Université de Neuchâtel
Institut de Microtechnique

Experimental investigation of optical neural networks and learning systems

Thèse

Présentée à la Faculté des sciences
pour obtenir le grade de docteur ès sciences

par

Kenneth J. Weible

Neuchâtel, Novembre 1993

IMPRIMATUR POUR LA THÈSE

Experimental investigation of optical neural networks and learning systems

de Monsieur Kenneth J. Weible

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

Le Faculté des sciences de l'Université de Neuchâtel
sur le rapport des membres du jury,

Messieurs R. Dändliker, F. Pellandini,

X. Arreguit (CSEM Neuchâtel) et A.C. Walker

(Edinburgh)

autorise l'impression de le présente thèse.

Neuchâtel, le 7 décembre 1993

Le doyen:

H.-H. Nägeli

Table of Contents

I. Introduction

The domain of optical information processing has evolved over the past thirty years from its introduction in the area of synthetic aperture radar in the early 1960's to the recent attempts of realizing an optical digital computer. Between these two extremes, the techniques and characteristics observed and exploited in the field of optical sciences have contributed in a large spectrum of applications. The earlier applications such as pattern recognition using matched filtering, treatment of electrical signals with acousto-optic correlators, matrix-vector multipliers, and matrix-matrix multipliers, have led to the more recent applications of optical neural networks and optical learning systems.

Biological neural systems are characterized by a very large number of neurons which are massively interconnected. The response time of the individual neurons is slow, on the order of tens of milliseconds. Yet the massive parallelism of the neural system allows for enormous computational capacity. Each neuron is interconnected with thousands of other neurons and is continuously updating its level of activation. The neurons communicate by sending activation or inhibition through the neural connections. The knowledge of the neural system is contained in the weighting of the interconnections rather than in the neurons themselves. Therefore, the system is able to adapt to varying input stimuli by modifying the weights of the neural interconnections. The neural system's learning capability allows it to optimize its performance.

Optics is an attractive technology for the implementation of neural network systems, because it exploits the third dimension to interconnect planes of neurons, such that the parallel operation of the network is maintained. There are two main technologies that can be used to realize the adaptation of the interconnection weights for learning in optical neural systems, photorefractive crystals and liquid crystal devices. The response time of the nematic liquid crystal devices is on the order of tens of milliseconds, which is quite slow but corresponds to the response observed in biological neurons. The emphasis of the optical implementation is focused on mass parallelism and not on the speed of the system.

While the optical devices available today for modulating and thresholding the optically encoded information are not ideal, they allow the preliminary realization of prototype systems. The experimental realization and evaluation of these prototype systems provides valuable feedback on the problems and the promises of optically implementing neural networks and other learning systems. The experimental results allow the analysis of the benefits and the detriments of various devices and architectures, as well as different neural models and learning algorithms. Also, the prototype systems allow a real appreciation of the constraints associated with the system alignment and system size. The device related limitations imposed on the system provide guidelines for the development of future devices, while shedding some light on the realizability of large parallelism in optical information processing systems.

The work presented here is based on the experimental investigation of three different types of optical learning systems. Each of the three systems employs liquid crystal devices for either intensity or phase modulation. Diffractive phase gratings are another similar feature to each of the systems. The last of the three systems uses an optical parallel processor to optimize the performance of binary phase gratings, while the first two systems use binary phase gratings to replicate a coherent input source or image. In each of the chapters a different type of neural model or learning algorithm is implemented. Chapter 2 presents the realization of an associative memory with a fully-interconnected single layer neural network. An adapted version of the Hopfield model is used to calculate the desired interconnection weights for a fixed interconnection system and a Hebbian type of learning algorithm is used to actively adapt the interconnection weights for the neural system with learning. Chapter 3 discusses the results of the optical implementation of a self-organizing neural network using the Kohonen model. The system is used to optimize the well-known Traveling Salesman Problem. Chapter 4 deals with the optical optimization of binary phase gratings using an optical parallel processor to implement the simulated annealing of the grating design. Finally, Chapter 5 discusses some global conclusions concerning the experimental realization of optical neural networks and learning systems.

II. Neural network auto-associative memory

The Hopfield Model

1. Single layer neural networks - The Hopfield model

In a single layer network all neurons act as both input and output. Each neuron is defined as having two levels of excitation, the "off" state $s_i = 0$ and the "on" state $s_i = 1$.[2.1] The single layer of neurons is fully interconnected, i.e. each neuron is interconnected with all of the other neurons of the network. The interconnection between neuron i and neuron j has a weight of T_{ij} which can be positive, negative, or zero. The state of each neuron is determined by the nonlinear response function of the neuron which thresholds the summation of inputs to the neuron, i.e.

$$s_i = f\left[\sum_{j=1}^{N} T_{ij}s_j\right], \tag{2.1}$$

where N is the total number of neurons and $f[x]$ is the nonlinear neural response function defined as

$$f[x] \equiv \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}, \tag{2.2}$$

and illustrated in Fig. 2.1. The neural response function is a simplistic approximation to the sigmoid function observed in biological neurons.[2.2]

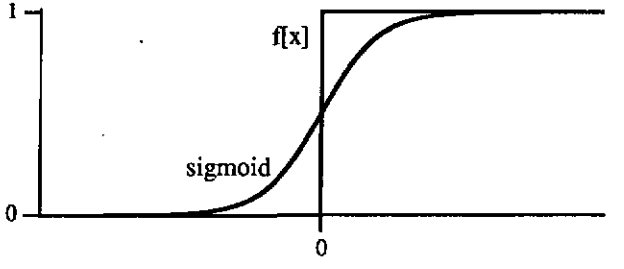


Fig. 2.1 The nonlinear threshold function f[x].

The state of each neuron depends on the states of all the other neurons of the network. The fully interconnected system is described by its dynamic evolution towards a steady-state configuration of the neural excitation states.

The feedback of the neuron states to the other neurons is the impetus for the evolution of the network. The evolution of the single layer network towards a steady-state configuration can be exploited to serve as an associative memory.

The Hopfield model is applicable to the associative memory task for binary images made up of separable pixel elements.[2.2] The weights of the interconnections between neurons is calculated for M memory patterns according to

$$T_{ij} = \begin{cases} \sum_{m=1}^{M} (2s_i^{(m)} - 1)(2s_j^{(m)} - 1) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \qquad (2.3)$$

where $s_i^{(m)}$ is the ith neuron of the mth stable configuration. The state of each neuron is randomly and asynchronously evaluated and updated according to Eq. (2.1). Updating the state of all of the neurons in parallel can lead to oscillatory behavior where one configuration systematically produces another and the network never reaches a stable state condition. Asynchronous evaluation of the neurons helps to eliminate the unstable oscillations associated with synchronous updating of the neurons.

The Hopfield model provides a means to calculate *a priori* the interconnection weights of the network. However, the Hopfield network has some major limitations. The storage capacity of the network is severely limited. The number of patterns that can be stored and successfully recalled is about *0.15 N* before error in recall becomes severe.[2.2] If too many patterns are stored the net may converge to a spurious configuration which is not one of the memory patterns. A further requirement is placed on the stored memory patterns. To avoid an unstable memory pattern, the stored patterns must be separated by a large Hamming distance, i.e. they must not share too many common pixels.[2.3] The Hopfield network is best applied using orthogonal memory patterns.

The implementation of a fully interconnected neural network requires enormous parallelism. Memory patterns of N neurons require an interconnection matrix of $N \times N$ individual weights. In many implementations optical techniques have been employed to provide parallel, crosstalk free interconnections. In the original optical implementation of the Hopfield model it was necessary to use electronic thresholding in order to subtract the bipolar results before feedback.[2.4] Later implementations have manipulated the Hopfield equations in order to remove the necessity of bipolar weighting on

the neural interconnections.[2.5,2.6] However, as a result of these modifications the nonlinear threshold for each neuron became dependent upon the number of neural interconnections.

In another approach, the interconnection weights of the Hopfield model are adapted to be unipolar, thus facilitating its optical implementation.[2.7] The variable threshold which results is compensated through the normalized weighting of each neuron's interconnections. In this approach only the inhibitory interconnections of the Hopfield model are used. All constructive (positive) weights are set to zero and the modified weights T'_{ij} are represented by

$$T'_{ij} = \begin{cases} T_{ij} & \text{if } T_{ij} < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.4}$$

where T_{ij} is taken from Eq. (2.3). A uniform threshold value is maintained for all neurons by normalizing the resulting inhibitory interconnections so that the sum for one neuron is always equal to 1. The resulting normalized inhibitory interconnection weights T''_{ij} are given by

$$T''_{ij} = \frac{T'_{ij}}{\sum_{j=1}^{N} T'_{ij}}. \tag{2.5}$$

The resulting neural response function is also modified and is given by the NOR function

$$f'[x] \equiv \begin{cases} 1 & \text{for } x < 0.5 \\ 0 & \text{for } x \geq 0.5 \end{cases}, \tag{2.6}$$

which is illustrated in Fig. 2.2. As a result of using only the inhibitory interconnection weights the memory capacity of the network is reduced from

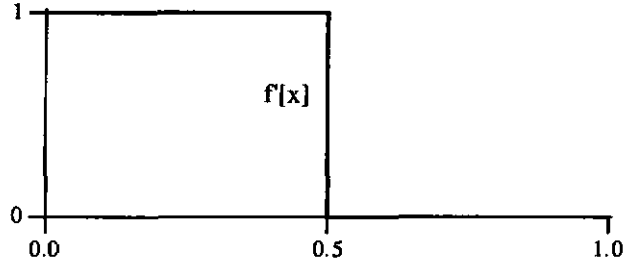


Fig. 2.2 The nonlinear response function f'[x].

0.15 N for the original Hopfield model to *0.11 N*.[2.8] Thus only a small percentage of the memory capacity of the network is sacrificed by adapting the Hopfield model to the unipolar inhibitory weighting scheme.

Further effort to adapt the neural model for optical implementation has led to the development of a new model employing only binary inhibitory interconnection weights. The weights are calculated according to the Inverted Neural Network (INN) model developed especially for optically implemented neural systems, where the weights are unipolar and binary.[2.9] This model was developed by the same research group that proposed the inhibitory modifications to the Hopfield model. The advantage of the INN model is that precise binary interconnection weights are more easily obtained than gray level weights. However, the memory capacity is slightly reduced, from *0.11 N* to *0.10 N*, as a result of the sacrificed dynamic range.[2.8]

In the following, our investigations of the performance of both the inhibitory Hopfield model and the INN model are reported. For both models the neural network is single layer and fully interconnected as in the case of the original Hopfield model. Their performance is investigated both experimentally and through simulation results. The capability of the optically realized associative memories to recall memory patterns and remove pattern noise is tested for binary discrete pixel images.

The simulation results of the neural models allow the comparison of the theoretically ideal performance of each model. However, the experimental results of the two models are greatly affected by any imperfections of the optical system used in their implementation. To minimize the influence of these imperfections the neural models must be able to naturally compensate for their presence. The system must be able to modify the weighting of the neural interconnections so as to better adapt to the actual characteristics of the optical system and its imperfections. A fully interconnected neural network with modifiable interconnection weights is demonstrated using an electrically addressable spatial light modulator. An adapted learning algorithm is employed to adjust the interconnection weights during an iterative learning phase. The learning algorithm which has been developed is similar in nature to the Hebb algorithm.[2.10] The performance of the neural model with learning ability is compared to the performance of the two models with fixed interconnection weights.

Finally, the simulated and experimental results of the inhibitory Hopfield model and the INN model are compared to those of a discrete binary

correlator, also referred to as a Hamming network.[2.3] The correlator serves as a reference by which to judge the efficiency of the neural models in recognizing the discrete images. For all results presented, the recall capability is tested only for noise removal. The special cases of translation, rotation, or scaling of the memory patterns is not considered.

2. Optical implementation

The inhibitory Hopfield model and the INN model discussed above were originally realized using holographic optical elements (HOEs) to produce the unipolar weighted interconnections between neurons.[2.7,2.9] While providing a highly efficient means of weighting the interconnections, the HOEs require considerable effort to obtain precise gray levels. Also, since the HOEs are fixed elements, any modifications to the stored memory patterns requires the creation of a new HOE.

The work presented here describes an alternative system architecture which allows the use of gray scale photographic masks to achieve the desired weighting of the interconnections between neurons. The photographic masks have precise gray levels and are fabricated by photographing a printed copy of the desired interconnection, where the weights are represented by gray level shading on the printed image. Consequently, a simple diffraction grating may be used to reproduce the input image, since the weighting is performed by the gray scale photographic mask. The memory matrices can be easily modified by replacement of the photographic mask.

The system architecture that has been developed is based on the uniform intensity distribution obtained in the diffraction orders of binary phase gratings. A pair of crossed phase gratings is used to produce the initial input neural matrix and to fan-out the matrix to the weighted interconnection matrix. The sum of each neuron's interconnections are thresholded and fedback through the system to complete the iteration. The system continues to iterate until a steady-state configuration is achieved.

Generation of input neural matrix

A schematic of the optical system used to generate the input neural matrix is illustrated in Fig. 2.3. A Helium-Neon laser (λ = 633 nm) illuminates a pair of crossed phase gratings yielding $N \times N$ diffraction orders of approximately equal intensity. The diffraction orders are collimated by a lens to produce an array of equally spaced points, Fig. 2.4a. The input neural matrix is produced

by independently rotating the polarization of the desired "on" pixels 90° by means of an electronically addressed liquid crystal display (LCD) placed in the Fourier plane of the lens. The LCD we are using is an Epson EG-Y84320AT graphic display which has been modified to operate in transmission.[2.11] A polarizer positioned after the LCD converts the polarized data into intensity, Fig. 2.4b. The separation Δx of the pixels of the input neural matrix is given by the focal length f_0 of the lens and the period d of the diffraction grating and is written as

$$\Delta x = \frac{\lambda f_0}{d}, \tag{2.7}$$

where λ is the wavelength of the laser source.

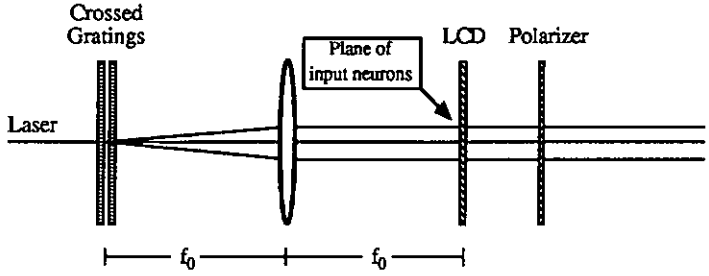


Fig. 2.3 Schematic of the optical implementation of the input neural matrix using an LCD and a polarizer.

(a) (b)

Fig. 2.4 The N x N diffraction orders of equal intensity (a) and the resulting input neural matrix after modulation (b).

Fan-out of neural matrix

The neural matrix is iterated in a series of matrix-matrix multiplications, as originally represented in Eq. (2.1). Figure 2.5 illustrates the desired fan-out of the input neural matrix to the $N^2 \times N^2$ interconnection matrix to achieve the matrix-matrix multiplication. Optically this fan-out is produced using a second pair of crossed diffraction gratings. These gratings are displaced a distance b from the focal plane of the lens following the input neurons to provide a

separation of the output neurons equal to that of the input neurons.[2.12] This condition is expressed as

$$\Delta x = \frac{f_3}{f_2}\frac{\lambda b}{D}, \tag{2.8}$$

where D is the period of this second pair of diffraction gratings. Thus, after substitution of Eq. (2.7), the displacement b of the gratings may be determined by

$$b = \frac{D}{d}\frac{f_2}{f_3} f_0 . \tag{2.9}$$

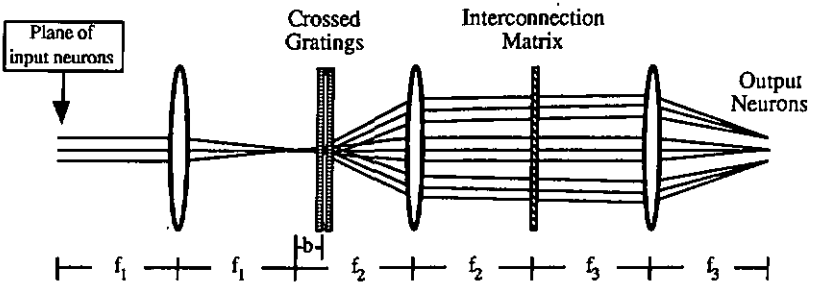


Fig. 2.5 Schematic of the fan-out of the input neural matrix to the interconnection matrix.

The separation ΔX of the diffraction orders produced by the gratings must be large enough to separate the images of the input neural matrix at the interconnection matrix, Fig. 2.6. The separation of the diffraction orders is given by

$$\Delta X = \frac{\lambda f_2}{D} . \tag{2.10}$$

The pixel separation $\Delta x'$ in the image of the input neural matrix is given by the ratio of the focal lengths of the telescopic imaging system as

$$\Delta x' = \frac{f_2}{f_1}\Delta x . \tag{2.11}$$

Consequently, to ensure that the neighboring images at the interconnection plane do not overlap, the focal lengths f_0 and f_1 must be chosen so that

$$\Delta X \geq N\,\Delta x' . \tag{2.12}$$

After combining the above equations, the focal length of the lens following the gratings f_2 falls out and thus only has an influence on the displacement of the gratings b.

Fig. 2.6 The optically generated N x N replication of the input neural matrix at the interconnection plane.

