called a soft thresholding function.

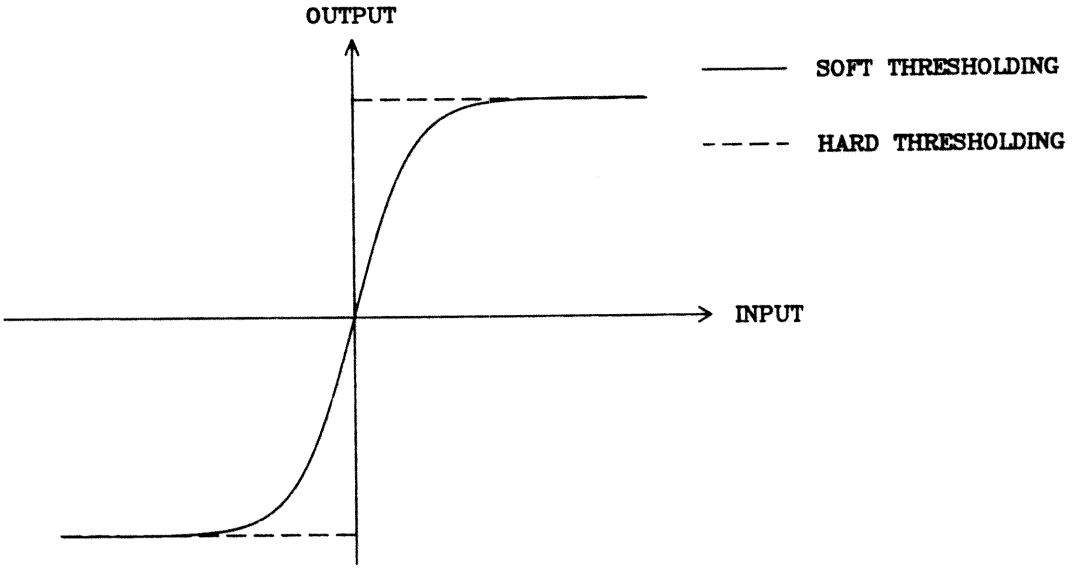


Fig.1.1.2 Thresholding functions.

Many models assume that neurons are divided into different layers, as shown in Fig.1.1.3. A neuron receives signals only from neurons in the previous layer and sends signal only to neurons in the next layer. The simplest network is a single layer network. It contains one layer of interconnections and two layers of neurons. The neurons before and after the interconnections are called the input and output neurons respectively. Multilayer networks are needed to achieve complicated functions.

One simple neural network is a classifier. It contains two layers of neurons, N input neurons and only one output neuron. The task is to classify many objects

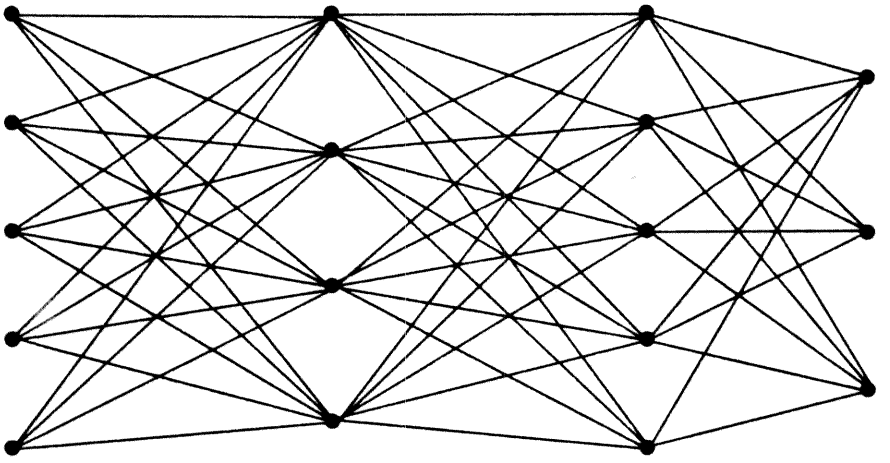


Fig.1.1.3 Multilayer neural network.

into two different classes. An object can be specified by a given set of values for all N neurons in the first layer. The class which the given object belongs to corresponds to the output value of the neuron in the second layer. This neuron uses a hard thresholding function to give an output value of either high or low, usually 1 or -1. All objects corresponding to a final output value 1 are classified into class 1, otherwise class 2. By selecting the proper weight matrix $\mathbf{w}$ and the threshold value w_0, a number of objects can be classified into two desired classes. As N becomes very large, the maximum number of objects that can be correctly classified into any possible dichotomy is defined as the capacity of this neural network. The capacity of a linear classifier is $2(N+1)$ [1].

More complicated neural networks can be derived from the simple classifier. A general single layer network can be generated by adding more output neurons, each of those being a classifier. Multilayer networks can be created by cascading

a single layer network. In general, the set of functions that are implementable by a neural network depends on its structure and interconnections.

1.1.2 Formation of Interconnections

Once the structure of a neural network is given, the interconnections need to be specified. The structure determines the capacity of the neural network. The interconnections depend on the particular task. For example, let us assume that a neural network is designed to recognize several faces. How many faces can be recognized depends on the structure. Whose faces can be recognized depends on the interconnections, i.e., the weight matrix. The procedure that the interconnections are specified according to given samples is called training.

A number of training methods have been found. Some treat neural networks as dynamical systems, and they are known as dynamical training algorithms. Some calculate the weight matrix directly from the given samples, known as direct calculations.

Dynamical training typically specifies an energy function in weight space. Different training methods define different energy functions. The global minimum of the energy function is the point corresponding to the correct weight matrix. A training procedure starts from any point in the weight space, where the energy function is usually not at a minimum value. Then the first point is moved along the energy surface according to the steepest descent until the global minimum point is reached.

An example of dynamical training is the *Perceptron.* Consider only a single output neuron with a hard thresholding function. The weight matrix becomes a

weight vector $\mathbf{w}$. The goal is to find a weight vector such that

$$\begin{cases} \mathbf{w} \cdot \mathbf{x}^i > 0 & \forall\, \mathbf{x}^i \in \text{Class 1}, \\ \mathbf{w} \cdot \mathbf{x}^i < 0 & \forall\, \mathbf{x}^i \in \text{Class 2}, \end{cases} \tag{1.3}$$

where $\mathbf{x}^i$ is a sample vector whose elements specify the values of the first-layer neurons. (Here the threshold value w_0 was absorbed into the weight vector assuming $x_0^i = 1$ for all sample vectors.) The total number of arbitrary vectors to be classified, M, must not exceed the capacity of the neural network. The Perceptron algorithm starts from an arbitrary weight vector $\mathbf{w}_1$. Try to classify the sample vectors with $\mathbf{w}_1$. If a sample vector is correctly classified, go to the next sample. If a sample vector is misclassified, change the weight vector according to

$$\mathbf{w}_2 = \mathbf{w}_1 + \sigma^i \mathbf{x}^i, \tag{1.4}$$

where

$$\begin{cases} \sigma^i = 1 & \text{if } \mathbf{x}^i \in \text{Class 1}, \\ \sigma^i = -1 & \text{if } \mathbf{x}^i \in \text{Class 2}. \end{cases} \tag{1.5}$$

Then try to classify the sample vectors with $\mathbf{w}_2$. Change the weight vector to $\mathbf{w}_3$ when a misclassified sample is encountered. Follow the same procedure in the kth step,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \sigma^k \mathbf{x}^k, \tag{1.6}$$

whenever there is a misclassified sample vector $\mathbf{x}^k$, until all the sample vectors are correctly classified. The final weight vector is not unique but can always be found as a result of this training, assuming a solution exists. The convergence proof can be found in reference [1].

In some cases, the weight matrix can be easily calculated and direct calculation methods can be used.

An example of direct calculation is the *Simple Sum*. Given the same set of sample vectors as before, the Simple Sum algorithm calculates the weight vector as

$$\mathbf{w} = \sum_{i=1}^{M} \sigma^i \mathbf{x}^i. \tag{1.7}$$

To verify that this weight vector can correctly classify a sample vector $\mathbf{x}^l$, consider the product $\mathbf{w} \cdot \mathbf{x}^l$.

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}^l &= \sum_{i=1}^{M} \sigma^i \mathbf{x}^i \cdot \mathbf{x}^l \\ &= \sigma^l \mathbf{x}^l \cdot \mathbf{x}^l + \sum_{i \neq l} \sigma^i \mathbf{x}^i \cdot \mathbf{x}^l. \end{aligned} \tag{1.8}$$

When two different sample vectors are statistically independent, the first term in the last expression will dominate. The product will be greater than 0 if σ^l is 1 and less than 0 if σ^l is -1. Therefore the weight vector so calculated will correctly classify the sample vectors.

The above methods and many others used in training a classifier also apply to the training of neural networks with more than one output neuron. Each neuron can be trained as a classifier, and the weight matrix can be decomposed into weight vectors with respect to each output neuron. For each input vector, the neural network will give a series of classifications, resulting in an output vector.

Direct calculation and dynamical training are used in different situations. For multilayer networks, dynamical training can obtain interconnections without knowing the outputs of the middle-layer neurons. The most commonly used

method is Backward Error Propagation (BEP) [2]. For single layer networks, direct calculation can reach the optimal solution without changing the weights iteratively.

1.1.3 Associative Memory

The interconnections are stored in the network as memories, once they are established. When the input to the network is one of the sample vectors, the network will recall the related output stored during the training procedure.

The memory used in neural networks is associative memory, or content addressable memory. It can recall the related output by inputting the content of a stored vector, even partially correct. The content of the output can be the same as (auto-associative memory) or different from (hetero-associative memory) that of the input used in the training procedure.

The outer product scheme is one of the associative memory models. Suppose the input vectors $\{\mathbf{x}^k\}$ are to be associated with the output vectors $\{\mathbf{y}^k\}$, $(k = 1, 2, \cdots, M)$, respectively. The weight matrix, according to the outer product scheme, is a sum of the outer product of the input and output vectors

$$\mathbf{w} = \sum_{k=1}^{M} \left|\mathbf{y}^k\right\rangle \left\langle\mathbf{x}^k\right|. \tag{1.9}$$

The element w_{ij} is

$$w_{ij} = \sum_{k=1}^{M} y_i^k x_j^k. \tag{1.10}$$

The outer product scheme can implement both auto-associative memory ($\mathbf{y}^k = \mathbf{x}^k$) and hetero-associative memory ($\mathbf{y}^k \neq \mathbf{x}^k$).

The outer product scheme works in the same way as the Simple Sum algorithm, as discussed in the last subsection. The Simple Sum algorithm is a special case when y_i^k is either 1 or -1.

When an input vector is used to recall the corresponding output vector, the network performs an inner product of the weight matrix and the vector. Suppose the first input vector, $\mathbf{x}^1$, is presented to the network, the inner product

$$\mathbf{w} \cdot \mathbf{x}^1 = \left|\mathbf{y}^1\right\rangle \left\langle \mathbf{x}^1 \middle| \mathbf{x}^1 \right\rangle + \sum_{k=2}^{M} \left|\mathbf{y}^k\right\rangle \left\langle \mathbf{x}^k \middle| \mathbf{x}^1 \right\rangle \tag{1.11}$$

consists of two terms. The first term is the desired output $\mathbf{y}^1$ multiplied by $\left\langle \mathbf{x}^1 \middle| \mathbf{x}^1 \right\rangle$. The second term represents the noise resulting from the cross correlation between different input vectors. When the number of memories, M, does not exceed the capacity, the thresholding function of the output neurons will give the correct answer without the cross talk noise.

The capacity is the maximum number of associative memories that can be stored and correctly recalled by the network. It can be calculated under the assumption that the input and output vectors are independent random vectors with each component a random variable with equal probability of being 1 and -1. In this case, each component in the vector equation, Eq.(1.11), is a random variable with Gaussian probability distribution, according the Central Limit Theorem. The threshold value is set at 0. The probability of correctly classifying each output element is required to approach 1 when the number of input neurons approaches infinity. The maximum number of associated pairs of random vectors

can be stored and recalled is

$$M_{max} = \frac{N}{2(logN + logN_1)}, \tag{1.12}$$

where N_1 is the number of neurons at the output layer [3].

The outer product scheme can tolerate errors in the input vectors, that is, if the input vector used in the recall process is not exactly that used in the training process, the output vector will still be the desired one. For example, an input vector used in the recall process is x^1_{error}, which is close to the input vector x^1 used in the training process but different in ρN bits. The inner product

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}^1_{error} &= \left|\mathbf{y}^1\right\rangle \left\langle \mathbf{x}^1 \middle| \mathbf{x}^1_{error}\right\rangle + \sum_{k=2}^{M} \left|\mathbf{y}^k\right\rangle \left\langle \mathbf{x}^k \middle| \mathbf{x}^1_{error}\right\rangle, \\ &= \left|\mathbf{y}^1\right\rangle (1-2\rho)N + \sum_{k=2}^{M} \left|\mathbf{y}^k\right\rangle \left\langle \mathbf{x}^k \middle| \mathbf{x}^1_{error}\right\rangle, \end{aligned} \tag{1.13}$$

under the same conditions for random binary vectors. The signal term is decreased by a factor $(1 - 2\rho)$. As long as the signal term still dominates, the thresholded output will be $\mathbf{y}^1$, the desired output vector.

However, tolerating errors in the input vectors decreases the maximum number of vectors that can be correctly recalled, i.e., the capacity. For the same conditions, the capacity of a network capable of tolerating any input vector with ρN bits errors is

$$M_{max} = \frac{(1-2\rho)^2 N}{2(logN + logN_1)}. \tag{1.14}$$

The Hopfield model modifies the outer product scheme and introduces feedback to realize an auto-associative memory [4]. The weight matrix, in this model,

is a sum of outer products of the training vectors without the diagonal terms,

$$w_{ij} = \begin{cases} \sum_{k=1}^{M} x_i^k x_j^k & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{1.15}$$

Any input to the network will produce some output at the second layer. This output is returned to the first layer as new input. The iterations result in a flow in the space specifying the vectors. As the feed-back continues, the flow converges to a stable state corresponding to the stored vector which is nearest to the input vector. This model has been implemented both electronically [5] and optically [6].

1.2 Optical Implementations

1.2.1 Advantages of Optics

Optics provides large capacity in neural network modeling. The human brain contains more than 1,000 million [7] cells and most of its volume is occupied by interconnections. It is difficult for electronics to implement such massive interconnections in a reasonable volume. For optics, the short wave length of light enables it to accomplish huge number of interconnections in a relatively small volume. As will be discussed in the following chapters, the storage capacity of an optical system with volume V is proportional to V/λ^3, where λ is the wavelength of light.

Optics is a convenient technology to realize parallel processing. For example, a serial digital computer does a Fourier transform by digitizing the input into a sequence of discrete values, doing the integral by sequentially handling these

discrete values, and finally giving a display containing a finite number of pixels. Optically, the Fourier transform can be simply done by a lens. Putting the input pattern at the front focal plane of a converging lens and illuminating it with a coherent plane wave, its Fourier transform appears at the back focal plane. The intrinsic parallelism of optics is very useful in simulating parallel processes in neural networks.

Optics is also capable of implementing interconnections in 3-dimensional space. While building many layers in a VLSI chip is difficult, optical interconnections can be stored in 3-dimensional crystals. Further more, the interconnections in neural network models must be modifiable in order to implement learning. Photorefractive crystals provide a medium for changing the interconnections in real time.

1.2.2 Optical Neural Networks

In the recent years, many experiments have demonstrated the power of optical implementation of neural networks [8–14].

An example is the optical auto-associative loop with planar holograms [15], performed at Caltech. A Liquid Crystal Light Valve (LCLV) is used in this experiment to simulate a plane of neurons. The memories are stored in two thin holograms. Four pictures are stored in the loop. When an image is presented to the system, it forms correlations with all the stored pictures. The diffracted beam coming out of the two holograms contain a linear combination of all four pictures with the strength of each picture determined by the correlation between the input and the picture. The output is again directed to the neural plane, forming a

feed-back loop. The LCLV does the thresholding. After several iterations, only the picture with the strongest correlation with the input image is left as the final output. This experiment is an optical implementation of the Hopfield model.

Another example is the optical perceptron using photorefractive crystal [16, 17]. Two identical images are exposed to the crystal forming a weight vector with each of its element a volume hologram recorded in certain spatial location of the crystal. By either writing or erasing the hologram according to the training patterns, the weight vector is changed iteratively as described in Eq.(1.6). When the hologram is exposed to an input sample image, it diffracts the beam to a detector which integrates the output image and forms a single thresholded output signal. The output signal is either high or low corresponding to the two states of a neuron. If the input gives a correct output, the hologram will not be changed. If the input corresponding to a high output gives a low signal, there is a feed-back system controlling the training beams to write the hologram of the input pattern. Otherwise, if the input corresponding to a low output gives a high signal, the feed-back system will turn on a piezoelectric mirror so that the two images forming on the crystal are not coherent, therefore the hologram of the input pattern will be erased. The amount of interconnection strength to be recorded or erased depends on the time the crystal is exposed to the training beams. This experiment demonstrates an implementation of dynamical training process.

In this thesis, the investigation will be concentrated on the optical implementation of interconnections using Fourier holograms in thick medium. The input and output patterns are Fourier transformed and combined to form an

interference pattern. This interference pattern is then recorded as the spatial variation of the refractive index inside a photorefractive crystal. During the reconstruction, the input pattern is used as the reference beam, the reconstructed object beam is the stored output pattern. The use of Fourier holograms enables the partial input pattern to reconstruct the whole output pattern. Therefore, it is convenient to implement the associative memories of neural networks. In the optical system, neurons are implemented by points, interconnections are implemented by sinusoidal gratings superimposed in the photorefractive crystal. The use of volume holograms allows the storage of information in 3-dimensional space, therefore the optical system will have higher storage density than the system using planar holograms [18, 19]. Unfortunately, sinusoidal gratings in a 3-dimensional crystal can not independently interconnect input and output neurons both located at 2-dimensional planes [20, 21]. Details of the problems and solutions will be discussed in the following chapters.

1.3 Summary of Thesis

The basic mechanism of 3-dimensional storage of interconnection weights will be introduced in Chapter 2. The Vander Lugt system is used to store and read Fourier holograms written in the photorefractive crystal. The physical background for volume holography — the photorefractive effect and coupled wave analysis will also be reviewed.

In Chapter 3, a geometric **K**-space analysis will be used to discuss the storage of multiple gratings in a crystal with finite volume. Given a crystal, the total number of gratings can be stored in it, i.e., the storage capacity of the crystal,

is proportional to V/λ^3, where V is the volume of the crystal and λ is the wavelength of light used to record and read the gratings. This capacity can not be fully utilized due to the finite dimensions of the practical system. Among the practically accessible gratings, each grating can interconnect more than one pair of input-output pixels, causing degenerate interconnections. **K**-space analysis can also calculate the angular separation required between two distinguishable pixels. The result of this calculation is compared with that obtained from the coupled wave theory. The comparison indicates that the **K**-space analysis is complementary to coupled wave theory. The combination of both analyses provides a better understanding of volume holograms.

In Chapter 4, fractal sampling grids are derived to solve the degeneracy problem. To implement independent interconnections, the locations of neurons at both input and output planes are selected so that each grating connects only one pair of input-output neurons. A systematic way of designing sampling grids is described. Fractal mathematics is applied to the generation of higher order sampling grids while keeping the fractal dimension unchanged. Different kinds of sampling grids are derived.

Experiments demonstrating both global and local connectivities are described in Chapter 5. The resulting interconnections implement the outer product scheme. Hetero-associative memories are stored in the crystal and the desired output patterns are recalled by the corresponding input patterns, with the help of fractal sampling grids.

In conclusion, Chapter 6, volume hologram and planar hologram are compared. It is shown that the volume of the system is much smaller when volume

holograms are used.

In the Appendix, the optical implementation of neural networks using planar holograms is discussed. The **K**-space analysis is extended to the planar holograms. Fractal sampling grids for planar holograms are also derived.

2. OPTICAL HOLOGRAPHIC INTERCONNECTIONS

To implement a neural network, the transfer function of the system must be modifiable according to the given input and output. Consider a network containing a layer of input neurons, a layer of output neurons, and interconnections between the two layers. The weight matrix specifying the interconnections works as a transfer function which maps the input vector sent by the input layer to the output vector received by the output layer. The task of the training process is to find such a transfer function that accomplishes a desired mapping.

Optical holography provides a method to generate a transfer function between the desired input and output. The transfer function is recorded as a hologram by using the desired input as the reference beam and the desired output as the object beam. The reconstruction of this hologram diffracts the reference beam to the object beam, i.e., maps the input to the desired output. The conventional hologram recording takes two steps: 1) exposure of a film to the interference pattern; 2) film development. The second step makes the dynamic modification of a hologram inconvenient.

The discovery of photorefractive materials enables real time holography, since the development is not required. In addition, the photorefractive material allows the recording of holograms in 3-dimensional crystals — volume holograms. In the following sections, the recording and reconstruction of volume holograms, and the problem associated with 3-dimensional storage of interconnection weights will be discussed.

2.1 VANDER LUGT SYSTEM

The basic system used throughout this thesis is the Vander Lugt system, as shown in Fig.2.1.1. It consists of two Fourier transforming lenses with the same focal length f, separated by $2f$; the input plane located at the front focal plane of the first lens; and the output plane located at the back focal plane of the second lens. The input plane is illuminated by a collimated plane wave. The first lens takes the Fourier transform of the field distribution at the input plane. The resulting Fourier transform pattern appears at the common focal plane, the back focal plane of the first lens and the front focal plane of the second lens. At the common focal plane, the light distribution can be modified to realize Fourier domain processing. The second lens Fourier transforms this modified light distribution back to the spatial domain, resulting in a processed pattern at the output plane.

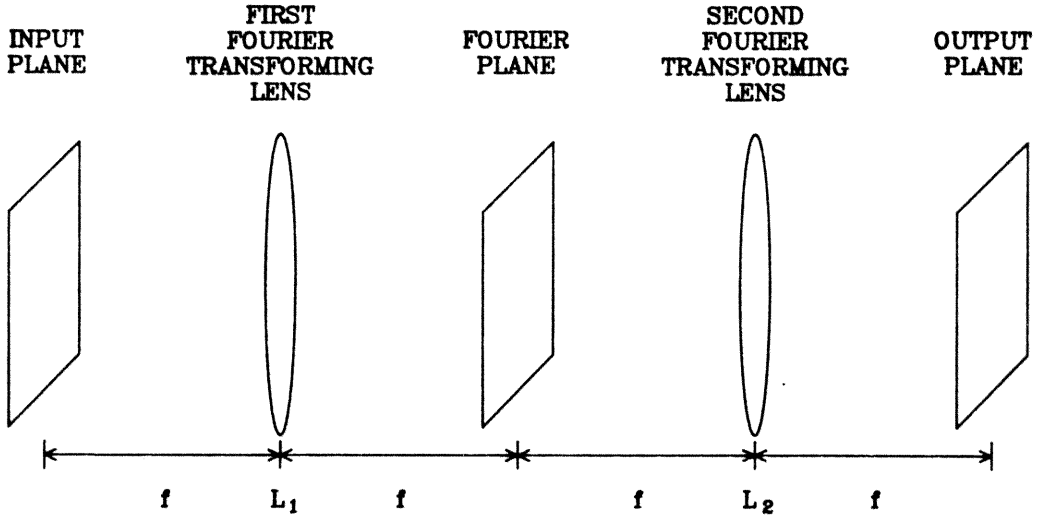
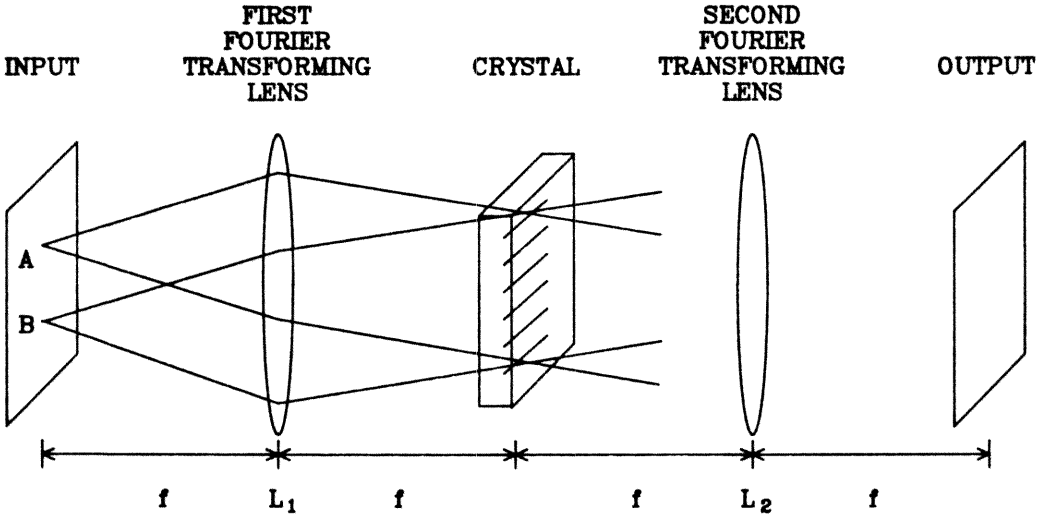


Fig.2.1.1 The Vander Lugt system.

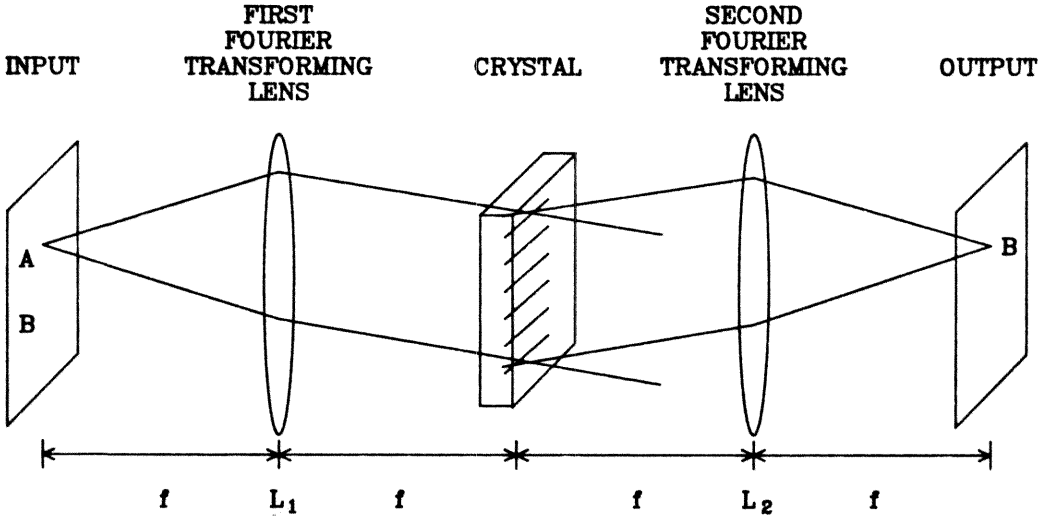
Consider two separate points at the input plane, A and B respectively. If nothing is done at the common focal plane (or Fourier plane), the output will simply be the images of the input points. Light originating from the input point A will not propagate to the output point B. In other words, the input point A is not connected to the output point B.

The input point A and the output point B can be connected by using a hologram in two steps. The first step is shown in Fig.2.1.2(a). The first lens converts the spherical waves, coming from the two points at the input plane, to two plane waves interfering with each other at the Fourier plane. When a crystal is placed at the Fourier plane, it can record this interference pattern as a hologram. This step corresponds to the training process of neural networks. The second step is shown in Fig.2.1.2(b). While the input point B is blocked, the plane wave coming from the input point A will be diffracted by the hologram to another plane wave, which is in turn converted by the second lens to a light spot at the place where the output point B used to be. This step corresponds to the recall process in neural networks. The input point B is only needed during the training process.

To set up a complete system to implement a neural network, the crystal is placed at the Fourier plane to record the interference pattern between the input and the training patterns. The front focal plane of the first lens is divided into two separate parts. One of them is called the input plane (from now on, only this part of the front focal plane is regarded as the input plane), and the other the training plane. During the training process, the input pattern is placed at the input plane, and the desired output pattern is placed at the training plane.



(a)



(b)

Fig.2.1.2 (a) The training process. (b) The recall process.

The back focal plane of the second lens is also divided into two separate parts. The part corresponding to the image of the training plane is the output plane.

The other part corresponding to the image of the input plane is not important in the discussion. Fig.2.1.3 shows the input plane, the training plane, and the output plane. During the recall process, the training plane is blocked. The light beam coming from the input plane will be partially diffracted by the crystal to the output plane.

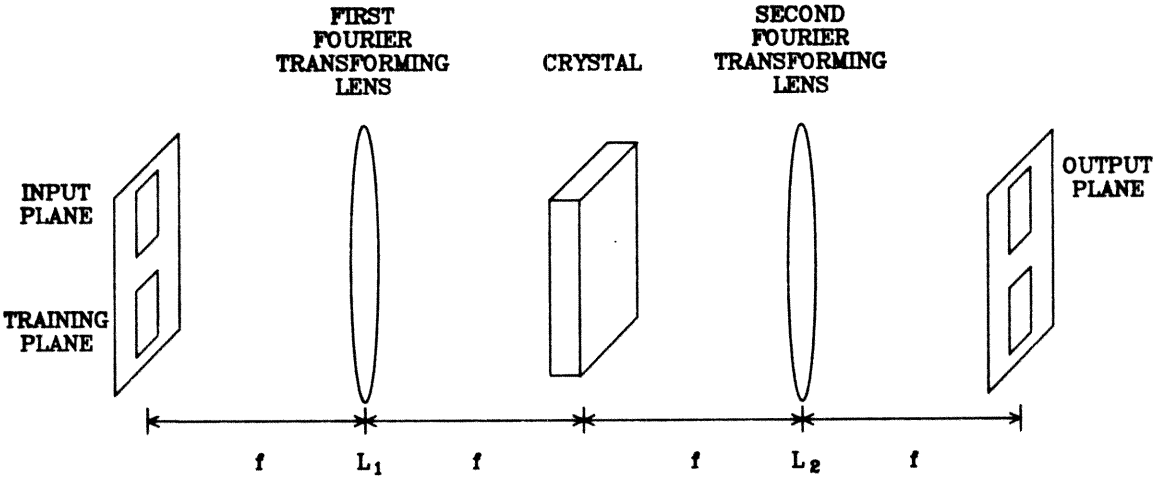


Fig.2.1.3 The input plane, the training plane and the output plane.

This system simulates a single layer neural network. Input and output neurons are implemented by points at the input and output planes respectively. The activation of a neuron is represented by the intensity at the location of that neuron. By setting a threshold intensity, hard thresholding neurons can be implemented. The output of a neuron is either high, if the intensity at that point is above the threshold, or low, if below.

In this system, interconnections are implemented by sinusoidal gratings inside

the crystal. Since the crystal is placed between two lenses which convert points to plane waves, the interference pattern generated by two points is sinusoidal. Many gratings can be superimposed inside the crystal.

2.2 Volume Hologram

A volume hologram is an interference pattern recorded in a 3-dimensional medium. A thick crystal can record more information than a thin film, because of the additional dimension. For example, compare the interference pattern of two plane waves recorded by a thin (planar) hologram and a thick (volume) hologram. The interference pattern is a sinusoidal grating

$$I(x, y, z) = 1 + \cos(K_x x + K_y y + K_z z), \tag{2.1}$$

where $\mathbf{K} = (K_x, K_y, K_z)$ is a grating vector whose magnitude reflects the fringe spacing and whose direction represents the orientation of the fringe normal. The volume hologram records these fringes in a finite volume with the information of both spacing and orientation preserved. In contrast, the planar hologram records only a 2-dimensional cross section of these fringes. For instance, a film placed at the $z = 0$ plane records only

$$I(x, y) = 1 + \cos(K_x x + K_y y). \tag{2.2}$$

The K_z component of the grating vector is not recorded in the film. Therefore, it is impossible to distinguish whether the fringes are tilted in the z-direction or not. Fig.2.2.1 shows two different gratings recorded by planar and volume holograms.

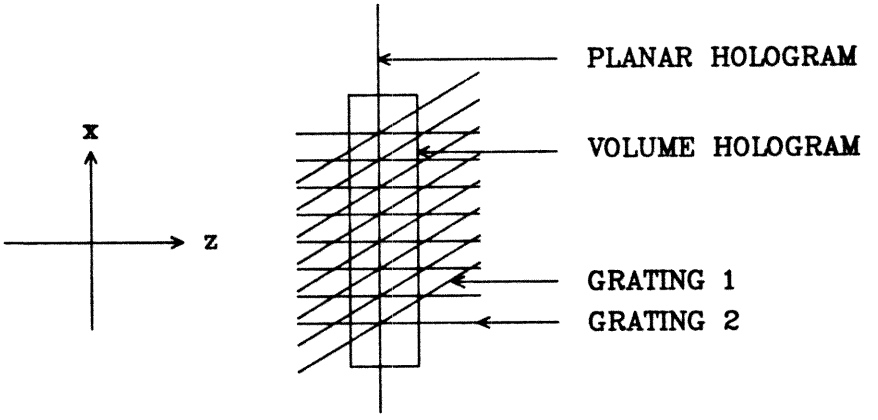


Fig.2.2.1 Two different gratings recorded by planar and volume holograms.

While the volume hologram records two distinct gratings, the recording in the planar medium is the same for both gratings.

The reconstruction of a volume hologram is also different from that of a planar hologram. For the same example as above, the grating recorded by the planar hologram can diffract any plane wave resulting in a diffracted plane wave (more details discussed in the Appendix). The grating recorded in the volume hologram will produce a diffracted plane wave only if the in-coming plane wave is incident at the Bragg angle, which will be discussed in subsection 2.2.2. The diffraction efficiency can reach 100%, if the thickness of the crystal is properly chosen.

Photorefractive crystals are widely used for volume holographic recording. In the following subsection, the properties of these materials will be briefly reviewed.

2.2.1 The Photorefractive Effect

The photorefractive effect has been discovered in many electro-optic crystals [22–27], such as Lithium Niobate ($LiNbO_3$), Barium Titanate ($BaTiO_3$), Strontium Barium Niobate (SBN) and others.

The recording of a holographic grating is the result of charge redistribution inside the crystal. A photorefractive crystal contains donors and empty electron traps (ionized donors). The donors are photosensitive electron traps with an absorption band in the visible region corresponding to the excitation of the electrons into the conduction band. When the crystal is exposed to two coherent laser beams with interference intensity

$$I(x,z) = I_0[1 + m\cos(Kx)], \tag{2.3}$$

the density of electrons excited to the conduction band is higher in the bright regions than in the dark regions. The excited free electrons are redistributed due to diffusion, drift and the photovoltaic effects, and they recombine with empty electron traps. The net charge density will be positive in the bright regions and negative in the dark regions, as shown in Fig.2.2.2. The electric field generated by the space charge density changes the refractive index of the crystal due to the electro-optic effect. The resulting index grating can be erased by heating or illuminating the crystal by a uniform laser beam [26].

Several theories for the photorefractive effect have been developed [28, 29, 30]. The Kukhtarev model, which is commonly accepted, describes the formation of

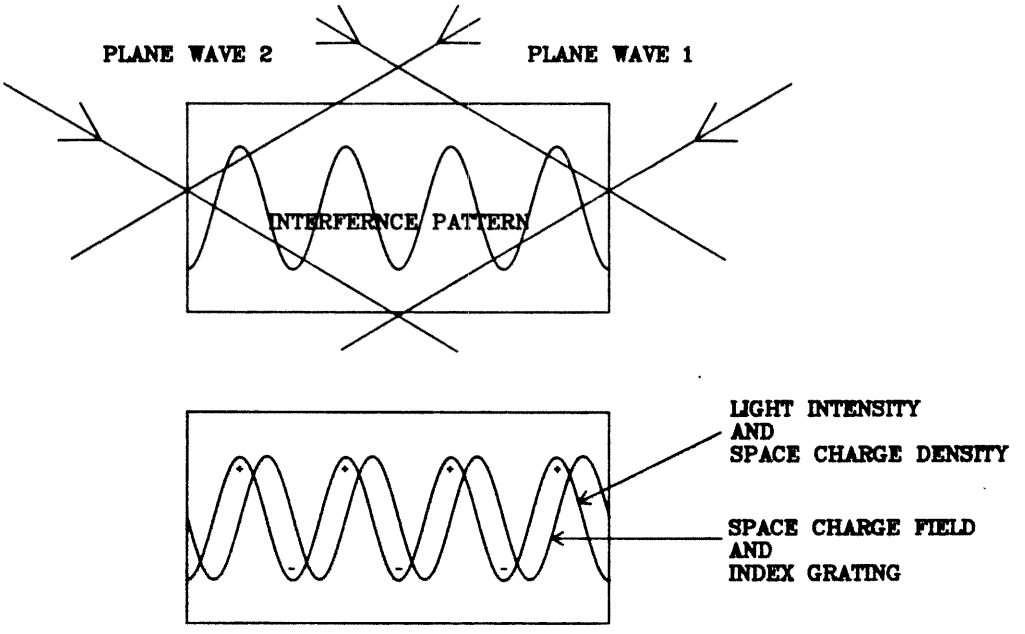


Fig.2.2.2 The charge density and the space charge field with respect to the interference pattern produced by two plane waves.

space charge field by the following equations:

$$\frac{\partial N_D^+}{\partial t} = (sI + \beta)(N_D - N_D^+) - \gamma_R n N_D^+, \tag{2.4}$$

$$\mathbf{j} = e\mu n(\mathbf{E} - \frac{k_B T}{e} \bigtriangledown \log n) + pI\mathbf{e}_c, \tag{2.5}$$

$$\frac{\partial n}{\partial t} = \frac{\partial N_D^+}{\partial t} - \frac{1}{e} \bigtriangledown \cdot \mathbf{j}, \tag{2.6}$$

$$\bigtriangledown \cdot (\epsilon_0 \mathbf{E}_{sc}) = 4\pi e(n + N_A - N_D^+). \tag{2.7}$$

Where N_D is the density of donors, N_D^+ is the density of ionized donors, n is

the density of charge carriers (usually electrons), s is the cross section of photo-ionization, I is the light intensity, β is the rate of thermal excitation, γ_R is the recombination constant, $\mathbf{j}$ is the current density, μ is the mobility of charge carriers, $\mathbf{E}$ is the total electric field, k_B is the Boltzmann's constant, T is the temperature, p is the photovoltaic constant, $\mathbf{e}_c$ is the unit vector along the c-axis, $\mathbf{E}_{sc}$ is the space charge field generated by charge redistribution, N_A is the density of acceptors, ϵ_0 is the dielectric constant, e is the charge of an electron, and t is the time.