Gray scale Interconnection matrix

The stored memory matrices are encoded by the gray level weighting of the synapses within the interconnection matrix. The gray level of each neuron is based on the normalization of the total number of contributing inhibitory neurons, as described in Eq. (2.5). The fabrication of the interconnection matrix using a gray scale photographic mask is illustrated in Fig. 2.7. Once the interconnection weights have been calculated for the desired memory matrices, the fabrication of the interconnection matrix is achieved in a two step process. Initially the negative of the desired mask is designed using a computer. A printout is then imaged with the required magnification onto a millimask plate. The resulting gray level mask is used in transmission to attenuate the intensity of the impinging neurons. Alterations to the stored images are easily achieved by repeating this two step process.

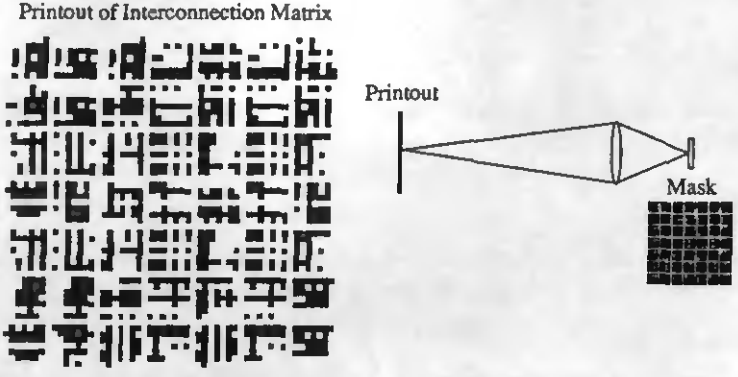


Fig. 2.7 The fabrication of the gray scale interconnection masks.

Nonlinear thresholding and feedback

In the output plane the intensity contributions from all the neurons multiplied by their synaptic weights are summed and an intensity threshold is performed using the neural response function of Eq. (2.6). The thresholded results are fedback as the input neurons to complete the cycle and the iteration continues. The thresholding and feedback can be realized either in a hybrid approach using electrical techniques or in an all-optical implementation where the nonlinear response of a liquid crystal light valve is exploited.

Hybrid implementation

The configuration of the hybrid system is illustrated in Fig. 2.8, which is essentially the combination of the previously discussed components. A detector array is used to measure the intensity of the weighted sum in the output plane. The results are thresholded and fedback to the LCD using a personal computer (PC). All neurons with an intensity above an established threshold are fedback in the "off" state and those with less intensity are fedback in the "on" state. Thus a NOR operation is realized and the iteration process continues. Best results are achieved when individual neurons are updated randomly, one after another, which simulates asynchronous operation.

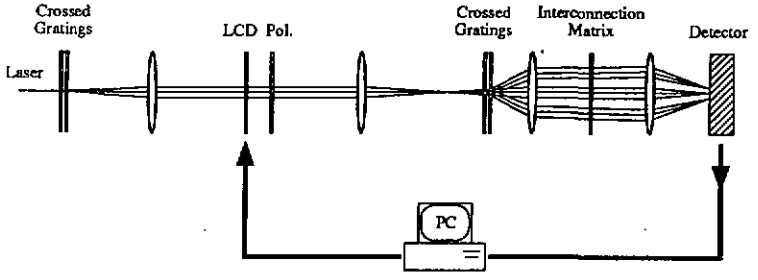


Fig. 2.8 Hybrid implementation of neural network.

Optical implementation

In the optical implementation, the detector of the hybrid approach is replaced by a liquid crystal light valve (LCLV) in the thresholding plane. The LCLV used is a prototype model which has a layer of nematic liquid crystal sandwiched between a photoconductive layer and a pair of transparent electrode layers, Fig. 2.9. A brief description of the LCLV is given here, whereas a detailed description of the device and its characteristics can be found in the literature.[2.13]

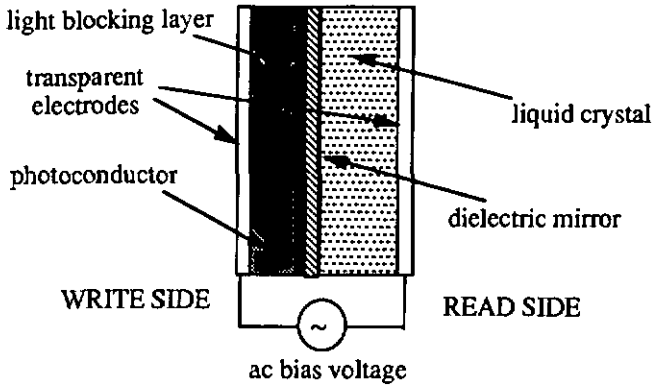


Fig. 2.9 Schematic of LCLV.

The molecules of the liquid crystal are initially aligned parallel with the surfaces of the crystal layer. When an electric field is applied, the molecules tend to align themselves with the field direction. The amount of rotation of the molecules depends on the magnitude of the electric field. The birefringence of the crystal layer is modified as the crystals are rotated and thus depends on the magnitude of the electric field. The transparent electrodes of the LCLV maintain a constant voltage across the liquid crystal and photoconductor layers. The magnitude of the intensity of the illumination on the *write* side of the valve directly influences the conductivity of the photoconductive layer and thus alters the magnitude of the electric field across the liquid crystal layer. The light that enters the LCLV on the read side and is reflected by the dielectric mirror has its polarization rotated by the liquid crystal layer. Consequently, the polarization of the light that exits the *read* side of the LCLV is a function of the intensity of the light impinging on the *write* side of the valve. Ideally the polarization would be rotated between 0° and 90° to yield a high contrast intensity modulation after passing through a polarizer. The intensity modulation of the LCLV is described in Fig. 2.10. As can be seen in the figure, the LCLV exhibits the NOR type of nonlinear response necessary to implement the neural response of the modified Hopfield model.

The optical implementation of the feedback using the LCLV is illustrated in Fig. 2.11. Initially, the input neural matrix is introduced into the system by opening shutter A while keeping shutter B closed. The polarization of the input source is vertical so as to be reflected by the polarization beamsplitter (PBS). A second PBS at the LCLV is used to separate the reflected *read* beam (horizontal polarization) from the input *read* beam (vertical polarization). The resulting thresholded intensity is fedback through the system until it is blocked

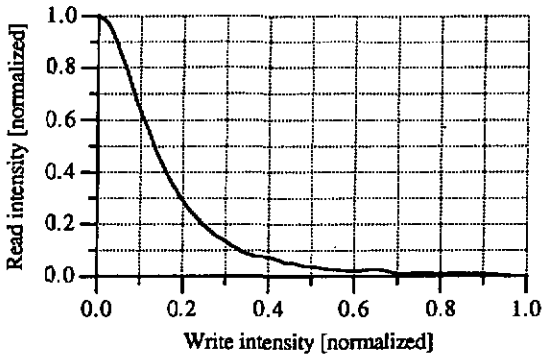


Fig. 2.10 Intensity modulation of LCLV.

at shutter B. The feedback of the system is completed by opening shutter B at the same time that shutter A is closed. The response of the LCLV is relatively slow with a response time on the order of hundred's of milliseconds. Therefore, there is enough time to manipulate the shutters without influencing the results. Once the feedback is complete, the system quickly stabilizes to a steady-state configuration of the neural matrix.

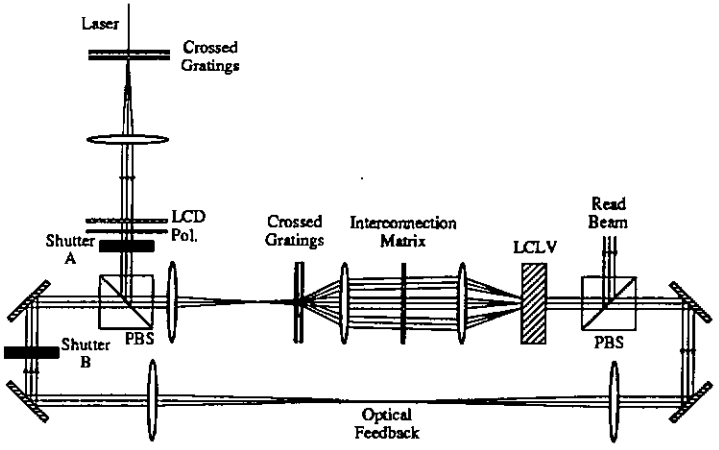


Fig. 2.11 All-optical implementation of neural network.

3. Experimental results

The all-optical implementation has been used to experimentally realize the inhibitory Hopfield model. The results however, have been limited to a very small number of neurons due the optical components available. Experimental

results for both the inhibitory Hopfield model and the INN model have been realized using the hybrid implementation with PC-controlled thresholding and feedback. The neural capacity of the hybrid implementation has been increased to as many as 49 neurons, represented in a 7 x 7 two-dimensional matrix. Whereas, the all-optical implementation has been realized for only 9 neurons.

Using the all-optical implementation the inhibitory Hopfield model has been applied in the recognition and noise removal of two memory patterns of 3 x 3 neurons. The two memory patterns, the letters "L" and "T", were selected to be very different so as to facilitate the pattern recognition, Fig. 2.12. To test the associative memory's performance discrete errors are introduced into the image. The number of discrete errors is often referred to as the Hamming distance. The recognition of the pattern "L" is shown for an initial input neural matrix with a Hamming distance of 2, Fig. 2.13a. While the shutter A of Fig. 2.11 is open and shutter B is closed the feedback matrix directly resembles the memory pattern "L", Fig. 2.13b. Once shutter A is closed and shutter B is opened the network converges immediately to the steady-state result, Fig. 2.13c. The recognition of the pattern "T" is observed as for the pattern "L". First the initial neural matrix with a Hamming distance of 2 is input into the system, Fig. 2.14a. The feedback results before modification of the shutters is shown in Fig. 2.14b. Finally, with the feedback loop closed the system converges to the steady-state result corresponding to the pattern "T", Fig. 2.14c.

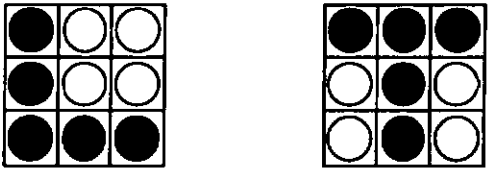
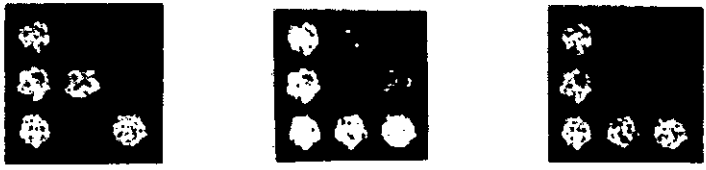
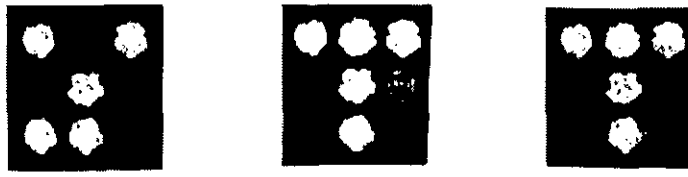


Fig. 2.12 The two 3 x 3 memory patterns.



(a) (b) (c)

Fig. 2.13 The input neural matrix for the pattern "L" with a Hamming distance of 2 (a), the feedback results (b), and the steady-state result (c).

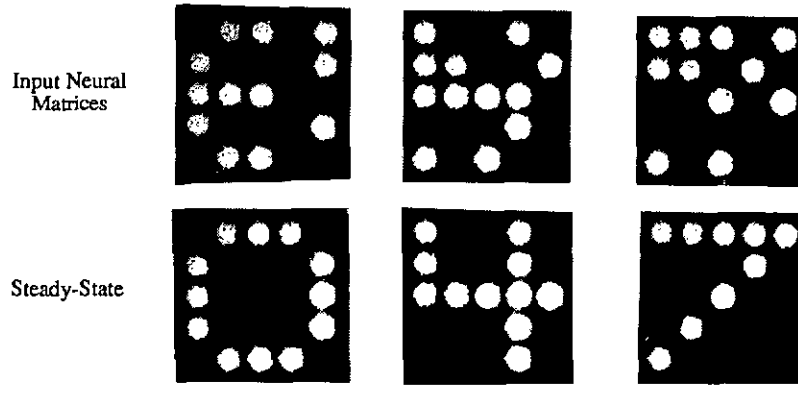


(a) (b) (c)

Fig. 2.14 The input neural matrix for the pattern "T" with a Hamming distance of 2 (a), the feedback results (b), and the steady-state result (c).

The performance of the LCLV used in the all-optical implementation limits the dimension of the neural matrices. Upon increasing the dimensions of the neural matrix to include more neurons and thus more detailed memory patterns, the sensitivity and uniformity of the LCLV greatly inhibit the system's performance. Thus to investigate the performance of the neural network for larger dimensions it was necessary to use the hybrid implementation.

The hybrid system has been used to implement the inhibitory Hopfield model for neural matrices of 5 x 5 neurons and 7 x 7 neurons. A CCD is used to measure the intensity of the output neurons, while the thresholding and feedback to the LCD are controlled by a personal computer. The interconnection weight matrix for the 5 x 5 neurons was calculated and fabricated for three memory patterns. Each of the memory patterns was able to be recovered from noisy input patterns, Fig. 2.15.



(a) (b) (c)

Fig. 2.15 The noisy input neural matrix and the steady-state result for the patterns "0" (a), "4" (b), and "7" (c).

The memory capacity of the neural network was increased to five memory patterns for the neural matrices of 7 x 7 neurons. Again the memory patterns were chosen to be very different from each other, with a Hamming distance of at least half of their total neurons from another memory pattern. The associative memory was able to recover the appropriate memory pattern even when presented with input patterns containing over 25% errors, i.e. a Hamming distance of greater than 12. An example of the degree of difficulty in recognizing the memory patterns for such high noise levels is illustrated in Fig. 2.16.

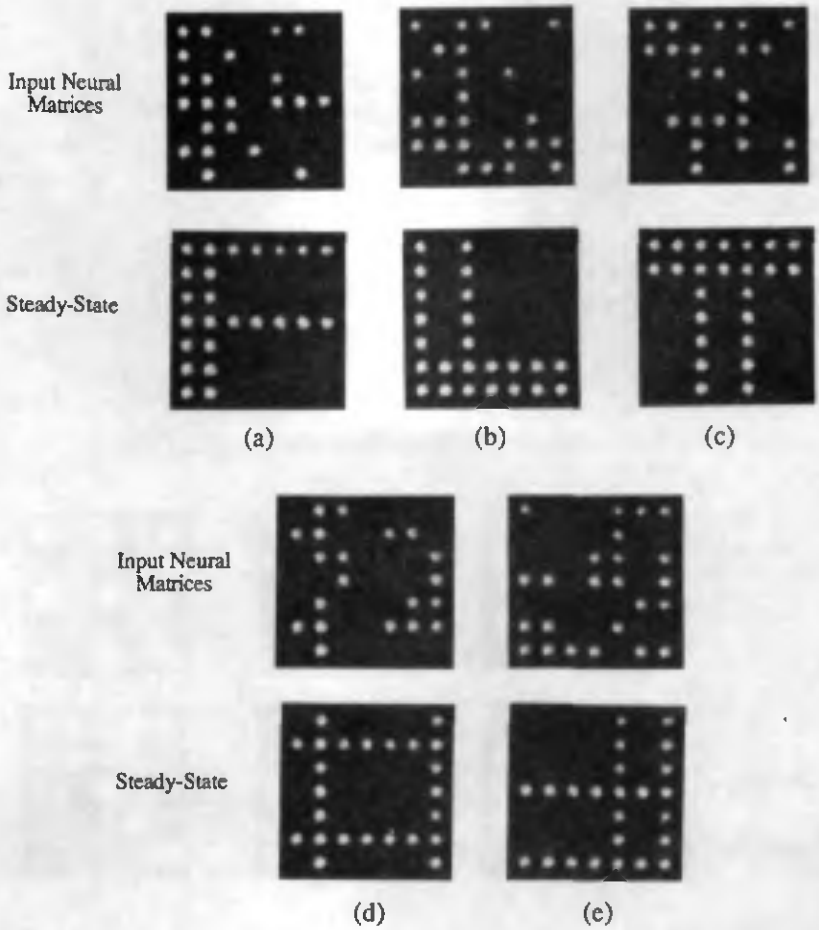
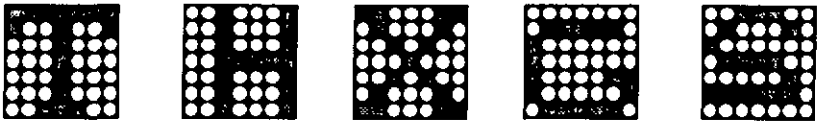


Fig. 2.16 The noisy input neural matrix and the steady-state result for the five 7 x 7 memory patterns.

Statistical performance

The results presented above are to give a feel for the nature of the recognition task as it is implemented in the optical and hybrid system implementations. As noted above, the neural network is capable of recovering even a highly deteriorated image. Yet, the consistency with which the correct memory pattern is recovered decreases as the noise level is increased. The system may succeed in recovering an input pattern with a high noise level, while failing to recover an input pattern with a lower noise level. However, the probability of recovering the input pattern decreases as the noise level increases. Therefore, to judge the performance of the network, it is necessary to study the statistical performance of the system over a large number of trials. The number of errors introduced into the input matrices is steadily augmented from the perfect memory pattern to approximately a 25% error level.

The statistical performance of the inhibitory Hopfield model and the INN model have both been experimentally verified for the case of five memory patterns of 7 x 7 neurons. The number of errors introduced was varied from 0 to 15, corresponding to error levels of from 0% to 30%, respectively. At each error level 10 trials were performed for each of the five memory patterns, yielding a total sampling of 50 trials for each error level. The inhibitory Hopfield model uses the five memory patterns previously presented in Fig. 2.16, which are mutually orthogonal and have approximately half of the neurons in the "off" state and half in the "on" state. These restrictions increase the performance of the model. The INN model, on the other hand, performs better with patterns containing a percentage of "on" neurons considerably less than 50%. Therefore, the INN model uses a set of memory patterns containing only 35% "on" neurons, Fig. 2.17. The patterns were selected so as to have only slight overlap of their "on" neurons, thus facilitating the recognition process.



(a) (b) (c) (d) (e)

Fig. 2.17 The five 7 x 7 memory patterns for the INN model.

The statistical results were generated using the hybrid system implementation. The system contains many sources of imperfections which yield a variation of the intensities of the output neurons. The largest

nonuniformity is introduced by the diffraction gratings whose orders have an intensity variation of approximately ±15%. The resulting nonuniformity of the intensity of the output neurons is easily measured. The interconnection mask could be reproduced to compensate for this nonuniformity. However, in the experimental results presented here, the variation of the intensity of the output neurons was compensated using a Look-Up-Table (LUT) to multiply the intensities measured by the detector before being thresholded. This is possible in the hybrid system, but in the all-optical implementation it would be necessary to modify the interconnection mask to compensate for the system defects.

The statistical performance of the inhibitory Hopfield model and the binary INN model are presented in Fig. 2.18. It can be seen in the results that the performance of the two neural models are comparable up to 7 input errors (14% error level). However, the performance of the binary neural model deteriorates more rapidly than that of the gray scale neural model, due to the information lost from the binarization of the interconnection weights. Both models maintain a recognition probability of approximately 100% for less than 8 errors (16% error level). With up to 12 errors (24%), there is a high percentage of correct convergence for the inhibitory Hopfield model, greater than 70%. The sharp reduction in percent convergence noted for the inhibitory Hopfield model for 13 errors is due to the limited sampling of the experimental data. All of the curves should be monotonically decreasing as the number of errors is increased.

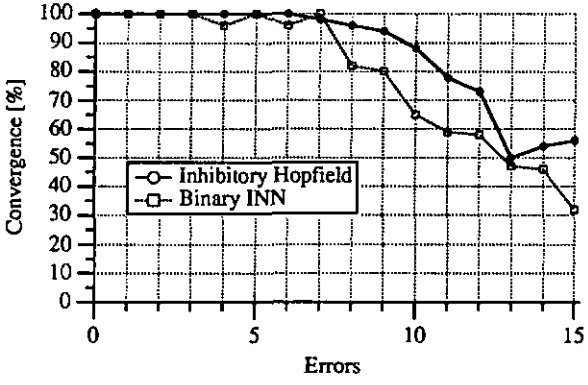


Fig. 2.18 The statistical performance of the 7 x 7 experimental results, percent convergence versus number of errors.

4. Improvement in neural performance with modifiable interconnections

For the experimental results presented above the interconnection weights were achieved using transparent photographic plates. All the interconnection weights were calculated *a priori* and did not take into account any nonuniformities resulting from system defects, such as nonuniform fan-out, integration errors at the detector, misalignments, and nonuniform detector responses. This section discusses the introduction of an electrically addressed spatial light modulator into the system as a modifiable interconnection matrix.

Replacement of the fixed interconnection matrix provides the system with the possibility to learn and correct some of the system defects mentioned above. The system nonuniformities may be directly compensated by actively varying the intensity transmission of the modulator. Furthermore, the memory patterns may be learned during an apprenticeship stage. An inherent result of the learning phase is that the system has directly compensated for some of the defects mentioned above. Another advantage is that the memory patterns can be easily modified without the need to fabricate another photographic mask. Thus rendering the associative memory much more versatile.

Electrically addressable spatial light modulator

An electrically addressable liquid crystal television (LCTV) is used to encode the intensity weights of the interconnection matrix. The LCTV is one of three active matrix twisted-nematic liquid crystal displays originally employed in the Seiko Epson Video Projector VPJ-2000. The display contains 480 x 440 pixels with a spacing of approximately 46 μm x 56 μm. The original electronics of the video projector are used to drive the LCTV, while the displayed image is generated by the PC using a framegrabber board with 8 bit gray scale resolution. The liquid crystal layer of the LCTV is used to rotate the polarization of the impinging beams. The displays are capable of rotating the polarization up to 90°, thus providing high contrast intensity modulation when a polarizer is placed after the display. The displays also exhibit very good gray scale performance, Fig. 2.19.

The transmission curve of Fig. 2.19 is normalized to one, yet the actual intensity transmission depends on the amount of reflection, absorption, and diffraction experienced by the beam when traversing the LCTV. In the case where the incident beam traverses a single pixel, approximately 44% transmission is observed. However, when the beam is larger in diameter and

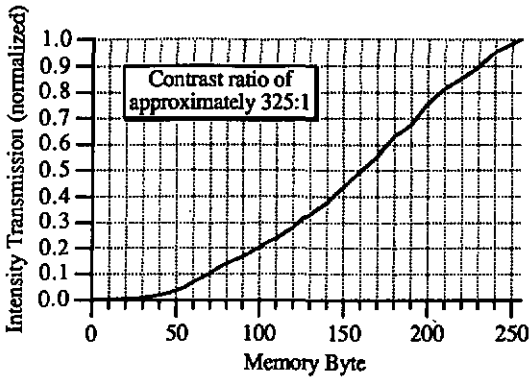


Fig. 2.19 Gray scale performance of LCTV.

passes through many pixels, diffraction effects due to the periodic array of pixels become significant. The transmitted intensity is diffracted into several diffraction orders. The zero order contains approximately 33% of the total transmitted intensity, corresponding to approximately 8% transmission. The normalized gray scale performance of the LCTV is independent of the incident beam diameter.

The architecture of the original hybrid system remains unchanged except for the replacement of the photographic plate by the modifiable LCTV, Fig. 2.20. Also, the LCD used in the initial experiments to produce the binary input pattern is replaced by one of the two remaining LCTVs. The decision process and information feedback are all controlled via the PC. To incorporate the learning capability of the LCTV in the all-optical implementation it would be necessary to employ a detector during the learning phase to supervise the network's performance. The PC controlled feedback is applied to the gray

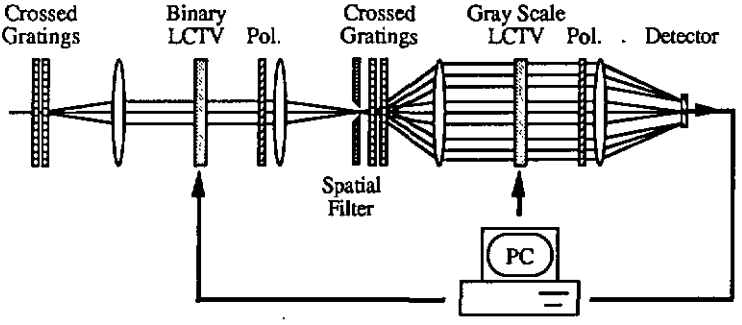


Fig. 2.20 Hybrid system for associative memory learning and recall results.

scale interconnection LCTV during the lear ning phase and to the binary input LCTV during the recall phase.

The optimum performance would be achieved by passing the impinging beams through individual pixels of the LCTV. This requires very accurate alignment and spacing of the beams with respect to the LCTV. An initial attempt produced an intensity variation across the display due to the Moiré interference resulting from the difference in spacing between the pixels and the beams. To avoid this effect the beam size was magnified until each beam was incident upon many pixels. The intensity variation was removed, but at the expense of a large decrease in the transmitted intensity due to absorption of the dead space between pixels and diffraction due to the periodic pixel structure. All but the zero order are spatially filtered in the Fourier plane of the lens following the binary input LCTV.

A learning algorithm has been developed to gradually adapt the interconnection weights. The algorithm is similar in nature to that of the Hebb model.[2.10] Before applying the learning algorithm the interconnection weights are initialized to values which are randomly varied between 40% to 60% transmission. The weights are initially set to values close to 50% transmission so that the algorithm is free to assert its influence over the interconnection weights, by selectively increasing or decreasing the amount of transmission. A wide spread of the initial transmission values is not desirable. It would result in situations where the algorithm is forced to reduce to low values the weights that are initially set high, or increase to high transmission those that are initially set low. Consequently, a wide spread of the initialization weights would result in an increase in the number of iterations required until convergence of the algorithm.

Once the interconnection weights have been initialized, one of the memory patterns is randomly selected and presented as the input to the neural system. The thresholded intensity of the output neurons $s_i^{(out)}$ is compared to the desired state of the corresponding neuron $s_i^{(m)}$ in the memory pattern m. If the state of the output neuron differs from the desired value, then the interconnection weights T_{ij} associated with the neuron in question are modified according to

$$T_{ij}(t+1) = \begin{cases} T_{ij}(t) - \varepsilon & \text{if } s_i^{(m)} = 1 \text{ and } s_j^{(m)} = 1 \\ T_{ij}(t) + \varepsilon & \text{if } s_i^{(m)} = 0 \text{ and } s_j^{(m)} = 1 \end{cases}, \qquad (2.13)$$

where $\varepsilon = 0.05$. Thus, the interconnection weights for the unrecognized neurons are adjusted by $\pm$ 5% transmission. If the neuron is recognized, i.e.

$s_i^{(out)} = s_i^{(m)}$, or the interconnection from the unrecognized neuron is to a neuron in the "off" state, i.e. $s_j^{(m)} = 0$, then the weight of the interconnection remains unchanged.

The learning algorithm as it is described above in Eq. (2.13) is not sufficient for the development of the interconnection weight matrix. Once all the memory patterns are correctly recognized there is no further improvement to the interconnection weights. The neural system is at that point capable of recognizing the memory pattern from an error free input, but is not well prepared to recognize the memory pattern from a noisy input. To allow the learning process to optimize the interconnection weights fully, an additional step is added to that of Eq. (2.13) and applied in parallel. The additional step improves the interconnection weights associated with the neurons which exhibit the worst performance, even if the state of the output neuron $s_i^{(out)}$ corresponds to the desired value $s_i^{(m)}$.

To determine the neurons with the worst performance, the neurons are first divided into two categories based on their desired state, i.e. whether the corresponding neuron of the memory pattern is "on" or "off", Fig. 2.21. At the presentation of each memory pattern there are always at least two neurons whose performance is classified as being the worst. Of the neurons that should be in the "on" state $s_i^{(m)} = 1$, the worst neuron is the one that has the highest intensity before application of the NOR function and is labeled $I_{max}(on)$. Similarly, for the neurons that should be in the "off" state $s_i^{(m)} = 0$, the worst neuron is that which has the lowest intensity before application of the NOR threshold function and is labeled $I_{min}(off)$. In the event that two neurons have

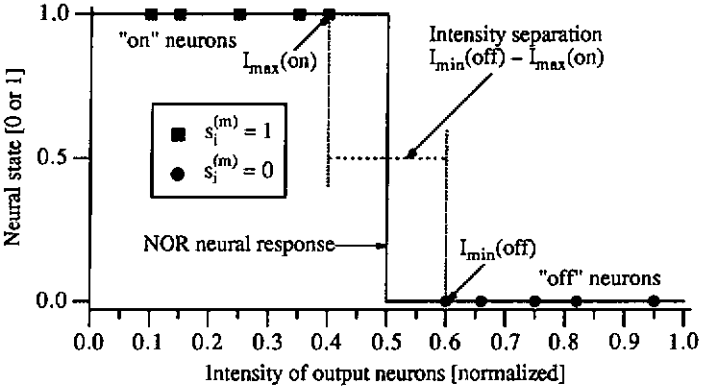


Fig. 2.21 Intensities of output neurons before applying the NOR threshold for the desired case where $I_{sep}^{(m)}$ is greater than zero.

the same output intensity which corresponds to the worst performance for the desired state, then the interconnection weights for both neurons are adapted.

The purpose of the additional step of the learning algorithm is to increase the intensity separation of the two neurons $I_{max}(on)$ and $I_{min}(off)$. This intensity separation varies for each memory pattern and is defined as

$$I_{sep}^{(m)} = I_{min}(off) - I_{max}(on) \; . \tag{2.14}$$

The case of $I_{sep}^{(m)} > 0$ corresponds to the desired situation where all of the "on" neurons have an output intensity less than all of the "off" neurons, as illustrated in Fig. 2.21. For $I_{sep}^{(m)} > 0$, the memory pattern m is correctly recognized by the neural system. If only Eq. (2.13) is used during the learning process, then the process stops when $I_{sep}^{(m)} \geq 0$, but no effort is made to maximize the intensity separation. In the case that $I_{sep}^{(m)} < 0$, then the memory pattern m is not yet recognized by the neural system, Fig. 2.22. The magnitude of $I_{sep}^{(m)}$ is a direct indication of the system's ability to recognize the pattern m. For positive values of $I_{sep}^{(m)}$, the amount of noise that the system is capable of tolerating and still recognize the memory pattern m depends directly on the magnitude of $I_{sep}^{(m)}$. For larger values of $I_{sep}^{(m)}$, the system is capable of accepting more noisy pixels in the input image. Therefore, to improve the system's performance, the learning algorithm attempts to increase the value of $I_{sep}^{(m)}$ for each memory pattern m presented. For each memory pattern, the output neuron corresponding to $I_{max}(on)$ has its interconnection weights reduced and the output neuron corresponding to $I_{min}(off)$ has its weights increased according to

$$T_{ij}(t+1) = \begin{cases} T_{ij}(t) - \varepsilon' & \text{for } I_{max}(on) \\ T_{ij}(t) + \varepsilon' & \text{for } I_{min}(off) \end{cases} \; , \tag{2.15}$$

where $\varepsilon' = 0.01$, thus corresponding to an adjustment of $\pm 1\%$ transmission. The result of this step of the learning procedure is an increase in the separation of the intensities of the "on" and "off" neurons. An average of the intensity separation $I_{sep}^{(avg)}$ is made for each of the M memory patterns to serve as the performance indicator of the system's ability to recognize all of the memory patterns. The average intensity separation is defined as

$$I_{sep}^{(avg)} = \frac{1}{M} \sum_{m=1}^{M} I_{sep}^{(m)} \; . \tag{2.16}$$

At the beginning of the learning process $I_{sep}^{(avg)}$ is negative, since the randomly initialized interconnection weights contain no information about the memory patterns to be learned. Each of the memory patterns is randomly

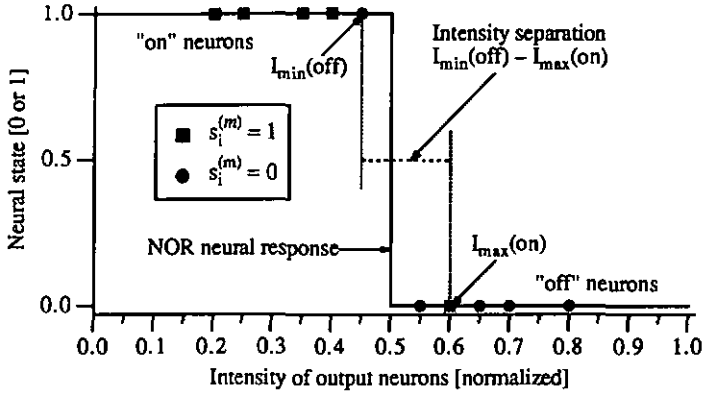
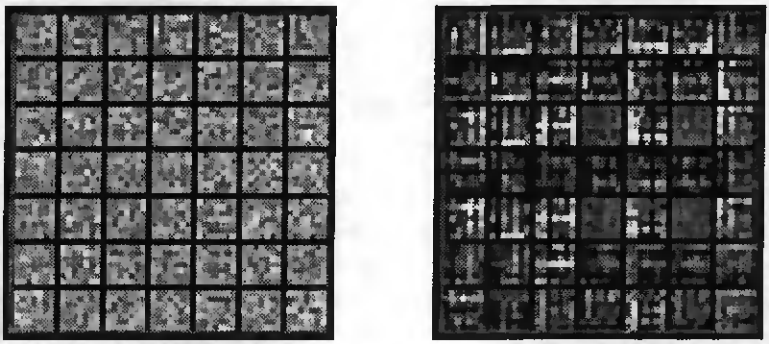


Fig. 2.22 Intensities of output neurons before applying the NOR threshold when $I_{sep}^{(m)}$ is less than zero.

selected and presented as the input to the neural system. If $I_{sep}^{(m)}$ is negative then Eq. (2.13) is applied to all the neurons i whose output state $s_i^{(out)}$ does not correspond to the desired state $s_i^{(m)}$. Independent of the value of $I_{sep}^{(m)}$, Eq. (2.15) is always applied to improve the performance of the interconnection weights of the neurons exhibiting the worst performance. Once the weights have been improved to the point that $I_{sep}^{(m)}$ is always greater than zero, i.e. the pattern m is correctly recognized, then Eq. (2.13) no longer influences the interconnection weights and only Eq. (2.15) still continues to adapt the weights. The learning algorithm continues to iterate through the memory patterns until the average of the intensity separation $I_{sep}^{(avg)}$ can no longer be increased. This results when the changes made during the presentation of one of the memory patterns are undone during the presentation of one of the remaining memory patterns. The learning process is terminated when no further increase in the average of the intensity separation is achieved during three consecutive iterations.