Eq.(2.4) describes the photo- and thermal-ionization of donors and retrapping of electrons. Eq.(2.5) expresses the current arising from electrons moving in the conduction band due to drift, diffusion and the photovoltaic effect [31, 32]. Eq.(2.6) relates the charge density changing rate with the current. It has been assumed that the acceptor levels are completely filled by electrons and are not involved in phototransitions. But the existence of acceptors allows part of the donors to be ionized even in the dark [30]. Eq.(2.7) represents the generation of the space charge field as a result of charge redistribution.

The solution of these non-linear equations have been obtained analytically for the steady state [30]. Suppose the light intensity of the interference pattern is given by Eq.(2.3). Define two parameters

$$\xi_T = (\frac{l_T}{\Lambda})^2 = \frac{\epsilon_0 \pi k_B T}{e^2 \Lambda^2 N_A}, \tag{2.8}$$

and

$$\xi_E = \frac{l_E}{\Lambda} = \frac{\epsilon_0 E_0}{2e\Lambda N_A}, \tag{2.9}$$

where $\Lambda = K/2\pi$ is the fringe spacing and E_0 is the external electric field. l_T and l_E represent charge transport lengths due to diffusion and drift. Assume the space charge field is along the c-axis, i.e.,

$$\mathbf{E}_{sc} = E\mathbf{e}_c. \tag{2.10}$$

The steady state solutions for the two extreme cases are:

1) For l_T, $l_E \gg \Lambda$:

$$E = \langle E \rangle - M_0 E_q \sin(Kx), \tag{2.11}$$

where

$$M_0 = m(\frac{sI_0}{\beta + sI_0}), \tag{2.12}$$

$$E_q = \frac{4\pi e N_A}{\epsilon_0 K}, \tag{2.13}$$

m is the modulation depth given in Eq.(2.3) and $\langle E \rangle$ is the spatially averaged electric field.

2) For l_T, $l_E \ll \Lambda$:

$$E = \sum_{n=-\infty}^{\infty} E_n \exp(inKx) \tag{2.14}$$

is a periodic function with fundamental period Λ and magnitude of each component [30]

$$E_n = \Delta E[\frac{\sqrt{1+M^2}-1}{M}]^n \exp(i\phi), \tag{2.15}$$

where

$$\Delta E = \{[E_0 + E_p(1 - m/M)]^2 + E_d^2\}^{1/2}, \tag{2.16}$$

$$\phi = \cot^{-1}[\frac{E_0}{E_d} + \frac{E_p}{E_d}(1 - m/M)], \tag{2.17}$$

$$M = M_0(1 - \xi_T - \xi_E), \tag{2.18}$$

$$E_p = \frac{pI_0}{e\mu\langle n\rangle}, \tag{2.19}$$

$$E_d = \frac{k_B T K}{e}. \tag{2.20}$$

E_p and E_d are the photovoltaic field and the diffusion field, respectively. The fundamental term giving rise to a sinusoidal phase grating is

$$E_1 = \Delta E[\frac{\sqrt{1+M^2}-1}{M}]\exp[i(Kx+\phi)]. \tag{2.21}$$

The time varying space charge field has been simulated numerically [28, 29]. In the short writing time limit, the electric field is sinusoidal with the same period as the interference pattern. As time changes, higher order terms will appear.

The space charge field changes the refractive index of the crystal through the electro-optic effect [33]. The index ellipsoid under the perturbation of an electric field becomes

$$\begin{aligned}(\frac{1}{n_1^2} + r_{1k}E_k)x^2 + (\frac{1}{n_2^2} + r_{2k}E_k)y^2 + (\frac{1}{n_3^2} + r_{3k}E_k)z^2 \\ +2r_{4k}E_k yz + 2r_{5k}E_k zx + 2r_{6k}E_k xy = 1,\end{aligned} \tag{2.22}$$

where repeated index k is supposed to be summed over 1, 2, 3, and

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \tag{2.23}$$

is the electro-optic tensor.

The change of refractive index of a Fe doped $LiNbO_3$ crystal is

$$\Delta n_3 = -\frac{1}{2} n_3^3 r_{33} E_3, \tag{2.24}$$

where suffix 3 represents the direction along the c-axis. For $LiNbO_3$, r_{33} is the largest electro-optic coefficient, and index gratings are written primarily along the c-axis.

In our experiment, a Fe doped $LiNbO_3$ crystal will be used. Some parameters of the $LiNbO_3$ crystal are summarized in Table 2.1.

Table 2.1 Parameters of $LiNbO_3$

Parameter	Notation	Value	Condition	Reference
Curie Temperature	T_c	1210^oC		[23]
Point Group		$3m$ or C_{3v}	$T < T_c$	[34]
		$\bar{3}2/m$ or D_{3d}	$T > T_c$	[34]
Refractive Index	n_o	2.38	at 4500 Å	[35]
	n_e	2.28	at 4500 Å	[35]
	n_o	2.34	at 5000 Å	[35]
	n_e	2.24	at 5000 Å	[35]
Non-zero Electro-optic Coefficient	$r_{13} = r_{23}$	8.6×10^{-12} m/V		[33]
	$r_{22} = -r_{12} = -r_{61}$	3.4×10^{-12} m/V		[33]
	r_{33}	30.8×10^{-12} m/V		[33]
	$r_{51} = r_{42}$	28×10^{-12} m/V		[33]

More details of the photorefractive effect can be found in the review papers [36, 37].

2.2.2 Coupled Wave Analysis

Coupled wave theory [38] is one of the methods [39–43] used to analyze the reconstruction of volume holograms.

The coupled wave analysis for a sinusoidal phase grating starts from the wave equation

$$\bigtriangledown^2 E + k^2 E = 0. \tag{2.25}$$

The optical field has been assumed to have the form

$$\mathbf{E} = E \exp(i\omega t)\mathbf{e}_y, \tag{2.26}$$

where $\mathbf{e}_y$ is the unit vector along the y-direction and ω is the angular frequency of the monochromatic light wave. The phase grating produces a small perturbation to the dielectric constant

$$\epsilon = \epsilon_0 + \epsilon_1 \cos(\mathbf{K} \cdot \mathbf{r}), \tag{2.27}$$

where $\mathbf{K}$ is the grating vector which is assumed to be in the (x, z)-plane, as shown in Fig.2.2.3, ϵ_1 is the magnitude of the spatial modulation,

$$\epsilon_1 \ll \epsilon_0, \tag{2.28}$$

and k^2 can be written as

$$\begin{aligned} k^2 =& \frac{\omega^2}{c^2}\epsilon, \\ =& \beta^2 + 2\kappa\beta[\exp(i\mathbf{K}\cdot\mathbf{r}) + \exp(-i\mathbf{K}\cdot\mathbf{r})], \end{aligned} \tag{2.29}$$

where c is the speed of light, β and κ represent

$$\begin{aligned} \beta =& \frac{2\pi\epsilon_0^{1/2}}{\lambda}, \\ =& \frac{2\pi n}{\lambda}, \end{aligned} \tag{2.30}$$

and

$$\begin{aligned} \kappa =& \frac{1}{4}\frac{2\pi}{\lambda}\frac{\epsilon_1}{\epsilon_0^{1/2}}, \\ =& \frac{\pi n_1}{\lambda}. \end{aligned} \tag{2.31}$$

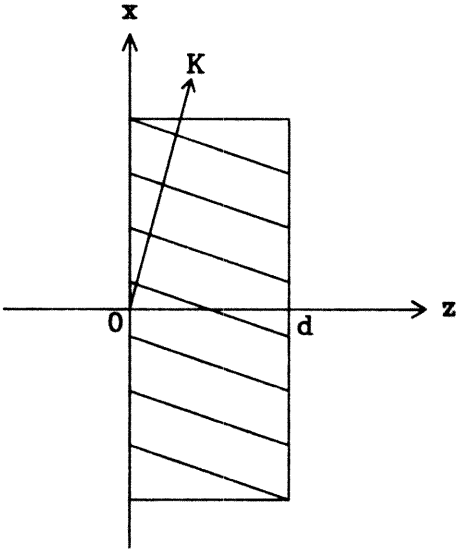


Fig.2.2.3 A sinusoidal grating inside a crystal of thickness d.

The electric field consists of two parts, the incident beam and the diffracted beam. E can be written as

$$E = R(z)\exp(-i\mathbf{k}_i \cdot \mathbf{r}) + S(z)\exp(-i\mathbf{k}_d \cdot \mathbf{r}), \tag{2.32}$$

where $\mathbf{k}_i$ and $\mathbf{k}_d$ are wave vectors assumed to be in the (x, z)-plane. Substitution of Eq.(2.32) into Eq.(2.25) reveals that when

$$\mathbf{k}_d = \mathbf{k}_i - \mathbf{K}, \tag{2.33}$$

the equation can be separated into two second order differential equations by matching coefficients of each of the two exponential terms in Eq.(2.32). The second order derivatives are neglected because $R(z)$ and $S(z)$ are changing slowly with z. The remaining first order differential equations are

$$\begin{aligned} c_R R' &= -i\kappa S, \\ c_S S' + i\delta S &= -i\kappa R, \end{aligned} \tag{2.34}$$

where

$$c_R = \frac{k_{iz}}{\beta}, \tag{2.35}$$

$$c_S = \frac{k_{dz}}{\beta}, \tag{2.36}$$

as shown in Fig.2.2.4, and

$$\delta = K\cos(\phi - \theta) - \frac{K^2\lambda}{4\pi n}. \tag{2.37}$$

The angles ϕ and θ are shown in Fig.2.2.4. The solutions to the above first order

equations have the form [38]

$$R(z) = t_1 \exp(\gamma_1 z) + t_2 \exp(\gamma_2 z),$$
$$S(z) = s_1 \exp(\gamma_1 z) + s_2 \exp(\gamma_2 z). \tag{2.38}$$

With given boundary conditions, the coefficients t_1, t_2, s_1, s_2, γ_1 and γ_2 can be solved by substituting Eq.(2.38) into Eq.(2.34). The solution for $\gamma_{1,2}$ is

$$\gamma_{1,2} = -i\frac{\delta}{2c_S} \pm \frac{1}{2}[(-i\frac{\delta}{c_S})^2 - \frac{4\kappa^2}{c_R c_S}]^{1/2}. \tag{2.39}$$

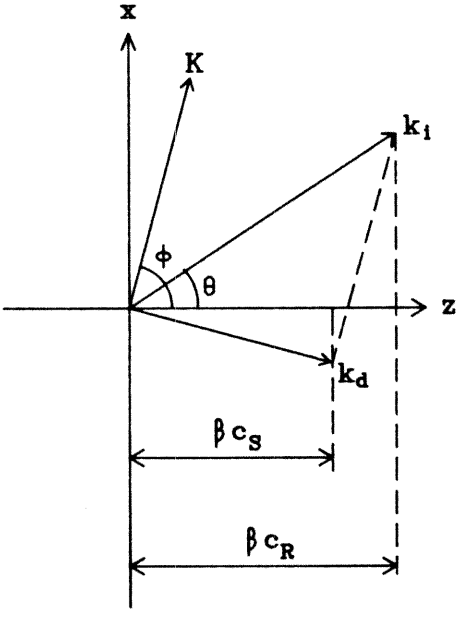


Fig.2.2.4 Variables used in the coupled wave theory.

For a transmission hologram with boundary conditions

$$S(z = 0) = 0,$$
$$R(z = 0) = 1, \tag{2.40}$$

the diffracted beam has the solution:

$$S(d) = -i(\frac{c_R}{c_S})^{1/2} e^{-i\xi} \frac{\sin(\nu^2 + \xi^2)^{1/2}}{(1 + \xi^2/\nu^2)^{1/2}}, \tag{2.41}$$

where d is the thickness of the crystal,

$$\nu = \frac{\pi n_1 d}{\lambda (c_R c_S)^{1/2}}, \tag{2.42}$$

and

$$\xi = \frac{\delta d}{2c_S}. \tag{2.43}$$

The diffraction efficiency of the grating is

$$\begin{aligned} \eta &\equiv \frac{|c_S|}{|c_R|} S(d) S^*(d), \\ &= \frac{\sin^2(\nu^2 + \xi^2)^{1/2}}{(1 + \xi^2/\nu^2)}. \end{aligned} \tag{2.44}$$

Fig.2.2.5 shows the variation of the diffraction efficiency versus the Bragg mismatch ξ.

The diffraction efficiency is maximum when $\xi = 0$, which occurs when

$$\cos(\phi - \theta_0) = \frac{K}{2\beta}. \tag{2.45}$$

The above is recognized as the Bragg condition.

If the incident beam is slightly away from the Bragg angle θ_0,

$$\theta = \theta_0 + \Delta\theta, \tag{2.46}$$

then the corresponding variables δ and ξ will not be zero. They are related to

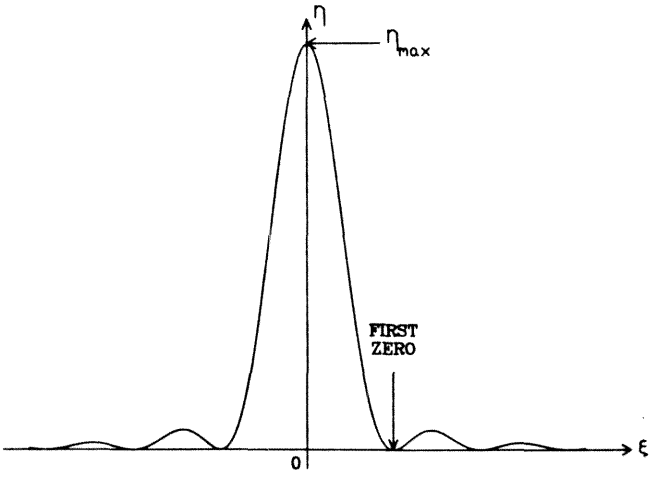


Fig.2.2.5 Diffraction efficiency versus the Bragg mismatch.

$\Delta\theta$ by:

$$\delta = \Delta\theta K \sin(\phi - \theta_0), \tag{2.47}$$

$$\xi = \Delta\theta \frac{Kd\sin(\phi - \theta_0)}{2c_S}. \tag{2.48}$$

The minimum angular deviation corresponding to the zero diffraction efficiency is

$$\delta\theta = \frac{2\pi\sqrt{c_S^2 - \frac{c_S}{c_R}\frac{n_1^2 d^2}{\lambda^2}}}{Kd\sin(\phi - \theta_0)}. \tag{2.49}$$

Any plane wave incident at an angle $\delta\theta$ away from the Bragg angle θ_0 will not be sufficiently diffracted. Therefore $\delta\theta$ represents the angular sensitivity of the Bragg condition.

A summary of different approaches for analyzing multiple gratings spatially superimposed inside a crystal can be found in the review paper [42]. The coupled

wave theory for multiple gratings in anisotropic crystals becomes very complicated. In Chapter 3, the storage of many sinusoidal gratings will be analyzed geometrically.

2.3 The Degrees of Freedom Argument

For a Vander Lugt system used to implement a neural network, neurons are located at the input and output planes, and interconnections are stored in a 3-dimensional crystal. The location of each neuron has 2 degrees of freedom.

Suppose neurons are distributed on a regular 2-dimensional grid. The position of a neuron can be specified by two indices (i, j), representing the ith column and the jth row on the grid. The signal sent by an input (output) neuron is an element of a 2-dimensional tensor x_{ij}^{input} (x_{kl}^{output}).

The weight tensor required for the mapping of two 2-dimensional tensors is 4-dimensional. The signal sent by an output neuron is related to the signals sent by the input neurons by

$$x_{kl}^{output} = g_{kl}(\sum_{i,j} w_{klij} x_{ij}^{input}), \tag{2.50}$$

where w_{klij} is the interconnection between the input neuron (i, j) and the output neuron (k, l), and $g_{kl}(x)$ is the thresholding function of the output neuron. Four indices are needed to specify an interconnection.

However, a stationary grating inside a 3-dimensional crystal has only 3 degrees of freedom. A grating vector can be described by its three components (K_x, K_y, K_z). It can not be used to represent a 4-dimensional tensor element.

The lack of 4-dimensional medium forces the dimension of either the input neurons or that of the output neurons to be reduced. For example, two kinds of mappings can be performed using a 3-dimensional weight tensor:

$$x_{kl}^{output} = g_{kl}(\sum_{i} w_{kli} x_{i}^{input}), \tag{2.51}$$

that is, 1-dimension to 2-dimension mapping; or

$$x_{k}^{output} = g_{k}(\sum_{i,j} w_{kij} x_{ij}^{input}). \tag{2.52}$$

that is, 2-dimension to 1-dimension mapping.

In general, if the input neurons are distributed d-dimensionally, the output neurons can be at most distributed $(3-d)$-dimensionally, because the gratings used to implement interconnections are limited within a 3-dimensional crystal. In the case of fractional dimension, the mapping can not be written in the above tensor form explicitly. The concept of fractals will be used to discuss the general neuron distributions in Chapter 4.

3. K-SPACE ANALYSIS

3.1 INTRODUCTION

K-space analysis is a geometric analysis of volume holographic gratings. With the help of the Fourier transform, the spatial superposition of multiple gratings in the real space is described as the distribution of points in the **K**-space. This distribution is in turn related to the characteristics of the crystal, such as its normal surface, shape and dimensions.

The following discussion involves two spaces — the **k**-space and the **K**-space. Here **k** and **K** represent the wave vector of a plane wave and the grating vector of a sinusoidal phase grating respectively. Mathematically, each point in the **k**-space (**K**-space) represents a plane wave (a set of sinusoidal fringes) extended to infinity in the real space.

Section 3.2 calculates the storage capacity of the crystal. First, a sinusoidal grating recorded by the crystal is mapped to the **K**-space with a finite **K**-space volume. Then, the storage of multiple gratings is considered. Two gratings are distinguishable only if their **K**-space uncertainty volumes do not overlap. The storage capacity of a crystal will be defined as the maximum number of distinguishable gratings that can be contained in the **K**-space.

Section 3.3 discusses the accessibility of the **K**-space. In practical systems, such as the Vander Lugt system shown in Fig.2.1.3, the dimensions of the input and the training planes limit the accessibility of the **K**-space. The part of the **K**-space which is practically reachable will be called the accessible grating space.

Section 3.4 gives the degeneracy condition in the **k**-space. The degeneracy arises when a grating is Bragg matched by more than one pair of incident-diffracted plane waves. The interconnections between such pairs of input-output neurons would be implemented by the same grating, and the corresponding weight tensor elements would not be independent. The degrees of freedom argument in the last chapter is related to the degeneracy in the **k**-space. Independent interconnections can not be implemented by simply reducing the dimensions of the neuron distributions. The positions of neurons, that is the directions of plane waves, will also be taken into consideration.

Section 3.5 evaluates the angular resolution of the Bragg diffraction. The wave vectors representing plane waves are related to the grating vectors by the Bragg condition. It is assumed that an ideal grating (a point in the **K**-space) can only diffract ideal plane waves (points in the **k**-space) incident exactly at the Bragg angle. However, since a real grating recorded in the crystal has a finite **K**-space volume, plane waves incident slightly off the Bragg angle will still be Bragg matched by some points inside this **K**-space volume. Therefore, the angular sensitivity of the Bragg condition is related to the geometry of the **K**-space volume of the grating.

The validity of the **K**-space analysis will be discussed at the end of this chapter. The comparison of the angular resolution obtained using the **K**-space analysis with that obtained using coupled wave theory indicates that for thick medium and low modulation depth, the **K**-space analysis and coupled wave theory are both valid. However, coupled wave theory is limited to only Bragg scattering or *thick* medium diffraction [38], but the **K**-space analysis allows large Bragg

mismatch, therefore, is valid for both *thick* and *thin* media. On the other hand, the **K**-space analysis cannot give the diffraction efficiency without the help of coupled wave theory. Therefore, the **K**-space analysis and coupled wave theory are complementary to each other. The combination of the two analyses will provide a better understanding of both volume and planar holograms.

3.1.1 **k**-space and the Normal Surface

In an anisotropic medium, the commonly used eigen states of the wave equation are plane waves with their polarization specified and their wave vectors confined to the normal surface [33]. Each eigen state has its eigen value — the wave vector **k**.

The wave equation in an anisotropic medium can be written as

$$\mathbf{k} \times (\mathbf{k} \times \mathbf{E}) + \omega^2 \mu \epsilon \mathbf{E} = 0, \tag{3.1}$$

where **k** is the wave vector, **E** is the electric field, ω is the angular frequency of the electromagnetic wave, μ is the permeability tensor which is assumed to be a constant, and ϵ is the dielectric tensor. This equation is derived from Maxwell's equations by assuming that the electric field and the magnetic field have the form

$$\begin{aligned} &\mathbf{E} \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)], \\ &\mathbf{H} \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]. \end{aligned} \tag{3.2}$$

Eq.(3.1) will have nontrivial solutions if and only if

$$\det \begin{vmatrix} \omega^2 \mu \epsilon_x - k_y^2 - k_z^2 & k_x k_y & k_x k_z \\ k_y k_x & \omega^2 \mu \epsilon_y - k_x^2 - k_z^2 & k_y k_z \\ k_z k_x & k_z k_y & \omega^2 \mu \epsilon_z - k_x^2 - k_y^2 \end{vmatrix} = 0. \tag{3.3}$$

Where the coordinate system has been chosen such that the dielectric tensor is

diagonal, i.e.,

$$\epsilon = \begin{pmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_y & 0 \\ 0 & 0 & \epsilon_z \end{pmatrix}. \tag{3.4}$$

Eq.(3.3) specifies a surface in the 3-dimensional **k**-space, which is called the normal surface. In general, the normal surface consists of two shells crossing at four points. These four points specify two optical axes. Each optical axis passes through two of the four points and the origin. In a special case, two of the principal dielectric constants are equal and the normal surface becomes two tangential surfaces, one sphere and the other ellipsoid. Such a crystal is called a uniaxial crystal. Plane waves with their wave vectors confined on the sphere and the ellipsoid are called ordinary and extraordinary waves respectively.

In the following discussion, the ordinary and extraordinary waves will be considered separately with the assumptions that the crystal is uniaxial and there is no polarization change during diffraction.

In a uniaxial crystal, the normal surface can be separated into a sphere,

$$\frac{k_x^2}{n_o^2} + \frac{k_y^2}{n_o^2} + \frac{k_z^2}{n_o^2} = k_0^2, \tag{3.5}$$

and an ellipsoid,

$$\frac{k_x^2}{n_e^2} + \frac{k_y^2}{n_o^2} + \frac{k_z^2}{n_e^2} = k_0^2, \tag{3.6}$$

where $\mathbf{k} = (k_x, k_y, k_z)$ is the wave vector, $n_o^2 = \epsilon_x/\epsilon_0 = \epsilon_z/\epsilon_0$, $n_e^2 = \epsilon_y/\epsilon_0$, $k_0 = \omega/c$, ϵ_0 and c are the dielectric constant and the speed of light in vacuum respectively.

It is sufficient to consider only the diffraction of extraordinary waves, since the sphere can be regarded as a special case of an ellipsoid with $n_e = n_o$. The coordinate system is chosen such that the optical axis is in the y-direction.

3.1.2 K-space

The **K**-space distribution of gratings is confined to a finite volume, if the plane waves used to write and to read these gratings are monochromatic. The boundary of this distribution can be found with the help of the normal surface containing the wave vectors.

Consider a grating written in a crystal by two plane waves. When two different plane waves $\exp(i\mathbf{k}_i \cdot \mathbf{r})$ and $\exp(i\mathbf{k}_d \cdot \mathbf{r})$ form an interference pattern $\cos[(\mathbf{k}_d - \mathbf{k}_i) \cdot \mathbf{r}]$ inside a crystal, the resulting dielectric phase grating will have a first order sinusoidal component $\cos(\mathbf{K} \cdot \mathbf{r} + \phi)$, where $\mathbf{K}$ is the grating vector and ϕ is the phase shift [44] between the grating and the interference pattern. The relation $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$ is the Bragg condition.

Fig.3.1.1 shows a grating vector $\mathbf{K}$ and a normal surface depicted inside the **K**-space. The origin of the wave vectors, $\mathbf{k}_i$ and $\mathbf{k}_d$, is at the center of the normal surface. The origin of the grating vector, $\mathbf{K}$, is at the origin of the **K**-space, $K_x = K_y = K_z = 0$. The Bragg condition $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$ constrains the three vectors $\mathbf{K}$, $\mathbf{k}_i$ and $\mathbf{k}_d$ to form a triangle. Therefore, the center of the normal surface is displaced to the point $\mathbf{K} = -\mathbf{k}_i$.

Given an input plane wave, different gratings can be written by different output plane waves. Consider a given input wave vector $\mathbf{k}_i$. The corresponding normal surface is centered at the point $\mathbf{K} = -\mathbf{k}_i$, which will be referred to as

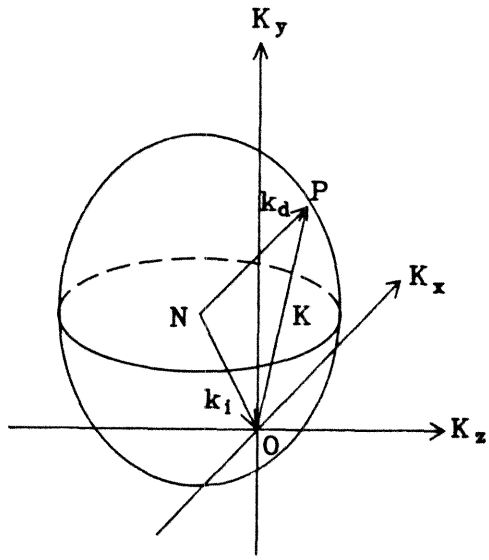


Fig.3.1.1 The **K**-space and the normal surface.

point N. Consider another point P on this normal surface. The vector from point N to point P represents a possible output wave vector $\mathbf{k}_d$. The vector from point O, which is the origin of the **K**-space, to point P is $-\mathbf{k}_i + \mathbf{k}_d$. Thus, point P represents a grating vector $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$. Since point P can be any point on the normal surface, all possible grating vectors are distributed on this 2-dimensional normal surface.

Another set of grating vectors can be found by using another input plane wave. Having a different input wave vector $\mathbf{k}_i'$ is equivalent to displacing the center of the normal surface to a different point N', which is given by $-\mathbf{k}_i'$, in the **K**-space. All possible grating vectors $\mathbf{K}' = \mathbf{k}_d' - \mathbf{k}_i'$ will be distributed on the 2-dimensional normal surface centered at point N'.

The two normal surfaces (centered at N and N' respectively) have the same orientation, which is determined by the given orientation of the crystal.

All possible locations of the centers of the normal surfaces are confined to an ellipsoid described by Eq.(3.6), since all $-\mathbf{k}_i$'s are confined to the normal surface described by Eq.(3.6). This ellipsoid is centered at the origin of the **K**-space and will be called the center ellipsoid.

By using all possible input plane waves, the gratings are 3-dimensionally distributed in a confined volume of the **K**-space. All possible grating vectors can be found by continuously moving the normal surface in the **K**-space. Since the center of the normal surface is restricted to the center ellipsoid, all possible gratings are distributed inside a bounded grating space.

The boundary of the grating space is an ellipsoid with its axes twice that of the normal surface ellipsoid. The equation for that boundary is

$$\frac{K_x^2}{(2n_e)^2} + \frac{K_y^2}{(2n_o)^2} + \frac{K_z^2}{(2n_e)^2} = k_0^2. \tag{3.7}$$

Only half of the gratings inside the above bounded **K**-space are necessary to describe independent gratings. Due to the sinusoidal nature of the grating, grating vectors **K** and $-\mathbf{K}$ represent the same fringe spacing and orientation. In the following discussion, the top half of the **K**-space bounded by Eq.(3.7) will be used. Fig.3.1.2 shows the relationship between the normal surface, the center ellipsoid and the bounded grating space.

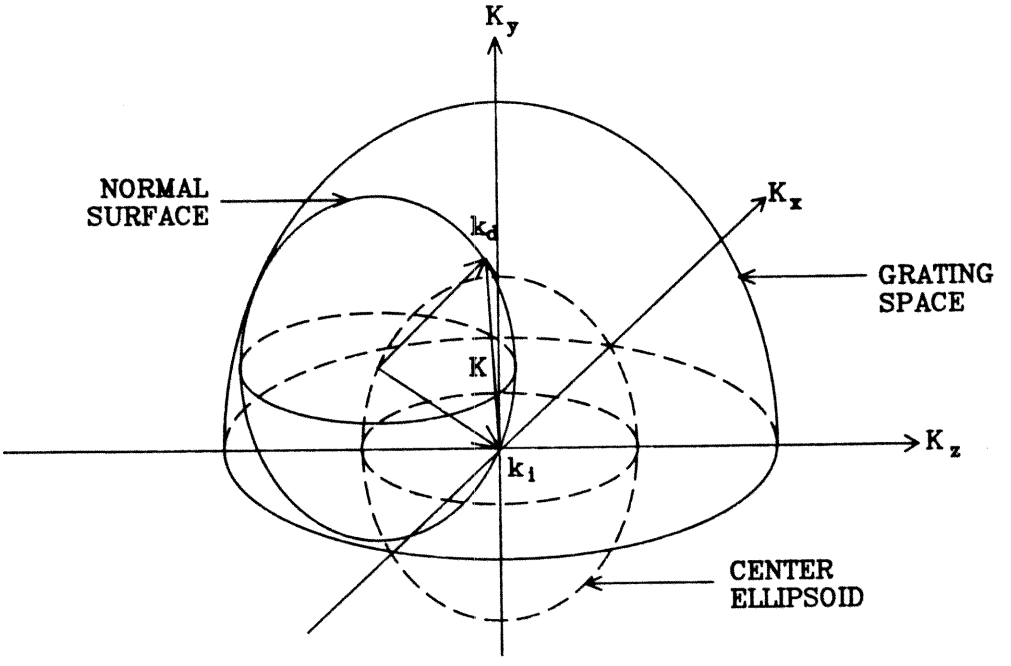


Fig.3.1.2 The big half ellipsoid is the bounded grating space. The dashed ellipsoid is the center ellipsoid. The input and output wave vectors are shown on the normal surface centered at a point on the center ellipsoid.

3.2 The Storage Capacity of a Crystal

The storage capacity of a crystal is the maximum number of distinguishable gratings that can be stored. The value can be obtained from the **K**-space representation of gratings. By calculating the whole volume of the grating space and the **K**-space volume of each grating, the storage capacity of a crystal is simply the ratio of these two volumes.

3.2.1 **K**-space Volume of a Grating

The **K**-space volume of each grating is related to the dimensions of the recording crystal. The uncertainty principle gives an estimate of this **K**-space volume of a grating, which will also be called the uncertainty volume of a grating.

Suppose the crystal has a rectangular shape with its sides L_x, L_y and L_z along the x, y and z directions. The rectangular crystal is chosen to simplify the following analysis. Other crystal geometries can be analyzed similarly.

The uncertainty values of the a grating vector are related to the sides of the crystal L_i by $\delta K_i L_i \approx 2\pi$, where $i = x, y, z$. The uncertainty values are explicitly written as:

$$\delta K_x \approx \frac{2\pi}{L_x}, \tag{3.8}$$

$$\delta K_y \approx \frac{2\pi}{L_y}, \tag{3.9}$$

$$\delta K_z \approx \frac{2\pi}{L_z}. \tag{3.10}$$

The uncertainty volume, v_g, of a grating is simply a multiplication of the three uncertainty values, that is,

$$v_g = \frac{(2\pi)^3}{L_x L_y L_z}. \tag{3.11}$$

The strength of each grating inside the uncertainty volume can be calculated from the Fourier transform of the dielectric modulation,

$$\Delta\epsilon(x, y, z) = \text{rect}(x/L_x)\text{rect}(y/L_y)\text{rect}(z/L_z)(\Delta\epsilon_0)\exp(i\mathbf{K}_g \cdot \mathbf{r}). \tag{3.12}$$

Where $\text{rect}(x/L_x)\text{rect}(y/L_y)\text{rect}(z/L_z)$ represents the rectangular crystal, and

$\exp(\mathbf{K}_g \cdot \mathbf{r})$ represents the interference pattern. The Fourier transform of $\Delta\epsilon(x,y,z)$ is:

$$\begin{aligned}\Delta\tilde{\epsilon}(K_x,K_y,K_z) &= \int\int\int \Delta\epsilon(x,y,z)\exp[-i(K_x x + K_y y + K_z z)]dxdydz \\ &= \Delta\epsilon_0 L_x L_y L_z \mathrm{sinc}(\frac{K_x - K_{gx}}{2\pi/L_x})\mathrm{sinc}(\frac{K_y - K_{gy}}{2\pi/L_y})\mathrm{sinc}(\frac{K_z - K_{gz}}{2\pi/L_z}).\end{aligned} \tag{3.13}$$

The strength of each grating $\mathbf{K}$ is represented by $\Delta\tilde{\epsilon}(K_x,K_y,K_z)$. Eq.(3.13) indicates that the grating $\mathbf{K} = \mathbf{K}_g$ has the maximum grating strength. The gratings with $K_x = K_{gx} \pm 2\pi/L_x$, or $K_y = K_{gy} \pm 2\pi/L_y$, or $K_z = K_{gz} \pm 2\pi/L_z$ have zero grating strength.

A rectangular box specified by $|K_x - K_{gx}| \leq \pi/L_x$, $|K_y - K_{gy}| \leq \pi/L_y$ and $|K_z - K_{gz}| \leq \pi/L_z$ is chosen to include ideal gratings with significant strength. The center of this box is at the point $\mathbf{K}_g$, and the sides are $2\pi/L_x$, $2\pi/L_y$ and $2\pi/L_z$ along x-, y- and z-direction respectively. For any grating outside the box, its grating strength will be reduced by sinc(1/2) or more, according to Eq.(3.13).

The volume of this box is defined as the $\mathbf{K}$-space volume of the nominal grating $\mathbf{K}_g$. It can be recognized that the sides of this box are the same as the uncertainty values given by Eq.(3.8), Eq.(3.9) and Eq.(3.10), and the $\mathbf{K}$-space volume of a grating is the same as its uncertainty volume given by Eq.(3.11).

The uncertain volume will also serve as the criterion for distinguishing two gratings. Two gratings written in a finite crystal are distinguishable only if their uncertainty volumes do not overlap.

3.2.2 Maximum Number of Distinguishable Gratings

The storage capacity of a crystal is defined as the maximum number of distinguishable gratings that can be contained by the grating space. It is a theoretical upper limit based upon the geometrical considerations. To reach this upper limit, gratings have to be stacked in the **K**-space, so that two adjacent gratings are barely distinguishable. Since the **K**-space volume of each grating is independent of the grating vector, as shown in Eq.(3.11), the total number of distinguishable gratings is the ratio of the volume of the grating space to the **K**-space volume of each grating.

The volume of the grating space, i.e., the upper half of the ellipsoid bounded by Eq.(3.7), can be evaluated as

$$V_K = \frac{1}{2}\frac{4\pi}{3}(2n_e k_0)(2n_o k_0)(2n_e k_0). \tag{3.14}$$

It can be written in the form

$$V_K = \frac{16\pi}{3} n_e^2 n_o k_0^3. \tag{3.15}$$

Therefore, the storage capacity is the ratio $C = V_K/v_g$, which can be expressed as

$$C = \frac{16\pi}{3} n_e^2 n_o \frac{V_{xtal}}{\lambda^3}. \tag{3.16}$$

Where $V_{xtal} = L_x \times L_y \times L_z$ is the volume of the crystal and λ is the wavelength of light in vacuum.

For example, the storage capacity of a $1cm^3$ $LiNbO_3$ crystal, with $\lambda = 0.5\mu m$, can be as high as approximately 10^{15}.

3.3 ACCESSIBILITY OF THE K-SPACE

Practically, the full storage capacity of a crystal is not reachable by using the Vander Lugt system, as shown in Fig.2.1.3. The finite dimensions of the input (output) plane limit the range of the input (output) wave vectors. The gratings which can be written and read by these wave vectors are constrained in a much smaller space than that given by Eq.(3.7).

To find the accessible grating space, it is necessary to specify the accessible normal surface. Suppose the input (output) wave vectors are limited within the upper (lower) right area of the normal surface, as shown in Fig.3.3.1. The portion of the normal surface selected for input (output) wave vectors is part of the ellipsoid cut by four planes, two perpendicular to the k_y-direction and two perpendicular to the k_x-direction. The input and output wave vectors are chosen symmetrically with respect to the plane $k_y = 0$. The accessible normal surface of the input (output) wave vectors will be called the input (output) normal surface.

The input normal surface gives rise to a symmetric part of the center ellipsoid, since each input wave vector $\mathbf{k}_i$ corresponds to a point $-\mathbf{k}_i$ on the center ellipsoid, as in Fig.3.1.2. This part of the center ellipsoid will be referred to as the partial center ellipsoid. Fig.3.3.1 shows the partial center ellipsoid in the **K**-space.

Given an input wave vector $\mathbf{k}_i$, a set of gratings can be found on the output normal surface, as shown in Fig.3.3.2(a). In the **K**-space, the center of the normal surface is at the point $-\mathbf{k}_i$. Each output wave vector $\mathbf{k}_d$ is related to a grating vector $\mathbf{K}$ by $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$.

A different input wave vector $\mathbf{k}_i'$ will give another normal surface centered at

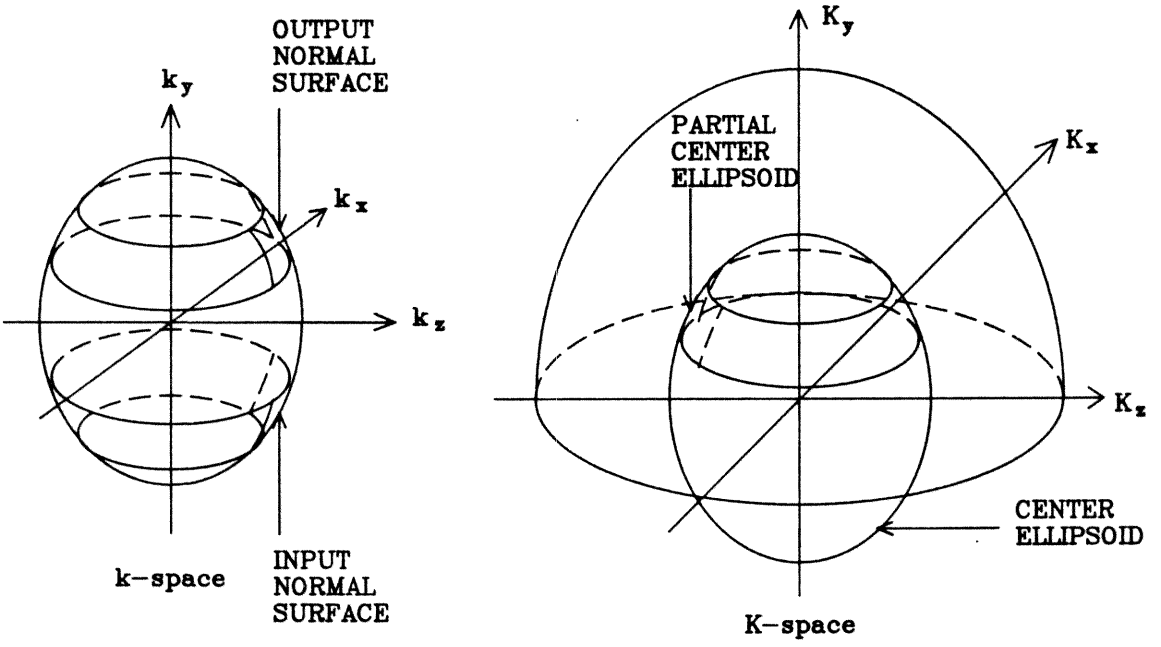
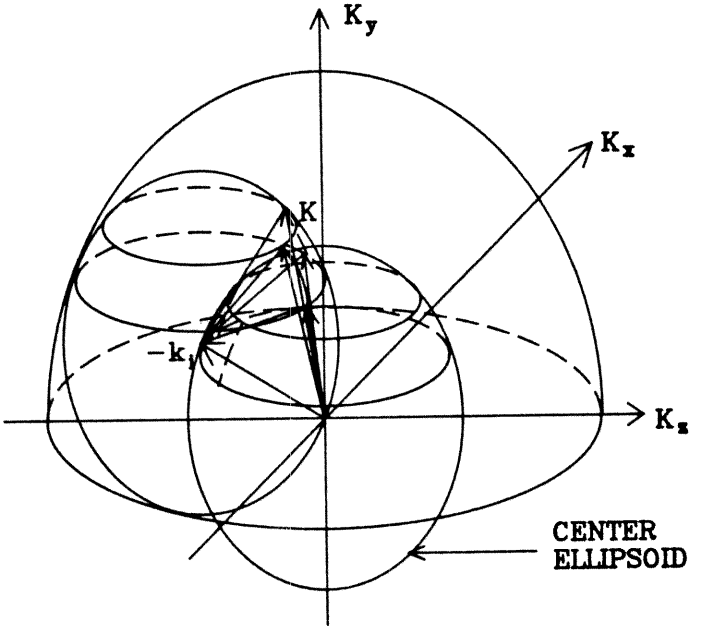


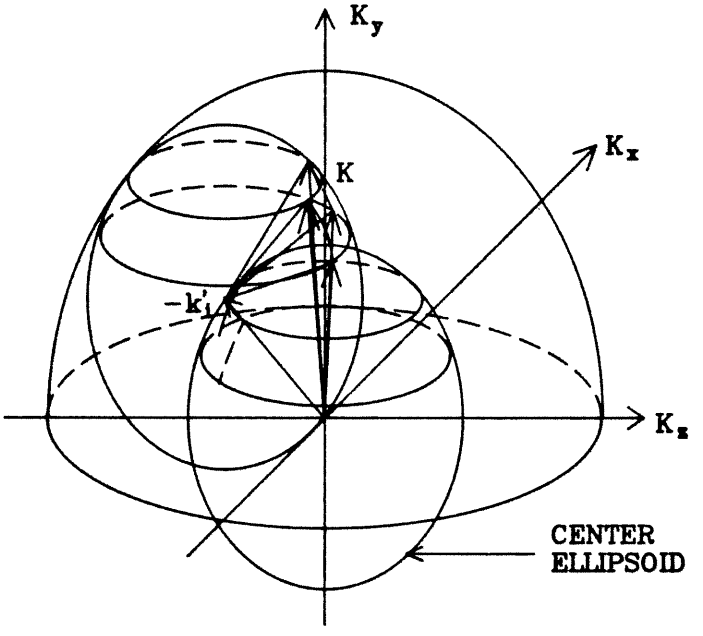
Fig.3.3.1 The input, output normal surfaces in the **k**-space and the partial center ellipsoid in the **K**-space.

the point $-\mathbf{k}_i'$ and another set of possible gratings, as shown in Fig.3.3.2(b). Since the center of the normal surface has been displaced from point $-\mathbf{k}_i$ to point $-\mathbf{k}_i'$, every point on the normal surface is also displaced by $(-\mathbf{k}_i') - (-\mathbf{k}_i) = \mathbf{k}_i - \mathbf{k}_i'$. The set of gratings given by the input wave vector $\mathbf{k}_i'$ and all possible output wave vectors will be the displaced version of the previous set of gratings. The displacement is again $\mathbf{k}_i - \mathbf{k}_i'$.

The 3-dimensional accessible grating space can be found by changing the input wave vector within its accessible values. As the input wave vector $\mathbf{k}_i$ moves on the input normal surface, the center of the normal surface $-\mathbf{k}_i$ moves on the partial center ellipsoid correspondingly, as shown in Fig.3.3.2(a) and (b). The set of gratings associated with the input wave vector $\mathbf{k}_i$ and all the possible



(a)



(b)

Fig.3.3.2 (a) A set of gratings formed by the input wave vector $\mathbf{k}_i$ and all the possible output wave vectors. (b) A new input wave vector $\mathbf{k}_i'$ gives a displaced version of the previous set of gratings.

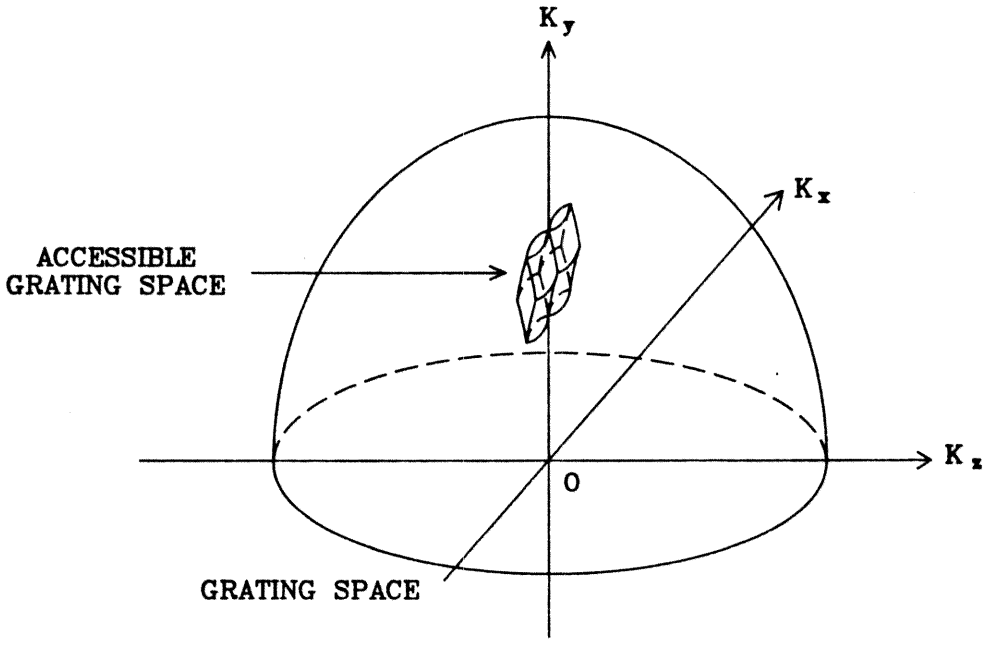


Fig.3.3.3 A plot of the accessible grating space determined by the input and output normal surfaces shown in Fig.3.3.1.

output wave vectors will also move in the **K**-space following the displacement of the center of the normal surface $-\mathbf{k}_i$. When input wave vector $\mathbf{k}_i$ moves continuously through all its possible values, the center of the normal surface $-\mathbf{k}_i$ also moves continuously through the partial center ellipsoid. During the continuous displacement, the output normal surface will sweep out a 3-dimensional accessible grating space in the **K**-space. Fig.3.3.3 shows a plot of the accessible grating space determined by the input and output normal surfaces shown in Fig.3.3.1.

The magnified accessible grating space and cross sections cut by planes are

shown in Fig.3.3.4. The accessible grating space is bounded by two planes perpendicular to the K_y-direction, two planes perpendicular to the K_x-direction, and the curved surfaces determined by the input and the output normal surfaces. The two boundary planes perpendicular to the K_y-direction represent the maximum and minimum values of the grating vector component K_y. The maximum K_y can be reached when k_{iy} is its minimum and k_{dy} is its maximum, i.e., $K_{ymax} = k_{dymax} - k_{iymin}$. Since the input (output) normal surface in Fig.3.3.1 is obtained by cutting the normal surface with two planes perpendicular to the k_y-direction and two planes perpendicular to the k_x-direction, any input wave vector $\mathbf{k}_i$ along the bottom arc of the input normal surface and any output wave vector $\mathbf{k}_d$ along the top arc of the output normal surface in Fig.3.3.1 will form a grating with $K_y = K_{ymax}$. All the grating vectors with $K_y = K_{ymax}$ and different values of K_x and K_z are on the top boundary of the accessible grating space. Similarly, the bottom boundary of the accessible grating space represents all grating vectors with minimum K_y value, and the two planar boundaries perpendicular to the K_x-direction represent all grating vectors with maximum and minimum K_x values respectively.

To determine the curved boundaries of the accessible grating space, it is sufficient to move the output normal surface vertically following the move of the center of the normal surface along one of the edges of the partial center ellipsoid. As will be seen in the following discussion, this procedure gives two of the four curved surfaces on the right side of the accessible grating space. Reflecting these two curved surfaces with respect to the $K_x = 0$ plane will result in the other two curved surfaces of the right side boundary, since both the input and the output

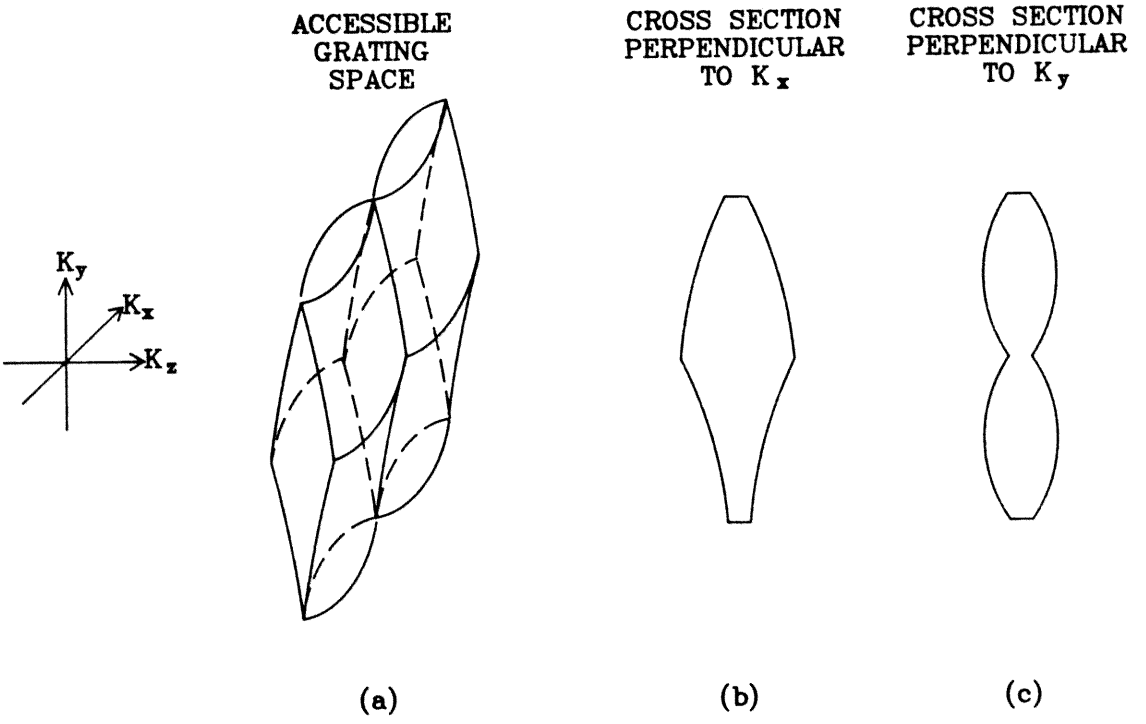


Fig.3.3.4 The magnified accessible grating space and its cross sections. (a) The accessible grating space. (b) The cross section cut by a plane perpendicular to the K_x-direction. (c) The cross section cut by a plane perpendicular to the K_y-direction.

normal surfaces are chosen symmetrically with respect to the $K_x = 0$ plane. The left side boundary is reflection symmetric to the right side boundary, since the partial center ellipsoid is reflection symmetric to the output normal surface. Where the reflection plane is the $K_z = 0$ plane and the comparison between the

partial center ellipsoid and the output normal surface is conducted on a common ellipsoid.

Consider the procedure of moving the output normal surface vertically following the move of the center of the normal surface along the front edge ($-k_{ix}$ is the minimum value) of the partial center ellipsoid, as shown in Fig.3.3.5. Start from the top point ($-k_{iy}$ is the maximum value) of this edge. The output normal surface itself gives the upper half of the right side boundary. This is because the output normal surface is tilted to the left and the partial center ellipsoid is tilted to the right, therefore moving the output normal surface from the top to the bottom will not exceed the original position. The lower half of the right side boundary is swept out by the bottom edge of the output normal surface during the moving, since the rest of the points on the output normal surface are all to the left of the bottom edge. The two curved surfaces are allocated on one side of the $K_x = 0$ plane. This is because the front edge of the partial center ellipsoid corresponds to $k_{ix} = k_{ixmax}$. Since it has been chosen that $k_{dxmax} = k_{ixmax}$, therefore, any grating vector $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$ is located on the $K_x \leq 0$ side. The boundaries of the accessible grating space will be discussed in Chapter 4 in more detail.

The volume of the accessible grating space can be calculated once its boundaries are found. This will be calculated in the next chapter, where the volume of the accessible grating space will be given according to the dimensions of the input and output planes.

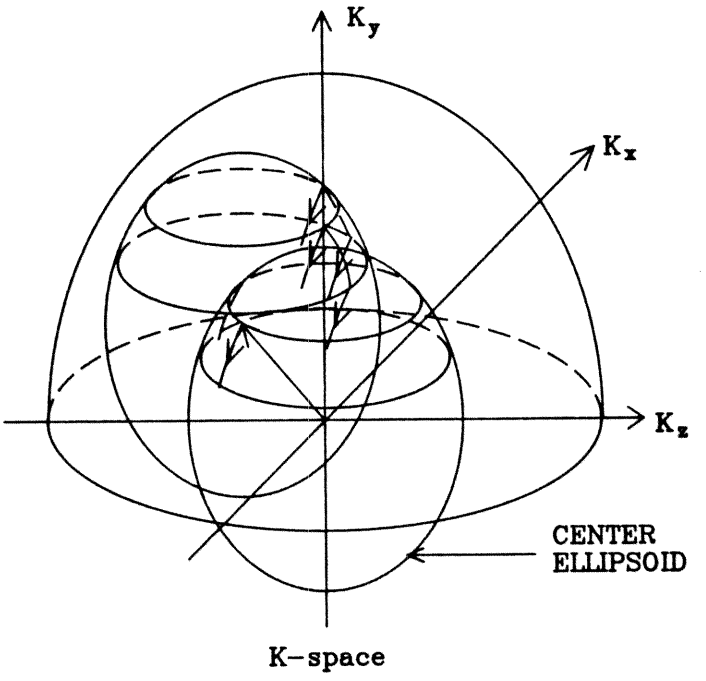


Fig.3.3.5 To obtain part of the curved boundaries of the accessible gratings space, move the output normal surface vertically following the move of the center of the normal surface along the front edge ($-k_{ix}$ is the minimum value) of the partial center ellipsoid.

3.4 Degeneracy in the k-space

The mapping between a grating vector and a pair of wave vectors is not a one to one mapping. A grating vector can be mapped to different pairs of input-output wave vectors as shown in Fig.3.4.1. More than one pair of input-output plane waves matched by the same grating vector are said to be degenerate. In this section, the degeneracy condition will be analyzed both in the **K**-space and in the **k**-space.

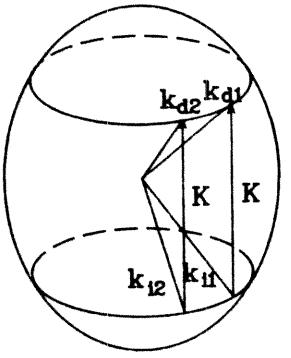


Fig.3.4.1 Degeneracy in the **k**-space.

3.4.1 **k**-space Degeneracy Ellipses

All pairs of degenerate input-output wave vectors matched by the same grating vector **K** are located on two ellipses, called degeneracy ellipses, on the normal surface. To find the two degeneracy ellipses in the **k**-space, consider a pair of input-output plane waves $(\mathbf{k}_i, \mathbf{k}_d)$. These two wave vectors must satisfy the following conditions:

$$\frac{k_{ix}^2}{n_e^2} + \frac{k_{iy}^2}{n_o^2} + \frac{k_{iz}^2}{n_e^2} = k_0^2, \tag{3.17}$$

$$\frac{k_{dx}^2}{n_e^2} + \frac{k_{dy}^2}{n_o^2} + \frac{k_{dz}^2}{n_e^2} = k_0^2, \tag{3.18}$$

and

$$\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i. \tag{3.19}$$

Eq.(3.17) and Eq.(3.18) indicate that $\mathbf{k}_i$ and $\mathbf{k}_d$ are confined to the normal

surface. Eq.(3.19) is the Bragg condition. Subtract Eq.(3.18) from Eq.(3.17) and substitute $\mathbf{k}_d$ by $\mathbf{K} + \mathbf{k}_i$. The resulting equations consist of only $\mathbf{k}_i$'s.

$$\begin{aligned} &\frac{k_{ix}^2}{n_e^2} + \frac{k_{iy}^2}{n_o^2} + \frac{k_{iz}^2}{n_e^2} = k_0^2, \\ &\frac{(2k_{ix} + K_x)K_x}{n_e^2} + \frac{(2k_{iy} + K_y)K_y}{n_o^2} + \frac{(2k_{iz} + K_z)K_z}{n_e^2} = 0. \end{aligned} \tag{3.20}$$

Similarly, the equations consisting of all $\mathbf{k}_d$'s are

$$\begin{aligned} &\frac{k_{dx}^2}{n_e^2} + \frac{k_{dy}^2}{n_o^2} + \frac{k_{dz}^2}{n_e^2} = k_0^2, \\ &\frac{(2k_{dx} - K_x)K_x}{n_e^2} + \frac{(2k_{dy} - K_y)K_y}{n_o^2} + \frac{(2k_{dz} - K_z)K_z}{n_e^2} = 0. \end{aligned} \tag{3.21}$$

It can be recognized that these two curves are two ellipses resulting from cutting the normal surface by two parallel planes. The direction of the vector normal to these two planes, $\hat{K}$, is given by

$$\hat{K} = \begin{pmatrix} K_x/n_e^2 \\ K_y/n_o^2 \\ K_z/n_e^2 \end{pmatrix}, \tag{3.22}$$

which, in general, is different from the $\mathbf{K}$ direction. The distances between the origin of the $\mathbf{k}$-space and these two planes are $\pm d$, where

$$d = \frac{\frac{1}{2}(\frac{K_x^2}{n_e^4} + \frac{K_y^2}{n_o^4} + \frac{K_z^2}{n_e^4})}{\sqrt{\frac{K_x^2}{n_e^4} + \frac{K_y^2}{n_o^4} + \frac{K_z^2}{n_e^4}}}. \tag{3.23}$$

The two degenerate ellipses are shown in Fig.3.4.2.

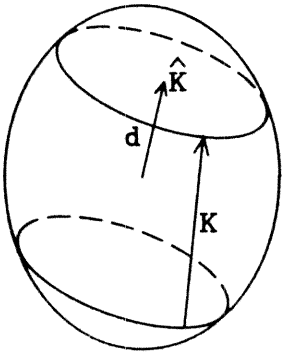


Fig.3.4.2 The **k**-space degeneracy ellipses for the grating vector **K**.

The degeneracy ellipses can be used to locate all pairs of input-output wave vectors which are matched by a given grating vector **K**. A pair of input-output wave vectors can be found by picking up one point on one ellipse, for instance $\mathbf{k}_i'$; then adding the vector **K** to $\mathbf{k}_i'$, resulting in a point on the second ellipse, $\mathbf{k}_d' = \mathbf{k}_i' + \mathbf{K}$. For each point on one degeneracy ellipse, there exist one and only one point on the other degeneracy ellipse such that these two points can be Bragg matched by the given grating vector **K**.

For isotropic crystals, degeneracy ellipses become degeneracy circles. The two circles are perpendicular to the grating vector **K**, since the $\hat{K}$ direction becomes the same as the **K** direction for $n_e = n_o$. The separation between each circle plane and the origin is half the magnitude of the grating vector, as given by Eq.(3.23).

3.4.2 K-space Degeneracy Ellipse

The degeneracy condition can also be discussed in the grating space. For a given grating vector **K**, an ellipse on the center ellipsoid can be found, such that when the normal surface is moved along this ellipse, the grating can always be matched by two points on the normal surface, as shown in Fig.3.4.3.

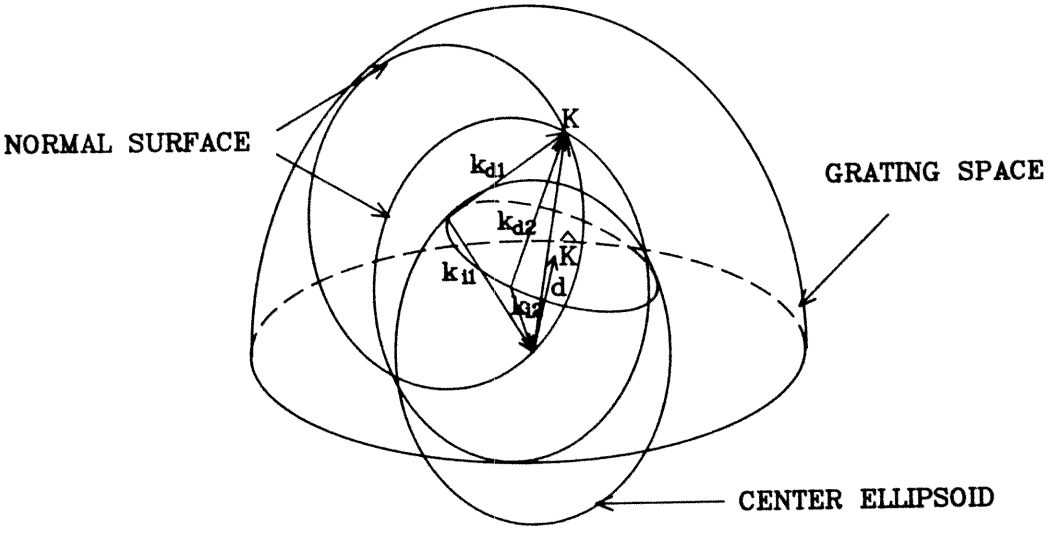


Fig.3.4.3 Degeneracy shown in the **K**-space.

Consider a grating vector **K** which is Bragg matched by the input and output wave vectors $\mathbf{k}_i$ and $\mathbf{k}_d$, as shown in Fig.3.4.3. Represent the origin of the wave vectors, that is the center of the normal surface, by $-\mathbf{k}_i = (x, y, z)$. Since $\mathbf{k}_i$ and $\mathbf{k}_d = \mathbf{K} + \mathbf{k}_i$ are confined to the normal surface. (x, y, z) must satisfy the

following conditions,

$$\frac{x^2}{n_e^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = k_0^2, \tag{3.24}$$

and

$$\frac{(K_x - x)^2}{n_e^2} + \frac{(K_y - y)^2}{n_o^2} + \frac{(K_z - z)^2}{n_e^2} = k_0^2. \tag{3.25}$$

Eq.(3.24) restricts the center of the normal surface to the center ellipsoid. Eq.(3.25) indicates that the point **K** is on the normal surface.

Subtract these two equations. The curve consisting of the centers of normal surfaces is found to be

$$\begin{aligned} &\frac{x^2}{n_e^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} = k_0^2, \\ &\frac{(2x - K_x)K_x}{n_e^2} + \frac{(2y - K_y)K_y}{n_o^2} + \frac{(2z - K_z)K_z}{n_e^2} = 0. \end{aligned} \tag{3.26}$$

This is an ellipse resulting from cutting the center ellipsoid by a plane. The normal direction of the plane is along the same direction as expressed by Eq.(3.22). The distance between the plane and the origin of the **K**-space is d, where d is the same as in Eq.(3.23).

In the case of isotropic crystals, the degeneracy ellipse becomes degeneracy circle. The degeneracy circle is perpendicular to the grating. And the distance between the circle plane and the origin of the **K**-space is half of the magnitude of the grating.

3.5 Angular Resolution in the k-space

The minimum angle between two wave vectors which can be used to distinguish two different gratings is defined as the angular resolution in the **k**-space. In this section, the angular resolution will be analyzed in terms of the uncertainty of a grating and compared with that obtained from coupled wave theory.

A grating written in a finite crystal can be matched by input-output wave vectors slightly off the Bragg angle determined by the nominal grating vector. This is because a grating vector $\mathbf{K} + \Delta\mathbf{K}$ within the uncertainty volume of $\mathbf{K}$ can diffract light from plane wave $\mathbf{k}_i + \Delta\mathbf{k}_i$ to plane wave $\mathbf{k}_d + \Delta\mathbf{k}_d$ if $\mathbf{K} + \Delta\mathbf{K} = (\mathbf{k}_d + \Delta\mathbf{k}_d) - (\mathbf{k}_i + \Delta\mathbf{k}_i)$, i.e.,

$$\Delta\mathbf{K} = \Delta\mathbf{k}_d - \Delta\mathbf{k}_i. \tag{3.27}$$

To distinguish two different gratings, the wave vectors are required to have a minimum separation angle. Since a plane wave off the Bragg angle can still be diffracted, when two adjacent gratings are recorded, e.g., $\mathbf{K}_a$ and $\mathbf{K}_b$, it is necessary for two input wave vectors to be separated by a sufficiently large angle so that the output wave can be recognized as a diffraction from one grating or the other. Similarly two output wave vectors have to be different from each other by certain angle so that diffractions from two different gratings can be separated.

The criterion for two pairs of input-output plane waves to distinguish two different gratings is that the difference of the difference between these two pairs of wave vectors, i.e., $(\mathbf{k}_{da} - \mathbf{k}_{ia}) - (\mathbf{k}_{db} - \mathbf{k}_{ib}) = \mathbf{K}_a - \mathbf{K}_b$, is outside the uncertainty volume of a grating.

3.5.1 k-space Calculation

It is only reasonable to discuss minimum separation in the direction perpendicular to the degeneracy curve, since any wave vector along the degeneracy curve can be matched by the same grating to a wave vector along the second degeneracy curve.

The discussion starts by defining a local coordinate system. Consider a pair of input-output wave vectors $\mathbf{k}_i$ and $\mathbf{k}_d$ connected by a grating vector $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$. Fig.3.5.1(a) shows the local coordinate systems. One of the directions chosen for a coordinate system located at the point $\mathbf{k}_i$ is the direction perpendicular to the normal surface, $\hat{k}_i$. Any wave vector $\mathbf{k}_i + \Delta\mathbf{k}_i$ has its increment $\Delta\mathbf{k}_i$ approximately tangential to the normal surface, that is perpendicular to $\hat{k}_i$. Here it has been assumed that the increment $\Delta\mathbf{k}_i$ is very small. Another two characteristic directions are the direction along the degeneracy curve, $\hat{n}_i$, and the direction perpendicular to it, $\hat{s}_i$. These unit vectors, $\hat{k}_i$, $\hat{n}_i$ and $\hat{s}_i$ can be expressed as

$$\hat{k}_i = \frac{1}{\sqrt{k_{ix}^2/n_e^4 + k_{iy}^2/n_o^4 + k_{iz}^2/n_e^4}} \begin{pmatrix} k_{ix}/n_e^2 \\ k_{iy}/n_o^2 \\ k_{iz}/n_e^2 \end{pmatrix}, \tag{3.28}$$

$$\hat{n}_i = \frac{\hat{k}_i \times \hat{k}_d}{|\hat{k}_i \times \hat{k}_d|}, \tag{3.29}$$

and

$$\hat{s}_i = \hat{k}_i \times \hat{n}_i, \tag{3.30}$$

where

$$\hat{k}_d = \frac{1}{\sqrt{k_{ix}^2/n_e^4 + k_{iy}^2/n_o^4 + k_{iz}^2/n_e^4}} \begin{pmatrix} k_{dx}/n_e^2 \\ k_{dy}/n_o^2 \\ k_{dz}/n_e^2 \end{pmatrix}. \tag{3.31}$$

It can be recognized that the normal direction of the degeneracy ellipse, $\hat{K}$ in Eq.(3.22), is the same as $\hat{k}_d - \hat{k}_i$. The vector $\hat{n}_i$, as expressed by Eq.(3.29), is perpendicular to $\hat{k}_i$, $\hat{k}_d$ and $\hat{k}_d - \hat{k}_i$. Therefore, $\hat{n}_i$ is in the plane tangential to the normal surface and also in the plane containing the degeneracy ellipse. In other words, $\hat{n}_i$ is along the degeneracy curve. The vector $\hat{s}_i$ formed by taking the cross product of $\hat{k}_i$ and $\hat{n}_i$ is in the plane tangential to the normal surface and perpendicular to the degeneracy curve.

Similarly, the coordinate system located at the point $\mathbf{k}_d$ can be defined by three unit vectors $\hat{k}_d$, $\hat{n}_d$ and $\hat{s}_d$. $\hat{k}_d$ has been given in Eq.(3.31). $\hat{n}_d$ can be chosen the same as $\hat{n}_i$, as in Eq.(3.29). Therefore, $\hat{s}_d$ is defined as

$$\hat{s}_d = \hat{k}_d \times \hat{n}_d. \tag{3.32}$$

It is the direction $\hat{s}_i$ along which the angular resolution will be discussed. The minimum separation of two wave vectors along the $\hat{s}_i$-direction for distinguishing two different gratings is denoted by δk_{is}. Here δ represents the minimum difference needed to distinguish two different gratings, and Δ represents arbitrary difference. Consider two pairs of input-output wave vectors, $(\mathbf{k}_i, \mathbf{k}_d)$ and $(\mathbf{k}_i + \Delta\mathbf{k}_i, \mathbf{k}_d + \Delta\mathbf{k}_d)$. The difference of these two pairs is $(\Delta\mathbf{k}_i, \Delta\mathbf{k}_d)$. The minimum value δk_{is} is defined such that for any value of $\Delta\mathbf{k}_d$, the difference $\Delta\mathbf{k}_d - \delta\mathbf{k}_i$ is just outside the uncertainty volume.

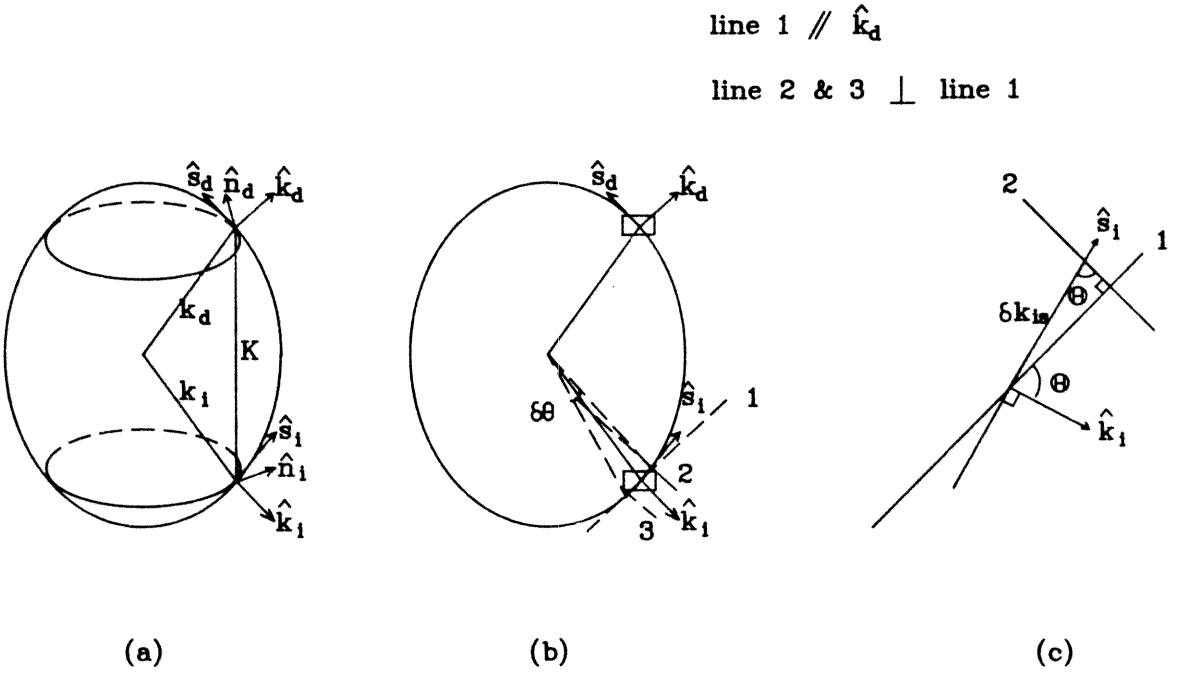


Fig.3.5.1 Geometrical explanation of the δk_{is}. (a) Local coordinates $(\hat{k}_i, \hat{n}_i, \hat{s}_i)$ and $(\hat{k}_d, \hat{n}_d, \hat{s}_d)$. (b) Line 1 is drawn parallel to the $\hat{k}_d$-direction. Line 2 and line 3 project $\delta\mathbf{K}$ on line 1 and cut the $\hat{s}_i$-axis by a segment of length δk_{is}. The points on the normal surface hit by line 2 and line 3 gives the angular resolution $\delta\theta_i$. (c) The area around point $\mathbf{k}_i$ is magnified. The angle Θ between line 1 (the $\hat{k}_d$-direction) and the $\hat{k}_i$-direction is the same as the angle between line 2 and the $\hat{s}_i$-direction.

The value of δk_{is} can be calculated as following. Write $\delta\mathbf{K} = (\delta K_x, \delta K_y, \delta K_z)$ with δK_x, δK_y and δK_z expressed in Eq.(3.8), Eq.(3.9) and Eq.(3.10) respectively. Consider $\delta\mathbf{K} = \Delta\mathbf{k}_d - \delta\mathbf{k}_i$, which is the extreme case that the vector $\Delta\mathbf{k}_d - \delta\mathbf{k}_i$ is on the boundary of the uncertainty volume. Since

$$\delta\mathbf{K} \cdot \hat{k}_d = (\Delta\mathbf{k}_d - \delta\mathbf{k}_i) \cdot \hat{k}_d \tag{3.33}$$

and

$$\Delta\mathbf{k}_d \cdot \hat{k}_d = 0 \tag{3.34}$$

for small $\Delta\mathbf{k}_d$, therefore

$$\delta\mathbf{K} \cdot \hat{k}_d = -\delta\mathbf{k}_i \cdot \hat{k}_d. \tag{3.35}$$

Similarly

$$\delta\mathbf{K} \cdot \hat{k}_i = \delta\mathbf{k}_d \cdot \hat{k}_i, \tag{3.36}$$

because

$$\Delta\mathbf{k}_i \cdot \hat{k}_i = 0. \tag{3.37}$$

Eq.(3.35) indicates that $\delta\mathbf{k}_i$ can be related to $\delta\mathbf{K}$ by projecting them in the $\hat{k}_d$-direction. By calculating the projection of $\delta\mathbf{k}_i = \delta k_{is}\hat{s}_i$ in the $\hat{k}_d$-direction, $\delta\mathbf{k}_i \cdot \hat{k}_d = \delta k_{is}(\hat{s}_i \cdot \hat{k}_d)$, the value δk_{is} can be found.

$$\delta k_{is} = \frac{\delta\mathbf{k}_i \cdot \hat{k}_d}{\hat{s}_i \cdot \hat{k}_d}. \tag{3.38}$$

From the definition of the unit vectors $\hat{n}_i$ and $\hat{s}_i$, Eq.(3.30) and Eq.(3.31), it can

be derived that

$$\hat{s}_i \cdot \hat{k}_d = \sin\Theta, \tag{3.39}$$

where Θ is the angle between $\hat{k}_i$ and $\hat{k}_d$. Substituting Eq.(3.35) and Eq.(3.39) into Eq.(3.38), the value δk_{is} is found to be

$$\delta k_{is} = -\frac{\delta\mathbf{K} \cdot \hat{k}_d}{\sin\Theta}. \tag{3.40}$$

Where the minus sign came from the relation $\Delta\mathbf{K} = \Delta\mathbf{k}_d - \Delta\mathbf{k}_i$.