The learning algorithm described above was used to iterate the randomly initialized interconnection weight matrix of Fig. 2.23a, to the final learned matrix of Fig. 2.23b for the five 7 x 7 memory patterns of Fig. 2.16. Initially the average of the intensity separation for the five memory patterns was equal to -35% of the threshold value, indicating that some of the output neurons that should be "on" have greater intensity than output neurons that should be "off". The learning procedure lasts for approximately 25 iterations. Once completed, the neural system is capable of recognizing all of the five memory patterns. The average of the intensity separation has been increased to +18% of the

threshold value. Additional iterations were unable to increase the separation any further.



(a) (b)

Fig. 2.23 The randomly initialized interconnection weight matrix (a) and the learned matrix (b).

The interconnection weight matrix of Fig. 2.23b determined according to the Hebbian type learning algorithm was experimentally tested to determine its statistical performance. The statistics are based on 10 iterations of each of the 5 memory patterns as the number of errors is increased from 0 to 15, Fig. 2.24. The experimental results with the learned matrix are significantly better than the earlier results achieved with the fixed weights of the photographic mask for a Hamming distance greater than 10, corresponding to an error level of greater than 20%. The results are essentially equivalent for the Hamming distances less than or equal to 10. Computer simulations were performed to

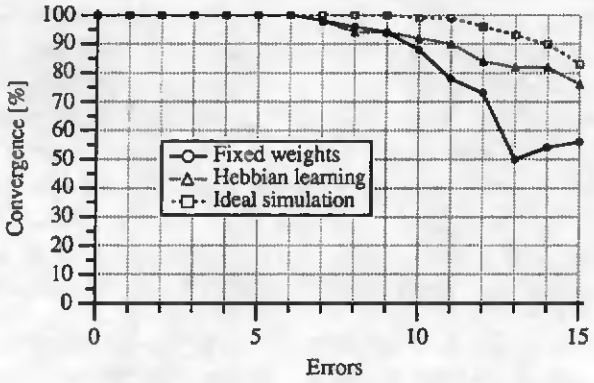


Fig. 2.24 Experimental performance of the inhibitory Hopfield model for fixed and learned weights compared to ideal simulation.

illustrate the limit anticipated for the inhibitory Hopfield model in the ideal case of no system defects or nonuniformities. The performance of the learning system remains below the ideal performance of the simulation results, yet is significantly improved over that of the fixed weights system.

5. Comparison of the performance of the neural approach to that of a discrete binary correlator

Although the emphasis of the work presented is on the experimental performance of inhibitory neural systems as associative memories, it serves as a worthwhile performance measure to compare the results of these systems to that of a discrete binary correlator, also referred to as a Hamming network.[2.3] The output results of the Hamming network differ from those of the associative memory neural systems in that its output is not a stored memory pattern as in a content addressable memory (CAM), but a decision as to which pattern the input image most closely resembles. The Hamming net could be adapted to function as a CAM by placing a pattern substitution subsystem after the correlator to regenerate the memory pattern chosen by the correlator. A system of this type has already been studied in various symbolic substitution applications.[2.12] The difference between the system's implementation as a CAM compared to that of symbolic substitution, is that the thresholding operation of the CAM must be modified so that only the strongest pattern recognized produces an output. Whereas in symbolic substitution, the thresholding operation can produce more than one output.

Since the memory patterns are encoded as binary variations of intensity, it is necessary to use dual-rail recognition of the patterns, i.e. both the pattern and its inverse are recognized within the correlator. There are two approaches to implementing the dual-rail recognition needed for the binary patterns. In the first method, the intensity resulting from the correlation of the input image with the inverse memory pattern is electronically subtracted from the intensity of the input image's correlation with the positive memory pattern. In the second method, the image and its inverse are replicated at the interconnection matrix, Fig. 2.25. The intensity resulting from the correlation of the input image with the positive memory pattern is electronically added to the intensity of the correlation of the inverse input image with the inverse memory pattern.

The second method is better adapted for optical implementation since it requires the addition of two intensities. By rotating the polarization of the intensity resulting from the inverse correlation 90° with respect to the

polarization of the intensity resulting from the positive correlation, the two intensities can be added incoherently at the detector plane. Therefore, the need to electronically add the measured intensities is avoided.

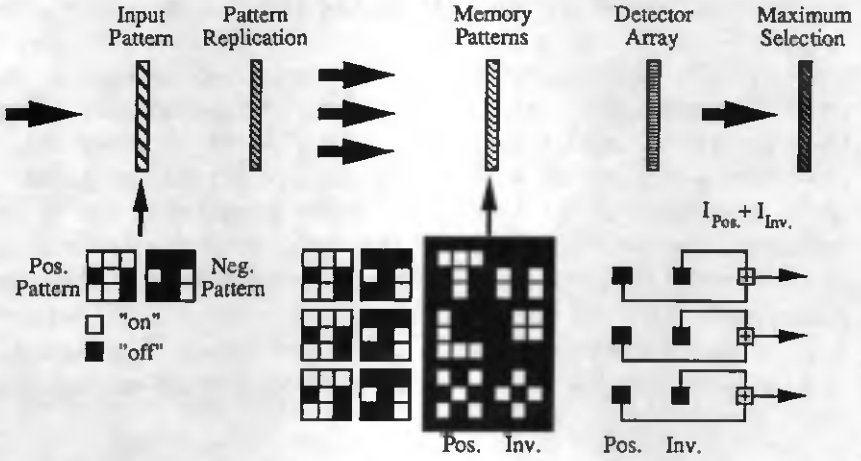


Fig. 2.25 Schematic of dual-rail correlator for recognition of binary intensity input images.

The same optical system used to realize the neural network hybrid implementation, Fig. 2.8, can be used for the Hamming network. The only modification necessary is the replacement of the photographic mask used for the interconnection weights by a mask adapted for the dual-rail correlation of the Hamming net, Fig. 2.26. Also, as discussed above, the input pattern and its inverse are presented as the input to the system. The maximum of the measured intensities must be determined, which represents a more complicated process than the simple thresholding function used for the neural responses in the neural implementations.

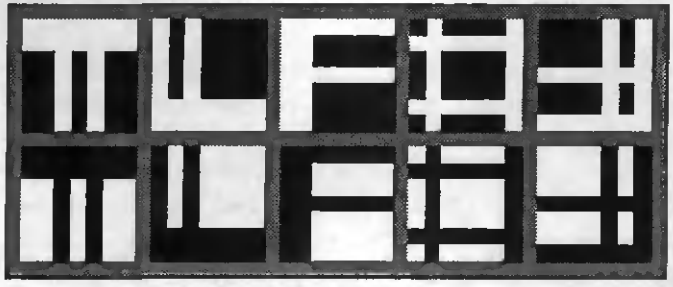


Fig. 2.26 The dual-rail correlator mask.

The statistical performance of the Hamming net was experimentally verified for the same set of memory patterns as used in the inhibitory Hopfield results. As before, the statistics are based on a sampling of 50 measurements per Hamming distance, i.e. 10 trials for each of the five memory patterns for each error level. The errors are increased from 0 to 15 and the results are compared to the earlier results obtained with the neural systems, Fig. 2.27. As clearly illustrated in the experimental results, the performance of the Hamming net is significantly superior to that of the inhibitory neural implementations. Even the ideal performance of the inhibitory Hopfield model, as determined by computer simulations, is inferior to that of the optically realized correlator system. The INN model, which like the Hamming net also employs binary weighting in the interconnection matrix, exhibits the poorest performance. The superior performance of the Hamming net suggests that it is much better suited for the recognition of discrete binary patterns. This same conclusion has already been drawn elsewhere, based on simulation results.[2.3,2.14]

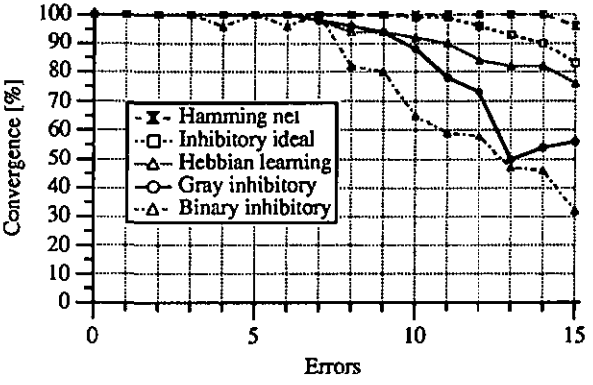


Fig. 2.27 Comparison of statistical performance of Hamming net to the results of the neural models.

The comparison of the experimental results illustrates the superior performance of the Hamming net for the recognition task. Besides the superior performance, the Hamming net is also much more versatile with respect to the number of "on" neurons included in the memory patterns. As discussed earlier, the inhibitory Hopfield model functions best for memory patterns containing approximately 50% "on" neurons and 50% "off" neurons. The INN model, on the other hand has its best performance for memory patterns containing considerably less "on" neurons. The Hamming net is however, not

at all influenced by the number of "on" neurons in the memory pattern when using the intensity addition method described earlier.

Clearly, as the number of memory patterns is increased the recall capability of an associative memory is reduced. The influence that increasing the memory size has on the statistical performance of the neural models and the Hamming net was analyzed using simulation studies performed on a personal computer. These results showed that the neural models were much more affected by the increasing memory capacity. The Hamming net showed a small decrease in recall for high error rates, Fig. 2.28, but remained relatively unaffected as the memory capacity was increased from 3 patterns to 7 patterns. The memory patterns were randomly generated and the same patterns were used for each of the neural models, except the INN model which had other memory patterns to accommodate its improved performance with less "on" neurons. The results of the inhibitory Hopfield model are included on the graph with the Hamming net to illustrate the model's reaction to the increasing number of patterns.

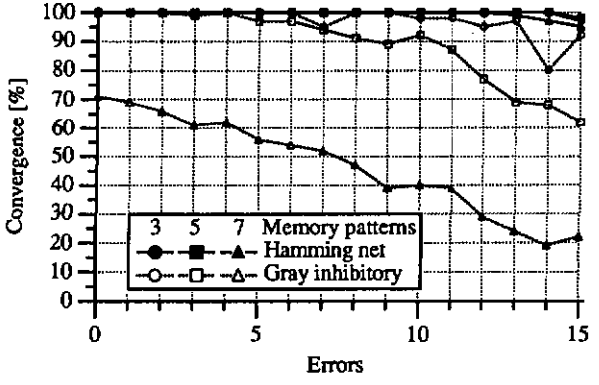


Fig. 2.28 Recall capacity as a function of memory size for the Hamming net and the inhibitory Hopfield model.

Based on the results of the experimental and simulation comparisons of the performance of the Hamming network to the inhibitory neural models, it is clear that the Hamming net is much better suited to the task of recognizing binary discrete patterns. However, the Hamming net requires a more complicated thresholding process, where the maximum intensity must be found amongst the set of output neurons. There are many possibilities to realize this function, e.g. Maxnet or Winner-take-all arrangements. An electro-optical realization of the winner-take-all arrangement has recently been proposed for

winner selection in an optical unsupervised learning neural network.[2.15] Another limitation of the Hamming net is that it lacks robustness with respect to interconnection errors. Yet, since the Hamming net has a much greater space-bandwidth-product (SBWP), it is possible to sacrifice a portion of the SBWP for added robustness.

6. Limitations of system architecture

The system architecture that has been exploited for the inhibitory neural networks and for the Hamming network is based on the use of diffraction gratings for the distribution of the intensity weighted information. The complete systems, as illustrated in Figs. 2.8 and 2.11, are limited in their computational capacity and physical dimensions. The two limitations are related and are both influenced by the optical components available. The speed with which the system can operate is determined by the liquid crystal devices used to encode input neurons and threshold the output neurons.

The theoretical limitation of the computational capacity of the system is based on the available space-bandwidth-product (SBWP). The SBWP is equivalent to the number of neurons that can be processed in parallel by the system. As can be seen in the schematics of the optical implementations, the last lens before the output plane represents the limiting aperture of the system. The maximum dimensions of the neural matrix can be derived in terms of the geometrical constraints of the system. Assuming a simple geometrical approach and ignoring the effects of point broadening due to diffraction, the minimum required diameter of the limiting clear aperture can be derived as a function of the dimension of the neural matrix N. The parameters used are taken from Figs. 2.3 and 2.5, and a few reasonable assumptions are made to simplify the equations. The image of the input neural matrix onto the interconnection matrix is assumed to be of unity magnification 1:1, thus stipulating that $f_1 = f_2$ and $\Delta x' = \Delta x$. Also, it is assumed that the same grating is used to produce the input matrix as well as its fan-out, therefore the period of the second grating is equal to the period of the first grating, $D = d$. Furthermore, it is assumed that the period of the gratings is chosen to satisfy the condition expressed in Eq. (2.12), which implies that

$$\frac{\lambda f_2}{D} = N\Delta x \,, \tag{2.17}$$

and thus the period of the gratings D can be written as

$$D = \frac{\lambda f_2}{N\Delta x} \,. \tag{2.18}$$

As a result of these assumptions, the condition expressed in Eq. (2.12) implies that $f_0 = f_2 / N$.

Referring to Fig. 2.5, the position of the limiting ray at the interconnection mask x_M is given by

$$x_M = \frac{1}{2} (N - 1) (N + 1) \Delta x , \tag{2.19}$$

and the position of the same ray at the lens before the interconnection mask x_L is given by

$$x_L = \frac{(N - 1)}{2} \Delta x + (f_2 - b) \frac{\lambda}{D} \frac{(N - 1)}{2} , \tag{2.20}$$

where b is defined in Eq. (2.9) and is equivalent to

$$b = \frac{f_2}{f_3} f_0 = \frac{f_2}{f_3} \frac{f_2}{N} . \tag{2.21}$$

Substituting Eqs. (2.18) and (2.21) into Eq. (2.20) and simplifying yields

$$x_L = \frac{(N - 1)}{2} \Delta x \, (N + 1 - \frac{f_2}{f_3}) . \tag{2.22}$$

The diameter of the limiting clear aperture at the last lens before the output plane is given by

$$\text{diameter} = 2 \sqrt{2} \left\{ \frac{(x_M - x_L)}{f_2} (f_2 + f_3) + x_L \right\} . \tag{2.23}$$

After substituting Eqs. (2.19) and (2.22) and simplifying, the limiting clear aperture can be expressed as

$$\text{diameter} = \sqrt{2} \, (N - 1) \, (N + 2) \, \Delta x . \tag{2.24}$$

Besides the constraints already discussed, it is also necessary to specify that the separation of the output neurons $\Delta x'$ must be equal to or greater than the diameter of the laser beam a' in the output plane so that there is no overlap of the neurons, i.e. $\Delta x' = \Delta x \geq a'$. The diameter of the laser beam in the output plane is described in terms of the magnification of the two telescopic imaging systems as

$$a' = \frac{f_3}{f_2} \frac{f_1}{f_0} a , \tag{2.25}$$

where a is the diameter of the laser beam input into the system. In addition to the previous assumptions that $f_1 = f_2$ and $f_0 = f_2 / N$, the limiting condition of $b = f_2$ can be assumed, thus from Eq. (2.21) $f_3 = f_2 / N$.

The result presented in Eq. (2.24) is only a first-order paraxial approximation. Diffraction effects can be taken into account by considering the diameter of the airy disk pattern that results from the Fraunhofer diffraction of the input laser beam at the focal plane of the first lens f_0. The diameter of the airy disk pattern for the input laser beam of diameter a must be smaller than the separation of the input neurons Δx to ensure that there is no crosstalk of the neighboring neurons. This condition is represented by

$$2.44 \frac{\lambda f_0}{a} \leq \Delta x \,. \tag{2.26}$$

The diffraction effects will have an effect on the period of the diffraction gratings needed to ensure the separation of the neurons. By substituting Eq. (2.7) into Eq. (2.26), the relation between the period of the gratings D and the diameter of the laser beam a can be given by

$$D \leq \frac{a}{2.44} \,. \tag{2.27}$$

To give a numerical example, it is assumed that the neurons are separated by 150 μm, which implies that the diameter of the laser beam must be reduced to $a \leq 150\ \mu m$ and that the gratings must have a period of $D \leq 60\ \mu m$. For a clear aperture of 60 mm, then the neural matrices may be as large as 16 x 16 neurons, thus 256 neurons.

The discussion above gives a first-order approximation to the computational capacity of the fully interconnected neural network implemented with the architecture based on diffraction gratings. In the above analysis, the diameter of the clear aperture is used to define the limits. However, the length of the system is also directly influenced by the assumptions made and the analysis that followed. The length of the optical system is defined as the sum of the lengths of the optical systems which make up the generation of the input neural matrix and its fan-out. Thus from Figs. 2.3 and 2.5, the system length can be defined as

$$\text{length} = 2\,(f_0 + f_1 + f_2 + f_3) \,. \tag{2.28}$$

From the previous analysis we can assume that $f_1 = f_2 = f$ and $f_0 = f_3 = f / N$. The total length of the optical system can then be written as

$$\text{length} = \frac{4f\,(N+1)}{N} \approx 4f \,. \tag{2.29}$$

The result of this analysis indicates that the length of the optical system is dominated by the focal lengths f_1 and f_2, and that the contributions from f_0 and f_3 can be neglected. Therefore, the value of f_2, which is determined by the fan-

out condition of Eq. (2.17), is the parameter that governs the total length of the optical system. The parameters of the numerical example used for the computational capacity are reapplied for the determination of the system length, i.e. $\Delta x = 150\ \mu m$ and $N = 16$. The period of the gratings is limited to 60 μm, as previously discussed, and the wavelength λ is taken as 500 nm. Depending on the fabrication techniques used to produce the grating, a period of 60 μm is already difficult to achieve for a uniform fan-out of 16 orders. As a result of these assumptions the focal length of the lens before the interconnection matrix is constrained to

$$f = \frac{D}{\lambda} N\Delta x = 288 \text{ mm}, \qquad (2.30)$$

which according to Eq. (2.29) would correspond to a total system length of 1224 mm. Thus the length of the system is over one meter and folding techniques must be used to reduce the packaging dimensions associated with the complete system.