It is the magnitude that limits the angular resolution. So the minimum separation required between two input wave vectors along the $\hat{s}_i$-direction is

$$\delta k_{is} = \left|\frac{\delta\mathbf{K} \cdot \hat{k}_d}{\sin\Theta}\right|. \tag{3.41}$$

Similarly the minimum separation required for two output wave vectors along the $\hat{s}_d$-direction is

$$\delta k_{ds} = \left|\frac{\delta\mathbf{K} \cdot \hat{k}_i}{\sin\Theta}\right|. \tag{3.42}$$

The value of δk_{is} is geometrically shown in Fig.3.5.1(b) and (c). Vectors $\hat{k}_i$ and $\hat{k}_d$ are drawn together with their common starting point at the tip of $\mathbf{k}_i$. Two lines, line 2 and 3, perpendicular to $\hat{k}_d$ and across the edges of the uncertainty volume hits the normal surface at two points separated by δk_{is}.

Finally, the angular resolution $\delta\theta_i$ is related to δk_{is} by

$$\begin{aligned}\delta\theta_i &= \frac{\delta k_{is}}{|\mathbf{k}_i|}\cos\alpha_i, \\ &= \frac{\delta\mathbf{K} \cdot \hat{k}_d}{|\mathbf{k}_i|\sin\Theta}\cos\alpha_i,\end{aligned} \tag{3.43}$$

where α_i is the angle between $\mathbf{k}_i$ and $\hat{k}_i$. Similarly

$$\begin{aligned}\delta\theta_d =& \frac{\delta k_{ds}}{|\mathbf{k}_d|}\cos\alpha_d, \\ =& \frac{\delta\mathbf{K}\cdot\hat{k}_i}{|\mathbf{k}_d|\sin\Theta}\cos\alpha_d,\end{aligned} \tag{3.44}$$

for the output plane wave.

3.5.2 Comparison with Coupled Wave Theory

In this subsection, the **K**-space analysis and coupled wave theory will be compared. The comparison starts from the angular resolution obtained from each analysis.

From coupled wave theory, the angular resolution is given by Eq.(2.49). Using the new notation, Eq.(2.49) can be rewritten as

$$\delta\theta_i = \frac{2\pi\sqrt{(\frac{k_{dz}}{n_o k_0})^2 - \frac{k_{dz}}{k_{iz}}\frac{n_1^2 d^2}{\lambda^2}}}{Kd\sin((\pi-\Theta)/2)}, \tag{3.45}$$

where $K = |\mathbf{K}|$; d is the thickness of the crystal, i.e., L_z; n_1 is the coupling index. The crystal was assumed to be isotropic with infinite L_x and L_y.

In the case of weak coupling,

$$\frac{k_{dz}}{k_{iz}}\frac{n_1^2 d^2}{\lambda^2} \ll (\frac{k_{dz}}{n_o k_0})^2. \tag{3.46}$$

This angular resolution is given by

$$\delta\theta_i = \frac{2\pi\frac{k_{dz}}{n_o k_0}}{Kd\sin((\pi-\Theta)/2)}.$$

For isotropic crystal $K = 2n_o k_0 \sin(\Theta/2)$. Substitute d by L_z,

$$\delta\theta_i = \frac{\frac{2\pi}{L_z}\frac{k_{dz}}{n_o k_0}}{n_o k_0 \sin\Theta}. \tag{3.47}$$

This is the same result obtained from Eq.(3.43).

The difference between the angular resolution obtained from the **K**-space analysis (Eq.(3.43)) and that obtained from coupled wave analysis (Eq.(3.45)) arises from the different approximations made in these two analyses. In the **K**-space analysis, the normal surface has been assumed to be approximately unchanged by the modulation of the refractive index. This approximation is legitimate if the perturbation of refractive index caused by the holographic grating is much smaller than the average refractive index of the crystal, which is practically satisfied since the electro-optic coefficient is usually very small [33]. This approximation has also been used in coupled wave theory, since the input wave vector has been assumed to be on the unperturbed normal surface. In addition, coupled wave theory has assumed that the condition $\mathbf{K}_g = \mathbf{k}_d - \mathbf{k}_i$ is always satisfied, even if the input plane wave is not incident exactly at the Bragg Angle. When the incident plane wave is off the Bragg angle, the diffracted plane wave $\mathbf{k}_d$ will not be confined to the normal surface, as shown in Fig.3.5.2. But the assumption of *Thick* medium in coupled wave theory [38] implies that the Bragg mismatch is very small, therefore, the diffracted wave $\mathbf{k}_d$ will be approximately on the normal surface. This approximation indicates that coupled wave theory is more accurate in analyzing diffractions exactly satisfying the Bragg condition than in analyzing diffractions with Bragg mismatch.

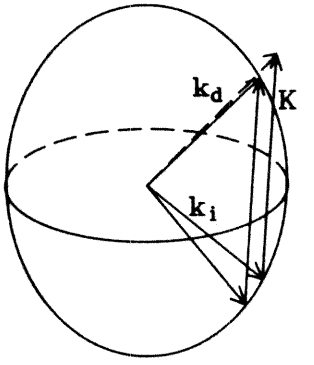


Fig.3.5.2 When the incident plane wave is off the Bragg angle, coupled wave theory does not confine the diffracted plane wave to the normal surface.

In the case of large Bragg mismatch, the angular resolution should be calculated according to Fig.3.5.3. Eq.(3.43) was derived with the assumption that the increment $\Delta\mathbf{k}_i$ is approximately tangential to the normal surface. When $\Delta\mathbf{k}_i$ is large, this approximation is no longer valid. However, the **K**-space analysis can still be used to calculate the angular resolution. Fig.3.5.3 shows the geometry for calculating the angular resolution. When the input wave vector is changed by certain value $\delta\mathbf{k}_i$, the uncertainty volume associated with the grating barely touches the normal surface. The angle between the two input wave vectors $\mathbf{k}_i$ and $\mathbf{k}_i + \delta\mathbf{k}_i$ is the angular resolution $\delta\theta_i$ for the input plane wave. Since the uncertainty volume may touch the normal surface by different points on its boundary depending on particular input and output wave vectors, therefore, the angular

resolution may have different expressions for different pairs of input-output wave vectors.

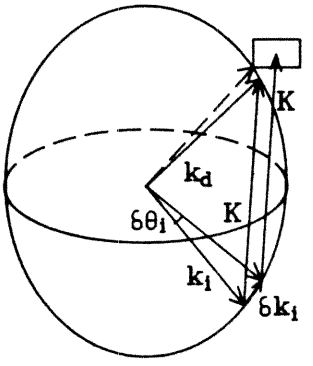


Fig.3.5.3 When the incident plane wave is off the Bragg angle, the **K**-space analysis confines the diffracted plane wave to the normal surface. The angular resolution is calculated according to the condition that the uncertainty volume barely touches the normal surface.

The combination of coupled wave theory and the **K**-space analysis gives diffraction efficiency for diffractions with large Bragg mismatch. For a grating vector **K** which is exactly Bragg matched by a pair of input-output wave vectors $(\mathbf{k}_i, \mathbf{k}_d)$, the diffraction efficiency can be obtained from coupled wave theory, Eq.(2.44). The strength of the grating **K** can be evaluated by using Fourier transform, Eq.(3.13). Therefore, the diffraction efficiency associated with the input wave vector $\mathbf{k}_i$, the output wave vector $\mathbf{k}_d$ and the grating vector **K** can be

found from the combination of these two analyses. For example, for a nominal grating $\mathbf{K}_g$ stored in a crystal with $L_x, L_y \rightarrow \infty$ and $L_z = d$, the modulation index n_1 can be written as

$$n_1(K_x, K_y, K_z) = n_{1max} d\delta(K_x - K_{gx})\delta(K_y - K_{gy})\text{sinc}(\frac{K_z - K_{gz}}{2\pi/d}). \qquad (3.48)$$

Since $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$, there is no Bragg mismatch in the calculation using coupled wave theory. Suppose the crystal is isotropic, Eq.(2.42) and Eq.(2.44) give the diffraction efficiency

$$\begin{aligned}\eta(\mathbf{k}_i, \mathbf{k}_d) &= \sin^2(\frac{\pi n_1(\mathbf{k}_i, \mathbf{k}_d)d}{\lambda}\frac{n_0 k_0}{\sqrt{k_{iz}k_{dz}}}), \\ &= \sin^2\{\frac{\pi d^2 n_0 k_0 n_{1max}}{\lambda\sqrt{k_{iz}k_{dz}}} \times \\ &\quad \delta(k_{dx} - k_{ix} - K_{gx})\delta(k_{dy} - k_{iy} - K_{gy})\text{sinc}(\frac{k_{dz} - k_{iz} - K_{gz}}{2\pi/d})\},\end{aligned} \qquad (3.49)$$

where n_0 is the average refractive index and $k_0 = 2\pi/\lambda$. In the above expression, $\mathbf{k}_i$ and $\mathbf{k}_d$ are not necessarily Bragg matched by the nominal grating vector $\mathbf{K}_g$, unlike the approximation used in coupled wave theory [38].

In conclusion, **K**-space analysis is complementary to coupled wave theory. It is valid for both *thick* and *thin* media. The extension of the **K**-space analysis to the planar holograms is discussed in the Appendix. When the storage medium is anisotropic with finite dimensions, coupled wave theory becomes very complicated, the **K**-space analysis is especially helpful. The combination of coupled wave theory and the **K**-space analysis provides a better understanding of both volume and planar holograms.

4. FRACTAL SAMPLING GRIDS

4.1 INTRODUCTION

4.1.1 Introduction

In this chapter, fractal sampling grids are used to select the locations of neurons at the input and output planes so that independent interconnections can be implemented and the storage capacity of the crystal can be reached. To implement independent interconnections, the fractal sampling grid for the input plane and that for the output plane are derived simultaneously to avoid degeneracy between any two pairs of input-output neurons. To reach the storage capacity of the crystal, the sum of the dimensions of a pair of input-output sampling grids is designed to be 3, which is limited by the dimension of the crystal.

Section 4.2 relates the degeneracy condition for the locations of neurons in the input and output planes to the degeneracy ellipses in the **k**-space. The conversion between the wave vectors and the positions at the input and output planes are performed by the Fourier transforming lenses. The corresponding degeneracy curves at the input and output planes will be calculated.

Section 4.3 gives the optimal configuration for the system set up so that the storage capacity of the crystal can be reached. The maximum number of accessible gratings that can be sufficiently used, when the separation of pixels at the input and output planes and the shape of the crystal are properly chosen, is derived. The selection of the aperture and the focal length of the lenses will also be discussed.

In Section 4.4, fractal sampling grids are systematically designed. Different kinds of fractal sampling grids are found for mapping d-dimensionally distributed input neurons to $(3-d)$-dimensionally distributed output neurons for $1 \leq d \leq 2$. Different fractal sampling grids can be used for different purposes.

The system used in the following discussions is a modified Vander Lugt system, as shown in Fig.4.1.1. The input plane and the training plane are separated by a large angle. Two lenses L_1 and L_2' are used for the input plane and the training plane, respectively. Lenses L_2 and L_2' have the same optical axis. The angle between the optical axis of lens L_1 and that of lens L_2 is chosen to yield high diffraction efficiency.

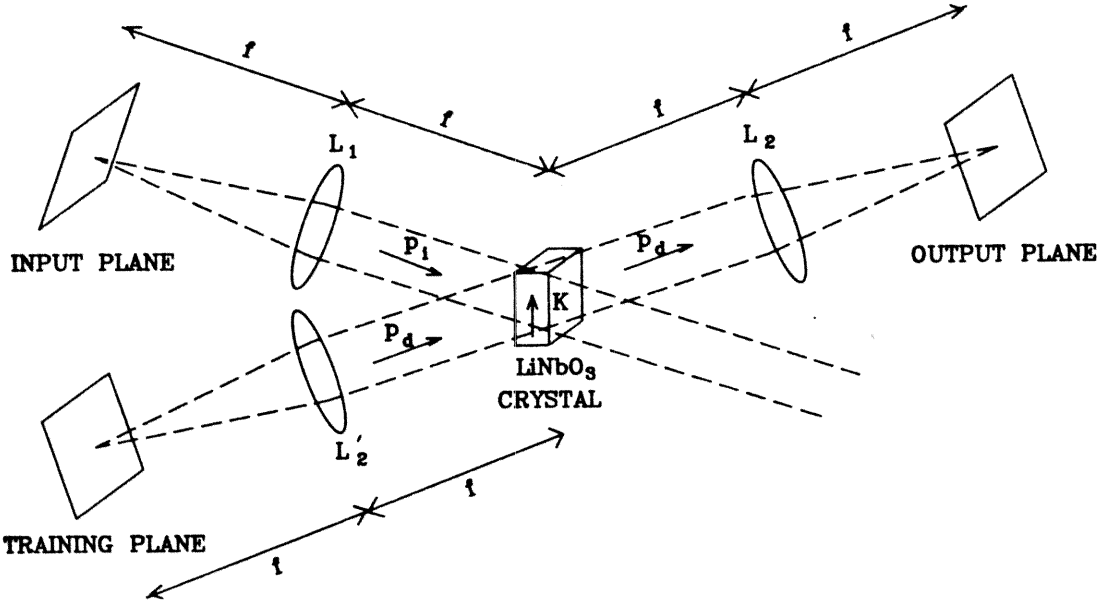


Fig.4.1.1 A modified Vander Lugt system.

4.1.2 Fractals

Two important mathematical properties of fractals [45, 46] are the fractal dimension and self similarity.

The fractal dimension describes the non-integer dimensions. The definition for fractal dimension can be expressed in different ways. The definition used in this thesis is as follows. In the 2-dimensional Euclidian space, the input plane has a total of $N \times N$ pixels and N_1 of those will be used as locations of neurons. The fractal dimension of the sampling grid is defined as $d_1 = \log N_1 / \log N$, in other words, $N_1 = N^{d_1}$. Similarly, the dimension of the sampling grid for output neurons, d_2, is defined as $d_2 = \log N_2 / \log N$, or $N_2 = N^{d_2}$. N_2 is the number of output neurons.

The self similarity is the scaling invariance of the geometrical characteristics of fractals. Mathematically, a fractal should have infinite orders, i.e., a fractal keeps the same geometrical structure even it is enormously magnified.

The self similar property will be used to generate higher order fractal sampling grids based upon first order ones while keeping the fractal dimensions unchanged. Examples will be given in Section 4.4. In reality, the fractal sampling grids always have finite orders. Typically, only the first or the second order fractal sampling grids are used.

4.2 The Degeneracy Conditions at the Input (Training) Plane

4.2.1 Mapping Between Neuron Positions to Wave Vectors

The position of a neuron at the input (training) plane is mapped to a wave vector $\mathbf{k}_i$ ($\mathbf{k}_d$) through two steps. First, a point source at the input (training) plane is converted to a plane wave propagating in the air by the Fourier transforming lens L_1 (L_2'), as shown in Fig.4.1.1. Then, this plane wave is refracted at the surface of the crystal, resulting in a plane wave propagating inside the crystal.

The position of neurons at the input plane will be expressed by the local coordinate system (x_i', y_i', z_i'). The optical axis of lens L_1 is chosen as the z_i'-direction. The input plane is chosen as the (x_i', y_i')-plane, where the x_i'-direction is along the x-direction of the global coordinate system and the y_i'-direction is perpendicular to both the x_i' and z_i' directions. Similarly, the local coordinate system for the training plane is (x_d', y_d', z_d'). Fig.4.2.1 shows the global and local coordinate systems.

The relationship between the local coordinates of a point source at the input plane and the wave vector of the corresponding plane wave propagating in the air can be found by using the paraxial approximation. Consider a point (x_i', y_i') at the input plane. The wave vector of the plane wave after lens L_1 is represented by $\mathbf{p}_i$. Assume that the area of the input plane is small compared with the aperture and the focal length of the lens so that the paraxial approximation is valid. The

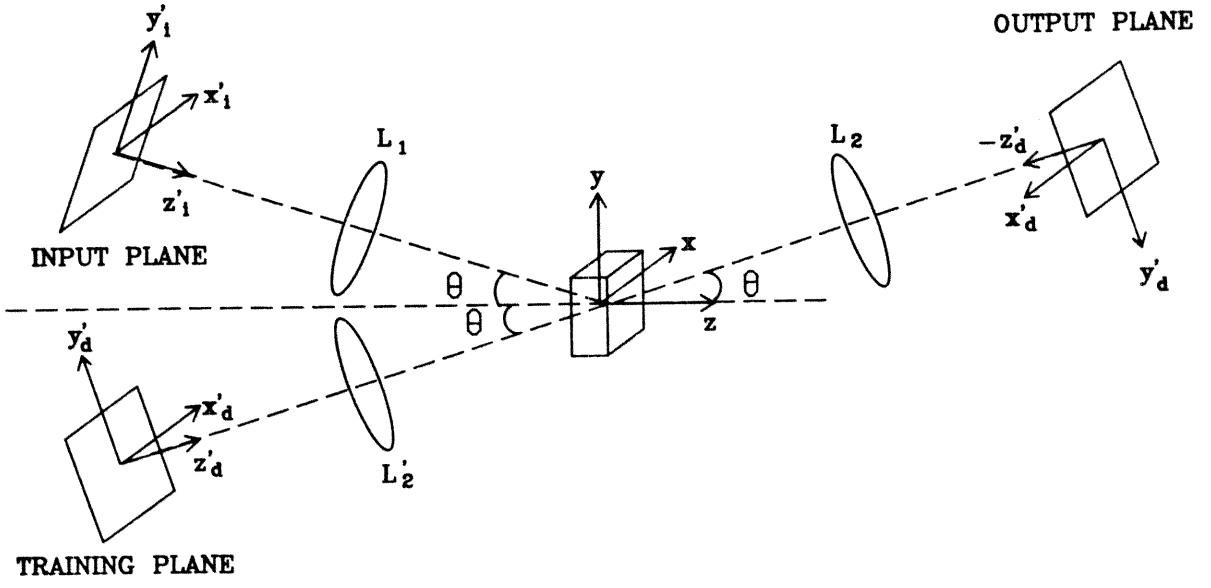


Fig.4.2.1 The global and local coordinate systems.

components of $\mathbf{p}_i$ can then be written in the global coordinate system as

$$\begin{aligned} p_{ix} &\approx -\frac{k_0}{f}x_i', \\ p_{iy} &\approx \cos\theta(-\frac{k_0}{f}y_i') - \sin\theta k_0, \\ p_{iz} &\approx \sin\theta(-\frac{k_0}{f}y_i') + \cos\theta k_0. \end{aligned} \tag{4.1}$$

Where f is the focal length of the lens, θ is the angle between the optical axis of lens L_1 and the z-direction of the system, as shown in Fig.4.2.1.

Similarly, the wave vector representing a plane wave coming from the point

(x'_d, y'_d) at the training plane is given by

$$\begin{aligned} p_{dx} &\approx -\frac{k_0}{f} x'_d, \\ p_{dy} &\approx \cos\theta(-\frac{k_0}{f} y'_d) + \sin\theta k_0, \\ p_{dz} &\approx -\sin\theta(-\frac{k_0}{f} y'_d) + \cos\theta k_0. \end{aligned} \tag{4.2}$$

Here the angle between the optical axis of lens L'_2 and the z-direction of the system is chosen to be the same as that between the optical axis of lens L_1 and the z-direction.

The wave vector of the refracted plane wave can be related to the wave vector of the incident plane wave by the boundary conditions. Suppose the plane wave hits the crystal at the surface parallel to the (x, y) plane. Denote the wave vectors of the input and the training plane waves inside the crystal by $\mathbf{k}_i$ and $\mathbf{k}_d$ respectively. The boundary condition requires that $k_{ix} = p_{ix}$, $k_{iy} = p_{iy}$ and $k_{dx} = p_{dx}$, $k_{dy} = p_{dy}$. Inside the crystal, the wave vectors are confined to the normal surface. Therefore, the z-components, k_{iz} and k_{dz}, can be calculated from Eq.(3.6). The components of an input wave vector inside the crystal are related to the local coordinates of a point at the input plane by

$$\begin{aligned} k_{ix} &\approx -\frac{x'_i}{f} k_0, \\ k_{iy} &\approx \cos\theta(-\frac{y'_i}{f} k_0) - \sin\theta k_0, \\ k_{iz} &\approx n_e k_0 \sqrt{1 - \frac{\sin^2\theta}{n_o^2}} - \frac{\sin\theta\cos\theta k_0 \frac{n_e}{n_o^2} \frac{y'_i}{f}}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}}. \end{aligned} \tag{4.3}$$

Where the approximation is made to the first order of x'_i/f or y'_i/f, which is assumed to be small under the paraxial approximation. Similar expressions can

be found for the training wave

$$
\begin{aligned}
k_{dx} &\approx -\frac{x'_d}{f}k_0, \\
k_{dy} &\approx \cos\theta(-\frac{y'_d}{f}k_0) + \sin\theta k_0, \\
k_{dz} &\approx n_e k_0\sqrt{1-\frac{\sin^2\theta}{n_o^2}} + \frac{\sin\theta\cos\theta k_0\frac{n_e}{n_o^2}\frac{y'_d}{f}}{\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}.
\end{aligned}
\tag{4.4}
$$

The grating vector of a grating written by an input point located at (x'_i, y'_i) and a training point located at (x'_d, y'_d) is $\mathbf{K} = \mathbf{k}_d - \mathbf{k}_i$, i.e.,

$$
\begin{aligned}
K_x &\approx k_0(x'_i - x'_d)/f, \\
K_y &\approx \cos\theta k_0(y'_i - y'_d)/f + 2k_0\sin\theta, \\
K_z &\approx \frac{\sin\theta\cos\theta k_0\frac{n_e}{n_o^2}}{\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}(y'_i + y'_d)/f.
\end{aligned}
\tag{4.5}
$$

4.2.2 Degeneracy Lines at the Input and the Training Planes

To find the degeneracy condition for two input neurons and two output neurons, the degeneracy ellipses in the **k**-space are mapped to the input and the training planes. The result, to the linear approximation, are two lines, one at the input plane and the other at the training plane.

The equations for the degeneracy curves at the input and the training planes can be obtained by expressing the degeneracy ellipses in terms of the local coordinates. Substitute Eq.(4.3) into Eq.(3.20). The first part of Eq.(3.20) is automatically satisfied. The second part of Eq.(3.20) gives the degeneracy curve at the input plane. To the first order approximation, the degeneracy curve is

described as a line at the input plane, in terms of the local coordinates (x_i', y_i'). The equation for this line is

$$\begin{aligned}&(\frac{2K_x k_0}{fn_e^2})x_i' + (\frac{2\cos\theta K_y k_0}{fn_o^2} + \frac{2\sin\theta\cos\theta K_z k_0}{fn_o^2 n_e\sqrt{1-\frac{\sin^2\theta}{n_o^2}}})y_i' + \\ &(-\frac{K_x^2}{n_e^2} - \frac{K_y^2}{n_o^2} - \frac{K_z^2}{n_e^2} + \frac{2\sin\theta K_y k_0}{n_o^2} - \frac{2K_z k_0\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}{n_e}) = 0,\end{aligned} \tag{4.6}$$

where $\mathbf{K} = (K_x, K_y, K_z)$ is a given grating.

Similarly, the degeneracy line at the training plane is

$$\begin{aligned}&(\frac{2K_x k_0}{fn_e^2})x_d' + (\frac{2\cos\theta K_y k_0}{fn_o^2} - \frac{2\sin\theta\cos\theta K_z k_0}{fn_o^2 n_e\sqrt{1-\frac{\sin^2\theta}{n_o^2}}})y_d' + \\ &(\frac{K_x^2}{n_e^2} + \frac{K_y^2}{n_o^2} + \frac{K_z^2}{n_e^2} - \frac{2\sin\theta K_y k_0}{n_o^2} - \frac{2K_z k_0\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}{n_e}) = 0.\end{aligned} \tag{4.7}$$

The above linear approximation is valid when the maximum deviation between these straight lines and the actual curves is much less than the separation between two adjacent pixels. More accurate expressions for the degeneracy curves at the input and training planes can be obtained by keeping higher order terms of x_i'/f, y_i'/f, x_d'/f and y_d'/f in Eq.(4.3) and Eq.(4.4).

Fig.4.2.2 shows two degeneracy lines drawn on regular 2-dimensional grids at the input and the training planes. The two lines are almost parallel, since the difference between the slopes of the two lines is very small when $K_z \ll K_y$ which is true if the separation angle between the input plane and the training plane, 2θ, is much bigger than the angle between two adjacent input or training wave vectors.

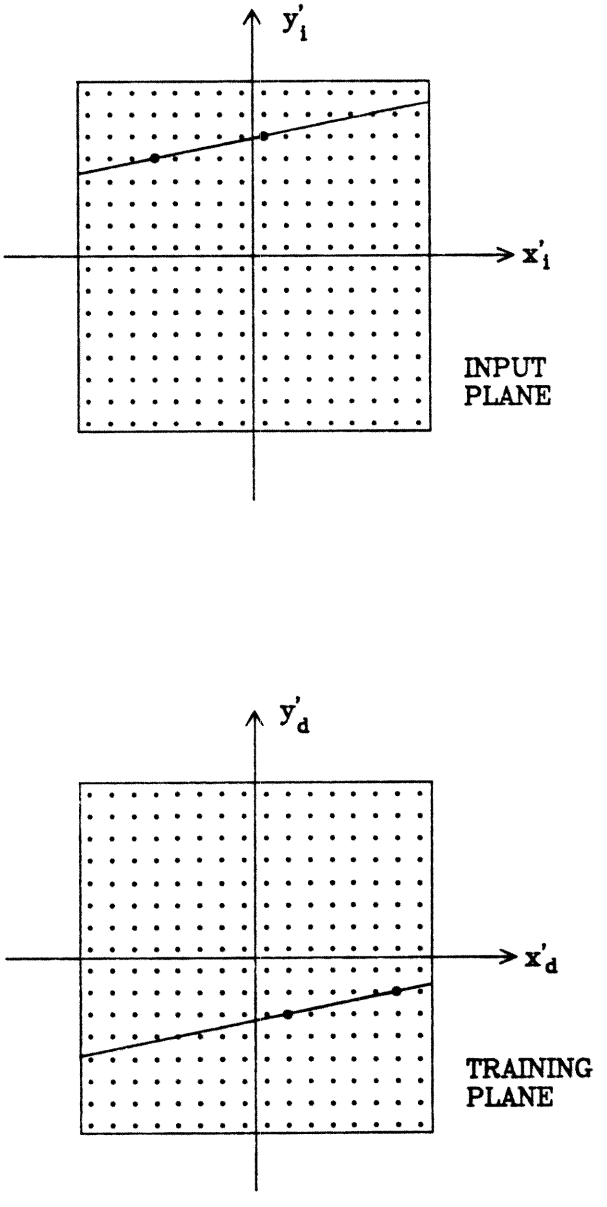


Fig.4.2.2 Two degeneracy lines drawn on regular 2-dimensional grids at the input and the training planes.

Once the degeneracy lines are found for a grating, all the degenerate interconnections corresponding to the same grating can be found. If a neuron on the degeneracy line at the input plane and a neuron on the degeneracy line at the training plane are separated by $x_i' - x_d' = K_x f/k_0$ in the x-direction, these two neurons are connected by the grating specifying the two lines. When two or more pairs of input-training neurons satisfy the above condition, the interconnections between those pairs of neurons are degenerate.

4.3 Optimal Configuration

Due to the degeneracy, the total number of independent gratings is less than the product of the number of pixels at the input plane and that at the output plane. In the last chapter, the accessibility of the gratings were discussed in the K-space. In the following discussion, the total number of accessible gratings will be calculated in terms of the dimensions of the input and the output planes. The result will show that for a 2-dimensional input (output) plane with $N \times N$ resolvable pixels, the maximum number of accessible gratings is in the order of N^3.

The optical system has to satisfy certain conditions in order to utilize the maximum number of accessible gratings. The aperture of the system, the dimensions of the crystal and the separation between two nearest pixels will be considered to achieve the optimal accessibility.

4.3.1 Total Number of Accessible Gratings

To find the relationship between the accessible grating space and the dimensions of the input and the training planes, it is necessary to express the grating vector in terms of (x_i', y_i') and (x_d', y_d'). Eq.(4.5) relates the local coordinates of a point at the input plane and a point at the training plane to the the grating vector which is Bragg matched by plane waves coming out of these two points.

To specify the finite dimensions of the input and the training planes, suppose the points at the input plane, (x_i', y_i') plane, are confined within a square, $-a/2 \leq x_i' \leq a/2$ and $-a/2 \leq y_i' \leq a/2$, and the points at the training plane, (x_d', y_d') plane, are confined within $-a/2 \leq x_d' \leq a/2$ and $-a/2 \leq y_d' \leq a/2$.

The minimum and maximum values of K_x, K_y and K_z, according to Eq.(4.5), are

$$\begin{aligned}
K_{xmin} &= -k_0 a/f, \\
K_{xmax} &= k_0 a/f, \\
K_{ymin} &= 2k_0 \sin\theta - k_0 \cos\theta a/f, \\
K_{ymax} &= 2k_0 \sin\theta + k_0 \cos\theta a/f, \\
K_{zmin} &= -k_0 \frac{\sin\theta \cos\theta \frac{n_e}{n_o^2}}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}} \frac{a}{f}, \\
K_{zmax} &= k_0 \frac{\sin\theta \cos\theta \frac{n_e}{n_o^2}}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}} \frac{a}{f}.
\end{aligned} \tag{4.8}$$

The shape of the accessible grating space is a parallelepiped bounded by edge points given above. This result can be obtained by considering the cross sections of the accessible grating space. In Eq.(4.5), the value of K_x does not depend

on the values of either K_y or K_z. Therefore, the cross section of the accessible grating space cut by any plane perpendicular to the K_x-direction is the same for all values of K_x. For a constant K_x, the minimum K_y is reached only when $y_i' = -a/2$ and $y_d' = a/2$, which results in $K_z = 0$. Likewise, the maximum K_y is reached only when $y_i' = a/2$ and $y_d' = -a/2$, and also $K_z = 0$. The minimum (maximum) K_z is given by $y_i' = y_d' = -a/2$ ($y_i' = y_d' = a/2$). When $K_z = K_{zmin}$ or $K_z = K_{zmax}$, K_y will be $2k_0 \sin\theta$, which is half way between its maximum and minimum values. The boundaries of the cross section are straight lines, since K_y and K_z are changing linearly with both y_i' and y_d'. The cross section at a constant K_x can be obtained by connecting the four points, $(K_x, K_{ymin}, 0)$, $(K_x, 2k_0 \sin\theta, K_{zmax})$, $(K_x, K_{ymax}, 0)$ and $(K_x, 2k_0 \sin\theta, K_{zmin})$. The resulting shape is a rhombus, as shown in Fig.4.3.1(a). The accessible grating space can be obtained by moving the rhombus in the **K**-space from K_{xmin} to K_{xmax}. The final result is a parallelepiped, as shown in Fig.4.3.1(b).

Comparison of Fig.4.3.1 with Fig.3.3.4 reveals that the linear approximation has replaced all curved lines in Fig.3.3.4 by straight lines. This approximation is valid when the area of the input (output) normal surface is very small so that the curvature of the normal surface can be neglected. When the area becomes large, higher order terms need to be added in Eq.(4.3) and Eq.(4.4).

The total number of accessible gratings can be calculated in the accessible grating space. The area of the rhombus in Fig.4.3.1(a) is $\frac{1}{2}(K_{ymax} - K_{ymin})(K_{zmax} - K_{zmin})$. The side of the parallelepiped in the K_x-direction is $K_{xmax} - K_{xmin}$. Therefore, the volume of the parallelepiped shaped

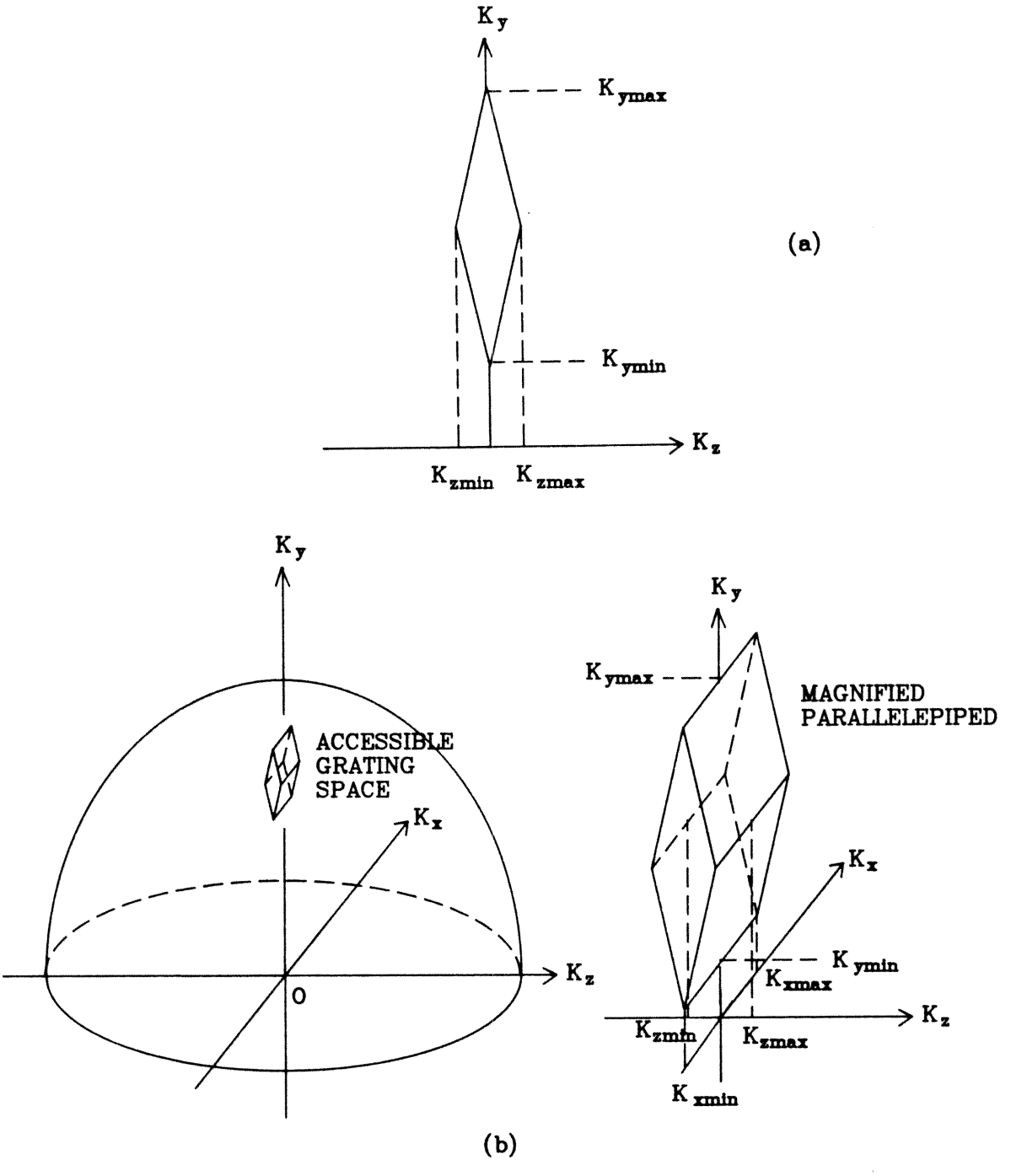


Fig.4.3.1 (a) The cross section of the accessible grating space cut by a plane perpendicular to the K_x-direction. (b) The accessible grating space, to the linear approximation.

accessible grating space is

$$V_a = \frac{1}{2}(K_{ymax} - K_{ymin})(K_{zmax} - K_{zmin})(K_{xmax} - K_{xmin}). \tag{4.9}$$

Substitution of Eq.(4.8) into Eq.(4.9) obtains

$$V_a = 2k_0^3(\frac{a}{f})^3 \frac{n_e}{n_o^2} \frac{\sin(2\theta)\cos\theta}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}}. \tag{4.10}$$

The ratio between the volume of accessible grating space and the volume of the whole grating space is V_a/V_K, i.e.,

$$\rho = \frac{3}{8\pi} \frac{1}{n_e n_o^3} \frac{\sin(2\theta)\cos\theta}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}} (\frac{a}{f})^3. \tag{4.11}$$

The total number of accessible gratings is just $N_a = V_a/v_g$, where v_g is the uncertainty volume of a grating, or $N_a = \rho C$. The expression in terms of the parameters of the optical system is

$$N_a = 2\frac{V_{xtal}}{\lambda^3}(\frac{a}{f})^3 \frac{n_e}{n_o^2} \frac{\sin(2\theta)\cos\theta}{\sqrt{1 - \frac{\sin^2\theta}{n_o^2}}}. \tag{4.12}$$

This result shows that the total number of accessible gratings is proportional to the storage capacity, Eq.(3.16), and depends on the numerical aperture of the system, a/f, and the geometry of the optical set up, θ.

4.3.2 Maximum Number of Pixels at the Input (Output) Plane

The separation between two nearest pixels is determined by the angular resolution of two adjacent plane waves, as discussed in the last chapter. Due to the refraction at the crystal surface, the separation angle inside the crystal, $\delta\theta_i$, and the separation angle outside the crystal, $\delta\theta_i'$, are different.