Liquid crystal devices are employed in all of the implementations presented. A liquid crystal display (LCD) was used to produce the binary input neural matrix in both the optical and hybrid systems. Liquid crystal televisions (LCTVs) are used in the hybrid implementation of the learning system for encoding both the binary input matrix and the gray scale interconnection matrix. The threshold device used in the all-optical implementation is achieved with a liquid crystal light valve (LCLV). All of these devices use nematic liquid crystal which has a response time on the order of tens of milliseconds. The LCD and LCTVs are driven by their original electronics which operate at video rates.

The all-optical approach is the least inhibited by this slow response time. The LCD is used only to introduce the initial input matrix. Afterwards, the optical feedback is completed and the iterations of all the neurons is performed in parallel. In this respect, it takes approximately 30 - 50 milliseconds to complete one full iteration of the network. The hybrid implementation, on the other hand, is much more affected by the response time of the liquid crystal devices. To achieve best results the neurons are updated in an asynchronous manner, thus avoiding oscillatory behavior. This stipulates that the response time needed to modify the state of a neuron in the input matrix using the LCD must be repeated N^2 times for one full iteration of a N x N neural matrix. For large neural matrices, e.g. 16 x 16, one full iteration will take approximately 20 - 30 seconds.

7. Conclusions

A system architecture based on the uniform distribution of intensity in the diffraction orders resulting from phase gratings has been experimentally investigated for the optical realization of a neural network associative memory. The advantage of this implementation is that the memory matrices are stored in an easily fabricated gray scale intensity mask. The experimental results of the optical implementation of the inhibitory Hopfield model and the INN model have been compared. The statistical results demonstrate that the binarization of the interconnection weights only slightly deteriorates the performance of the associative memory. Computer simulation of the two models predicts a much larger difference in their recall performance. According to the simulation, the experimental performance of the gray scale model should be better than what was measured. The difference between the experimental and simulation results is due to the inaccuracy of the gray levels of the interconnections in the inhibitory mask. These variations have a much smaller effect on the binary model.

The presented architecture has the flexibility of allowing the replacement of the fixed interconnection mask by an electrically addressable spatial light modulator, as realized using an LCTV. The incorporation of the LCTV provides for total freedom in the modification of the interconnection weights. The memory patterns can be modified using either *a priori* calculations or through the iterative apprenticeship of an adapted learning algorithm. The performance of the neural system has been greatly improved through the system's ability to compensate for the imperfections of the optical system, such as nonuniform fan-out and nonuniform detector response. Still the performance of the system remains inferior to that predicted in the simulation results. This is due to the system's inability to compensate for the misalignments in the interconnection matrix. The learning system is unable to correct the system error due to the misalignment of the LCTV with respect to the impinging optical beams. To address this system error it is necessary to adopt a learning algorithm where not only the magnitude of the interconnection weights are learned, but also their position within the LCTV display. Such algorithms are feasible, but are time consuming. This approach would perhaps be better adapted to system architectures employing photorefractive crystals, where an active adaptation of the interconnection weights can be achieved. These architectures could employ a supervised learning algorithm to modify the interconnections within the photorefractive

material. Consequently, both the magnitude and position could be adapted in a more parallel fashion than is possible using a LCTV.

The architecture of the neural system was used to implement a Hamming network. While the Hamming net requires a more complicated thresholding procedure, which inhibits its all-optical realization, its performance is much superior to that of the neural systems. The function of the Hamming net differs from that of a content addressable memory (CAM) achieved with the neural systems, but pattern substitution subsystems could be adjoined to realize a CAM function. The Hamming net has many advantages over the neural systems with respect to the number of patterns stored as well as the type. The Hamming net has a much larger memory capacity, but lacks the neural systems' robustness with respect to interconnection errors. The experimental performance of the optical Hamming net was shown to be significantly superior to that of the ideal performance of the inhibitory Hopfield model as predicted by computer simulations. Therefore, the conclusion must be made that for the task of recognizing discrete binary images the Hamming net is the much better adapted approach. This conclusion has already been drawn elsewhere, based on simulation results.[2.3,2.14]

The architecture of the optical implementation, whether applied as a Hamming net or a neural system, was analyzed to see where the limits lie. While the architecture is capable of handling large neural matrices, e.g. 16 x 16 neurons, the physical dimensions of the system become quite large. The length of the system can be as large as one meter, depending on the number of neurons and the available optical components. The long system length implies that to take the system from the optical bench to the market place, a folding of the system is required. The alignment of the beams with the interconnection matrix will become a serious limitation for large numbers of neurons. In the numerical example given with 16 x 16 neurons and a neuron spacing of 150 μm, the precision needed in the separation of the diffraction orders will be better than 1 μm. Distortion of the impinging image by the lens before the interconnection plane will limit the size of the neural matrices which can be imaged onto the interconnection matrix.

Finally, the results of this experimental investigative research have led to the overall conclusion that the system architecture using diffractive gratings is not the best approach to optically implementing the fully interconnected neural networks. The architecture is much more attractive for the Hamming net, where the fan-out in the interconnection plane is much smaller than in the

neural systems. Still, the system architecture allows for the experimental study of the system's performance in the presence of imperfections due to the optical realization. The architecture allows for the easy installation of a modifiable interconnection matrix, thus allowing for the study of the possible improvements that learning algorithms may provide in the compensation of the system's nonuniformities and other imperfections. It is clear, that the big advantage of the neural systems is their ability to learn and adapt, and that this capability should be exploited as much as possible.

References

2.1 W.S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous system," Bulletin of Mathematical Biophysics **5**, 115-133 (1943).

2.2 J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. USA **79**, 2554-2558 (1982).

2.3 R.P. Lippmann, "An introduction to computing with neural nets," IEEE ASSP Magazine, April, 4-22 (1987).

2.4 D. Psaltis and N. Farhat, "Optical information processing based on an associative-memory model of neural nets with thresholding and feedback," Opt. Lett. **10**, 98-100 (1985).

2.5 J.-S. Jang, S.-W. Jung, S.-Y. Lee, and S.-Y. Shin, "Optical implementation of the Hopfield model for two-dimensional associative memory," Opt. Lett. **13**, 248-250 (1988).

2.6 H.J. White, "Experimental results from an optical implementation of a simple neural network," *Optical Computing 88*, Proc. SPIE **963**, 570-575 (1988).

2.7 I. Shariv and A.A. Friesem, "All-optical neural network with inhibitory neurons," Opt. Lett. **14**, 485-487 (1989).

2.8 H. Shouval, I. Shariv, T. Grossman, A.A. Friesem, and E. Domany, "An all-optical Hopfield network: Theory and experiment," Intl. J. of Neural Systems **1**, 355-360 (1991).

2.9 I. Shariv, T. Grossman, E. Domany, and A.A. Friesem, "All-optical implementation of the inverted neural network model," *Optics in Complex Systems*, Proc. SPIE **1319**, 194-195 (1990).

2.10 D.O. Hebb, *The organization of behavior*, John Wiley & Sons, New York (1949).

2.11 B. Acklin and R. Thalmann, "Liquid crystal displays as transmission spatial light modulators," IMT Internal Report **245**, (1988).

2.12 R. Thalmann, G. Pedrini, and K.J. Weible, "Optical symbolic substitution using diffraction gratings," Appl. Opt. **29**, 2126-2134 (1990).

2.13 N. Collings and W. Xue, "Characterization of optically addressed SLMs for recurrent optical neural networks," Int. J. of Opt. Comp. **2**, 97-107 (1991).

2.14 L. Tarassenko, B.G. Seifert, J.N. Tombs, J.H. Reynolds, and A.F. Murray, "Neural network architectures for associative memory", First IEE International Conference on Artificial Neural Networks, Publication No. 313, 17-22 (1989).

2.15 K. Wagner and T.M. Slagle, "Optical competitive learning with VLSI/liquid-crystal winner-take-all modulators," Appl. Opt. **32**, 1408-1435 (1993).

III. Self-organizing neural networks

The Kohonen Model

1. Self-organizing neural networks - The Kohonen model

Self-organizing neural networks were derived from the observations of a topological correspondence between the governing cortex and the sensory regions of the body. It was confirmed that the organization was not genetic, but was developed during a learning phase. From these observations a new algorithm was derived that reproduces an ordered feature map from a set of applied stimuli.[3.1] The construction of the feature map is based on the stimulus data presented to the network and is not based on a supervised response. The Kohonen model is described by this self-organization of the learning process.

The principle function of the algorithm is to transform a set of continuous input stimuli into a discrete set of quantified output data. A relationship between the neighboring output neurons is specified, based on the assumption of an existing relationship between neighboring input stimuli. As observed in biological neurons, a lateral interaction between the artificial neurons is implemented, where the closest neighbors excite each other, while neighbors that are further away interact inhibitively. This form of local interaction can be described by what is known as the Mexican hat function, Fig. 3.1. As a result of these local interactions a clustering effect is witnessed within the network.

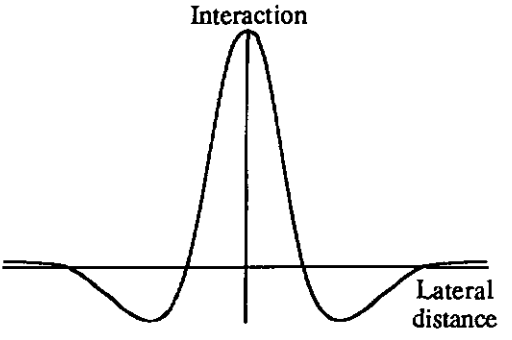


Fig. 3.1 Local interconnection between output neurons.

The network basically consists of two layers of interconnections and one layer of neurons. The interconnections between the input stimuli and the output neurons are modifiable and represent the feature map of the network. While the interconnections between the output neurons are based on the Mexican hat function described above. During the learning phase the weights of the interconnections between the input stimuli and the output neurons are self-adjusted based on the state of the output neurons. The excitatory and inhibitory lateral interconnections between the neurons are fixed and do not vary during the learning phase. The first layer of interconnections determines essentially the activity of each output neuron. Whereas the second layer provides the mechanism by which the topology of the input stimuli is transferred to the discrete output neurons.

The Kohonen model is well suited for the task of pattern classification. The algorithm has been effectively demonstrated in many classification and optimization tasks.[3.2,3.3] Although the calculations involved for each neuron are relatively simple, the number of calculations needed may become very time consuming in software simulations of the Kohonen model as the number of neurons increases. To reduce the calculation times involved a parallel processing implementation of the algorithm is desired. VLSI chips of this nature have already been demonstrated for the implementation of the Kohonen model.[3.4] In this paper an optical parallel processing system is used to implement the Kohonen model in solving the problem of the traveling salesman.

2. The Traveling Salesman Problem and its realization with the Kohonen model

The Traveling-Salesman Problem (TSP) is a well defined optimization problem employed to illustrate the computational capability of neural or computing networks. The problem can also be closely related with real life tasks such as circuit design and robot guidance. The TSP is described by a set of n cities, each of which must be visited once during a closed loop tour. The problem is to find the tour with the shortest path length. Finding the *best* solution to the TSP can become very computationally intensive. The calculation time necessary to find the *best* tour of the $(n\text{-}1)!/2$ possible tours, grows exponentially with the number of cities. To reduce the calculation times involved, non-deterministic algorithms have been applied to the TSP.[3.3,3.5] These neural algorithms yield *good* solutions to the TSP, but not necessarily the *best* solution.

Initially, a fully interconnected single layer neural solution was proposed to solve the TSP.[3.5] This approach is very strongly influenced by the choice of the parameter coefficients used. Also, the network requires an excessively large number of neurons, n^2. The application of the Kohonen algorithm to the TSP has been shown to yield very good results, while requiring much fewer neurons, $2n$.[3.3]

In realizing the TSP with a Kohonen network, the original orientation of the network structure has been somewhat modified, Fig. 3.2. The input stimuli to the network I_x, I_y are proportional to the x and y position of the presented city in a Cartesian coordinate system. The feature map of the network T_{ij} is represented by the interconnections between a constant input source and the set of N output neurons, where $i = 1 \,..\, N$ and $j = x, y$. Each output neuron O_{ij} consists of an x and y component. The similarity between the input stimuli and each output neuron defines each neuron's activity. The activity of the neuron is based on the Euclidean distance between the x and y components of the output neuron and the input stimuli. To determine the winning neuron, the x and y components of the input stimuli are subtracted from the x and y components of each output neuron. The differences of the two components are squared and summed to yield the Euclidean separation of the output neuron and the input

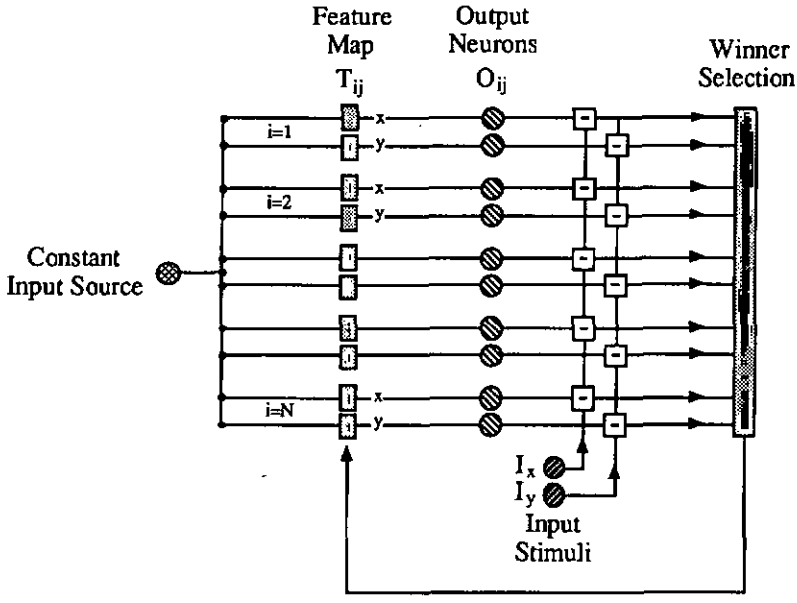


Fig. 3.2 Kohonen realization of the TSP.

stimuli, described by

$$\{ [O_{ix} - I_x]^2 + [O_{iy} - I_y]^2 \}_{\text{minimum}} \Rightarrow O_{winner} . \qquad (3.1)$$

The winning neuron O_{winner} is the one who's Euclidean separation is smallest, thus corresponding to the output neuron whose response is closest to the input stimuli.

The feature map is gradually developed by randomly presenting each city of the tour as the input stimuli and modifying the feature map of those neurons in the neighborhood of the winning neuron. The interconnection weights are modified according to

$$T_{ij}(t+1) = T_{ij}(t) + \alpha[I_j(t) - O_{ij}(t)] , \qquad (3.2)$$

where $i = 1..N$ and $j = x, y$, $T_{ij}(t)$ is the interconnection weight of the jth component of the ith neuron at the discrete time step t, $I_j(t)$ is the jth component of the input stimuli, and $O_{ij}(t)$ is the jth component of the ith output neuron. The gain α represents the sensitivity of the neurons to the input stimuli and may be varied during the learning process. The magnitude of the gain may also be varied over the neighborhood of the winning neuron, where its magnitude is decreased for neurons at a greater distance from the winning neuron. The extent of the neighborhood is another parameter that may vary during the learning process. Typically, the neighborhood is reduced to the nearest neighbor towards the end of the learning phase.

During one cycle of the learning phase, each city of the tour is presented. The order in which the cities are presented is selected at random for each cycle. The duration of the learning phase depends upon the initial values of the gain and neighborhood parameters and their variation with time. The learning phase may consist of as few as ten cycles or as many as tens of thousands of cycles. In general, best results are obtained using a long learning process with a small gain factor. The neighborhood is usually gradually reduced from the full tour to the nearest neighbor during the learning process. This allows for global organization at the beginning with finer adjustments at the end. It is evident that if the neighborhood is less than one then the network loses the ability to self-organize. Therefore, for best results it is necessary to have twice as many neurons as cities, so that at the end of the learning phase, when the winning neuron only influences his nearest neighbors, it is possible to have a neuron situated at each city, with its nearest neighbor between it and the next city on the tour, Fig. 3.3.

In most applications of the Kohenen network, the gain is reduced during

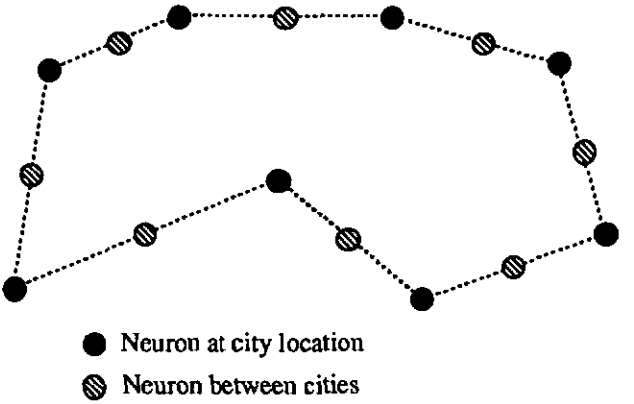


Fig. 3.3 Optimal arrangement of neurons at end of learning phase.

the successive iterations until its magnitude is quite small. Thus towards the end of the optimization process only slight modifications are made in the established neural states. In this arrangement, where the neighboring neurons have only very slight influence on each other towards the end of the optimization, it is possible to optimize city tours with only one neuron per city instead of two neurons per city as required in our realization using an increasing gain term.

The implementation of the Kohonen network presented here has been adapted to better exploit the the dynamic range of the device representing the feature map. Originally, the two input stimuli I_x, I_y were interconnected to the feature map instead of the constant input source as represented in Fig. 3.2. While the advantages of the optical implementation are admittedly reduced in the adapted implementation, the dynamic range of the feature map device is much better exploited. In the original implementation the interconnections corresponding to the winning neurons must reduce the input stimuli down to a common threshold value. This is a consequence of the fact that the device we are using has a maximum interconnection value of 1.0. Since the cities within a tour can take on values from 1 to 10, the corresponding interconnection weights vary from 1.0 to 0.1, respectively. The interconnection weights are determined by a 1/x function which is illustrated in Fig. 3.4. As can be seen in the figure, the interconnection weights corresponding to cities with coordinates between 4 and 10 are all tightly grouped about a narrow range between 0.25 and 0.1. Consequently, the selection of the winning neuron is inhibited due to the close grouping of the response of the output neurons. The adapted

implementation provides for a uniform distribution of the interconnection weights between 0.1 and 1.0.

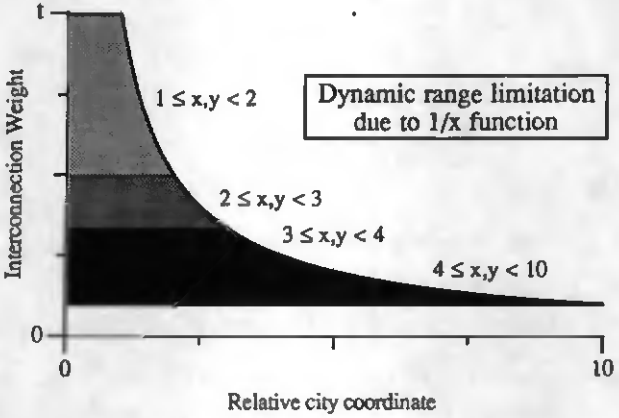


Fig. 3.4 Dynamic range limitation.

3. Optical implementation

The application of the Kohonen model to the TSP involves the comparison of the input stimuli, the x and y positions of each city, to each of the output neurons, so as to determine which of the output neurons is the most active, i.e. most similar to the input stimuli. This comparison is the basis of the modification of the feature map of the network during the learning process [Eq. (3.2)]. Since the input stimuli are the same for each neuron during the same time period it is possible to make the comparison to all the output neurons at the same time in a parallel processing system. Thus, the computational time involved, which grows quadratically with the number of cities, may be considerably reduced. The parallel distribution of information achieved with an optical system can be exploited in resolving the TSP using the Kohonen algorithm. However, it should be noted that since the Kohonen network involves only local interconnections, it lends itself well to a VLSI implementation, and is perhaps not the best illustration of where the benefits of optics can be exploited.

The optical system used to simulate the neural network may be realized using relatively simple optical components, Fig. 3.5. A laser diode (LD) serves as the constant input source to the network, refer to Fig. 3.2. Its laser beam is divided into many beams using a phase diffraction grating. The feature map of the network T_{ij} can be implemented using an electrically-addressed liquid

crystal television (LCTV) to modulate the intensity of the impinging beams. Each beam traverses the LCTV and is imaged onto an array of photodetectors (PDA). The intensity of the photodetectors corresponds to the response of the output neurons O_{ij}. A personal computer (PC) is used to read the intensities measured by the PDA and to compare the measured intensities with the input stimuli I_x and I_y. The PC determines the winning neuron O_{winner} and adjusts the individual pixels of the LCTV to realize the modification of the interconnection weights represented in Eq. (3.2) for the neurons in the neighborhood of O_{winner}.

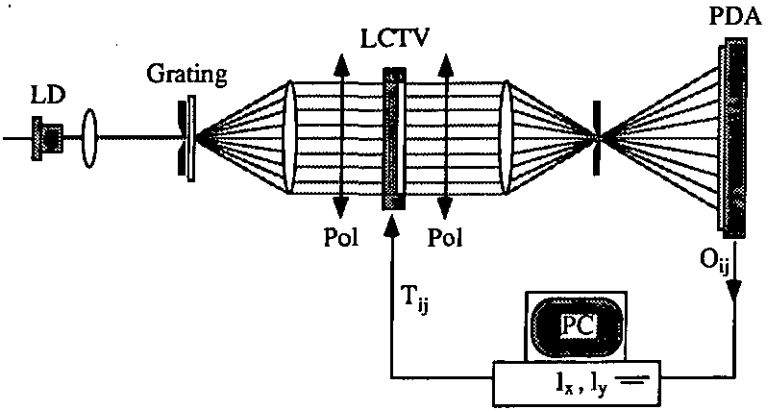


Fig. 3.5 Schematic of optical realization of Kohonen algorithm.

The two apertures of Fig. 3.5, are necessary to spatially filter the diffraction orders which result from the pixellated structure of the LCTV. The first aperture reduces the diameter of the impinging laser beam so that when the beam is imaged onto the second aperture there is no overlap of the neighboring diffraction orders. The first aperture has a diameter of approximately 5 mm and is 1:1 imaged onto the second aperture which has the same diameter.

The laser diode emits in the visible range, at the wavelength $\lambda = 670\ nm$. The emitted intensity I_D is controlled by the electrical current i_D of the diode, Fig. 3.6. The emitted intensity of the diode is highly sensitive to temperature variations. Therefore, either the temperature of the diode must be precisely maintained using a Peltier cooling element, or the temperature variation must be compensated by actively adjusting the threshold used in the selection of the winning neuron. The laser does not begin lasing until a forward current of approximately 71 mA is reached. Until this point, the diode is functioning as a light emitting diode (LED) and is emitting weak intensity nonpolarized light.

When the diode lases, a strong intensity polarized beam is superimposed on the LED beam. A polarizer is placed before the LCTV, Fig. 3.5, to remove the LED component of the LD's intensity which does not have the same polarization as the lasing mode, thus improving the performance of the LCTV.

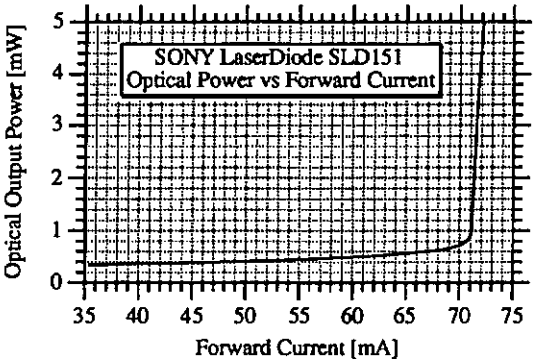


Fig. 3.6 Output intensity characteristic of laser diode.

A binary phase diffraction grating is placed after the LD to fan-out the impinging beam into 13 orders of approximately uniform intensity, Fig. 3.7. The grating has a period of $d = 87\ \mu m$, thus producing an angular separation of 0.44° for the diffracted beams of the LD at $\lambda = 670\ nm$. The intensity of the diffraction orders varies by approximately ±16%, while the efficiency of the grating, i.e. the intensity of desired diffraction orders with respect to the incident intensity, is approximately 66%.

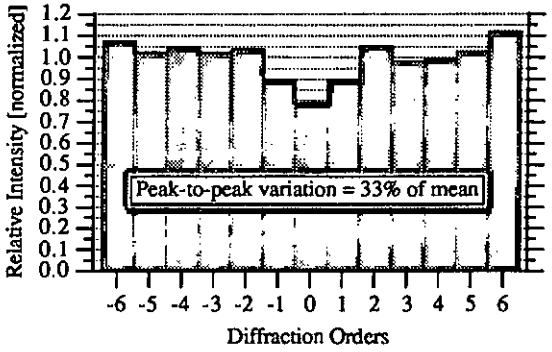


Fig. 3.7 Intensity distribution of the diffraction orders.

The LCTV is one of three active matrix twisted-nematic liquid crystal displays originally employed in the Seiko Epson Video Projector VPJ-2000. The same display was used in the experimental realization of the associative

memory neural network discussed in Chapt. 2. The reader is referred to Chapt. 2 Sect. 4 for the details oo the performance of the LCTV.

The intensity of the beams transmitted by the LCTV is measured using a photodetector array (PDA) containing 38 photodetectors. The photodetectors are spaced 1 mm apart in a linear array. A lens is placed after the LCTV to image the transmitted beams onto the PDA with the correct magnification, so that the individual beams are accurately aligned with each detector element. This lens also allows for the spatial filtering of the diffraction orders. The response of the PDA is transferred to the PC and compared to the input stimuli, i.e. the relative x and y position of the presented city. The output neuron whose response most closely resembles the input stimuli is selected as the winner. Consequently, the intensity transmission of the LCTV is modified in accordance with Eq. (3.2) by changing the output video signal of the framegrabber board of the PC.

Two photodetectors are necessary to measure the x and y component of each output neuron. Due to the limited number of intense diffraction orders generated by the Dammann phase gratings, only 13 of the 38 available photodetectors are being used. This severely limits the capacity of the optical system to resolve city tours. The criteria of 2 neurons per city and 2 diffraction orders per neuron limits the number of cities included in a tour to

$$n \leq \frac{\text{(total diffraction orders at LCTV)}}{4} . \tag{3.3}$$

Unfortunately, another grating which generates more diffraction orders was not available. However, the number of orders achieved with the same grating can be doubled by introducing a second laser diode source (LD2). The angular alignment of the beam of LD2 is adjusted so that a small angle ϕ exists between the beams of the two laser diodes, Fig. 3.8. The angle ϕ is adjusted so that the diffraction orders of LD2 are displaced from the orders of LD1 by half the separation Δy of the orders. The angle ϕ is determined by the period of the grating d, and the wavelength of the laser diodes λ, namely

$$\phi = \sin^{-1}\left[\frac{\lambda}{2d}\right] . \tag{3.4}$$

The beams of LD1 and LD2 are combined using a nonpolarizing beamsplitter BS. LD2 is aligned so that the two beams are superimposed at the plane of the grating. This is necessary so that the two beams will equally be superimposed when the diffraction orders caused by the LCTV are spatially filtered.

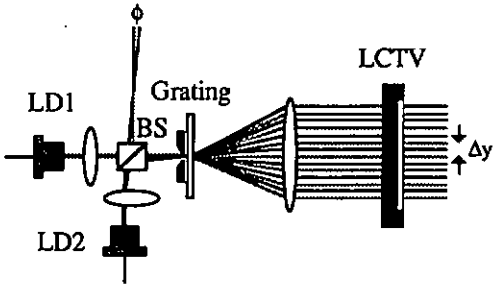


Fig. 3.8 Alignment of the two laser diode sources.

As a result of introducing a second source, the number of diffraction orders are increased to 26, corresponding to 13 output neurons. Therefore city tours of up to 6 cities can be effectively resolved by the optical system. It is possible to apply the system to tours with more cities. The maximum number of cities feasible is given by the total number of neurons. However, the resolution of the system will be reduced and thus the performance will suffer.

4. Experimental results

The results of the Kohonen algorithm are of a statistical nature. The algorithm finds a *good* solution for the presented tour, although not necessarily the *best* solution. Therefore, the maximum and average values of the dispersion of the experimental results are used to quantitatively describe the performance of the optical system. The maximum dispersion is defined as the difference between the Euclidean distance of the *best* tour and the distance of the longest tour of a series of trials. The average dispersion is the difference between the distance of the *best* tour and the average of the all the tours found in the series of trials. The actual magnitude of these differences depends on the cities within the tour. Therefore, the dispersion terms are expressed as a percentage of the *best* solution, so that the results of various city tours may be compared quantitatively. The maximum and average dispersion are described by

$$\delta_{max} = \frac{d_{longest} - d_{best}}{d_{best}} \tag{3.5}$$

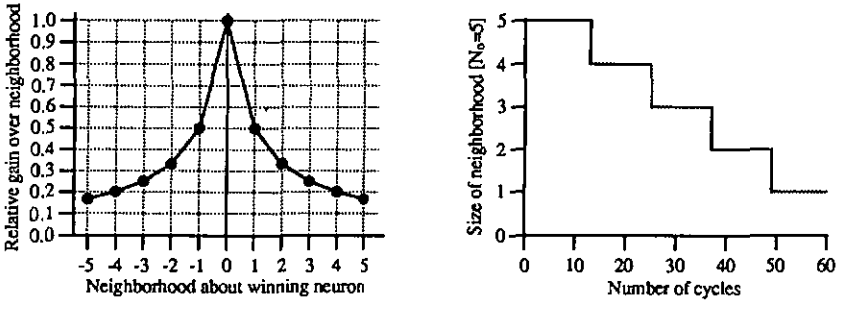
and

$$\delta_{avg} = \frac{d_{avg} - d_{best}}{d_{best}},$$

where $d_{longest}$, d_{avg}, and d_{best} represent the Euclidean distances of the longest and the average tours of the batch, and the best tour, respectively.

A computer simulation of the Kohonen algorithm is implemented to serve as a reference for the performance of the optical implementation. By comparing the results of the optical system with those of the computer simulation it is possible to analyze the performance of the optical implementation independently of the algorithm. Without the computer simulation, the non-deterministic nature of the Kohonen algorithm would make it difficult to evaluate the benefits of the optical architecture.

The same randomly initialized feature map is used in both the computer simulation and the optical implementation. The gain α is initially set to α_o and then increased each cycle by $\Delta\alpha$. During one cycle every city of the tour is presented once for modification of the feature map. The order in which the cities are presented is selected at random. The magnitude of the gain is varied over the neighborhood as described in Fig. 3.9a. The neighborhood about the winning output neuron is initially set to $\pm N_o$. The neighborhood is reduced throughout the learning process as illustrated in Fig. 3.9b. The parameters and starting conditions of the Kohonen algorithm are identical for the computer simulation and the optical implementation. For all of the experimental results presented here the initial gain is $\alpha_o = 0.5$, the delta gain per cycle is $\Delta\alpha = +0.005$, the initial neighborhood is $N_o = 5$, the number of cycles is 60, and the feature map is randomly initialized.



(a) (b)

Fig. 3.9 The variation of the gain over the neighborhood (a), and the variation of the neighborhood during the optimization process (b).

It should be pointed out that in most applications the gain α is decreasing during the optimization process, i.e. $\Delta\alpha$ is negative. The selection of the

parameters for the experimental results presented here was based on trial and error observations. The parameters chosen performed well for the given examples, but may be less performant for more complex city tours. The effect of increasing the gain term serves to strongly pull the output neurons to the city locations in only a few iterations, but also reduces the finesse of the distinction process.

There exists for the optical algorithm an additional parameter C_{LCTV}, which is not needed for the computer simulation. This parameter represents the relation between the LCTV transmission and the 8 bit value sent from the framegrabber board. In the ideal case, where the gray scale response of the LCTV was totally linear, C_{LCTV} would equal 255 due to the 8 bit dynamic range of the framegrabber board. However, it is easily observed by comparison of the optical performance and the simulation performance that C_{LCTV} is better represented by a value of 110. This is due to the nonlinear response of the LCTV seen in Fig. 2.19. For larger values of C_{LCTV} the response of the output neurons of the optical system is more brusque than in the computer simulation. As a result, Eq. (3.2) is modified for the optical system to include the C_{LCTV} constant, namely

$$T_{ij}(t+1) = T_{ij}(t) + \alpha(I_j(t) - O_{ij}(t))C_{LCTV} , \qquad (3.6)$$

where $T_{ij}(t)$ now represents the 8 bit value describing the LCTV transmission.

Two types of arrangements of the city patterns are used to describe the performance of the optical system in finding *good* solutions to the TSP. For part of the measurements, the arrangement of the city patterns is randomly generated and the number of cities is varied from 5 to 9 cities. For each number of cities, 5 arrangements are randomly selected, to give a total of 25 arrangements. For the remaining measurements, the arrangement of 6 cities is selected to test the resolution of the system. The position of one city within the arrangement is slightly adjusted, causing a slight modification in the *best* tour. Thus, the presentation of the experimental results is broken up into two parts, "Random tours" and "Resolution tours".

Random tours

The random configuration of the cities is to validate the application of the Kohonen algorithm to any arrangement of cities. As stated in Eq. (3.3), the algorithm achieves best results for city patterns containing n cities, where n is equal to or less than half the number of output neurons. The algorithm continues to perform for city tours containing more cities, but the resolution

of the results is reduced, as witnessed by an increase in the dispersion of the results. Five random arrangements of city patterns were selected for tours of 5, 6, 7, 8, and 9 cities. The city coordinates are normalized to fall within a unit square as shown in Fig. 3.10.