To decide the separation between two nearest pixels on a regular 2-dimensional grids with equal separation for all adjacent pixels, consider the point at the origin of the input plane ($x_i' = y_i' = 0$) and the point at the origin of the training plane ($x_d' = y_d' = 0$). It can be calculated from Eq.(3.31) that $\hat{k}_d = (0, \sin\frac{\Theta}{2}, \cos\frac{\Theta}{2})$, where Θ is the angle between $\hat{k}_i$ and $\hat{k}_d$. According to Eq.(3.40),

$$\delta k_{is} = (\frac{2\pi}{L_y}\sin\frac{\Theta}{2} + \frac{2\pi}{L_z}\cos\frac{\Theta}{2})\frac{1}{\sin\Theta}. \tag{4.13}$$

The value of δk_{is} is related to the value of $\delta\theta_i$ by Eq.(3.43). Since $\sin\theta_i' = n\sin\theta_i$, where n is the refractive index for the plane wave with incident angle θ_i', the relation between $\delta\theta_i'$ and $\delta\theta_i$ is

$$\cos\theta_i'\delta\theta_i' = n\cos\theta_i\delta\theta_i. \tag{4.14}$$

Therefore, the separation angle outside the crystal is

$$\delta\theta_i' = \frac{\cos\theta_i}{\cos\theta_i'}\frac{\delta k_{is}}{k_0}\cos\alpha_i. \tag{4.15}$$

Where the relation $|\mathbf{k}_i| = nk_0$ has been used, and the angle θ_i can be expressed

in terms of θ_i',

$$\cos\theta_i = \sqrt{1 - \frac{\sin^2\theta_i'}{n^2}}. \tag{4.16}$$

To obtain the minimum spatial separation $\delta y_i'$ in the y_i'-direction, use $\delta\theta_i' \approx \delta y_i'/f$. The expression for $\delta y_i'$ is

$$\delta y_i' \approx \frac{\sqrt{1 - \frac{\sin^2\theta}{n^2}}}{\cos\theta}\left(\frac{2\pi}{L_y}\sin\frac{\Theta}{2} + \frac{2\pi}{L_z}\cos\frac{\Theta}{2}\right)\frac{1}{\sin\Theta}\frac{f}{k_0}\cos\alpha_i. \tag{4.17}$$

The spatial separation between two nearest pixels in the x_i'-direction is calculated according to the uncertainty value of the grating vector in the x-direction. In order to separate the degeneracy from the angular resolution, it is assumed that when considering the minimum separation of the two input plane waves, the same output plane wave is concerned, and vice versa. In general, the minimum separation of two input plane waves, along the degeneracy curve, must be

$$\delta k_{in} = \delta\mathbf{K}\cdot\hat{n}_i \tag{4.18}$$

in order to be diffracted by two distinguishable gratings. In the case of two beams coming out of the origins of the input and the training planes, $\hat{n}_i$ is the x-direction. The difference between two input plane waves have to be at least $\Delta k_x = \delta K_x = 2\pi/L_x$ along the x-direction in order to distinguish two different gratings. Since $\delta p_{ix} = \delta k_{ix}$ and $\delta p_{ix} \approx \delta x_i' k_0/f$ according to Eq.(4.1), the minimum separation of two nearest pixels along the x_i' direction is

$$\delta x_i' \approx \frac{2\pi}{L_x}\frac{f}{k_0}. \tag{4.19}$$

A regular 2-dimensional grid with pixels equally separated in both directions

can be obtained when $\delta x_i' = \delta y_i'$. For Eq.(4.17) and Eq.(4.19) to be equal, the dimensions of the crystal and the angle between the input plane and the output (training) plane have to satisfy

$$\frac{1}{L_x} = \frac{\sqrt{1-\frac{\sin^2\theta}{n^2}}}{\cos\theta}\left(\frac{1}{L_y}\sin\frac{\Theta}{2} + \frac{1}{L_z}\cos\frac{\Theta}{2}\right)\frac{1}{\sin\Theta}\cos\alpha_i. \tag{4.20}$$

In this case, the maximum number of resolvable pixels at the input plane is $N \times N$, where

$$\begin{aligned} N &= a/\delta x_i' = a/\delta y_i', \\ &= \frac{a}{f}\frac{L_x}{\lambda}. \end{aligned} \tag{4.21}$$

The number of pixels along each direction, N, can be expressed in terms of the maximum number of accessible gratings, N_a, as in Eq.(4.12). It can be derived that

$$\begin{aligned} N^3 &= \left(\frac{a}{f}\frac{L_x}{\lambda}\right)^3, \\ &= N_a \frac{L_x^2\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}{2L_yL_z\sin(2\theta)\frac{n_e}{n_o^2}\cos\theta}. \end{aligned} \tag{4.22}$$

This relation shows that the total number of accessible gratings is of the order of magnitude of N^3.

4.3.3 Optimal Optical Setup Condition

To use the maximum number of accessible gratings, the dimensions of the crystal should be chosen such that the factor L_x^2/L_yL_z in Eq.(4.22) is maximized under the condition of Eq.(4.20). Substitute L_x in Eq.(4.20) into the factor

L_x^2/L_yL_z,

$$\begin{aligned}\frac{L_x^2}{L_yL_z} &= \frac{\cos^2\theta\sin^2\Theta}{(1-\frac{\sin^2\theta}{n^2})[\sin\Theta+\frac{L_z}{L_y}\sin^2(\frac{\Theta}{2})+\frac{L_y}{L_z}\cos^2(\frac{\Theta}{2})]\cos^2\alpha_i}, \\ &\leq \frac{\sin\Theta\cos^2\theta}{2(1-\frac{\sin^2\theta}{n^2})\cos^2\alpha_i}.\end{aligned} \tag{4.23}$$

The equality holds when

$$L_z\sin(\frac{\Theta}{2}) = L_y\cos(\frac{\Theta}{2}). \tag{4.24}$$

This gives the maximum number of N^3,

$$N_{max}^3 = N_a\frac{1}{4}\frac{n_o^2}{n_e}\frac{\sin\Theta}{\sin(2\theta)}\frac{\cos\theta\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}{(1-\frac{\sin^2\theta}{n^2})\cos^2\alpha_i}. \tag{4.25}$$

Expressing n and Θ in terms of n_o, n_e and θ, N_{max}^3 can be expressed as

$$N_{max}^3 = N_a\frac{1}{4}\frac{1-\frac{\sin^2\theta}{n_o^2}}{1-\frac{\sin^2\theta}{n^2}}\frac{1}{(1-\frac{\sin^2\theta}{n_o^2}+\frac{n_e^2}{n_o^4}\sin^2\theta)\cos^2\alpha_i}, \tag{4.26}$$

where

$$n = \sqrt{\sin^2\theta(1-\frac{n_e^2}{n_o^2})+n_e^2}. \tag{4.27}$$

Summarize the conditions for maximum number of independent interconnections,

$$\begin{aligned}L_x &= \frac{\cos\theta}{\sqrt{1-\frac{\sin^2\theta}{n^2}}}\frac{\cos(\frac{\Theta}{2})}{\cos\alpha_i}L_y, \\ L_y\cos(\frac{\Theta}{2}) &= L_z\sin(\frac{\Theta}{2}).\end{aligned} \tag{4.28}$$

Through the discussion it is assumed that the lens aperture is large enough such that the crystal is fully covered by any plane wave coming out of a point

at the input (training) plane, but the lens aperture is not too large to fit in the system. As shown in Fig.4.3.2, these restrictions are satisfied by the following two conditions:

$$\begin{aligned} a &< A - (L_y \cos\theta + L_z \sin\theta), \\ a &< A - L_x, \\ \frac{A}{2} &< f \tan\theta, \end{aligned} \tag{4.29}$$

where a is the linear dimension of the input (training) plane, A is the aperture of lenses.

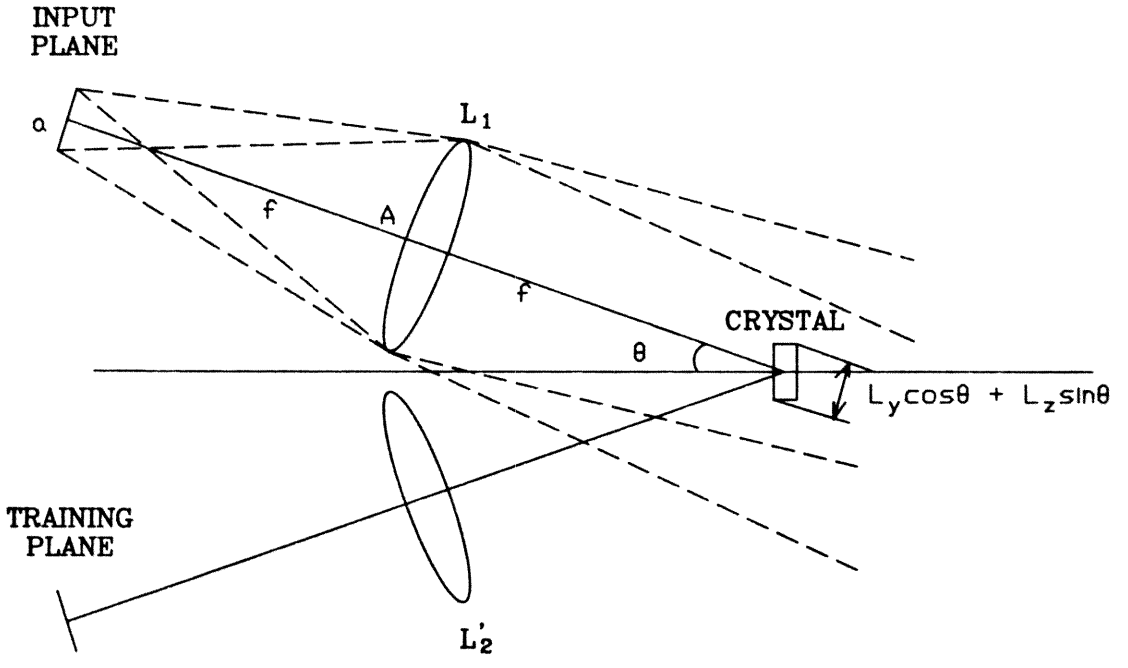


Fig.4.3.2 Geometrical optics considerations for the restrictions on the apertures of the system.

For the system to satisfy the paraxial approximation, it is also required that

$$a \ll 2f, \tag{4.30}$$

where f is the focal length of lenses which is assumed to be the same for all three lenses in the system shown in Fig.4.1.1.

To design an optimal optical system, the parameters of the system can be chosen as in the following example. 1) the angle θ between the two optical axes is chosen to yield high diffraction efficiency. 2) The relative dimensions of the crystal are chosen according to Eq.(4.28) to satisfy the optimal condition. 3) The numerical aperture a/f and the volume of the crystal V_{xtal} are chosen to implement the desired number of interconnections, according to Eq.(4.12) and Eq.(4.26). 4) The focal length of lenses f is chosen to match the separation between two adjacent pixels of the device used to implement neurons, such as SLM or LCLV, according to Eq.(4.19). 5) The dimension of the input (output) planes a is chosen as the numerical aperture times the focal length, $a = (a/f) \times f$. 6) The aperture of the lenses A is chosen to satisfy Eq.(4.29). An example of optimal set up parameters is given in Table 4.1.

In practical systems, the size of neurons must also be taken into consideration. The finite size of neurons will affect the effective separation between two adjacent pixels and the width of plane waves. Due to the finite width of plane waves, the crystal is not necessarily fully illuminated. Therefore, effective crystal dimensions, that is the dimensions of the part of crystal illuminated by plane waves, shall be used. The finite size of neurons will be considered in Chapter 5 when the experimental system is designed.

Table 4.1 An Example of Optimal Optical System		
Parameter	Notation	Value
Number of Pixels	N	100
Wave Length	λ	4880 Å
Refractive Index	n_o	2.34
of $LiNbO_3$	n_e	2.24
Angle	θ	20^o
Numerical Aperture	a/f	0.2
Dimensions	L_x	0.3 mm
of the	L_y	0.3 mm
Crystal	L_z	2 mm
Pixel Resolution	$\delta x'$	0.1 mm
Focal Length	f	60 mm
Dimension of the	a	12 mm
Input (Training) Plane		
Aperture of Lenses	A	30 mm

4.4 Fractal Sampling Grids

A set of locations of input (output) neurons without degenerate interconnections is called a fractal sampling grid. The word fractal implies that the sampling grid has a fractional dimension and usually higher order sampling girds can be generated according to the self similarity of fractals. The locations of input neurons and output neurons must be considered simultaneously so that any two pairs of input-output neurons are connected by two different gratings.

4.4.1 Dimensions of the Sampling Grids

The dimension of the input sampling grid, d_1, and the dimension of the output sampling grid, d_2, are related to each other and can be chosen arbitrarily within some range. The number of neurons at the input (output) plane, N_1 (N_2), determines the dimension of the fractal sampling grid, d_1 (d_2), since $N_1 = N^{d_1}$ ($N_2 = N^{d_2}$). The sum of d_1 and d_2 must be less than or equal to 3, because the number of independent interconnections connecting N_1 input neurons to N_2 output neurons, $N_1 \times N_2$, must be no more than N^3, which is the maximum number of accessible gratings. On the other hand, the dimension of a sampling grid embedded in a plane cannot exceed the dimension of the Euclidian space, which is 2 in this case.

Therefore, the dimensions of the input and the output sampling grids must satisfy

$$\begin{aligned} 0 \leq d_1 \leq 2, \\ 0 \leq d_2 \leq 2, \end{aligned} \tag{4.31}$$

and

$$d_1 + d_2 \leq 3. \tag{4.32}$$

Fractal sampling grids are often designed to reach the upper bound of the accessible interconnections. To use all N^3 independent gratings, the product of the number of points on the input sampling grid and that on the output sampling grid should be N^3. The equality sign in Eq.(4.32) is desired. Let d be the dimension of the input sampling grid, i.e., $d_1 = d$. The corresponding dimension of the output sampling grid is $d_2 = 3 - d$.

4.4.2 Different Kinds of Fractal Sampling Grids

Fractal sampling grids are usually designed for the input plane and the training plane. Since the output plane is the image plane of the training plane, the fractal sampling grid for the training plane will also serve for the output plane. If the local coordinate system at the output plane is chosen such that the $x_d'^{(output)}$-direction and the $y_d'^{(output)}$-direction are opposite to the $x_d'^{(training)}$-direction and the $y_d'^{(training)}$-direction respectively, as shown in Fig.4.2.1, the distributions of neurons at the training plane will be the same as that at the output plane.

To systematically avoid degenerate interconnections, one of the methods is to arrange the input (training) neurons along vertical columns, i.e., in the y_i' (y_d') direction, and separate the columns such that the horizontal distance between any two columns at the input plane is different from that between any two columns at the training plane. The resulting fractal sampling grids will be free from degeneracy. This can be seen from the following two sufficient conditions for avoiding degenerate interconnections.

The first sufficient condition for avoiding degenerate interconnections can be found from Eq.(4.5). Consider two neurons at the input plane, (x_{i1}', y_{i1}') and (x_{i2}', y_{i2}'), and two neurons at the training plane, (x_{d1}', y_{d1}') and (x_{d2}', y_{d2}'). It is sufficient for the two corresponding grating vectors, $\mathbf{K}_1 = \mathbf{k}_{d1} - \mathbf{k}_{i1}$ and $\mathbf{K}_2 = \mathbf{k}_{d2} - \mathbf{k}_{i2}$, to be different if their x components are not equal, i.e., $K_{1x} \neq K_{2x}$. According to Eq.(4.5) this sufficient condition can be expressed in terms of x_{i1}', x_{i2}', x_{d1}' and x_{d2}' as

$$x_{i1}' - x_{d1}' \neq x_{i2}' - x_{d2}', \tag{4.33}$$

or equivalently,

$$x'_{i1} - x'_{i2} \neq x'_{d1} - x'_{d2}. \tag{4.34}$$

The second sufficient condition for avoiding degenerate interconnections can be found with the help of the degeneracy lines at the input and the training planes. If the degeneracy lines corresponding to the grating $\mathbf{K}_1 = \mathbf{k}_{d1} - \mathbf{k}_{i1}$ does not pass through both neuron (x'_{i2}, y'_{i2}) at the input plane and neuron (x'_{d2}, y'_{d2}) at the training plane, the grating $\mathbf{K}_2 = \mathbf{k}_{d2} - \mathbf{k}_{i2}$ will definitely be different from $\mathbf{K}_1$. Therefore, it is sufficient that either point (x'_{i2}, y'_{i2}) or point (x'_{d2}, y'_{d2}) avoids the degeneracy lines, which are determined by point (x'_{i1}, y'_{i1}) and point (x'_{d1}, y'_{d1}).

Now, consider any two neurons belonging to two different columns at the input plane and any two neurons belonging to two different columns at the training plane. The first sufficient condition, Eq.(4.34), is satisfied, since the horizontal distance between the two columns at the input plane, $x'_{i1} - x'_{i2}$, is chosen to be different from the horizontal distance between the two columns at the training plane, $x'_{d1} - x'_{d2}$. Even if the two input neurons belong to the same column, the first sufficient condition will still be satisfied as long as the two output neurons do not belong to the same column, and vice versa.

Consider two neurons belonging to the same column at the input plane and two neurons belonging to the same column at the training plane. Since $x'_{i1} = x'_{i2}$ and $x'_{d1} = x'_{d2}$, the first sufficient condition is violated. However, the second sufficient condition is satisfied, since the two degeneracy lines determined by $\mathbf{K}_1 = \mathbf{k}_{d1} - \mathbf{k}_{i1}$ will not be both vertical as long as the angle θ is not 0^o or 90^o. This can be seen by considering the coefficient in front of y'_i in Eq.(4.6) and the

coefficient in front of y'_d in Eq.(4.7). The two coefficients will be both zero only if $\theta = 90^o$, or $\theta = 0^o$ and $K_y = 0$. By choosing the angle θ within the range $0^o < \theta < 90^o$, the degeneracy lines will not pass through both neuron (x'_{i2}, y'_{i2}) and neuron (x'_{d2}, y'_{d2}). Therefore, these two pairs of input-output neurons will not be connected by degenerate interconnections. In conclusion, this type of fractal sampling grids are free of degeneracy.

In the following discussion on fractal sampling grids, the unit of distance is chosen as the spatial separation of two adjacent pixels, and the first column to the left of the input (training) plane will be represented by $x = 1$.

Fig.4.4.1 shows an example of this kind of fractal sampling grid with $N = 16$. There are 4 columns of neurons at both the input plane and the training plane. The neurons at the input plane are distributed column by column, i.e., the separation between two adjacent columns is $\Delta x'_i = 1$. The neurons at the training plane are distributed uniformly with the separation between two adjacent columns $\Delta x'_d = 4$. Since the maximum separation between two columns at the input plane is less than the minimum separation between two columns at the training plane, the horizontal distance between any two columns at the input plane will not be equal to that at the training plane. Since this fractal sampling grid consists of $N^{3/2} = 64$ neurons at both the input plane and the training plane, the fractal dimensions are $d_1 = d_2 = 3/2$.

Fractal Sampling Grids for $N^d \mapsto N^{3-d}$ Mappings

The fractal sampling grids for $N^d \mapsto N^{3-d}$ mapping consist of N^d neurons at the input plane and N^{3-d} neurons at the training plane, where d is between

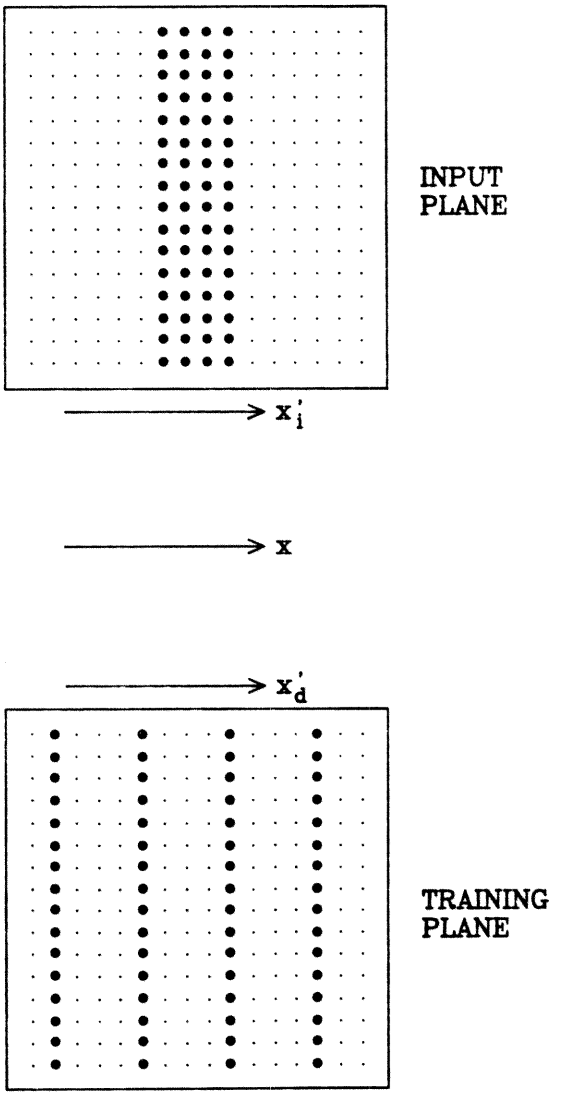


Fig.4.4.1 An example of fractal sampling grids with $N = 16$.

1 and 2. The neurons at the input (training) plane are arranged along n_1 (n_2) columns, with n_1 (n_2) defined by $N^d = N \times n_1$ ($N^{3-d} = N \times n_2$). Therefore, n_1 and n_2 are related by $n_1 \times n_1 = N$. Assume $n_1 \leq n_2$ in the following discussion. A family of fractal sampling grids can be systematically designed following the next three steps.

1) Select the n_2 neuron positions along one row at the training plane. Label the neurons as 1, 2, $\cdots$, n_2. Separate the n_2 neurons uniformly by n_1, i.e.,

$$x_k^{(d)} = (k-1) \times n_1 + x_1^{(d)}, \tag{4.35}$$

where k represents the kth neuron with $1 \leq k \leq n_2$ and $x_1^{(d)}$ represents the position of the first neuron. There are n_1 different distributions in this case, for $x_1^{(d)}$ can take one of the values from 1 through n_1. These n_1 distributions are called n_1 row patterns denoted by $\{B_1, B_2, \cdots, B_{n_1}\}$.

2) Select the n_1 neuron positions along one row at the input plane. Label the neurons as 1, 2, $\cdots$, n_1. Suppose that there are M different row patterns. For the jth row pattern, select the position of the kth neuron to be

$$x_k^{(i)} = (l_k^j - 1) \times n_1 + k, \tag{4.36}$$

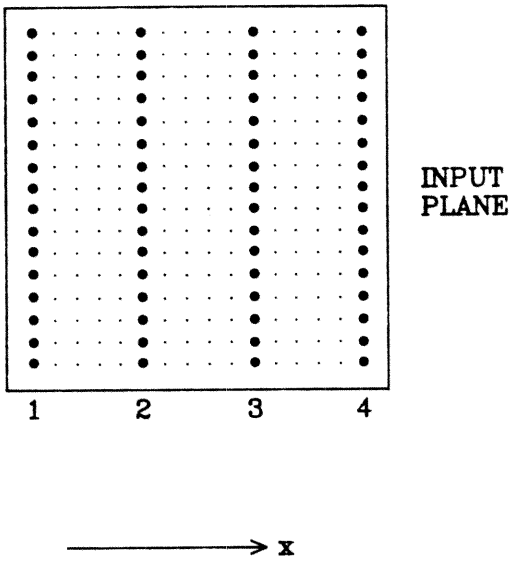
where l_k^j is an integer within the range $1 \leq l_k^j \leq n_2$. For each neuron, l_k^j can be any of the n_2 values from 1 through n_2. Since there are n_1 neurons, there exist $n_2^{n_1}$ different combinations for l_1^j, l_2^j, $\cdots$, $l_{n_1}^j$. Thus the number of different row patterns is $M = n_2^{n_1}$. Represent these M different row patterns by $\{A_1, A_2, \cdots, A_M\}$.

3) Design a pair of fractal sampling grids for the input and the training planes. For a row at the input plane, choose any of the $\{A_1, A_2, \cdots, A_M\}$ row patterns. For a row at the training plane, choose any of the $\{B_1, B_2, \cdots, B_{n_1}\}$ row patterns. Duplicate other rows at the input (training) plane according to the same row pattern. A total number of $(n_1 \times n_2^{n_1})$ pairs of different fractal sampling grids can be generated in this way. Fig.4.4.2 shows three pairs of different fractal sampling grids.

The degeneracy is avoided by using any pair of these fractal sampling grids. Consider the mth and the nth column at the input plane and the m'th and the n'th column at the training plane. The horizontal distance between the two columns at the input plane $x_m^{(i)} - x_n^{(i)} = (l_m^i - l_n^i) \times n_1 + (m - n)$, which is never a multiple number of n_1 for $m \neq n$. The horizontal distance between the two columns at the training plane $x_{m'}^{(d)} - x_{n'}^{(d)} = (m' - n') \times n_1$, which is always a multiple number of n_1 for $m' \neq n'$. So, the two distances can not be equal, i.e., $x_m^{(i)} - x_n^{(i)} \neq x_{m'}^{(d)} - x_{n'}^{(d)}$. According to the argument of sufficient conditions for avoiding degenerate interconnections, the fractal sampling grids are free of degeneracy.

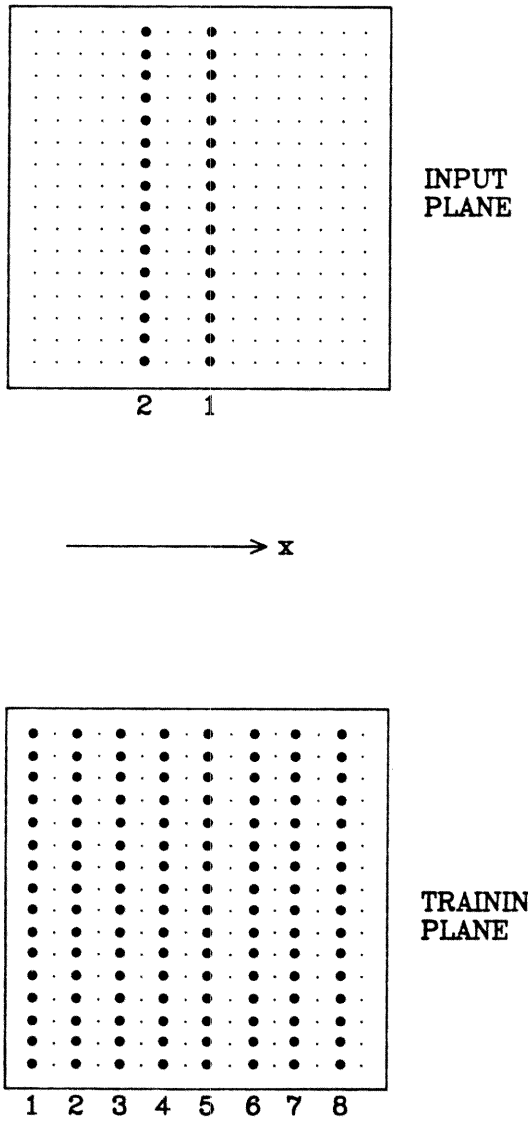
Higher Order Fractal Sampling Grids

Higher order fractal sampling grids can be generated from the first order fractal sampling grids if $\sqrt{n_1}$ and $\sqrt{n_2}$ are both integers. The higher order fractal sampling grids, which will be given below, are not necessarily scaling invariant, however the fractal dimension will be kept. The following three steps are used to generate the second-order fractal sampling grids.



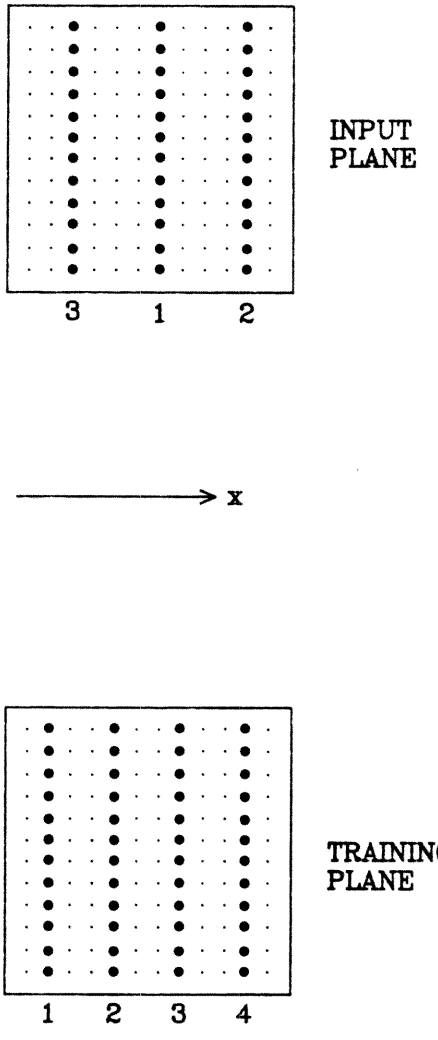
(a)

Fig.4.4.2 Three examples of fractal sampling grids for $N^d \mapsto N^{3-d}$ mappings. The neurons are labeled as 1, 2, 3, etc. Positions of neurons at the input and training planes are given by Eq.(4.36) and Eq.(4.35) respectively. (a) $N = 16$, $d_1 = d_2 = 3/2$, $n_1 = n_2 = 4$; in Eq.(4.35), $x_1^{(d)} = 1$; in Eq.(4.36), $l_1^j = 1$, $l_2^j = 2$, $l_3^j = 3$ and $l_4^j = 4$.



(b)

(b) $N = 16$, $d_1 = 5/4$, $d_2 = 7/4$, $n_1 = 2$, $n_2 = 8$; in Eq.(4.35), $x_1^{(d)} = 1$; in Eq.(4.36), $l_1^j = 5$, $l_2^j = 3$.



(c)

(c) $N = 12$, $d_1 = \log 36/\log 12 \approx 1.44$, $d_2 = \log 48/\log 12 \approx 1.56$, $n_1 = 3$, $n_2 = 4$; in Eq.(4.35), $x_1^{(d)} = 2$; in Eq.(4.36), $l_1^j = 3$, $l_2^j = 4$, $l_3^j = 1$.

1) Construct the first-order row patterns at the input and the training planes. Consider a row of $\sqrt{N}$ pixels at the input (training) plane and select $\sqrt{n_1}$ ($\sqrt{n_2}$) neuron positions. Using the steps 1) and 2) discussed above to design $\sqrt{n_1}$ row patterns at the output plane and $\sqrt{n_2}^{\sqrt{n_1}}$ row patterns at the training plane.

2) Generate a set of second-order row patterns for the fractal sampling grids with N total pixels per row. To generate the second-order row patterns at the input plane, choose any two (could be the same) first-order row patterns for the input plane. Replace each neuron of one of the two first-order row patterns by the other first-order row pattern, and replace each blank pixel by $\sqrt{N}$ blank pixels. That gives

$$X^{(i)}_{k1,k2} = [(l^{(j1)}_{k1} - 1)\sqrt{n_1} + k1 - 1]n_1 + (l^{(j2)}_{k2} - 1)\sqrt{n_1} + k2. \qquad (4.37)$$

Here, X is used to denote the horizontal neuron location on the second-order fractal sampling grid. The subscripts $k1$ and $k2$ indicate that the $k1$th neuron of the $j1$th first-order row pattern is replaced by the $j2$th first-order row pattern, and $k2$ represents the $k2$th neuron of the $j2$th first-order row pattern. The superscript (i) represents the input neuron. The location at the training plane will be represented by a superscript (d). Since $1 \leq j1, j2 \leq (\sqrt{n_2}^{\sqrt{n_1}})$, the total number of second-order row patterns for the input plane is $(\sqrt{n_2}^{\sqrt{n_1}})^2$. Similarly, the second-order row patterns for the training plane can be generated by using two first-order row patterns for the training plane.

$$X^{(d)}_{k1,k2} = [(k1 - 1)\sqrt{n_1} + x^{(d1)}_1 - 1]n_1 + (k2 - 1)\sqrt{n_1} + x^{(d2)}_1. \qquad (4.38)$$

Since $1 \leq x_1^{(d1)}, x_2^{(d2)} \leq \sqrt{n_1}$, the total number of second-order row patterns for the training plane is $(\sqrt{n_1})^2$, i.e., n_1.

3) Generate the second-order fractal sampling grids by using the second-order row patterns. Choose one of the second-order row patterns for a row at the input (training) plane. Duplicate other rows at the input (training) plane according to the same row pattern. The total number of pairs of different fractal sampling grids is $(\sqrt{n_2}^{\sqrt{n_1}})^2 \times (\sqrt{n_1})^2$, i.e., $n_1\sqrt{n_2}^{2\sqrt{n_1}}$. Fig.4.4.3 shows two examples of second-order fractal sampling grids.

To show that the second-order fractal sampling grids are free of degeneracy, it is sufficient to prove that the horizontal distance between two columns at the input plane is different from that at the training plane. Consider two columns at the input plane,

$$\begin{aligned} X^{(i)}_{k1,k2} &= [(l^{(j1)}_{k1} - 1)\sqrt{n_1} + k1 - 1]n_1 + (l^{(j2)}_{k2} - 1)\sqrt{n_1} + k2, \\ X^{(i)}_{m1,m2} &= [(l^{(j1)}_{m1} - 1)\sqrt{n_1} + m1 - 1]n_1 + (l^{(j2)}_{m2} - 1)\sqrt{n_1} + m2, \end{aligned} \tag{4.39}$$

and two columns at the training plane,

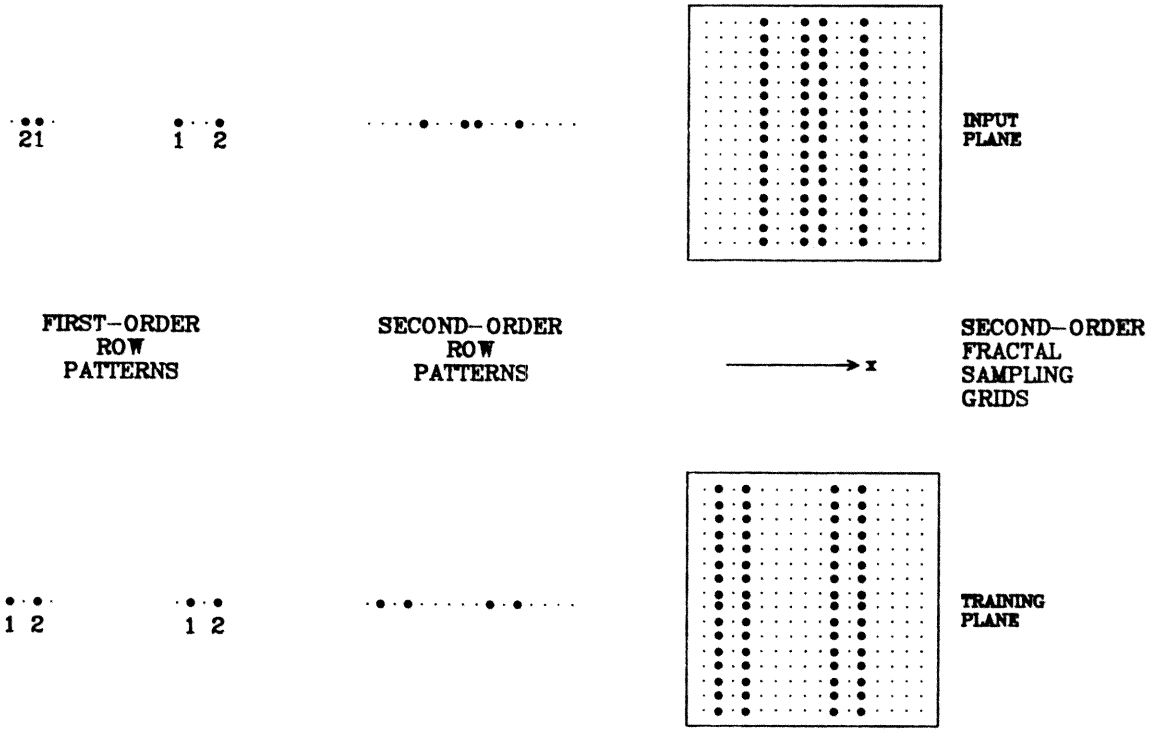
$$\begin{aligned} X^{(d)}_{k1',k2'} &= [(k1' - 1)\sqrt{n_1} + x_1^{(d1)} - 1]n_1 + (k2' - 1)\sqrt{n_1} + x_1^{(d2)}, \\ X^{(d)}_{m1',m2'} &= [(m1' - 1)\sqrt{n_1} + x_1^{(d1)} - 1]n_1 + (m2' - 1)\sqrt{n_1} + x_1^{(d2)}. \end{aligned} \tag{4.40}$$

The horizontal distances are

$$\begin{aligned} X^{(i)}_{k1,k2} - X^{(i)}_{m1,m2} =& [(l^{(j1)}_{k1} - l^{(j1)}_{m1})\sqrt{n_1} + (k1 - m1)]n_1 \\ &+ (l^{(j2)}_{k2} - l^{(j2)}_{m2})\sqrt{n_1} + (k2 - m2), \end{aligned} \tag{4.41}$$

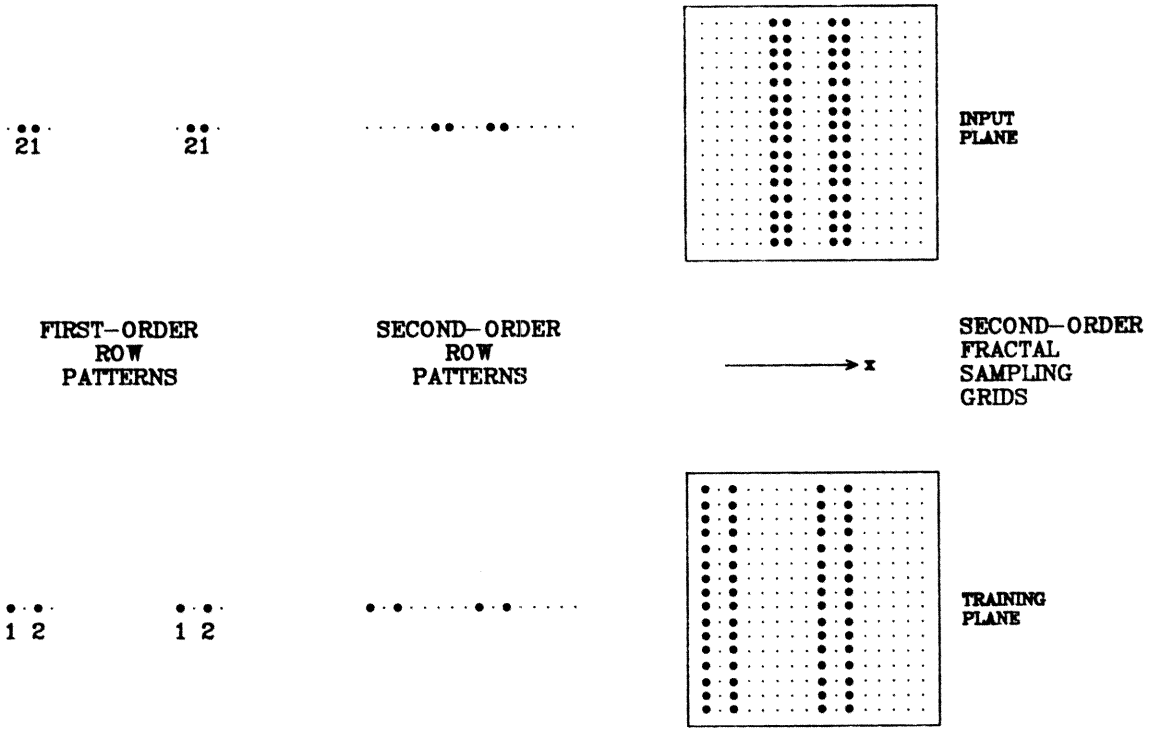
and

$$X^{(d)}_{k1',k2'} - X^{(d)}_{m1',m2'} = [(k1' - m1')\sqrt{n_1}]n_1 + (k2' - m2')\sqrt{n_1}. \tag{4.42}$$



(a)

Fig.4.4.3 Two examples of second-order fractal sampling grids for $N = 16$ and $d_1 = d_2 = 3/2$. (a) The second-order row pattern at the input (training) plane is generated by replacing each neuron of the first-order row pattern on the left by the first-order row pattern on the right, and replacing each empty pixel by 4 empty pixels.