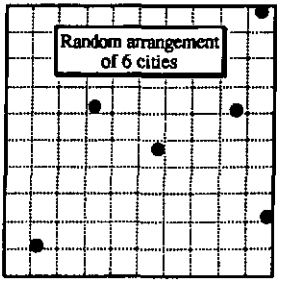


Fig. 3.10 Random arrangement of 6 cities.

For each of the random city configurations, a batch of 10 trials are performed. In each trial both the optical system and the software simulation are applied. The Euclidean distances of the optimized solutions for each trial are recorded. Upon completion of the batch, the performance of the optical implementation and the software simulation can be compared based on their statistical results. The results are quantitatively described by the Euclidean distance of the optimized tour, as shown in Table 3.1, where each distance corresponds to a specific tour of the 6 cities presented in Fig. 3.10. The maximum and average dispersion are then calculated following Eq. (3.5).

	Euclidean distance	
Trial #	Optical system	Simulation
1	3.118	3.141
2	3.118	3.118
3	3.118	3.141
4	3.141	3.141
5	3.118	3.180
6	3.141	3.141
7	3.118	3.180
8	3.180	3.141
9	3.118	3.141
10	3.141	3.118

Table 3.1 Statistics of batch run.

The example in Table 3.1 has a maximum dispersion of $\delta_{max} = 2.0\%$. The average dispersion is $\delta_{avg} = 0.4\%$ for the optical system, and $\delta_{avg} = 0.8\%$ for the simulation. The results shown in Table 3.1 are more informative when

presented graphically as in Fig. 3.11. The optimized solutions are all very close in length, as confirmed by the small values of the dispersion terms. The actual tours associated with the Euclidean distances are illustrated in Fig. 3.12.

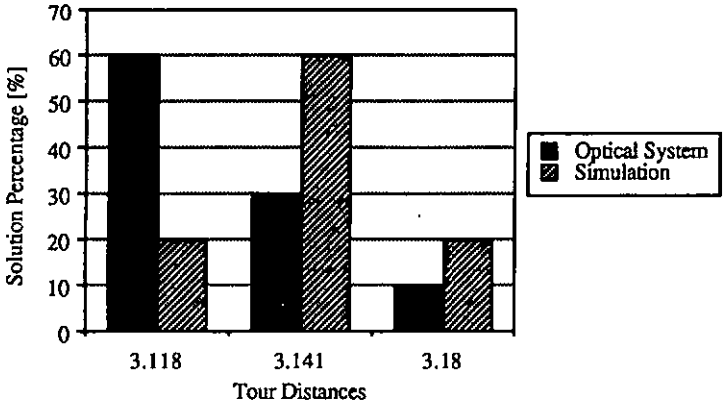


Fig. 3.11 Graphical representation of batch results.

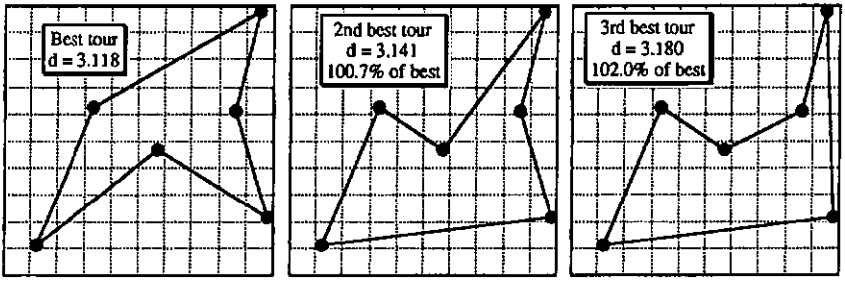
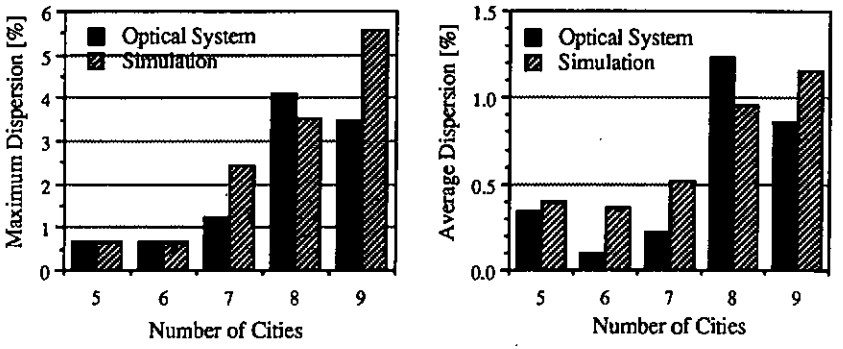


Fig. 3.12 The three optimized tours.

In the example tour described above, it can be seen that the optical system performed as well as the software simulation, if not better. Although, it is clear that the sampling of these measurements limits the precision of the statistical results, a qualitative comparison of the results is possible. There is also an influence of the results depending on the specific arrangement of the cities. To somewhat remove this influence, five different randomly generated city patterns were analyzed and compared for each number of cities. The accumulated results of the 10 trials for each city arrangement are used to determine the average response of the Kohonen algorithm for the specified number of cities. By averaging the results over various arrangements, the individual effects of a specific city arrangement are removed. The batch

results for the five arrangements of 6 cities yield average values of 0.7% for the maximum dispersion, 0.1% for the average dispersion of the optical system, and 0.4% average dispersion of the software simulation.

As the number of cities is increased from 5 to 9, the dispersion of the experimental results also increases. This is to be expected, due to the decreasing number of output neurons per city. For the results presented here, using 13 output neurons, the system is well adapted for up to 6 cities. As the number is increased beyond 6 cities, the average values for the dispersion terms also increase accordingly, as seen from Fig. 3.13.



(a) (b)

Fig. 3.13 Dispersion results for city patterns of 6 cities. The batch values of the maximum dispersion (a) and the average dispersion (b).

The fact that the average dispersion does not increase dramatically as the number of cities is increased is an indication that the Kohonen algorithm continues to find *good* solutions which are close to the *best* solution. The results also indicate that there exists no significant difference between the performance of the optical implementation and the software simulation of the Kohonen algorithm. It can be concluded by these results that the optical implementation does not degrade the performance of the algorithm. The software simulation confirms that the deviation of the results from the *best* tour are due to the application of the algorithm itself, and are not due to the optical system used to implement it.

Resolution tours

It was seen in the experimental results of the random tours, that the performance of the algorithm can be greatly influenced by the configuration

of the cities in the tour. This effect was, to some extent, removed from the results of the random tours by averaging the results of many configurations. Some of the random tours tested produced large deviations from the *best* tour, while in other tours the only solution found during the series of trials was the *best* solution.

To test the algorithm's ability in selecting the *best* tour, a specific arrangement of cities shown in Fig. 3.14(a) is used. This configuration has two equidistant *best* solutions, Figs. 3.14(b) and (c). When the city indicated by the * is displaced, the two tours indicated in Figs. 3.14(b) and (c), are no longer equidistant. One of the tours is preferred over the other, depending on the direction of the displacement of the city. The resolution of the system, for this specific tour, can be defined as the percent difference needed between the *best* tour and the *second best* tour, so that the system is able to always find the *best* solution.

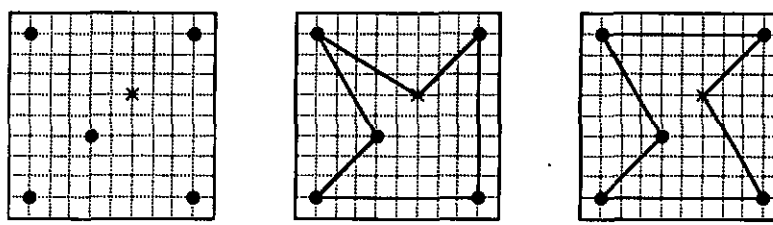
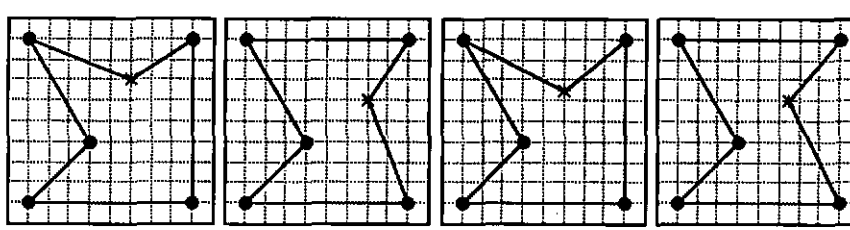


Fig. 3.14 Initial arrangement of 6 cities for resolution measurements (a) and the 2 equidistant solutions (b) and (c).

Initially, the city * is displaced vertically a relative distance of 0.1, as shown in Fig. 3.15(a). The length of the *second best* tour is 3.8% greater than the *best* tour. Both the optical system and the software simulation find the *best* tour for every trial in a series of 25 trials. If similarly, the city * is displaced horizontally the same relative distance of 0.1, as shown in Fig. 3.15(b), the results are identical to those for the vertical displacement. Thus, neither the



(a) (b) (c) (d)

Fig. 3.15 The 4 modified city arrangements and their best tours.

optical system nor the simulation have a preference for one of the two tours of Figs. 3.14(b) and (c).

When the displacement of the city is reduced to 0.05, Figs. 3.15(c) and (d), the software simulation still is able to find the *best* solution for all of the 25 trials. However, the optical system fails to find the *best* solution once out of the 25 trials for both the vertical and horizontal displacements, corresponding to a success rate of 96%. With the displacement of 0.05, the difference between the lengths of the *best* and the *second best* tours is 1.9% of the length of the *best* tour. It can be concluded from these results that the resolution of the algorithm in the optical implementation is approximately 2.0% for city configurations of 6 cities. This value is confirmed in the results of the random tour, where the maximum dispersion for all of the batch runs with 6 cities is less than 2.0%. Thus, in both the random tours and the resolution tours, the optical system never deviated to a solution with a length that was 2% longer than the *best* solution. This value can be assumed to be the resolution of the optical implementation of the Kohonen algorithm for the traveling salesman problem.

5. Limitations of optical implementation

The favorable comparison of the experimental results of the optical implementation with the software simulation suggests that the performance of the algorithm is in no way deteriorated by the optical system. However, the computational capacity of the optical demonstration is severely limited in its temporal response. Also, the experimental realization demonstrated only a limited number of output neurons, and therefore the number of cities per tour were severely limited.

The speed at which each city can be presented and the appropriate modifications of the feature map achieved is limited by the response time of the LCTV and the PC control of the PDA. The twisted-nematic liquid crystal has a limited response time on the order of tens of milliseconds. After the feature map on the LCTV has been modified, it is necessary to allow enough time for the liquid crystal to respond before reading the intensity at the photodetectors. Also, the readout of the PDA and the modifications of the feature map on the LCTV are slow due to their control by the PC.

To improve the feedback of the information from the PDA to the LCTV it is necessary to remove the PC from the loop. An array of electronic comparators at the detector level could be used to select the winning neuron in

a winner-take-all circuitry, and the resulting modifications of the feature map could be directly connected to the memory buffer of the LCTV. The LCTV described in this work is a serial device which is addressed by a video signal, therefore the display is refreshed only at video rates. Yet, since the response time of the nematic liquid crystal is comparable to video rates, then there is no net gain by individually addressing the pixels of the display in parallel. An arrangement where a liquid crystal device which has fewer pixels, but which are individually addressable, would be of interest if a ferro-electric liquid crystal is used. The ferro-electric liquid crystal has a much faster response time than the nematic liquid crystal. The problem is that, for the moment, ferro-electric displays are not as far advanced as the nematic displays and their commercial availability is only on the prototype level. Still, by directly connecting the memory buffer of the LCTV with the output of a winner-take-all circuitry, the speed of the system could be greatly improved by a factor of ten, from the 500 msec/update of our system to the limiting 50 msec/update of the LCTV response.

The computational capacity of the optical system was limited in its experimental realization due to the availability of uniform diffraction gratings. Thus, the parallelism of the optical implementation was poorly exploited. The capacity of the system can be increased simply by replacing the grating used with another grating yielding more diffraction orders of uniform intensity. As the capacity is increased, the system is able to handle more complicated tours.

The extent to which the system capacity can be increased in its one-dimensional structure is finally governed by the limitations associated with fabricating gratings with a large number of uniform diffraction orders. However, the capacity of the optical system can be greatly increased by the straightforward modification of the system to a two-dimensional implementation. The LCTV, with its 480 x 440 pixels, is a two-dimensional display with an extremely large capacity when applied to the TSP. The one-dimensional approach has not fully exploited the capacity of the LCTV nor the parallelism of the system architecture. The modification of the fan-out element to a two-dimensional grating or two one-dimensional gratings, one rotated 90° with respect to the other, to produce a two-dimensional distribution at the LCTV, will greatly increase the computational capacity of the optical implementation. Already with 2 crossed gratings of 13 diffraction orders which produce 169 diffraction orders, the 84 output neurons could handle tours of up to 42 cities. If two input sources are used as was the case in these experiments, then the number of cities could be increased to 84. Thus,

although the results presented were of a very limited computational complexity, the architecture of the optical system can be easily exploited to handle tours of much greater complexity.

6. Conclusions

The experimental results achieved with the optical implementation of the Kohonen algorithm compare very favorably with the results of the software simulation. The results of the optical system, and their comparison with the simulation results, indicate that the performance of the optical system is in no way inferior to that of the simulation. It remains probable, that the performance of the optical system and the simulation could both be improved by a more carefully studied selection of the operating parameters. However, the concern of this work was with the optical implementation of the algorithm and not with the optimization of the operating conditions associated with the algorithm itself. Thus, only a limited effort has been made to optimize the parameters used for the experimental results, the main concern being the comparison of the performance of the optical system with that predicted by the simulation of the algorithm.

The experimental results were achieved for only a very limited number of cities per tour. The extension of the optical system to a two-dimensional architecture is straightforward, thus allowing a major increase in the system's computational capacity. The results of the system will finally be limited by the Kohonen algorithm itself. As seen in the simulation results, the algorithm is very sensitive to the arrangement of the cities. Thus, as the complexity of the city tours increases with the growing number of cities, the application of the algorithm will have to be modified. It will be necessary to decrease the gain parameter and increase the number of cycles so that the algorithm will be able to resolve the more complicated tours. However, the ability of the optical implementation of the algorithm will not be modified, only the application of the algorithm itself.

The fact that the modifications of the feature map are exact in the simulation but approximate in the optical implementation made almost no significant difference in the experimental results. Using a constant value C_{LCTV} to represent the nonlinear response of the LCTV, provided an approximate variation to the weights of the feature map. Yet, due to the iterative nature of the algorithm any over or under modifications of the weights were subsequently corrected in later iterations. The optical implementation did not

in any way inhibit the performance of the algorithm. Therefore, although the exact aspect in which the Kohonen algorithm is used to resolve the TSP may need to be reexamined, the optical system will remain an interesting parallel implementation of the algorithm and a good approach to handle problems involving many iterations, such as the traveling salesman problem.

Finally, as discussed in the section concerning the system limitations, the main improvement of the proposed optical implementation would be to remove the PC from the feedback loop. A more efficient technique should be exploited to realize the winner selection at the output plane. Also, the resulting modifications of the feature map should be connected in parallel from the winner selection device to the LCLV. The winner-take-all function can be easily realized with VLSI techniques. Recently, an electro-optical realization of the winner-take-all operation has been proposed.[3.6] This device is a combination of VLSI circuitry and ferro-electric liquid crystal technology. An integrated device of this type, where electronics is used to realize the complex winner selection operation, will be the most likely solution for these types of neural systems. Optical techniques can be exploited to widely distribute the information throughout the system, but all-optical devices capable of realizing the complex thresholding operations are not available. The best altemative will be integrated devices, such as smart pixels, where optical detectors and modulators address the optical information and electronic circuitry realizes the complex thresholding functions.

References

3.1 T. Kohonen, *Self-Organisation and Assaciative Memory, Vol. 8* (Springer-Verlag, Berlin, 1984).

3.2 T. Kohonen, "The 'Neural' Phonetic Typewriter", IEEE Computer Magazine (1988).

3.3 J. C. Fort, "Solving a combinatorial problem via self-organizing process: An application of the Kohonen algorithm to the Traveling Salesman Problem", Biol. Cybern., Vol. **59** (1988).

3.4 O. Landolt, "An analog CMOS implementation of a Kohonen network with learning capability", International Workshop on VLSI for Neural Networks & Artificial Intelligence, Univ. of Oxford, (Sept. 1992).

3.5 J. J. Hopfield and D. W. Tank, "'Neural' computation of decisions optimization problems", Biol. Cybern., Vol. 52 (1985).

3.6 K. Wagner and T.M. Slagle, "Optical competitive learning with VLSI/liquid-crystal winner-take-all modulators," Appl. Opt. 32, 1408-1435 (1993).

IV. Design optimization using an optical parallel processor

Simulated Annealing

1. Combinatorial optimization - Simulated annealing

Many optimization problems can be defined by the minimization or the maximization of a function of numerous independent variables. This function is described as the cost function of the problem and is a quantitative representation of the performance of some complex system. The actual value of the cost function depends on the detailed arrangement of the multiple parts of the complex system. The number of variables may range up to many thousands, depending on the system and its resolution.