(b)

(b) The self similar second-order row pattern for each plane is generated from two identical first-order row patterns.

If $k2 \neq m2$, the right hand side of Eq.(4.41) is not a multiple number of $\sqrt{n_1}$, since $1 \leq k2, m2 \leq \sqrt{n_1}$. However, the right hand side of Eq.(4.42) is always a multiple number of $\sqrt{n_1}$. So the two distances can not be equal. If $k2 = m2$, Eq.(4.41) becomes

$$X^{(i)}_{k1,k2} - X^{(i)}_{m1,m2} = [(l^{(j1)}_{k1} - l^{(j1)}_{m1})\sqrt{n_1} + (k1 - m1)]n_1. \tag{4.43}$$

The right hand side of Eq.(4.43) is always a multiple number of n_1. However, the right hand side of Eq.(4.42) is not a multiple number of n_1, since $1 \leq k2', m2' \leq \sqrt{n_1}$. So the two distance can not be equal either. Therefore, it has been proven that for the second-order fractal sampling grids generated above, the horizontal distance between two columns at the input plane is different from that at the training plane.

If $\sqrt{\sqrt{n_1}}$ and $\sqrt{\sqrt{n_2}}$ are still integers, third-order fractal sampling grids can be generated in a similar way. If not, higher order fractal sampling grid can still be designed by using the closest integers, but only a portion of the final sampling grid within $N \times N$ pixels will be utilized to locate input (training) neurons.

Special Fractal Sampling Grids for $N^{3/2} \mapsto N^{3/2}$ Mappings

The $N^{3/2} \mapsto N^{3/2}$ mappings are specially interesting because they allow the same number of neurons at the input and the training planes.

A special family of fractal sampling grids, which cannot be obtained using the method previously described, is found for $N^{3/2} \mapsto N^{3/2}$ mappings. The procedure for deriving these fractal sampling grids can also be described in three steps.

1) Select the $\sqrt{N}$ neuron positions along one row at the training plane. Label the neurons as 1, 2, $\cdots$, $\sqrt{N}$. Separate the $\sqrt{N}$ neurons uniformly by $\sqrt{N}+1$, i.e.,

$$x_k^{(d)} = (k-1) \times (\sqrt{N}+1) + 1, \tag{4.44}$$

where k represents the kth neuron with $1 \leq k \leq \sqrt{N}$. There is only one row pattern for the training plane.

2) Select the $\sqrt{N}$ neuron positions along one row at the input plane. Label the neurons as $1, 2, \cdots, j-1, j+1, \cdots, \sqrt{N}+1$. Notice that there is not jth neuron, and j can be any integer between 1 and $\sqrt{N+1}$. Suppose there are M different row patterns. For the ith row pattern, select the position of the kth neuron to be

$$x_k^{(i)} = (l_k^i - 1) \times (\sqrt{N}+1) + k, \tag{4.45}$$

where $1 \leq l_k^i \leq \sqrt{N}-1$ and $k \neq j$. For each neuron, l_k^i can be one of the $\sqrt{N}-1$ values from 1 through $\sqrt{N}-1$. Since there are $\sqrt{N}$ neurons, there exist $(\sqrt{N}-1)^{\sqrt{N}}$ different combinations for l_1^i, l_2^i, $\cdots$, l_{j-1}^i, l_{j+1}^i, $\cdots$, $l_{\sqrt{N}+1}^i$. And because j can be 1, 2, $\cdots$, $\sqrt{N+1}$, there are $\sqrt{N}+1$ possibilities of choosing j. Thus the number of different row patterns is $M = (\sqrt{N}+1)(\sqrt{N}-1)^{\sqrt{N}}$. Denote these M different row patterns by $\{A_1, A_2, \cdots, A_M\}$.

3) Design a pair of fractal sampling grids for the input and the training plane. For a row at the input plane, choose any of the $\{A_1, A_2, \cdots, A_M\}$ row patterns. For a row at the training plane, use the only row pattern derived

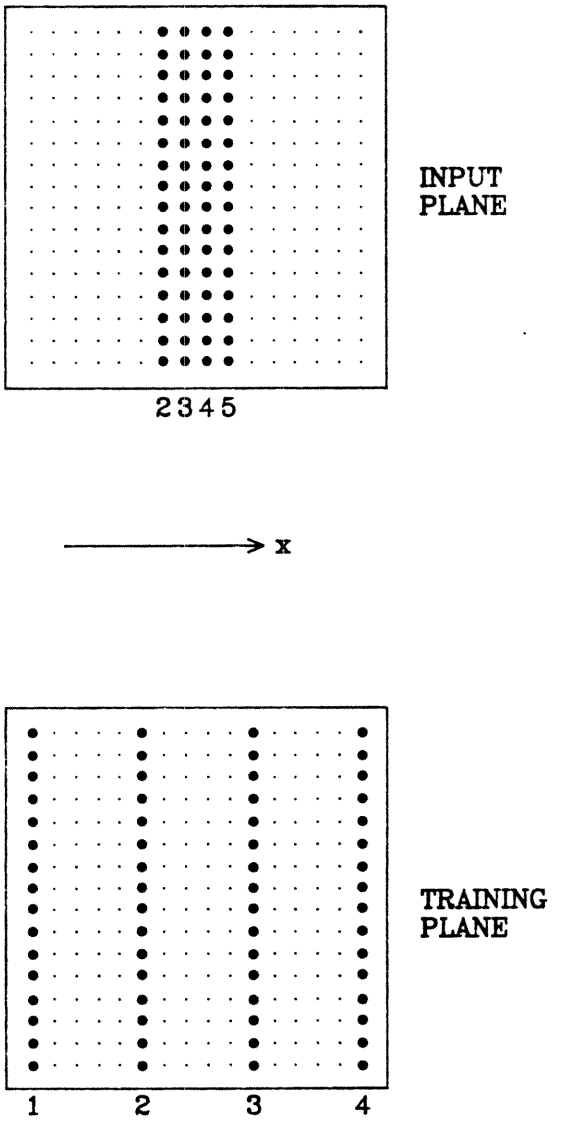
in step 1). Duplicate other rows at the input (training) plane according to the same row pattern. A total number of $[(\sqrt{N}+1)(\sqrt{N}-1)^{\sqrt{N}}]$ pairs of different fractal sampling grids can be generated in this way. Fig.4.4.4 (a) and (b) show two pairs of fractal sampling grids for $N^{3/2} \mapsto N^{3/2}$ mappings.

The fractal sampling grids are free from degeneracy, since they can be regarded as partial sampling grids for a bigger input (training) plane with $(\sqrt{N}+1) \times (\sqrt{N}+1)$ pixels.

Other Fractal Sampling Grids

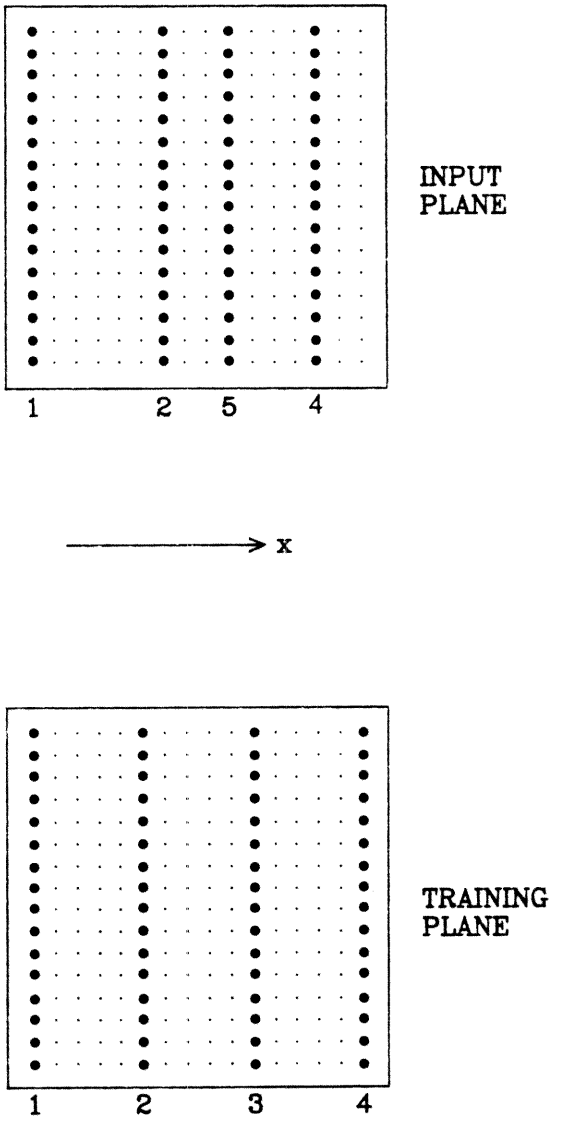
Neurons are not necessarily aligned along columns. Fig.4.4.5 shows a pair of fractal sampling grids with tilted lines. The neurons shown in Fig.4.4.5 are distributed more uniformly than those grouped in columns.

For this kind of fractal sampling grids to implement independent interconnections, the angle θ, which is the angle between the optical axis of lenses and the z-direction, has to be properly chosen. The slopes of degeneracy lines at the input and the training planes depend on the angle θ, as in Eq.(4.6) and Eq.(4.7). By changing the angle θ, the degeneracy line at the input (training) plane can be tilted so that it will not pass through certain neurons which are possible to cause degeneracy. Since the neurons are aligned along lines titled by the same angle, all degeneracy lines can be adjusted simultaneously to avoid any degeneracy. The degeneracy lines whose slopes do not change with the angle θ are lines along the x-direction. For each row at the input (training) plane, the positions of neurons can be given by Eq.(4.36) (Eq.(4.35)). It has been proven that the row patterns given by Eq.(4.35) and Eq.(4.36) can avoid degeneracy. Therefore, the



(a)

Fig.4.4.4 Two examples of fractal sampling grids for $N^{3/2} \mapsto N^{3/2}$ mappings with $N = 16$. Positions of neurons at the input and training planes are given by Eq.(4.45) and Eq.(4.44) respectively. (a) $j = 1$; in Eq.(4.45), $l_2^i = l_3^i = l_3^i = l_5^i = 2$.



(b)

(b) $j = 3$; in Eq.(4.45), $l_1^i = 1$, $l_2^i = 2$, $l_4^i = 3$, and $l_5^i = 2$.

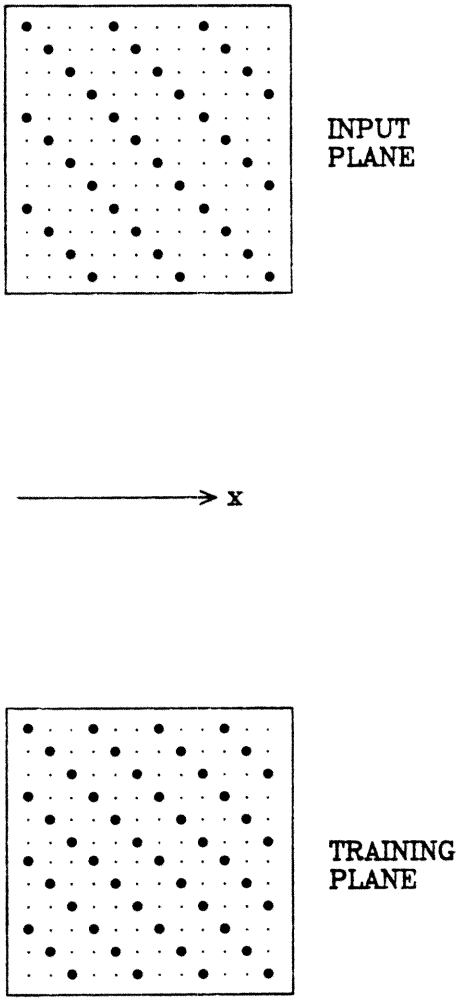


Fig.4.4.5 An example of fractal sampling grids with titled lines. $N = 12$, $d_1 = \log 36/\log 12 \approx 1.44$, $d_2 = \log 48/\log 12 \approx 1.56$.

fractal sampling grids, shown in Fig.4.4.5, can be used to avoid any degenerate interconnections.

Fig.4.4.6 shows another pair of fractal sampling girds. The only difference between the sampling grids shown in Fig.4.4.6 and that shown in Fig.4.4.5 is that one of the grids in Fig.4.4.6 contains broken lines of neurons. For a $N^{3/2} \mapsto N^{3/2}$ mapping as in Fig.4.4.6, the breaking of tilted lines makes the distribution of neurons more uniform than continuous lines. The arguments for degeneracy free conditions are similar as above.

The discussions in this chapter gives sufficient and necessary conditions for a pair of fractal sampling grids. The sufficient and necessary conditions can be described as: *any two pairs of input-training neurons must be related to two different pairs of degeneracy lines.* Here a pair of input-training neurons represents two neurons, one at the input plane and the other at the training plane. The related pair of degeneracy lines indicates two lines, one at the input plane and the other at the training plane, give by Eq.(4.6) and Eq.(4.7) respectively. The grating vector $\mathbf{K}$ in Eq.(4.6) and Eq.(4.7) is related to the positions of the given pair of neurons by Eq.(4.5).

The systematically designed fractal sampling grids given in this chapter are not complete. There exist other possible fractal sampling grids which cannot be generated using the formulas in Section 4.4. The complete set of fractal sampling grids still needs to be found.

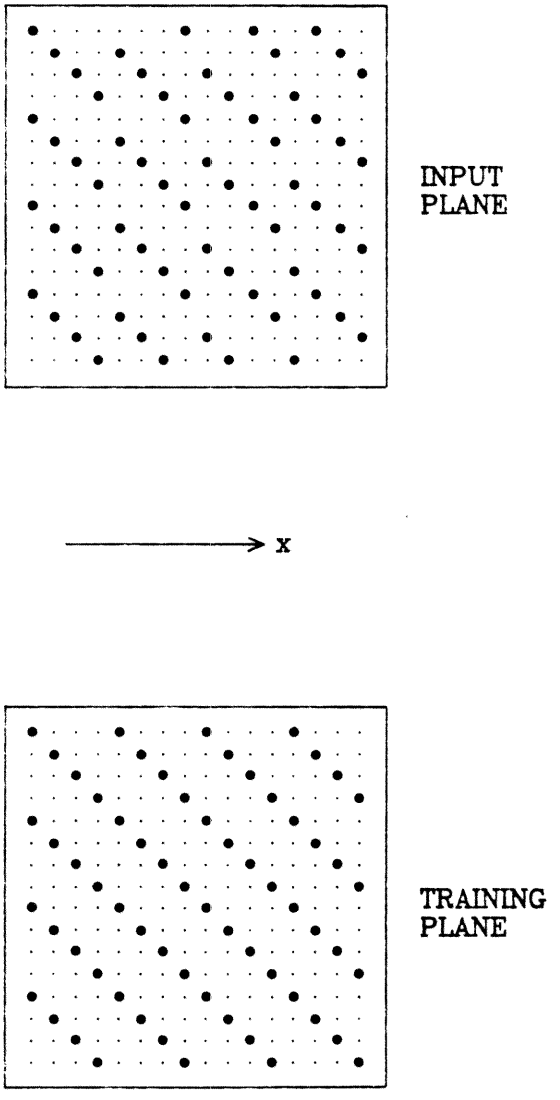


Fig.4.4.6 An example of fractal sampling girds with broken lines of neurons. $N = 16$, $d_1 = d_2 = 3/2$.

5. EXPERIMENTS

The optical implementations of neural networks with global and local connectivities are experimentally investigated by using the photorefractive crystals. For global connectivity, each neuron at the input plane is connected to all the neurons at the output plane. Whereas for local connectivity, each neuron at the input plane is only connected to neurons within a local neighborhood at the output plane. Results of a sequence of experiments are given in the following sections.

Experiments in this chapter are conducted to implement interconnections by sinusoidal phase gratings inside volume hologram crystals. Section 5.1 describes the experimental procedures. Section 5.2 shows the degeneracy problem. Section 5.3 demonstrates that with the help of the fractal sampling grids, independent interconnections for the global connectivity can be implemented. Section 5.4 illustrates the implementation of holographic hetero-associative memories using the outer product scheme. In Section 5.5, the fractal sampling grids for local connectivities are derived and experimentally tested.

Both the Vander Lugt system shown in Fig.2.1.3 and that shown in Fig.4.1.1 are used in the experiments. The system shown in Fig.2.1.3 is used to implement global connectivity, when the light intensity is sufficient to write gratings with low spatial frequencies. The system shown in Fig.4.1.1 can be used to implement both global and local connectivities. By using two lenses for the input and the training planes respectively, the system can be used to increase the strength of gratings by increasing the spatial frequencies of grating vectors, and to realize

local connectivities by spatial multiplexing.

5.1 Experimental Procedures

The experiments are performed in two steps: a) the storage of gratings into the crystal; b) the reading out of the stored memories. The first step is called the training process and the second the recall process.

5.1.1 Training

During the training process, the crystal records the interference pattern of the desired input and output patterns. The desired input and output patterns are placed at the input plane and the training plane respectively. The input (training) plane is illuminated by a collimated laser light propagating along the direction of the optical axis of the Fourier transforming lens L_1 (L_2').

Multiple exposures are conducted to store more than one pair of hetero-associative patterns. After the interference pattern formed by the first pair of desired input-output patterns is recorded, the patterns at the input and the training planes are replaced by the second pair of desired input-output patterns. During the second exposure, the second interference pattern is superimposed upon the first one inside the crystal. While the second set of gratings are written in the crystal, the first set of gratings are partially erased. By controlling the exposure times, the second set of gratings can be written with the same strength as that of the first set which has been partially erased by the second exposure. Similarly, many exposures can be done before the strength of each grating is too weak to be detected [16].

5.1.2 Recall

During the recall process, each output pattern is read out by the associated input pattern used in the training process. At the training plane, the light is blocked. At the input plane, the input pattern is placed at the same position as in the training process. At the output plane, the desired output pattern associated with the input pattern will be read out.

Together with the desired output pattern, there will be some weak cross-talk points and degenerate points at the output plane. The cross-talk points arise from the overlapping between different input patterns. The locations of these cross-talk points will be on the output fractal sampling grid. The degenerate points come from the Bragg matching of gratings by more than one pair of input-output points and are off the output fractal sampling grid.

The photorefractive crystal used in the experiments is the Fe-doped $LiNbO_3$. The parameters of the $LiNbO_3$ crystal are given in Table 5.1.

Table 5.1 Parameters of the $LiNbO_3$ Crystal

Parameter	Notation	Value
Dimensions	L_x	25 mm
of the	L_y	25 mm
Crystal	L_z	5 mm
Doping Density		0.05 %

5.2 DEGENERACY IN THE k-SPACE

Two experiments are performed to show the degeneracy problem. The first experiment records a grating with two points one at the input plane and the other at the training plane, and reads out the grating by another point at the input plane. The second experiment records a pair of continuous patterns without using the fractal sampling grids. The read-out pattern consists of the desired output pattern and some ghost images produced by the degenerate interconnections.

5.2.1 The Degenerate Interconnections

The points used for recording are shown in Fig.5.2.1. The two points at the input plane, (x'_{i1}, y'_{i1}) and (x'_{i2}, y'_{i2}), are aligned along the x'_i-direction, i.e., $y'_{i1} = y'_{i2}$. The point at the training plane, (x'_{d1}, y'_{d1}), is aligned in the y'-direction with the first input point, i.e., $x'_{d1} = x'_{i1}$. The coordinate systems in Fig.5.2.1 have been rotated 90^o with respect to those used in the last chapter.

The grating to be Bragg matched by both of the two input points is $\mathbf{K}_1 = \mathbf{k}_{d1} - \mathbf{k}_{i1}$ formed by the points (x'_{i1}, y'_{i1}) and (x'_{d1}, y'_{d1}). Eq.(4.5) indicates that grating $\mathbf{K}_1$ has a zero K_x component, since $x'_{d1} = x'_{i1}$. Eq.(4.6) and Eq.(4.7) give two horizontal degeneracy lines for grating $\mathbf{K}_1$, one at the input plane and the other at the output plane. Since the two input points are on the degeneracy line, $y'_{i1} = y'_{i2}$, the point (x'_{i2}, y'_{i2}) will also be Bragg matched by the grating $\mathbf{K}_1$.

Fig.5.2.2 shows the result of the read-out pattern. The two bright spots are the two input points going straight through the crystal. The two weaker spots result from the two diffracted plane waves. The point at the position (x'_{d1}, y'_{d1})

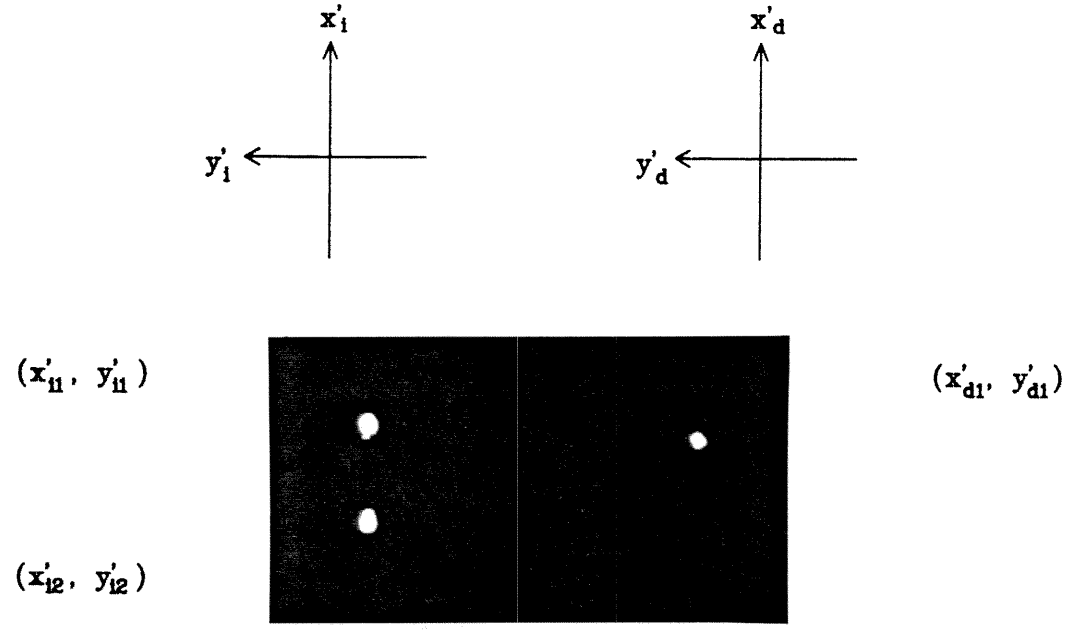


Fig.5.2.1 Three points used to show the degenerate interconnections.

is the desired output. The point at the position (x'_{d2}, y'_{d2}) is a degenerate spot read-out by the input point (x'_{i2}, y'_{i2}) which is degenerately Bragg matched by the grating $\mathbf{K}_1$. It can be seen that the degenerate point (x'_{d2}, y'_{d2}) is on the degeneracy line of $\mathbf{K}_1$, $y'_{d2} = y'_{d1}$.

5.2.2 Recording without Fractal Sampling Grids

Another experiment showing the degeneracy problem involves more that one degenerate gratings. Fig.5.2.3 shows the patterns used for recording. During the training process, letter B is put at the input plane and letter A is put at the training plane. The Fourier transforms of letters A and B interfere inside the crystal forming many gratings superimposed upon one another.

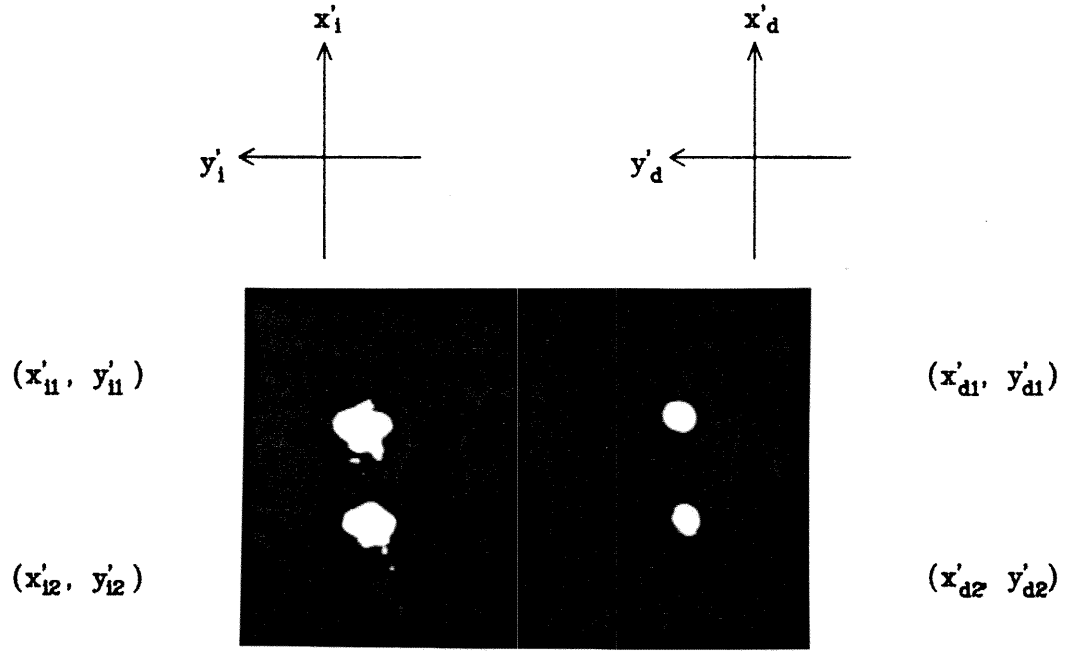


Fig.5.2.2 The input and the output points during the recall process.

During the recall process, letter A is blocked and the input letter B reads out the recorded volume hologram. Fig.5.2.4 shows the resulting pattern at the output plane. Besides letter A, the desired output, there are two weaker images of letter A shifted up and down. This is because letter B contains shift invariant patterns along the x_i'-direction. The upper half and the lower half of the letter B have similar structures. The ghost images of letter A are read out by more than one point of the letter B. There are other degenerate points. Those read out by fewer input points have lower intensities. Some of them can not be seen on the picture.

These experiments have demonstrated the necessity for using fractal sampling

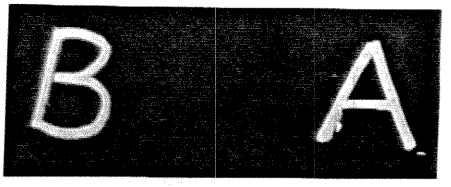


Fig.5.2.3 The 2-dimensional patterns to be recorded without using fractal sampling grids.

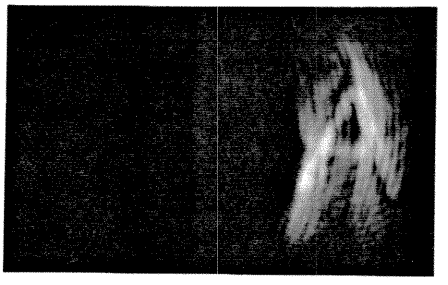


Fig.5.2.4 The pattern read out by the input letter B.

grids to implement independent interconnections. During the training process, the fractal sampling grids for the input and the training planes are used to record independent gratings. During the recall process, the fractal sampling grid for the output plane eliminates all the degenerate points, therefore only the desired output pattern will be the net output.

5.3 Global Connectivity

5.3.1 Optical System for Global Connectivity

The system used in the experiments demonstrating the global connectivity is the system with a single optical axis, as shown in Fig.2.1.3. The input and the training planes are separated by a distance s. Each neuron is represented by a spot of diameter $2R$. Since the spot size is not infinitesimal, the crystal is not entirely illuminated. Therefore, the gratings are written in an effective crystal volume with $L_{x(eff)}$, $L_{y(eff)}$ and $L_{z(eff)}$. By calculating the values of $\delta x'$ and $\delta y'$ according to Eq.(4.17) and Eq.(4.19), the distance between two adjacent pixels is chosen to be more than the required minimum value.

The effective crystal dimensions $L_{x(eff)}$, $L_{y(eff)}$ and $L_{z(eff)}$ can be calculated using the method of Fourier Optics [47]. If the electric field amplitude at the front focal plane of a Fourier transforming lens is $U(x_0, y_0)$, the field amplitude at the back focal plane will be

$$U(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0,y_0)\exp[-i\frac{2\pi}{\lambda f}(x_0x + y_0y)]dx_0dy_0. \qquad (5.1)$$

Where a constant factor has been neglected for simplicity. Eq.(5.1) is the Fourier transform of $U(x_0, y_0)$. Each circular spot at the input (training) plane is transformed into an Airy function [47] at the Fourier plane. Two different spots give rise to two overlapping Airy functions multiplied by different phase factors. The width of the Airy function is $1.22\lambda f/2R$. Therefore, each neuron spot is approximately converted to a plane wave with a circular cross section of diameter

$1.22\lambda f/2R$. The cross section is taken to be perpendicular to the z-direction and near the Fourier plane. The effective crystal volume illuminated by the plane waves, as shown in Fig.5.3.1, is given by

$$L_{x(eff)} \approx \frac{1.22\lambda f}{2R}, \tag{5.2}$$

$$L_{y(eff)} \approx \frac{1.22\lambda f}{2R}, \tag{5.3}$$

and

$$L_{z(eff)} = \min\{L_z, L_{y(eff)}/\tan(\theta)\}. \tag{5.4}$$

Where 2θ is the angle between the input and the training plane waves inside the crystal.

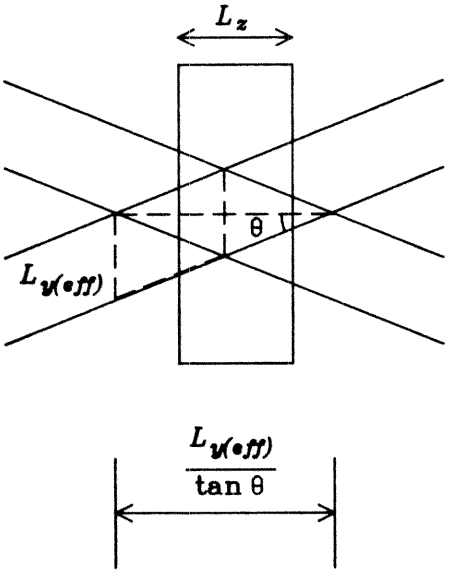


Fig.5.3.1 The effective crystal volume illuminated by the plane waves.

Table 5.2 Parameters of the System for Global Connectivities		
Parameter	Notation	Value
Focal Length	f	75 mm
Dimension of the Input (Training) Plane	a	5 mm
Separation between the Input and Training Planes	s	5 mm
Dimensions of the Fractal Sampling Grids	d_1 d_2	3/2 3/2
Distance between Two Nearest Pixels	$\Delta x'$ $\Delta y'$	$0.3mm$ $0.3mm$
Diameter of a Spot	$2R$	$0.15mm$
Wavelength of Ar-Laser	λ	$4880\mathring{A}$

The parameters of the system are given in Table 5.2. The effective crystal dimensions have been calculated as $L_{x(eff)} = L_{y(eff)} = 0.3mm$ and $L_{z(eff)} = L_z = 5mm$. According to Eq.(4.12) and Eq.(4.26), the maximum number of implementable interconnections is about $N^3_{max} \approx 3.3 \times 10^4$. The maximum number of pixels along each direction N_{max} is $N_{max} \approx 32$. Due to the finite size of neuron spots, N is chosen as $N = N_{max}/2 \approx 16$. With given spot size and focal length, the configuration given in Table 5.2 is close to optimal.

5.3.2 Sampling Grids and the Sampled Patterns

Different kinds of fractal sampling grids have been tested experimentally. In the last chapter, fractal sampling grids with neurons aligned along straight lines in the y'-direction and along tilted lines have been designed. Three pairs of them

will be used in the following experiments.

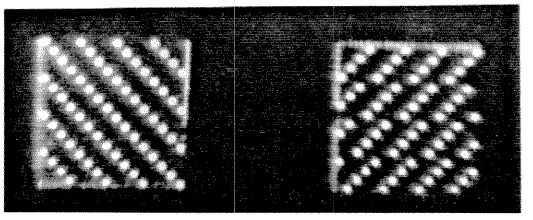
The fractal sampling grids and the sampled patterns are shown in Fig.5.3.2, Fig.5.3.3 and Fig.5.3.4. Fig.5.3.2(a) shows the fractal sampling grids with tilted lines. Fig.5.3.2(b) shows the sampled letters A and B. The total number of pixels at the input (training) plane is $N \times N = 16 \times 16 = 256$. The number of points on each fractal sampling grid is $N^{3/2} = 64$. To sample letters A and B, some of the points on the grids are illuminated and some are not. Fig.5.3.3(a) and Fig.5.3.4(a) show another kind of fractal sampling grids with neurons aligned along straight lines. Both of these grids have $N = 25$, $N_1 = N_1 = 125$ and $d_1 = d_2 = 3/2$. Fig.5.3.3(b) and Fig.5.3.4(b) show the corresponding patterns resembling letters A and B. The illuminated and the dark points represent neurons with output value 1 and 0 respectively. Each grating written by an input point and a training point represents an interconnection weight element w_{ij}.

Each pair of the sampled letters A and B will be used as input and training patterns in one of three different training processes. In each training process, a different location of the crystal will be used to record the volume hologram.

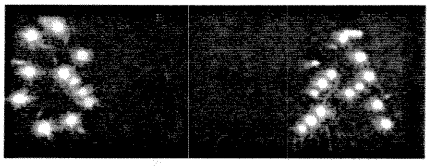
5.3.3 Experimental Results

Fig.5.3.2(c), (d) show the experimental results. Fig.5.3.2(c) is the read-out pattern obtained when the letter B at the training plane is blocked. Fig.5.3.2(d) is obtained by blocking letter A at the input plane. In fact, it is arbitrary to choose one of the letters A and B as the input pattern and the other as the training pattern, since the input plane and the training plane can be exchanged.

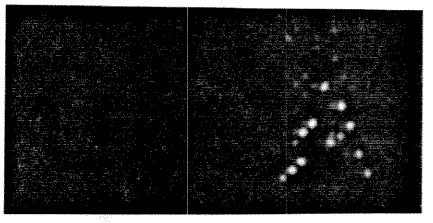
Fig.5.3.3(c), (d) and Fig.5.3.4(c), (d) show the experimental results when



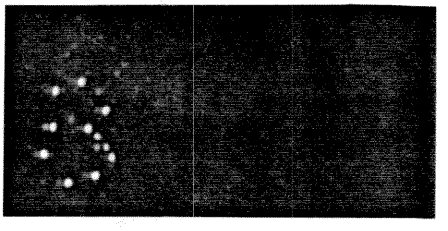
(a)



(b)

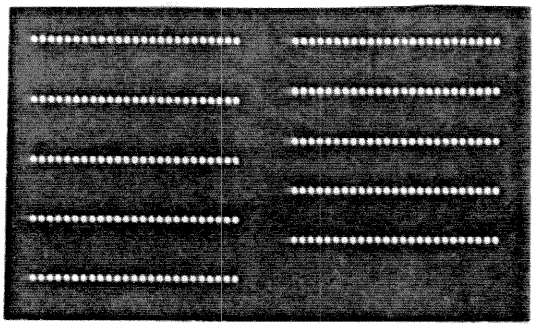


(c)

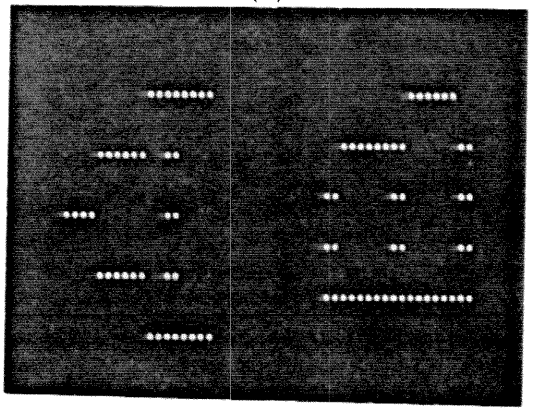


(d)

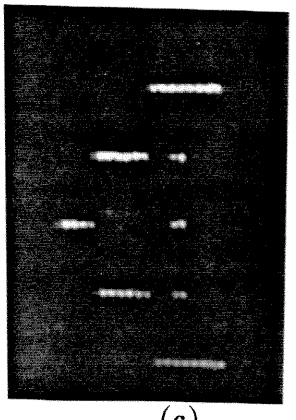
Fig.5.3.2 (a) The fractal sampling grids with tilted lines. (b) The sampled letters A and B. (c) Letter A read out by letter B. (d) Letter B read out by letter A.



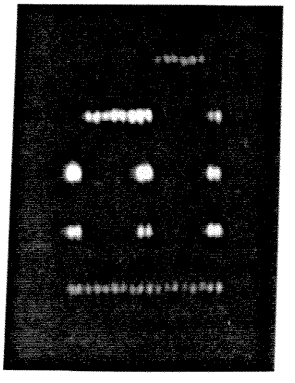
(a)



(b)

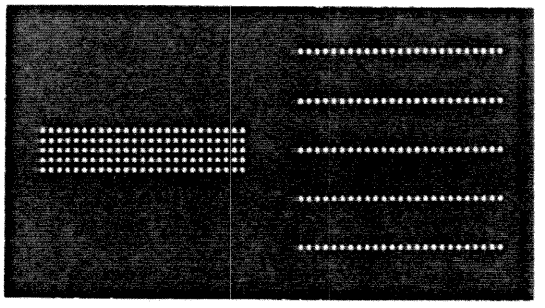


(c)

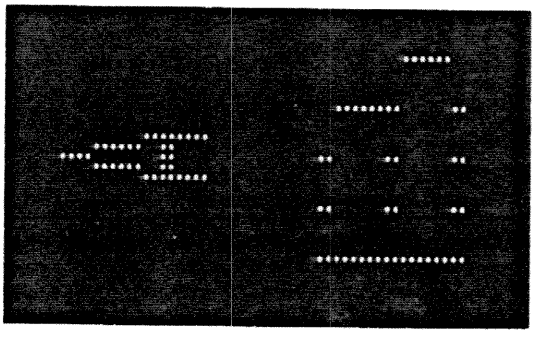


(d)

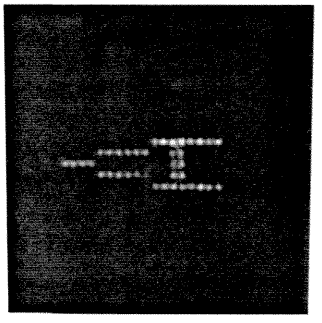
Fig.5.3.3 (a) The fractal sampling grids with neurons aligned along straight lines in the y'-direction. (b) The input and training patterns. (c) The output pattern A read out by the input pattern B. (d) The output pattern B read out by the input pattern A.



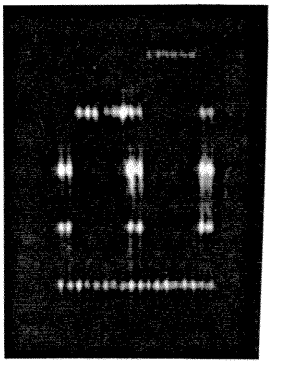
(a)



(b)



(c)



(d)

Fig.5.3.4 (a) Another pair of fractal sampling grids. (b) The input and training patterns. (c) The output pattern A read out by the input pattern B. (d) The output pattern B read out by the input pattern A.

using the other two pairs of fractal sampling grids.

The read-out patterns contain the desired output and some weak degenerate points. All the degenerate points are off the sampling grid. If the output fractal sampling grid is placed at the output plane, the net output pattern will contain only the desired output. Therefore, it has been demonstrated that the fractal sampling grids can eliminate the ghost images arising from the degenerate interconnections.

5.4 Holographic Hetero-associative Memories

With the help of the fractal sampling grids, the outer product scheme can be implemented using volume holograms. The following discussion will reveal the similarity between the outer product scheme and holographic memories. Hetero-associative memories with three different pairs of input-output patterns will be stored in the crystal and recalled sequentially.

5.4.1 Holographic Outer Product Scheme

The holographic gratings can be described as interconnection weight matrix elements. Suppose the N_1 input neurons and the N_2 training neurons have binary output values $(x_1, x_2, \cdots, x_{N_1})$ and $(y_1, y_2, \cdots, y_{N_2})$ respectively. The x_i's and y_j's are either 1 or 0. The value of x_i will be 1 if the neuron spot is illuminated, otherwise, 0. With the help of the fractal sampling grids, the grating connecting the input neuron i to the output neuron j can be formed only if both of these spots are illuminated, i.e., the product $x_i y_j = 1$. Therefore, the weight matrix

element w_{ji} can be written as

$$w_{ji} = y_j x_i = \begin{cases} 1, & \text{if the grating can be formed;} \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

The weight matrix can be written as an outer product of vectors $\mathbf{x}$ and $\mathbf{y}$, i.e.,

$$\mathbf{w} = |y\rangle \langle x| . \tag{5.6}$$

The storage of more than one pair of input-output patterns can be accomplished by multiple exposures. Multiple exposures will partially erase the holograms previously recorded in the crystal. The exposure times are controlled so that when the last hologram is recorded, all the holograms have the same strength. This can be done by exposing the first hologram with the longest time, then reducing the exposure time for the following exposures sequentially. The final weight matrix can then be written as

$$\mathbf{w} = \sum_{n=1}^{M} |y^n\rangle \langle x^n| . \tag{5.7}$$

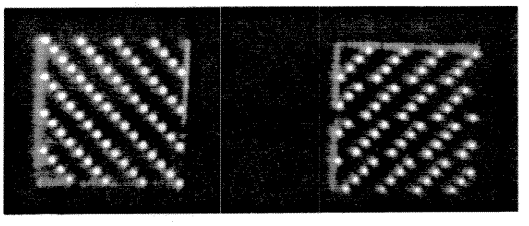
In the above equation, M is the number of exposures, or the number of associative input-output pairs. It can be noticed that Eq.(5.7) is the same as that given by the outer product scheme, Eq.(1.9).

5.4.2 Optical Implementation of Hetero-associative Memories

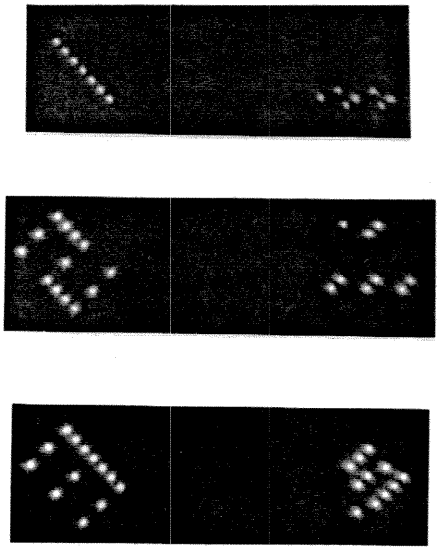
The following experiment associates the Chinese characters *one, two, three* with the Arabic numerals 1, 2, 3 respectively. The purpose is to recall the Arabic 1, 2, 3 by the corresponding Chinese *one, two, three* respectively. The optical system is the same as that described in the last section.

During the training process, the three pairs of input-output patterns are recorded by three exposures. Fig.5.4.1 shows three pairs of patterns to be associated. For the first exposure, Chinese *one* and Arabic 1 are placed at the input plane and the training plane respectively. For the second exposure, Chinese *two* and Arabic 2 are placed at the input and the training planes respectively. For the third exposure, Chinese *three* and Arabic 3 are placed at the input and the training planes respectively. The exposure times are reduced sequentially.

During the recall process, different input patterns are used to read out the corresponding output patterns. Fig.5.4.2 shows the Arabic numerals 1, 2 and 3 read out by the Chinese characters *one*, *two* and *three* respectively. Inspect the output number 2. Besides all the degenerate points off the output sampling grid, there are also some weak points resulting from the Arabic 1 and 3. This is because the Chinese character *two* has some common points with those of *one* and *three*. This cross-talk noise is the limitation of the capacity of the outer product scheme. Within the capacity of the outer produce scheme, the cross-talk noise can be eliminated by choosing a proper threshold value for the output neurons.



(a)



(b)

Fig.5.4.1 The fractal sampling grids and the sampled patterns used in the implementation of hetero-associative memories. (a) The fractal sampling grids. (b) The Chinese characters *one*, *two*, *three* to be associated with the Arabic numerals 1, 2, 3 respectively.

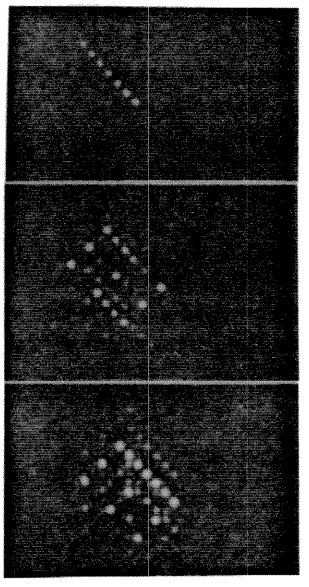


Fig.5.4.2 The Arabic numerals 1, 2 and 3 read out by the Chinese characters *one*, *two* and *three* respectively.

5.5 Local Connectivity

A network with local connectivity has a structure such that each neuron in the input layer is only connected to neurons within a local neighborhood in the output layer, as shown in Fig.5.5.1. The optical implementation of local connectivity requires that each beam coming from an input neuron intersects with only some of the beams coming from the training neurons.

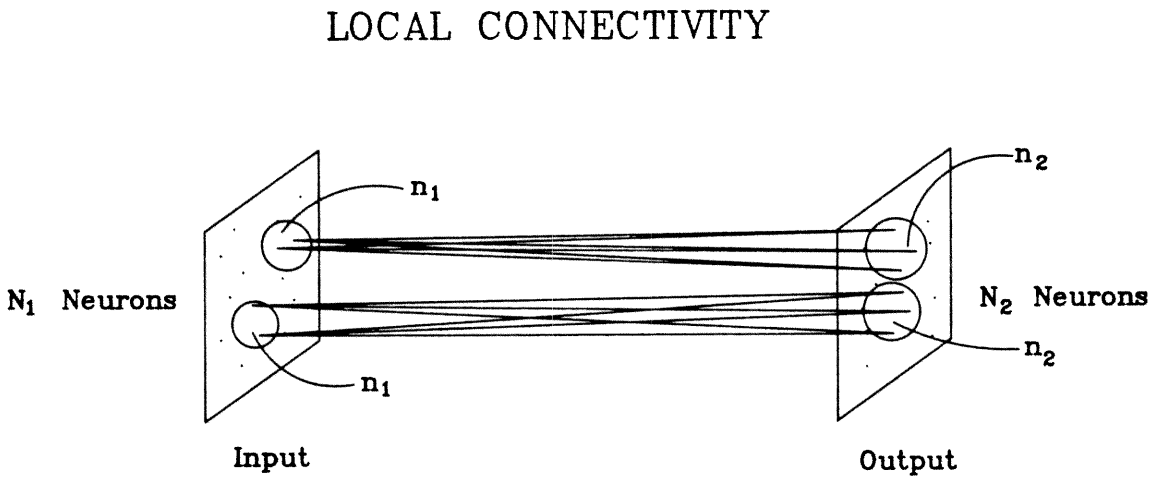
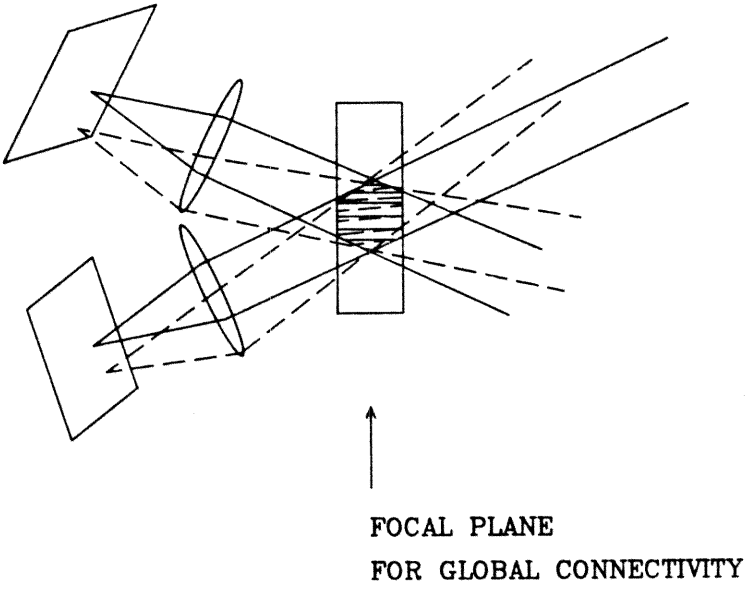


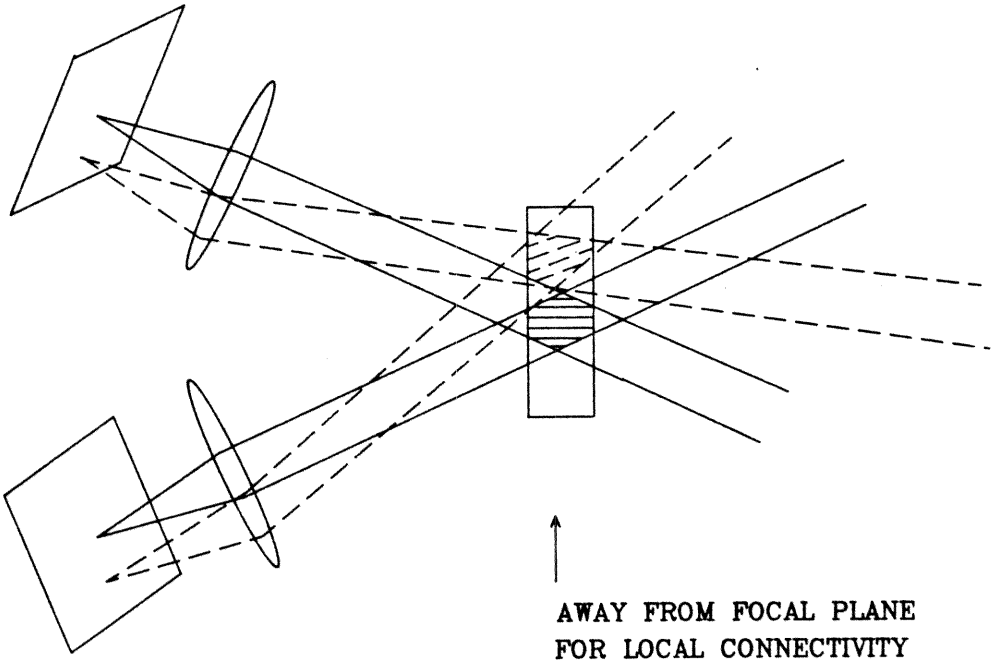
Fig.5.5.1 A neural network with local connectivity.

5.5.1 Optical System for Local Connectivity

The optical system used to implement local connectivities is a modified Vander Lugt system with two optical axes, similar to that shown in Fig.4.1.1. Fig.5.5.2(a) and (b) depict the geometrical optical path for the systems used for implementing global and local connectivities respectively. In Fig.5.5.2(a), the two optical axes cross each other at the common focal point of lenses L_1 and L_2'. All plane waves overlap at the same region where the crystal is placed. In Fig.5.5.2(b), the two optical axes cross each other at a point beyond the focal planes of lenses L_1 and L_2'. Inside the crystal, each plane wave overlaps with only some of the other plane waves.



(a)



(b)

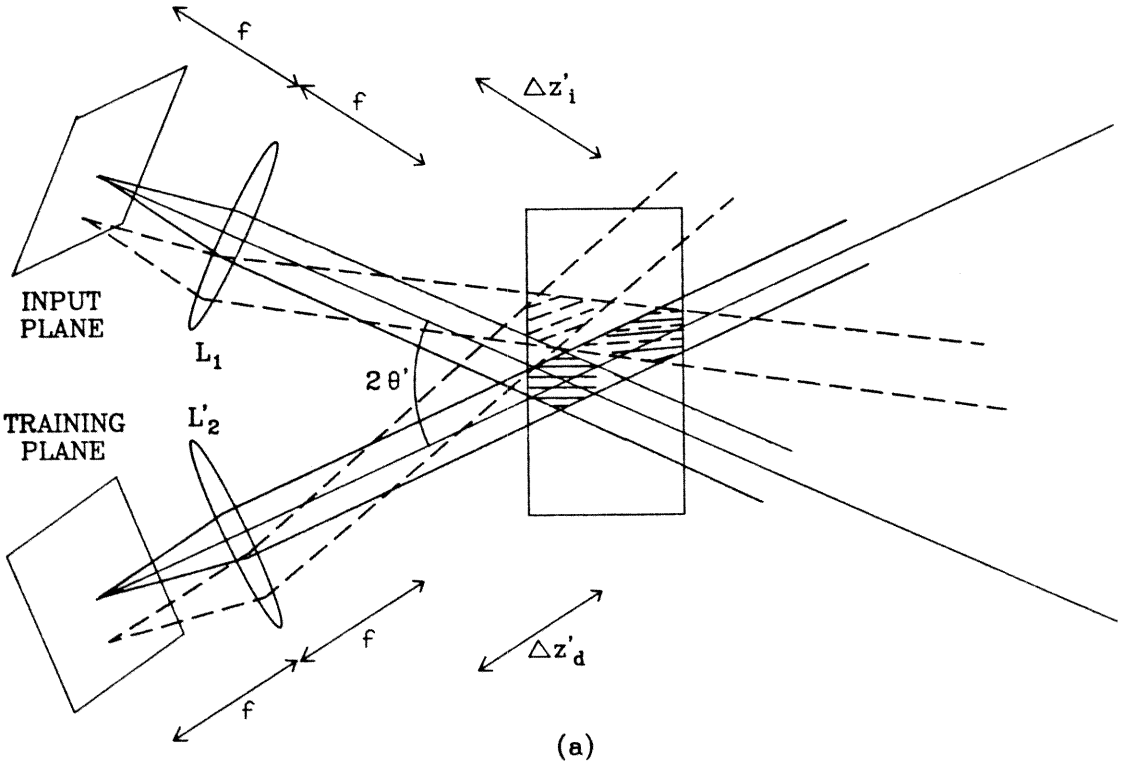
Fig.5.5.2 (a) The geometrical optical path for the systems used for implementing global connectivities. (b) The geometrical optical path for the systems used for implementing local connectivities.

The fact that each plane wave overlaps with only its near neighbors at a certain distance away from the focal plane can be used to implement local connectivity. Fig.5.5.3 shows the optical system for local connectivity. The crystal is placed at the cross point of the optical axes of lenses L_1 and L_2'. The two lenses are located at the same distance away from the crystal. The input (training) plane is located at the front focal plane of lens L_1 (L_2'). By controlling the distance between the crystal and the lens, the size of local neighborhood can be adjusted.

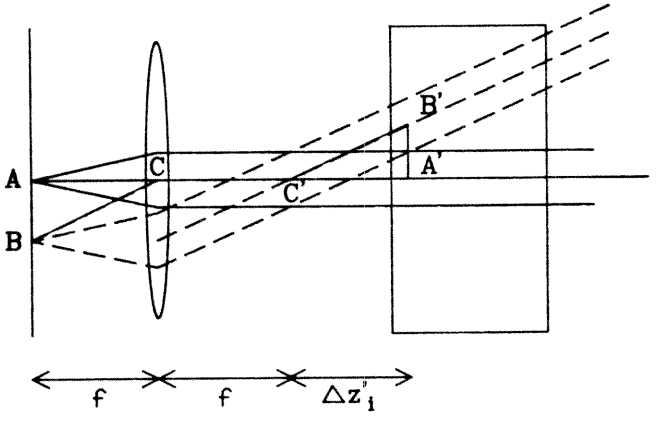
The distance between the crystal and the lens can be calculated, given the size of a neuron spot at the input (training) plane, the separation between two adjacent pixels, the focal length of the lens, the desired area of the local neighborhood and the wavelength of light. Suppose each neuron at the input plane is to be connected to output neurons within an area equal to $m \times m$ pixels. The distance between the crystal and the focal plane of lens L_1 (L_2') can be calculated according to Fig.5.5.3(b). Consider the two similar triangles ΔABC and $\Delta A'B'C'$. AB is the distance between two pixels at the input plane, $A'B'$ is the separation between the two beams at a distance $\Delta z'$ away from the focal plane. The value $\Delta z'$ is chosen so that when the two points A and B are m pixels apart, the separation $A'B'$ is the width of each plane wave b. Since $AB/AC = A'B'/A'C'$, $AC = f$ and $A'C' = \Delta z'$, therefore

$$\frac{m\Delta x'}{f} = \frac{b}{\Delta z'}. \tag{5.8}$$

$\Delta x'$ is the distance between two adjacent pixels, b is the width of each plane wave, $\Delta z'$ is the distance between the crystal and the focal plane of the input



(a)



(b)

Fig.5.5.3 The optical system for local connectivity. (a) The side view. (b) The top view.

(training) plane. The width of each plane wave b can be obtained from Eq.(5.2), which gives

$$b = \frac{1.22\lambda f}{2R}. \tag{5.9}$$

Where $2R$ is the diameter of each neuron spot. Therefore, $\Delta z'$ is related to the locality parameter m by

$$\Delta z' = \frac{1.22\lambda f^2}{2Rm\Delta x'}. \tag{5.10}$$

In the above calculation, the locality parameter m is determined according to the pixels along the x'-direction. It can be noticed that the x' and y' directions are not symmetric due to the large angle θ' between the two optical axes. Therefore, the number of pixels to be locally connected in the y'-direction, m_y, is usually larger than m. This will be seen in the following experiment. However, since the degeneracy occurs mostly along the x'-direction, the locality parameter m_y along the y'-direction will not affect the fractal sampling grids. In the following discussions on the fractal sampling grids, the local area will be taken as $m \times m$ pixels.

Table 5.3 lists the parameters of the optical system used to implement local connectivities.

Table 5.3 Parameters of the System for Local Connectivities		
Parameter	Notation	Value
Focal Length	f	75 mm
Angle between Two Optical Axes	2θ	45^o
Distance between Two Adjacent Pixels	$\Delta x'$ $\Delta y'$	$0.3mm$ $0.3mm$
Diameter of a Spot	$2R$	$0.15mm$
Wavelength of Ar-Laser	λ	$4880\mathring{A}$
Distance between Crystal Focal Plane	$\Delta z'$	$4mm$

5.5.2 Fractal Sampling Grids for Local Connectivity

It is necessary to design new fractal sampling grids for local connectivity. It is still necessary to use sampling grids for local connectivity, because the degeneracy problem is intrinsic in using holographic gratings. At the same space inside the crystal, a grating can always connect more than one pair of wave vectors. However, at different places inside the crystal two gratings with the same grating vector represent two independent gratings because of their spatial separation, similar to the spatial multiplexing of holograms. Therefore, the fractal sampling grids for local connectivities are different from those for global connectivities.

The sampling grids for local connectivity can be derived from those for global connectivity. First the sampling grids for global connectivity of neurons within input area of $m \times m$ pixels and output area of $m \times m$ pixels are designed, as in

Chapter 4. Then the same local input (output) sampling grid is duplicated and stacked to fill up the whole input (output) plane. Fig.5.5.4 shows an example of fractal sampling grids for local connectivity.

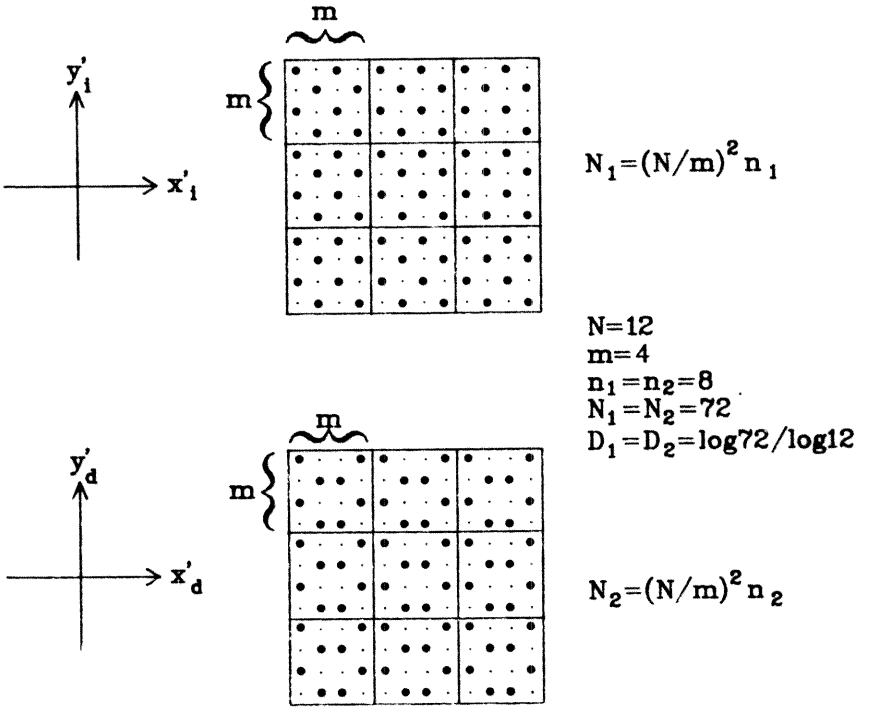


Fig.5.5.4 An example of fractal sampling grids for local connectivity.

The fractal dimension of the sampling grids for local connectivity depends on the size of the neighborhood and it is higher than those for global connectivity. As shown in Fig.5.5.4, the input (training) plane is divided into $(N/m) \times (N/m)$ blocks. Each neuron at the input plane is only connected to neurons within one

block at the output plane. Suppose each of the neurons at the input (output) plane is connected to n_2 (n_1) neurons at the output (input) plan. n_1 and n_2 are related by $n_1 \times n_2 = m^3$, where m is the number of pixels along each direction within a block. The total number of neurons at the input (output) plane is $N_1 = n_1 N^2/m^2$ ($N_2 = n_2 N^2/m^2$), where $N \times N$ is the total number of pixels. The fractal dimensions D_1 and D_2 of the sampling grids for the local connectivity are $D_1 = \log N_1 / \log N$ and $D_2 = \log N_2 / \log N$ respectively. The sum of D_1 and D_2 is

$$\begin{aligned} D_1 + D_2 =& \frac{\log(N_1 N_2)}{\log N}, \\ =& \frac{\log(N^4/m)}{\log N}, \\ =& 3 + \frac{\log(N/m)}{\log N}. \end{aligned} \tag{5.11}$$

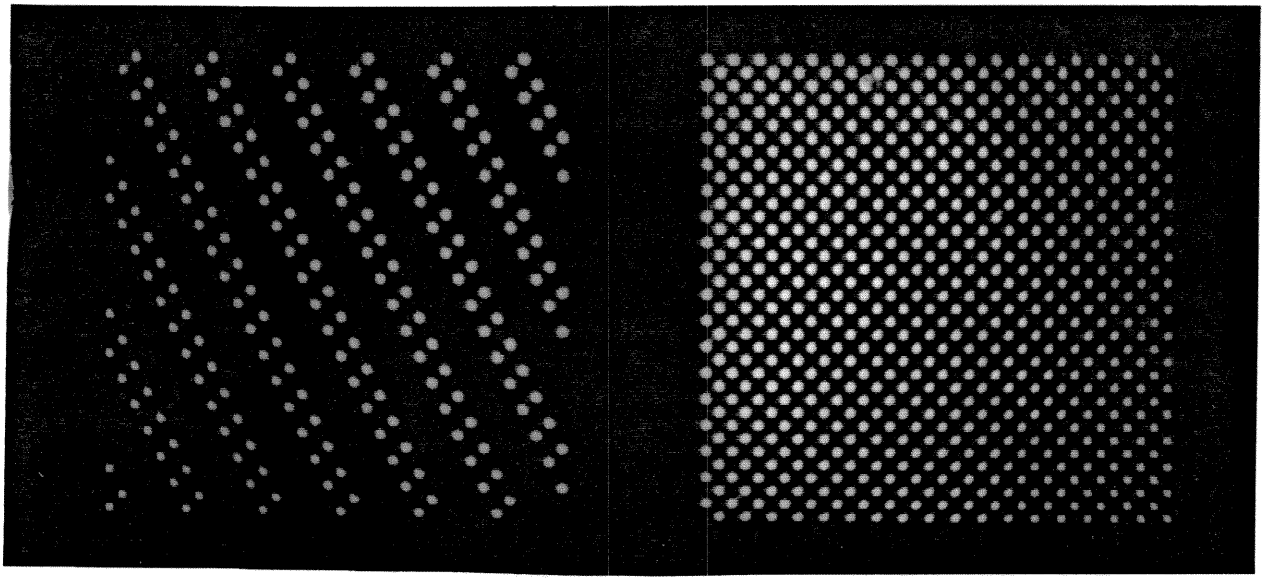
It can be noticed that

$$D_1 + D_2 \begin{cases} > 3, & \text{for local connectivity, i.e., } N > m; \\ = 3, & \text{for global connectivity, i.e., } N = m. \end{cases} \tag{5.12}$$

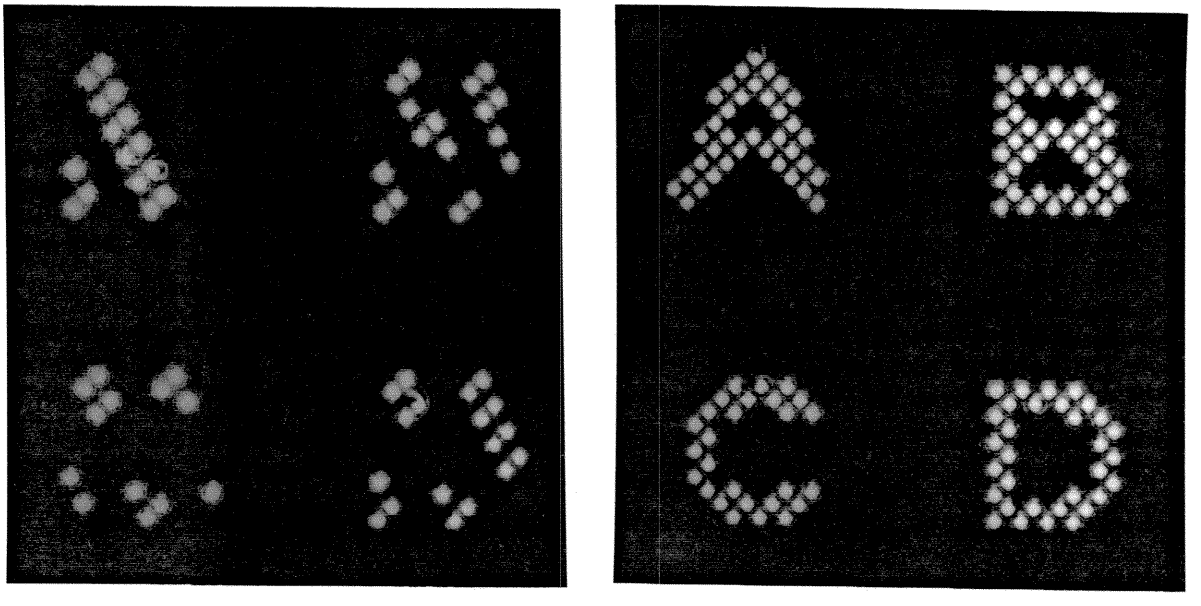
5.5.3 Experimental Demonstrations

The experiment implementing local connectivity is conducted with the fractal sampling grids and the sampled patterns shown in Fig.5.5.5. The fractal sampling grids have $n_1 = 24$, $n_2 = 72$, $m = 12$, $N = 36$, $N_1 = 216$, $N_2 = 648$, $D_1 \approx 1.5$ and $D_2 = 1.8$. The parameters of the optical system are given in Table 5.3.

During the training process, the crystal records gratings formed by overlapping plane waves. The pattern with lower density is placed at the input plane and the pattern with higher density is placed at the training plane. According

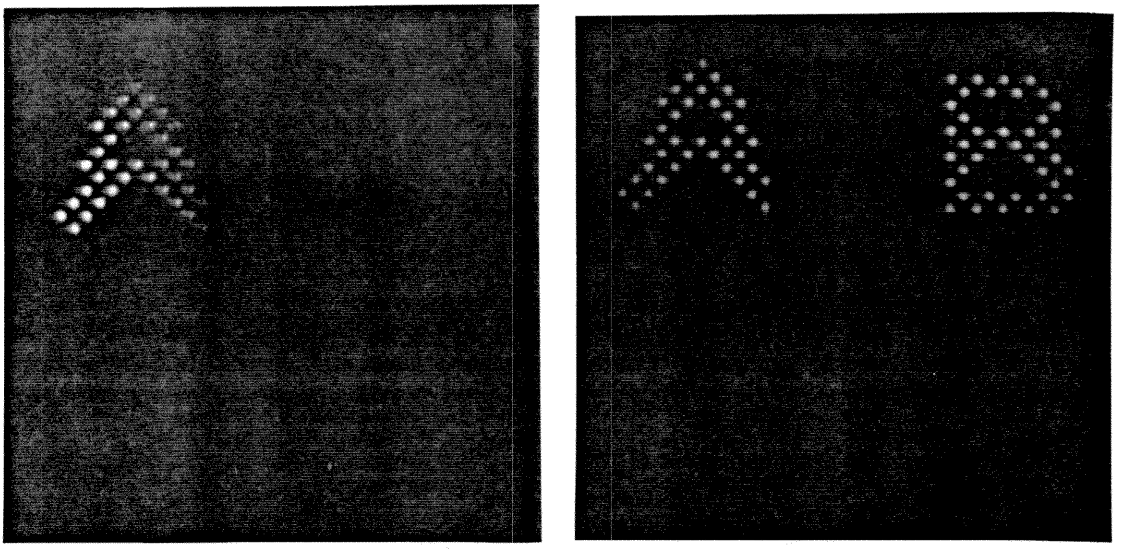


(a)

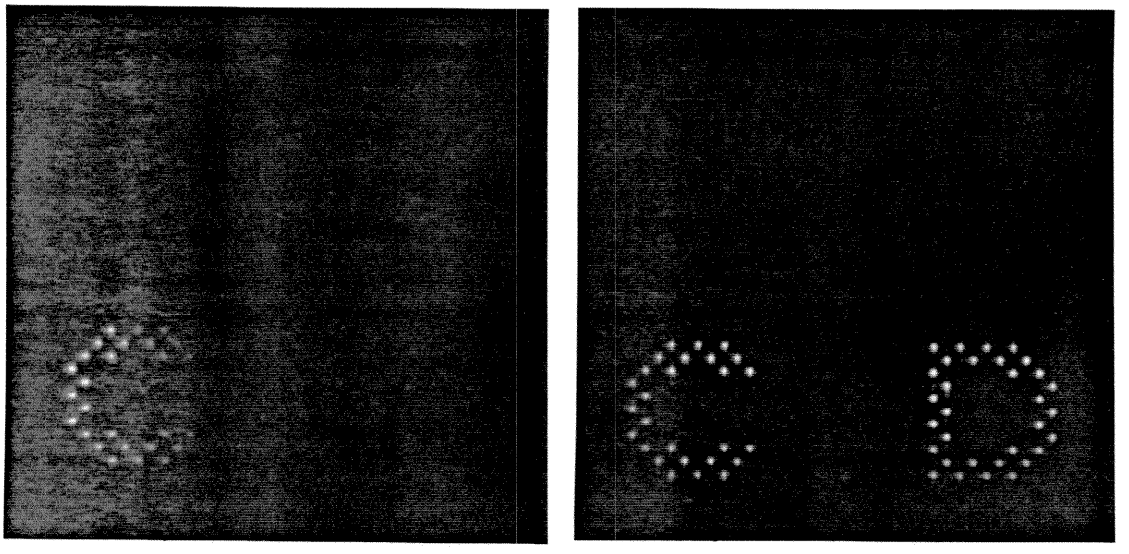


(b)

Fig.5.5.5 (a) The fractal sampling grids. (b) The sampled patterns.



(a) (b)



(c) (d)

Fig.5.5.6 (a) The output letter A read out by the input letter A. (b) The output letters A and B read out by the input letter B. (c) The output letter C read out by the input letter C. (d) The output letters C and D read out by the input letter D.

to the parameters chosen above, letter A at the input plane will be connected to letter A at the output plane only, and similarly for letter C. Due to the thickness of the crystal, the input pattern B (D) crosses both the output letters A and B (C and D) when the beams go through the crystal, as shown in Fig.5.5.3. Therefore, letter B (D) at the input plane will be connected to both letters A and B (C and D) at the output plane.

During the recall process, the training plane is blocked and different input patterns read out different output patterns. The experimental results are shown in Fig.5.5.6. When the letter A (C) is placed at the input plane, the output contains only letter A (C). When the letter B (D) is placed at the input plane, both letters A and B (C and D) are read out at the output plane.

The same system can be used to locally connect the input letter B (D) to the output letter B (D) only. This can be accomplished by using either multiple exposures or a thinner crystal. The method of multiple exposures involves two exposures, the first with the input letters A and C and the training letters A and C, the second with the input letters B and D and the training letters B and D. The method of using a thinner crystal leaves the overlapping between the input letter B (D) and the output letter A (C) outside the crystal, as shown in Fig.5.5.2(b). Therefore, each input neuron will be connected to the same number of output neurons.

Conclusion

The fractal sampling grids used in the optical holographic implementation of neural networks can prevent the degenerate interconnections. Both global

and local connectivities can be implemented with the help of the corresponding fractal sampling grids. The fractal dimensions of the sampling grids for the local connectivity are higher than those for the global connectivity.

6. CONCLUSION

In this chapter, the use of volume hologram and planar hologram will be compared to show that volume hologram provides higher storage density for the implementation of neural networks.

6.1 VOLUME OF THE SYSTEM

In order to compare the storage density of the optical system implemented with a volume hologram with that of a planar hologram system, it is necessary to calculate the volume of the optical system in terms of the number of interconnections. The optical system is shown in Fig.6.1.1, where the hologram can be either volume or planar. It has been discussed in Chapter 4 and the Appendix that for the same system, the number of interconnections is N^3 when using volume hologram and N^2 when using planar hologram. The volume of systems for the same number of interconnections can be calculated under certain conditions.

It is essential to keep the same accessible grating space, the separation between two adjacent pixels, the angle between two optical axes and the wave length so that the change of number of interconnections is only due to the change of system volume. Consider Eq.(4.10) and Eq.(4.12),

$$V_a = 2k_0^3(\frac{a}{f})^3\frac{n_e}{n_o^2}\frac{\sin(2\theta)\cos\theta}{\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}, \tag{4.10}$$

$$N_a = 2\frac{V_{xtal}}{\lambda^3}(\frac{a}{f})^3\frac{n_e}{n_o^2}\frac{\sin(2\theta)\cos\theta}{\sqrt{1-\frac{\sin^2\theta}{n_o^2}}}. \tag{4.12}$$

Eq.(4.12) indicates that the number of interconnections, which is proportional to N_a, depends on the numerical aperture of the system, a/f, the angle between the

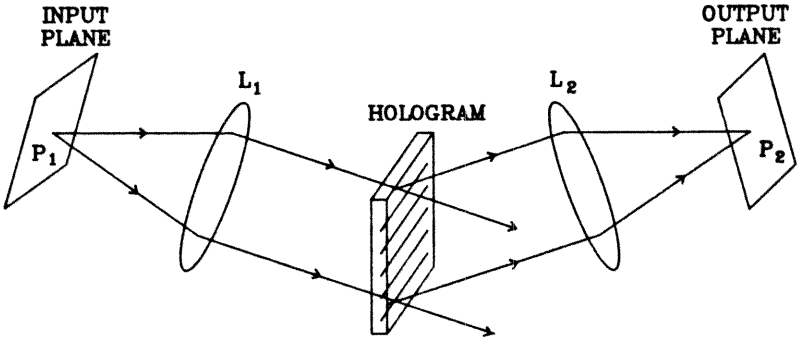


Fig.6.1.1 The optical system for the holographic implementation of neural networks.

two optical axes, θ, and the volume of the crystal, V_{xtal}. Eq.(4.10) indicates that any change of the numerical aperture or the angle between the two optical axes will change the volume of the accessible grating space V_a. To exclude the change of the number of interconnections due to the change of the accessible grating space so that the change of the number of interconnections is related only to the change of the system volume, the angle between the two optical axes θ and the numerical aperture of the system a/f are kept as constants. Write the numerical aperture as

$$\frac{a}{f} = \alpha, \tag{6.1}$$

where α is a constant, a is the linear dimension of the input (output) plane, f is the focal length of lenses L_1 and L_2. Eq.(4.12) also shows that for the fixed accessible grating space, the number of interconnections depends only on

the volume of the crystal which is in turn related to the uncertainty volume of a grating. When the dimensions of the crystal are changed, there are two possibilities to change the number of interconnections. First is to change the separation between two adjacent pixels so that there will be more (or less) neurons within the input (output) plane. Second is to change the focal length and the dimension of the input (output) plane so that the resolution of the system ($\delta x'$) will not be changed. In order to relate the number of interconnections only to the volume of the system, the resolution of the system will be kept as a constant. Therefore, the second possibility of changing the number of interconnections will be taken. Consider Eq.(4.19),

$$\delta x_i' \approx \frac{2\pi}{L_x}\frac{f}{k_0}. \tag{4.19}$$

The resolution of the system, that is the separation between two adjacent pixels $\delta x_i'$, depends on the ratio f/L_x. To exclude the change of the number of interconnections due to the change of resolution of the system, this ratio is kept as a constant, i.e.,

$$\frac{L_x}{f} = \beta, \tag{6.2}$$

where β is a constant.

The volume of the system can then be expressed in terms of the number of pixels along each direction N. The transverse area of the system shown in Fig.6.1.1 is the maximum among the area of the input (output) plane, the area of the lenses and the transverse area of the crystal. According to Eq.(6.1), Eq.(6.2), Eq.(4.28) and Eq.(4.29), these areas are proportional to one another. Since the

length of the system is proportional to the focal length of the lenses, the volume of the system is proportional to the transverse area of the input (output) plane times the focal length of the lenses, i.e., $V_{system} \propto a^2 f$. Because the numerical aperture of the system is kept as a constant (Eq.(6.1)), the volume of the system is then proportional to f^3, i.e.,

$$V_{system} = \gamma f^3, \tag{6.3}$$

where γ is a constant. According to Eq.(4.21), Eq.(6.1) and Eq.(6.2), the focal length f is related to the number of pixels along each direction at the input (output) plane, N, by

$$\begin{aligned} N =& \frac{a}{f}\frac{L_x}{\lambda}, \\ =& \frac{\alpha\beta}{\lambda} f. \end{aligned} \tag{6.4}$$

Therefore,

$$V_{system} = \gamma(\frac{\lambda}{\alpha\beta})^3 N^3, \tag{6.5}$$

i.e., the volume of the optical system is proportional to N^3. This result is the same for the volume hologram system and for the planar hologram system, since the above calculations do not involve the geometry of the storage medium.

The volume of system expressed in terms of the number of interconnections instead of the number of pixels will have different forms for the volume hologram system and the planar hologram system. Write the number of interconnections stored in the volume (planar) hologram as $N_c^{(3-D)}$ ($N_c^{(2-D)}$). According to the

discussions in Chapter 4 and the Appendix,

$$N_c^{(3-D)} = N^3, \tag{6.6}$$

and

$$N_c^{(2-D)} = N^2. \tag{6.7}$$

Substitute Eq.(6.6) and Eq.(6.7) into Eq.(6.5) respectively. The corresponding volume of system can be expressed as

$$V_{system}^{(3-D)} = \gamma(\frac{\lambda}{\alpha\beta})^3 N_c^{(3-D)}, \tag{6.8}$$

and

$$V_{system}^{(2-D)} = \gamma(\frac{\lambda}{\alpha\beta})^3 (N_c^{(2-D)})^{3/2}. \tag{6.9}$$

To store the same number of interconnections, i.e.,

$$N_c^{(3-D)} = N_c^{(2-D)} = N_c, \tag{6.10}$$

the ratio of the volume of the two systems ought to be

$$\frac{V_{system}^{(3-D)}}{V_{system}^{(2-D)}} = \frac{1}{\sqrt{N_c}}. \tag{6.11}$$

Therefore, the ratio of storage density, which is defined as $\rho = N_c/V_{system}$, is

$$\frac{\rho^{(3-D)}}{\rho^{(2-D)}} = \sqrt{N_c}. \tag{6.12}$$

For example, to implement 10^8 interconnections, the required volume of the planar hologram system is $10,000$ times that of the volume hologram system.

6.2 Comparison between Planar and Volume Holograms

Table 6.1 summarizes the comparison between the planar hologram system and the volume hologram system. The number of neurons at the input and the output planes are taken to be the same, i.e.,

$$N_n = N_1 = N_2 = \sqrt{N_c}. \tag{6.13}$$

All constant factors are neglected in order to show the change with respect to the number of neurons.

Table 6.1 Planar vs. Volume Holograms		
N_n = Number of Neurons		
2-D for Planar Hologram		
3-D for Volume Hologram		
	2-D	3-D
Linear Dimension	N_n	$N_n^{2/3}$
Area	N_n^2	$N_n^{4/3}$
Total System Volume	N_n^3	N_n^2
System Volume Ratio	$V^{(2-D)}/V^{(3-D)} = N_n$	
Storage Density Ratio	$\rho^{(2-D)}/\rho^{(3-D)} = 1/N_n$	

6.3 Conclusion

This thesis has theoretically analyzed and experimentally demonstrated the optical holographic implementations of independent interconnections. The **K**-space analysis gives the limit of the holographic storage capacity based upon

geometric considerations. By using the fractal sampling grids, the degenerate interconnections associated with Fourier holography are avoided, and the degrees of freedom of the hologram are fully utilized. The holographic hetero-associative memory is a direct implementation of the outer product scheme. The storage density of volume holograms is higher than the storage density of planar holograms.

There are several other factors, beyond the basic geometric constraints discussed in this thesis. In order to gain a complete understanding of the capabilities of volume holograms for implementing neural network interconnections, the physical limitations of the photorefractive crystal and the recording mechanism must be taken into consideration [16]. The effects of second and third order diffraction of gratings should also be addressed [48]. To effectively record a given set of gratings into the photorefractive crystal, the least number of exposures are desired [49]. These issues have been addressed by my colleagues, and together with the results of this thesis, they form the foundation for the practical realization of optical neural networks.

APPENDIX PLANAR HOLOGRAMS

The optical implementation of neural networks can also be accomplished by using planar holograms rather than volume holograms. Similar to the volume holographic implementation, the planar holographic implementation involves the training process and the recall process. The training process is implemented by the recording of the interference pattern of the reference beam and the object beam, and the recall process is implemented by the reconstruction of the object beam. In this appendix, the implementation of independent interconnections using planar holographic sinusoidal gratings will be analyzed.

Section A.1 discusses the shift invariant property of planar holograms. The diffraction condition of planar holographic gratings is different from that of volume holographic gratings. The Bragg condition, which must be satisfied during the diffraction of volume holographic gratings, is no longer a constrain during the diffraction of planar holographic gratings.

Section A.2 will extend the **K**-space analysis to the discussion of planar holographic gratings. The shift invariant diffraction of a planar holographic grating is related to the lack of information in the third dimension, as mentioned in Section 2.2. The difference in dimensionalities of the storage media will cause the difference in their storage capacities.

Section A.3 designs different fractal sampling grids for the planar holographic implementation of independent interconnections. The fractal dimensions of the sampling grids for planar holograms are less than those for volume holograms, due to the reduction of dimensions of the storage medium.

A.1 Shift Invariance

The shift invariant diffraction of a planar holographic grating can be understood by considering an example of an amplitude grating. Suppose a grating is recorded in a 2-dimensional medium as a transmittance function,

$$t(x_0, y_0) = Q[1 + \cos(K_x x_0 + K_y y_0)]. \tag{A.1}$$

Where Q is a constant, $\mathbf{K} = (K_x, K_y)$ is the 2-dimensional grating vector. A plane wave is incident upon the recorded hologram, which is located at the $z = 0$ plane. The field amplitude of the plane wave can be written as

$$U_0(x_0, y_0) = \exp[i(k_{ix} x_0 + k_{iy} y_0)]. \tag{A.2}$$

At a distance z after the hologram, the electric field amplitude $U_z(x, y)$ is given by [47]

$$\begin{aligned} U_z(x,y) = & \frac{\exp(ik_0 z)}{i\lambda z} \exp[i\frac{k_0}{2z}(x^2 + y^2)] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_0(x_0, y_0) t(x_0, y_0) \\ & \exp[i\frac{k_0}{2z}(x_0^2 + y_0^2)] \exp[-i\frac{2\pi}{\lambda z}(x x_0 + y y_0)] dx_0 dy_0. \end{aligned} \tag{A.3}$$

Where λ is the wavelength of light and $k_0 = 2\pi/\lambda$. Substitute Eq.(A.1) and

Eq.(A.2) into Eq.(A.3). The field amplitude $U_z(x,y)$ can be written as

$$\begin{aligned} U_z(x,y) =& Q\exp(ik_0z) \\ & \{\exp(-i\frac{k_{ix}^2+k_{iy}^2}{2k_0}z)\times\exp[i(k_{ix}x+k_{iy}y)] \\ & +\exp[-i\frac{(k_{ix}+K_x)^2+(k_{iy}+K_y)^2}{2k_0}z]\times \\ & \exp[i(k_{ix}+K_x)x+i(k_{iy}+K_y)y] \\ & +\exp[-i\frac{(k_{ix}-K_x)^2+(k_{iy}-K_y)^2}{2k_0}z]\times \\ & \exp[i(k_{ix}-K_x)x+i(k_{iy}-K_y)y]\}. \end{aligned} \tag{A.4}$$

Eq.(A.4) indicates that after the planar hologram, the optical field consists of three plane waves, i.e., the 0th, +1st and −1st order plane waves. The 0th order plane wave is the undiffracted plane wave with its wave vector $\mathbf{k}_i$. The ±1st order plane waves have their wave vectors

$$\begin{aligned} k_{dx(\pm1)} &= k_{ix}\pm K_x, \\ k_{dy(\pm1)} &= k_{iy}\pm K_y, \\ k_{dz(\pm1)} &= \sqrt{k_0^2-k_{dx(\pm1)}^2-k_{dy(\pm1)}^2}, \\ &\approx k_0-\frac{(k_{ix}\pm K_x)^2+(k_{iy}\pm K_y)^2}{2k_0}. \end{aligned} \tag{A.5}$$

It can be seen from Eq.(A.4) that a plane wave is not necessarily incident at the Bragg angle in order to be diffracted by the existing grating. To see the change of the diffracted plane wave with respect to the change of the incident plane wave, consider the first order diffraction. Suppose the incident plane wave $\exp(i\mathbf{k}_i\cdot\mathbf{r})$ gives rise to a diffracted plane wave $\exp(i\mathbf{k}_d\cdot\mathbf{r})$. By changing the incident angle, the new incident plane wave can be expressed as

$\exp\{i[(k_{ix}+\Delta k_{ix})x+(k_{iy}+\Delta k_{iy})y]\}$, where the z component of the wave vector is neglected for simplicity. According to Eq.(A.4), the new diffracted plane wave is $\exp\{i[(k_{dx}+\Delta k_{ix})x+(k_{dy}+\Delta k_{iy})y]\}$, i.e., the increments $\Delta k_{dx}=\Delta k_{ix}$ and $\Delta k_{dy}=\Delta k_{iy}$.

When the Vander Lugt system is used to implement neural network interconnections, the planar holographic diffraction makes the system shift invariant. Suppose the two input plane waves correspond to the two input points (x_i', y_i') and $(x_i'+\Delta x_i', y_i'+\Delta y_i')$, and the two output plane waves correspond to the two output points (x_d', y_d') and $(x_d'+\Delta x_d', y_d'+\Delta y_d')$. Since $\Delta k_{dx}=\Delta k_{ix}$ and $\Delta k_{dy}=\Delta k_{iy}$, it can be seen from Eq.(4.1) and Eq.(4.2) that the displacement of the output point, $(\Delta x_d', \Delta y_d')$, is the same as that of the input point, $(\Delta x_i', \Delta y_i')$, i.e., $\Delta x_d'=\Delta x_i'$ and $\Delta y_d'=\Delta y_i'$. If a pattern consisting of many points is reconstructed from the planar hologram by using a point at the input plane, the reconstruction by using a shifted point at the input plane will be the same pattern shifted at the output plane, since all points are shifted by the same displacement.

A.2 Extension of K-space Analysis to Planar Holograms

The **K**-space analysis discussed in Chapter 3 can be extended to the analysis of planar holograms. Fig.A.2.1 shows the **K**-space geometry when the recording medium becomes very thin. The uncertainty value $\delta K_z = 2\pi/L_z$ increases as the thickness L_z decreases.

When the thickness L_z becomes comparable with the wavelength λ, the diffraction becomes completely shift invariant. Fig.A.2.2 shows that when the incident plane wave $\mathbf{k}_i'$ is significantly off the Bragg angle, which is determined

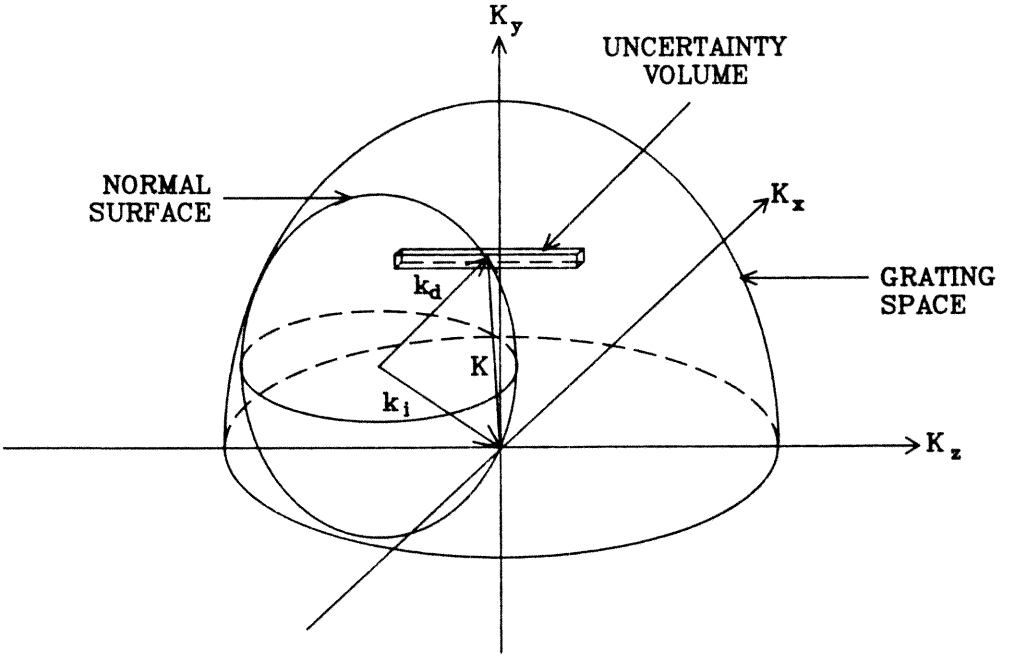


Fig.A.2.1 The **K**-space geometry for a thin recording medium.

by the nominal grating vector $\mathbf{K}_g$, the wave vector $\mathbf{k}_i'$ can still be Bragg matched by a grating vector $\mathbf{K}$ inside the uncertainty volume to produce a diffracted plane wave $\mathbf{k}_d'$. From Eq.(3.13), it can be noticed that the grating strength of $\mathbf{K}$ is close to that of $\mathbf{K}_g$ as long as $|K_z - K_{gz}| \ll 2\pi/L_z$. Therefore, shift invariant diffraction is a result of the uncertainty of the 2-dimensional recording.

The storage capacity of the 2-dimensional recording medium is proportional to A/λ^2, where A is the area of the recording medium. To derive an expression for the capacity, suppose the medium is isotropic, and the magnitude of any wave vector is k. Therefore, the bounded grating space is the top half of the sphere,

$$K_x^2 + K_y^2 + K_z^2 = (2k)^2. \tag{A.6}$$

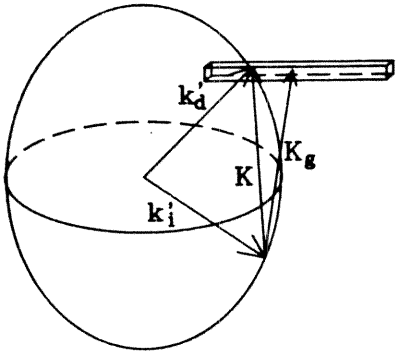


Fig.A.2.2 The shift invariant diffraction as a result of the uncertainty of the 2-dimensional recording.

For planar hologram, k is comparable with the uncertainty value $\delta K_z = 2\pi/L_z$. The dimension of the uncertainty volume along the K_z-direction is comparable with the dimension of the grating space. Different gratings correspond to non-overlapping uncertainty volumes separated in the K_x or K_y direction. The largest cross section of the grating space perpendicular to the K_x-direction is the circle in the (K_x, K_y)-plane. The area of this circle is $\pi(2k)^2$. The cross section of the uncertainty volume perpendicular to the K_z-direction is $\delta K_x \delta K_y$. Therefore, the storage capacity, which is defined as the maximum number of distinguishable gratings can be contained in the grating space, is given by

$$C = \frac{\pi(2k)^2}{\delta K_x \delta K_y}, \tag{A.7}$$

where $\delta K_x = 2\pi/L_x$, $\delta K_y = 2\pi/L_y$. Eq.(A.7) can be written as

$$C = 4\pi n^2 \frac{A}{\lambda^2}. \tag{A.8}$$

Where n is the refractive index of the material, and $A = L_x L_y$ is the area of the planar hologram.

The total number of accessible gratings N_a can be calculated using similar procedures as in Subsection 4.3.1. For the same input and output planes, the accessible grating space is the same as that shown in Fig.4.3.1. The cross section cut by the $K_z = 0$ plane is a rectangle. The area of the rectangle can be calculated from Eq.(4.8) as $A_a = (K_{xmax} - K_{xmin})(K_{ymax} - K_{ymin})$, which gives

$$\begin{aligned} A_a =& \frac{2k_0 a}{f} \times \frac{2k_0 \cos\theta a}{f}, \\ =& 4k_0^2 \cos\theta (\frac{a}{f})^2. \end{aligned} \tag{A.9}$$

The total number of accessible gratings is

$$\begin{aligned} N_a =& \frac{A_a}{(2\pi/L_x)(2\pi/L_y)}, \\ =& 4\cos\theta (\frac{a}{f})^2 \frac{A}{\lambda^2}. \end{aligned} \tag{A.10}$$

The number of pixels at the input (training) plane can be chosen to be $N \times N \propto N_a$. To determine N, consider the separation between two adjacent pixels $\delta x_i'$ and $\delta y_i'$. From Eq.(4.1), $\delta x_i'$ and $\delta y_i'$ can be related to δp_{ix} and δp_{iy} by

$$\begin{aligned} \delta p_{ix} &= \delta x_i' \frac{k_0}{f}, \\ \delta p_{iy} &= \delta y_i' \cos\theta \frac{k_0}{f}. \end{aligned} \tag{A.11}$$

According to Eq.(4.18),

$$\begin{aligned} \delta p_{ix} &= \frac{2\pi}{L_x}, \\ \delta p_{iy} &= \frac{2\pi}{L_y}. \end{aligned} \tag{A.12}$$

Substitute Eq.(A.12) into Eq.(A.11), the minimum separation between two adjacent pixels is given by

$$\begin{aligned} \delta x_i' &= \frac{2\pi}{L_x}\frac{f}{k_0}, \\ \delta y_i' &= \frac{2\pi}{L_y}\frac{f}{k_0 \cos\theta}. \end{aligned} \tag{A.13}$$

To have a regular 2-dimensional grid, select $\delta x_i' = \delta y_i'$, which gives

$$L_x = L_y \cos\theta. \tag{A.14}$$

The total number of pixels

$$\begin{aligned} N^2 =& \frac{a}{\delta x_i'} \times \frac{a}{\delta y_i'}, \\ =& \cos\theta (\frac{a}{f})^2 \frac{A}{\lambda^2}, \\ =& \frac{N_a}{4}. \end{aligned} \tag{A.15}$$

A.3 Fractal Sampling Grids for Planar Holograms

The shift invariant diffraction of planar holograms requires new fractal sampling grids for avoiding degenerate interconnections.

The fractal dimensions of the sampling grids and the allocations of the neurons are different from those discussed in Chapter 4. To reach the storage capacity, the number of neurons at the input (training) plane N_1 (N_2) is chosen such

that

$$N_1 \times N_2 = N^2. \tag{A.16}$$

Therefore, the fractal dimensions, d_1 and d_2, of the sampling grids for the input and the training planes are related by

$$d_1 + d_2 = 2. \tag{A.17}$$

It is required to avoid shift invariance in all directions in the case of 2-dimensional storage. The shift invariance condition for two neurons at the input plane, (x'_{i1}, y'_{i1}) and (x'_{i2}, y'_{i2}), and two neurons at the training plane, (x'_{d1}, y'_{d1}) and (x'_{d2}, y'_{d2}) is given by

$$x'_{i2} - x'_{i1} = x'_{d2} - x'_{d1}, \tag{A.18}$$

and

$$y'_{i2} - y'_{i1} = y'_{d2} - y'_{d1}. \tag{A.19}$$

A systematic method to design fractal sampling grids is to locate row patterns derived in Chapter 4 at proper vertical positions. The number of neurons within each row is $n_1 = \sqrt{N_1}$ at the input plane and $n_2 = \sqrt{N_2}$ at the training plane. The number of rows containing neurons is $n_1 = \sqrt{N_1}$ at the input plane and $n_2 = \sqrt{N_2}$ at the training. Since $N_1 N_2 = N^2$, n_1 and n_2 are related by $n_1 n_2 = N$.

In the following discussion on fractal sampling grids, the unit of distances is chosen as the spatial separation of two adjacent pixels. The first column to the

left of the input (training) plane will be represented by $x = 1$. The first row on the bottom of the input (training) plane will be represented by $y = 1$.

A family of fractal sampling grids can be systematically designed following the next five steps.

1) Select the n_2 neuron positions along one row at the training plane. Label the neurons as 1, 2, $\cdots$, n_2. Separate the n_2 neurons uniformly by n_1, i.e.,

$$x_k^{(d)} = (k-1) \times n_1 + x_1^{(d)}, \tag{A.20}$$

where the superscript (d) represents the training neurons, k represents the kth neuron with $1 \leq k \leq n_2$, and $x_1^{(d)}$ represents the position of the first neuron. There are n_1 different distributions in this case, for $x_1^{(d)}$ can be one of the values from 1 through n_1. These n_1 distributions are called n_1 row patterns denoted by $\{B_1, B_2, \cdots, B_{n_1}\}$.

2) Design fractal sampling grids for the training plane. Choose one of the n_1 row patterns, B_v. Duplicate this row pattern to obtain n_2 identical row patterns. Label these n_2 identical row patterns as 1, 2, $\cdots$, n_2. Separate the n_2 row patterns uniformly by n_1, i.e.,

$$y_k^{(d)} = (k-1) \times n_1 + y_1^{(d)}, \tag{A.21}$$

where k represents the kth row pattern with $1 \leq k \leq n_2$ and $y_1^{(d)}$ represents the vertical position of the first row pattern. For each chosen B_v, there are n_1 different possibilities to locate the row patterns vertically, since $y_1^{(d)}$ can be one of the values from 1 through n_1. Because B_v can be any of the n_1

row patterns derived in step 1, the total number of fractal sampling grids for the training plane is n_1^2.

3) Select the n_1 neuron positions along one row at the input plane. Label the neurons as 1, 2, $\cdots$, n_1. Suppose that there are M different row patterns. For the jth row pattern, select the position of the kth neuron to be

$$x_k^{(i)} = (l_k^j - 1) \times n_1 + k, \tag{A.22}$$

where the superscript (i) represents the input neurons, and l_k^j is an integer within the range $1 \leq l_k^j \leq n_2$. For each neuron, l_k^j can be any of the n_2 values from 1 through n_2. Since there are n_1 neurons, there exist $n_2^{n_1}$ different combinations for l_1^j, l_2^j, $\cdots$, $l_{n_1}^j$. Thus the number of different row patterns is $M = n_2^{n_1}$. Represent these M different row patterns by $\{A_1, A_2, \cdots, A_M\}$.

4) Design fractal sampling grids for the input plane. Choose one of the M row patterns, A_v. Duplicate this row pattern to obtain n_1 identical row patterns. Label these n_1 identical row patterns as 1, 2, $\cdots$, n_1. Suppose that there are M' different vertical distributions. For the jth distribution, select the position of the kth row pattern to be

$$y_k^{(i)} = (l_k^j - 1) \times n_1 + k, \tag{A.23}$$

where l_k^j is an integer within the range $1 \leq l_k^j \leq n_2$. For each row pattern, l_k^j can be any of the n_2 values from 1 through n_2. Since there are n_1 row patterns, there exist $n_2^{n_1}$ different combinations for l_1^j, l_2^j, $\cdots$, $l_{n_1}^j$. Thus the

number of different vertical distributions is $M' = M = n_2^{n_1}$. Because A_v can be any of the M row patterns derived in step 3, the total number of fractal sampling grids for the input plane is $M^2 = n_2^{2n_1}$.

5) Design pairs of fractal sampling grids for the input and training planes. A pair of fractal sampling grids can be obtained by selecting any one of the M^2 sampling grids derived in step 2 for the input plane and any one of the n_1^2 sampling grids derived in step 4 for the training plane. Therefore, the total number of pairs of fractal sampling grids is $n_1^2 n_2^{2n_1}$.

The shift invariance is avoided by using any pair of these fractal sampling grids. It has been proven in Chapter 4 that the horizontal distance between any two columns of neurons at the input plane is different from that at the training plane. Since the vertical distribution of neurons is designed in the same way as for the horizontal distribution of neurons, according to the same argument, the distance between any two rows of neurons at the input plane is different from that at the training plane. Therefore, the shift invariance condition, Eq.(A.18) and Eq.(A.19), is avoided by these fractal sampling grids.

Fig.A.3.1 shows an example of fractal sampling grids. In this example, $N = 12$, $N_1 = 9$, $N_2 = 16$, $n_1 = 3$, $n_2 = 4$, $d_1 \approx 0.88$ and $d_2 \approx 1.12$. The horizontal locations of neurons at the input plane are given by

$$\begin{aligned} x_1^{(i)} &= (3-1) \times 3 + 1 = 7, \\ x_2^{(i)} &= (2-1) \times 3 + 2 = 5, \\ x_3^{(i)} &= (2-1) \times 3 + 3 = 6. \end{aligned} \tag{A.24}$$

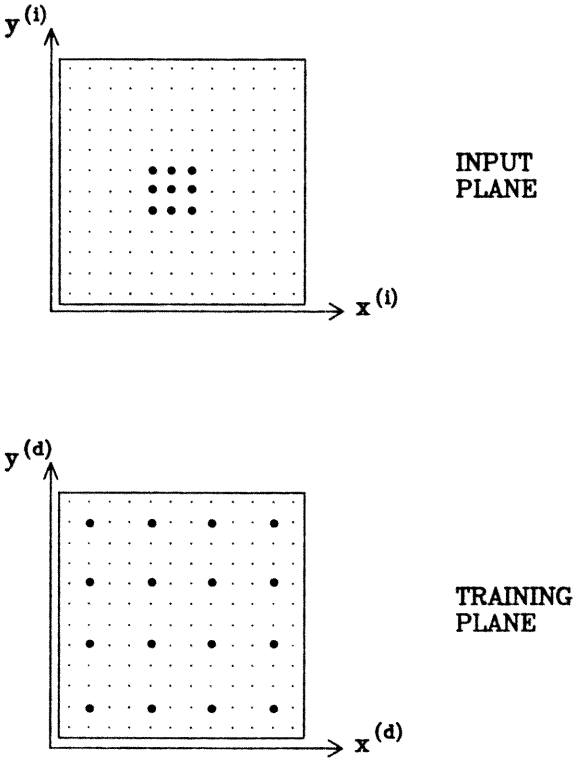


Fig.A.3.1 An example of fractal sampling grids for the planar hologram.

The horizontal locations of neurons at the training plane are given by

$$\begin{aligned} x_1^{(d)} &= 2, \\ x_2^{(d)} &= (2-1) \times 3 + 2 = 5, \\ x_3^{(d)} &= (3-1) \times 3 + 2 = 8, \\ x_4^{(d)} &= (4-1) \times 3 + 2 = 11. \end{aligned} \tag{A.25}$$

The vertical locations of neurons at the input plane are given by

$$\begin{aligned} y_1^{(i)} &= (3-1)\times 3+1 = 7, \\ y_2^{(i)} &= (2-1)\times 3+2 = 5, \\ y_3^{(i)} &= (2-1)\times 3+3 = 6. \end{aligned} \tag{A.26}$$

The vertical locations of neurons at the training plane are given by

$$\begin{aligned} y_1^{(d)} &= 2, \\ y_2^{(d)} &= (2-1)\times 3+2 = 5, \\ y_3^{(d)} &= (3-1)\times 3+2 = 8, \\ y_4^{(d)} &= (4-1)\times 3+2 = 11. \end{aligned} \tag{A.27}$$

Other fractal sampling grids, such as higher order fractal sampling grids, can be generated using the same strategy.

REFERENCES

1. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, (1973).

2. D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, (1986).

3. Course notes, *Pattern Recognition*, CNS/EE 124.

4. J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proc. Natl. Acad. Sci. USA*, **79**, 2554, (1982).

5. J. J. Hopfield, "Neurons with Graded Response Have Collective Computational Properties Like Those of Two-state Neurons," *Proc. Natl. Acad. Sci. USA*, **81**, 3088, (1984).

6. N. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical Implementation of the Hopfield Model," *Appl. Opt.*, **24(10)**, 1469, (1985).

7. S. W. Kuffler and J. G. Nicholls, *From Neuron to Brain*, Sinauer Associates, Massachusetts, (1976).

8. D. Psaltis and N. Farhat, "Optical Information Processing Based on an Associative-memory Model of Neural Nets with Thresholding and Feedback," *Opt. Lett.*, **10(2)**, 98, (1985).

9. D. Psaltis and J. Hong, "Shift-Invariant Optical Associative Memories," *Optical Engineering*, **26(1)**, 10, (1987).

10. K. Wagner and D. Psaltis, "Multilayer Optical Learning Networks," *Appl. Opt.*, **26(23)**, 5061, (1987).

11. P. Batacan, "Can Physics Make Optics Compute," *Computers in Physics*, **2(2)**, 9, (1988).

12. D. Psaltis, D. Brady, X. Gu and K. Hsu, "Optical Implementation of Neural Computers," *Optical Processing and Computing*, H. Arsenault, ed., Academic Press, Inc., (New York) 1989.

13. D. Psaltis, X. Gu and D. Brady, "Holographic Implementations of Neural Networks," *An Introduction to Neural and Electronic Networks*, S. F. Zornetzer, J. L. Davis and Clifford Lau eds., Academic Press, Inc., (New York) 1989.

14. D. Psaltis, A. Yamamura, K. Hsu, S. Lin, X. Gu, and D. Brady, "Optoelectronic Implementation of Neural Networks," to appear in *IEEE Communications Magazine* special issue on neural networks.

15. Y. S. Abu-Mostafa and D. Psaltis, "Optical Neural Computers," *Scientific American*, **256(3)**, 88, (1987).

16. D. Psaltis, D. Brady, and Kelvin Wagner, "Adaptive Optical Networks Using Photorefractive Crystals," *Appl. Opt.*, **27**. 1752, (1988).

17. D. Brady, X. Gu, and D. Psaltis, "Photorefractive Crystals in Optical Neural Computers," *SPIE Proceedings*, **882–20**, (1988).

18. D. Psaltis, J. Yu, X. Gu and H. Lee, "Optical Neural Nets Implemented with Volume Holograms," Topical Meeting on Optical Computing,

Technical Digest Series, Optical Society of America, (Washington, D.C.), **11**, 129, (1987).

19. X. Gu and D. Psaltis, "Local and Asymmetric Interconnections Using Volume Holograms," OSA Annual Meeting, 1988, *Technical Digest Series*, Optical Society of America, (Washington, D.C.), **11**, 148, (1988)

20. W. J. Burke and Ping Sheng, "Crosstalk Noise from Multiple Thick-Phase Holograms," *J. Appl. Phys.*, **48(2)**, (1977).

21. D. Psaltis, X. Gu and D. Brady, "Fractal Sampling Grids for Holographic Interconnections," *SPIE Proceedings*, **963–70**, (1988).

22. F. S. Chen, J. T. LaMacchia and D. B. Fraser, "Holographic Storage in Lithium Niobate," *Appl. Phys. Lett.*, **13(7)**, 223, (1968).

23. J. J. Amodei and D. L. Staebler, "Holographic Recording in Lithium Niobate," *RCA Review*, **33**, 71, (1972).

24. D. L. Staebler and W. Phillips, "Fe-Doped $LiNbO_3$ for Read-Write Applications," *Appl. Opt.*, **13(4)**, 788, (1974).

25. A. M. Glass, D. von der Linde, and T. J. Negran, "High-voltage Bulk Photovoltaic Effect and the Photorefractive Process in $LiNbO_3$," *Appl. Phys. Lett.*, **25(4)**, 233, (1974).

26. D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple Storage and Erasure of Fixed Holograms in Fe-Doped $LiNbO_3$," *Appl. Phys. Lett.*, **26(4)**, 182, (1975).

27. V. M. Fradkin and R. M. Magomadov, "Anomalous Photovoltaic Effect in

$LiNbO_3$:Fe in Polarized Light," *JETP Lett.*, **30(11)**, 686, (1979).

28. S. F. Su and T. K. Gaylord, "Unified Approach to the Formation of Phase Holograms in Ferroelectric Crystals," *J. Appl. Phys.*, **46(12)**, 5208, (1975).

29. M. G. Moharam, T. K. Gaylord, R. Magnusson and L. Young, "Holographic Grating Formation in Photorefractive Crystals with Arbitrary Electron Transport Lengths," *J. Appl. Phys.*, **50(9)**, (1979).

30. N. V. Kukhtarev, V. B. Markov, S. G. Odulov, M. S. Soskin and V. L. Vinetskii, "Holographic Storage in Electro-optic Crystals. I. Steady State," *Ferroelectrics*, **22**, 949, (1979).

31. W. Kraut and R. Baltz, "Anomalous Bulk Photovoltaic Effect in Ferroelectrics: A Quadratic Response Theory," *Phys. Rev. B*, **19(3)**, 1548, (1979).

32. R. Baltz and W. Kraut, "Theory of the Bulk Photovoltaic Effect in pure Crystals," *Phys. Rev. B*, **23(10)**, 5590, (1981).

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

34. N. Niizeki, et al., "Growth Ridges, Etched Hillocks and Crystal Structure of Lithium Niobate," *Japan. J. Appl. Phys.*, **6(3)**, 318, (1967).

35. G. D. Boyd, W. L. Bond and H. L. Carter, "Refractive Index as a Function of Temperature in $LiNbO_3$," *J. Appl. Phys.*, **38(4)**, 1941, (1967).

36. P. Günter, "Holography, Coherent Light Amplification and Optical Phase Conjugation with Photorefractive Materials," *Phys. Rep.*, **93(4)**, 199,

(1982).

37. T. J. Hall, R. Jaura, L. M. Connors and P. D. Foote, "The Photorefractive Effect—A Review," *Prog. Quant. Electr.*, **10**, 77, (1985).

38. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," *Bell Syst. Tech. J.*, **48(9)**, 2909, (1969).

39. L. Solymar, "A General Two-Dimensional Theory for Volume Holograms," *Appl. Phys. Lett.*, **31(12)**, 820, (1977).

40. D. J. Cooke, L. Solymar and C. J. R. Sheppard, "A Three-Dimensional Vector Theory for Volume Holograms," *INT. J. Electronics*, **46(3)**, 337, (1979).

41. W. E. Parry, D. J. Cooke and L. Solymar, "Solutions of the Vector Differential Equations of Volume Holography," *INT. J. Electronics*, **46(4)**, 357, (1979).

42. T. K. Gaylord, "Analysis and Applications of Optical Diffraction by Gratings," *Proceedings of the IEEE*, **73(5)**, 894, (1985).

43. F. Vachss and L. Hesselink, " Holographic Beam Coupling in Anisotropic Photorefractive Media," *J. Opt. Soc. Am. A*, **4(2)**, 325, (1987).

44. I. McMichael and P. Yeh, "Phase Shift of Photorefractive Gratings and Phase-conjugate Waves," *Opt. Lett.*, **12(1)**, 48, (1986).

45. B. B. Mandelbrot, *Fractals: Form, Chance and Dimension*, Freeman, San Francisco, (1977).

46. B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York,

(1982).

47. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, San Francisco, (1968).

48. H. Lee, X. Gu and D. Psaltis, "Volume Holographic Interconnections with Maximal Capacity and Minimal Cross Talk," *J. Appl. Phys.*, **65(6)**, 2191, (1989).

49. S. Hudson, D. Brady and D. Psaltis, "Properties of 3-D Imaging Systems," *Technical Digest Series*, **11**, Optical Society of America, Washington, DC, 74, (1988).