In most heuristic methods, an iterative approach is applied to the optimization problem. The system is initialized to a known configuration and a standard rearrangement of the system variables is effected, until a modified configuration which improves the cost function is found. The modified configuration then replaces the previous configuration, and the process continues until no further improvement can be found. The iterative approach searches through the variables for rearrangements which continuously reduce the cost function. These type of approaches are often referred to as downhill algorithms and are plagued by local minima, i.e. an arrangement of the system variables which corresponds locally to an optimum solution, yet is not the global optimum. To compensate for this possibility, the optimization process is carried out several times beginning from various randomly generated configurations, and the best solution is retained.

Simulated annealing is an optimization approach which provides a means to escape from a local minimum, although there is no guarantee of finding the global minimum.[4.1] The simulated annealing optimization algorithm is physically analogous to the temperature treatment of metals, in which the low temperature state of the metal is achieved by careful annealing. First the metal is heated until completely melted, then the system is slowly cooled allowing the metal to reorganize its structure at each temperature.

The simulated annealing algorithm is essentially a series of iterative modifications performed at a sequence of finite temperatures. The decision

whether the modification is accepted or rejected is based on the Metropolis algorithm, which is essentially an adaptation of the Monte Carlo scheme. The Metropolis criterion, which was originally used to describe the entropy of a system of atoms at a given temperature, provides the mechanism by which the system is capable of escaping from local minima.[4.2] In each step of the algorithm, a small random modification of the system configuration is made and the resulting change in the cost function ΔC is calculated. If $\Delta C \leq 0$, then the modification is accepted and another modification is made. If however $\Delta C > 0$, then the probability that the modification is accepted is defined as

$$P[\Delta C] = \exp\left(\frac{-\Delta C}{T}\right), \tag{4.1}$$

where T is the effective temperature. The implications of the probability term are that the optimization procedure always has the possibility of escaping from a local minimum at nonzero temperatures. The downhill algorithms are a subset of simulated annealing with an effective temperature $T = 0$.

In applying the simulated annealing algorithm to the combinatorial optimization of a complex system, the effective temperature T serves as a reference to control the optimization procedure and is defined in the same units as the cost function. Typically, the temperature is initially set to a value a few times larger than the average increase of the cost function due to a standard system modification. This is normally sufficient to achieve a "melting" of the system configuration. The temperature is then cooled exponentially, with a large number of attempted changes at each temperature. When a minimum number of acceptances does not occur for a series of three successive temperatures, then the optimization is complete and the system is considered to be "frozen". The global organization of the system is achieved at higher temperatures, whereas the finer local details are developed at lower temperatures.

To yield good results, the simulated annealing process requires a large sampling of the parameter space at each temperature. Also, the cooling of the temperature should be very gradually to avoid a quenching of the parameters into a less optimum configuration. Consequently, the optimization process requires long calculation times even on powerful computers.

2. Binary diffractive fan-out elements and their optimization using simulated annealing

Optical fan-out elements that produce a regular two-dimensional replication of an input source or image have found application in optical parallel processing systems and various sensor applications. In most applications, the replicated images are required to be very uniform in intensity, while keeping the losses associated with the fan-out element to a minimum. The demand for high efficiency imposes a phase-only structure. A phase grating can be used to produce a space-invariant replication of a coherent input beam in the Fourier domain with little losses and good uniformity.

Continuous phase variation yields the best results. Continuous phase profiles, often called kinoforms, are theoretically capable of almost 100% efficiency with negligible uniformity deviation.[4.3] Consequently, the actual performance of the kinoform is limited by the fabrication techniques available. The fabrication of kinoform structures requires either direct writing of the structures in photoresist using e-beam or laser beam writing systems, or holographic realization using interferometric techniques. An approximation to the kinoform structure may be achieved using discrete phase steps. The performance of these structures, while inferior to the kinoforms, is still capable of very high efficiency, depending on the number of discrete phase levels. In the extreme case, where only two phase levels are used in the unit cell, it is still possible to achieve approximately 80% efficiency. Binary phase elements have the advantage that their realization does not involve the alignment process associated with multi-level phase elements, and thus their fabrication is relatively simple. However, the arrangement of the intensity distribution amongst the various diffraction orders is constrained to be symmetric about the zero order, i.e. $I_{-x,-y} = I_{x,y}$.[4.4,4.5] In situations where the desired intensity distribution is not compromised by the imposed symmetry, binary phase elements are a good solution due to their relatively high efficiency and ease of fabrication.

The grating is made up of the two-dimensional periodic repetition of a single unit cell pattern. The intensity distribution of the various diffraction orders is determined by the form of the binary phase variation within the unit cell. The unit cell is made up of many thousands of pixels of unity amplitude transmittance and either 0 or π phase delay, i.e.

$$\text{unit cell} = \sum_{m,n=1}^{M,N} A_{mn} \exp(i\phi_{mn})\, \delta(x-x_m, y-y_n)\,, \tag{4.2}$$

where $A_{mn} = 1$ and $\phi_{mn} = 0$ *or* π. When the grating contains a large number of the unit cells, its Fourier transform yields a 2-D array of delta functions, i.e.

$$\text{F.T.}\{\text{grating}\} = \sum_{u,v=1}^{U,V} A_{uv} \exp(i\phi_{uv})\, \delta(x-x_u, y-y_v)\,, \tag{4.3}$$

where A_{uv} and ϕ_{uv} correspond to the amplitude and phase, respectively, of the uth, vth diffraction order. The terms x_u and y_v correspond to the separation of the diffraction orders in the x and y directions, respectively. The Fourier transform is generated in the back focal plane of a positive lens when the grating is placed in the front focal plane. Using the small angle approximation that $\sin(\theta) \approx \theta$, the spacing of the diffraction orders can be represented as

$$x_u = \frac{\lambda f}{X}\,,\; y_v = \frac{\lambda f}{Y}\,, \tag{4.4}$$

where f is the focal length of the lens, λ is the wavelength of the illumination, and X and Y are the dimensions of the unit cell.

The repetition of the unit cell produces the δ functions, while the configuration of the pixels of the unit cell determines the envelope function which governs their intensities, see Fig. 4.1. By varying the phase of individual pixels of the unit cell, the form of the envelope function can be modified to accommodate the desired distribution of the input intensity amongst the various diffraction orders. The envelope generated by a binary phase element is symmetric about the zero order as illustrated in Fig. 4.2.

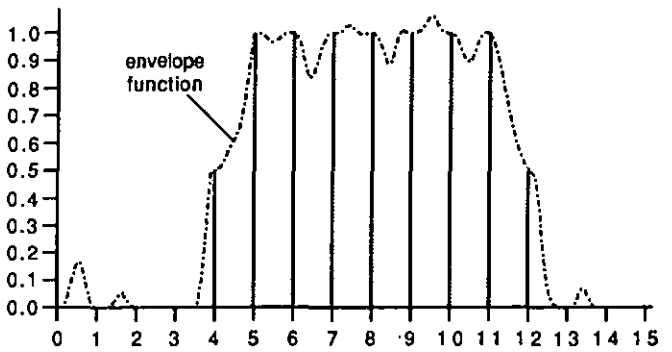


Fig. 4.1 Diffraction orders resulting from the 1-D repetition of a unit cell.

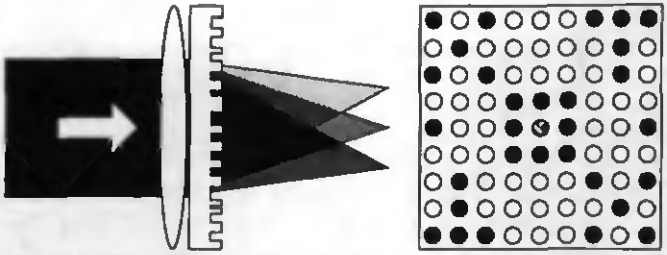


Fig. 4.2 Symmetrical intensity distribution resulting from 2-D binary phase grating.

Originally, a pair of orthogonally crossed I-D binary phase gratings were used to produce the desired 2-D intensity distribution.[4.6] The results were plagued by low efficiency and the imposed limitation of x and y symmetry of the intensity distribution. Each 1-D grating may have an efficiency of approximately 80%, thus yielding only about 65% efficiency for the crossed pair. If an inseparable design for the 2-D intensity distribution is used it is possible to increase the efficiency of the optimized element by 10 - 20%.[4.7]

The unit cell of an inseparable design may be effectively optimized using heuristic methods.[4.8] The desired intensity distribution of the Fourier transform defines the cost function which is used to optimize the unit cell using either downhill or simulated annealing algorithms. The individual pixels of a randomly generated unit cell are selectively modified, until the algorithm converges to an optimized solution. The initial cost function is calculated by taking a fast Fourier transform (FFT) of the random unit cell. The effects of subsequent modifications of individual pixels of the unit cell on the cost function can be calculated analytically.[4.9]

The values of A_{uv} and ϕ_{uv} of Eq. (4.3) may be written as the summation of the contribution of each pixel of the unit cell, namely

$$A_{uv}\exp(i\phi_{uv}) = \frac{1}{MN}\sum_{m,n=1}^{M,N} A_{mn}\exp(i\phi_{mn})\exp\left(2\pi i\left(\frac{mu}{M}+\frac{n\,v}{N}\right)\right), \qquad (4.5)$$

where N is the number of pixels of the unit cell in each direction. Since the optimization is concerned only with the intensity of the diffraction orders, the phase term is not controlled during the optimization process. The change in the amplitude ΔA_{uv} of the diffraction order u,v due to the modification of pixel m,n may then be given by

$$\Delta A_{uv} = \sum_{m,n=1}^{M,N} \frac{(A'_{mn} - A_{mn})}{MN} \exp\left(2\pi i\left(\frac{mu}{M} + \frac{n\,v}{N}\right)\right), \qquad (4.6)$$

where A'_{mn} is the new transmission term of the modified pixel. In the special case of binary phase elements, each pixel has unity transmission and phase values of 0 or π, corresponding to $A_{mn} = 1$ *or* -1, respectively. The modified cost function is calculated after the modification of each pixel.

In many cases a simulated annealing algorithm is employed to find a high efficiency design, and a downhill algorithm is then used to improve the uniformity of the optimized diffraction orders.[4.7,4.9] While the simulated annealing algorithm is very effective in avoiding local minima, it requires many iterations and is very time consuming. Thus, once a high efficiency design has been achieved a downhill algorithm is employed to further improve the uniformity of the optimized orders.

To begin the design process, each pixel of the phase structure is randomly initialized to values of 0 or π. The phase value of a randomly selected pixel or a group of pixels within the phase structure is inverted from 0 to π or from π to 0. The decision whether to accept the modification or not is based on the simulated annealing Metropole algorithm. The cost function of the annealing algorithm is based on the desired distribution of intensity, the overall efficiency, and the uniformity of the element. The definition of the cost function includes the sum of the total deviation of the actual intensities of the diffraction orders to be optimized from their desired values and the contrast of their relative intensities, namely

$$\text{Cost} = \sum_{\text{opt orders}} (I_{meas} - I_{desired})^2 + \text{Contrast}\,, \qquad (4.7)$$

where the *Contrast* term is defined as

$$\text{Contrast} = \text{Wgt}\,\frac{I_{max} - I_{min}}{I_{max} + I_{min}}\,. \qquad (4.8)$$

The first half of Eq. (4.7) optimizes for high efficiency based on the values of $I_{desired}$ and also influences the uniformity. The second half of the equation is the contrast of the intensity variation of the optimized orders and is a measure of the uniformity of the design. The contrast is multiplied by a relative weighting term *Wgt*, to adjust its contribution to the cost function. During the simulated annealing process, the contrast at first has little influence on the magnitude of the cost function. Towards the end of the optimization, when the efficiency of the design has been improved, the contrast becomes more important.

If the cost function is reduced the modification of the unit cell is accepted. However, when the modification produces an increase in the cost function, the probability that the change will be accepted is based on the magnitude of the cost increase and the current temperature of the algorithm and is given by

$$P[\text{accepted}] = \exp\left(\frac{-\Delta \text{Cost}}{\text{Temp}}\right). \tag{4.9}$$

The temperature associated with the annealing decision process is continuously reduced until the cost function reaches a minimum value and shows no further signs of decreasing. At this point the simulated annealing algorithm is terminated.

The second step of the optimization process applies a downhill algorithm to improve the uniformity of the intensity distribution in the Fourier transform of the already highly efficient design. This part of the optimization procedure involves relatively few iterations and does not need the long computational time associated with the simulated annealing algorithm. The purpose of the downhill algorithm is to improve the uniformity while maintaining the high efficiency of the design.

In this work an optical parallel processor is used to perform the simulated annealing stage of the optimization procedure. A liquid crystal television (LCTV) is used to actively modulate the phase of individual pixels within the base unit cell of the diffractive element. Once the element has been optimized by the optical processor for high efficiency, a downhill algorithm on the mainframe computer is employed to improve uniformity, as before. This last stage also serves to remove any imperfections in the design introduced during the optical optimization.

3. Optical implementation

The optimization of the unit cell of the phase fan-out element is realized using a relatively simple optical parallel processing system, Fig. 4.3. The most important element is the LCTV, which will be discussed in detail below. In a similar application these LCTVs have been demonstrated as efficient programmable phase kinoforms.[4.10] The Fourier transform of the phase information encoded within the television is achieved by traversing the LCTV with a collimated laser source (Argon laser at $\lambda = 488$ nm) and measuring the intensity in the focal plane of a Fourier transform lens with focal length $f = 380$ mm. Also in the Fourier plane, a diffraction pattern resulting from the pixellated structure of the LCTV is observed. The zero order of this

diffraction pattern is imaged onto a two-dimensional array of photodetectors using a 10x microscope objective with a numerical aperture of 0.25. The other diffraction orders are blocked by the spatial filter. The lateral magnification of the image produced by the microscope objective is adjusted to correctly scale the Fourier transform with respect to the photodetector array. The phase information of the Fourier transform is lost since only its intensity is being measured. Yet, for the optimization of fan-out elements, only the intensity variation of the diffraction orders is of interest and the phase information is not of concern.

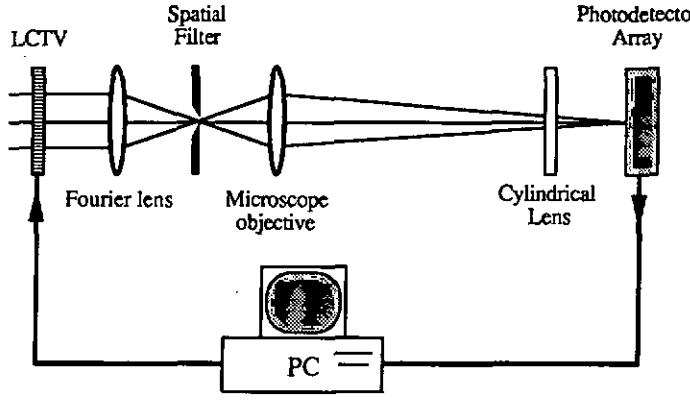


Fig. 4.3 Schematic of optical parallel processor.

It is critical that the lateral dimensions of the Fourier transform are correctly scaled with respect to the photodetector array, so that one photodetector corresponds to one diffraction order. The spacing of the LCTV pixels is 78 μm in the x direction and 90 μm in the y direction. Consequently, the unit cell is slightly longer than it is wide, which results in a difference between the separation of the diffraction orders in the x and y directions. Therefore a cylindrical lens of focal length 200 mm is positioned 1 cm before the photodetector array to slightly adjust the magnification in the x direction with respect to the y direction so that the spacing of the diffraction orders corresponds to the square spacing of the photodetectors. Thus, to optimize the intensity within a certain diffraction order, the intensity measured by the corresponding photodetector is incorporated into the cost function of the simulated annealing algorithm. The intensity impinging on the photodetector array is read out using an IBM compatible DX-386 personal computer (PC).

Besides reading the photodetector array, the PC also serves to calculate the cost function and to modulate the LCTV. The PC controls the data

input/output operations and the evaluation of the cost function for each iteration of the simulated annealing algorithm. First, the PC modifies the phase of a group of pixels of the unit cell by modifying the video signal sent to the LCTV. Then the photodetector intensity is read into the PC using the Direct Memory Access (DMA) mode of the Keithley MetraByte DAS-16G board. The measured intensities which correspond to the diffraction orders to be optimized are used to calculate the new value of the cost function resulting from the modification of the LCTV pixels. The Metropolis decision process on whether to accept the modification or not is realized by the PC. If the modification is accepted, then the PC records the new value of the cost function and begins the next iteration by modifying the phase of another group of pixels in the LCTV. If the modification is not accepted, then the PC resets the phase of the group of LCTV pixels that were modified and then begins the next iteration. The PC is also responsible for terminating the optimization process and for storing the optimized results.

After completion of the simulated annealing optimization with the optical processor, the resulting unit cell is stored to be used as the starting point for further improvement on a mainframe computer or a powerful personal computer using FFT algorithms. The application of the downhill algorithm improves the uniformity while maintaining the high efficiency of the design. It also serves to remove any noise from the design due to the optical system itself. Noise effects due to the optical variations of the beam and the electronic noise of our photodetector array will be removed during the downhill optimization using the FFT.

Phase modulation using a LCTV

The spatial light modulator used is one of three LCTVs originally employed in a Seiko Epson Video Projector VPJ-700. Each of the LCTVs contains 320 x 220 pixels, where each pixel measures 78 μm x 90 μm. The electronics of the projector are still used to control the television, while the video signal is generated by the PC using a Data Translation Quick Capture framegrabber board. The video signal is modified by writing to the memory buffer of the framegrabber board. The LCTVs are designed to perform as intensity modulators by effectively providing a 90° rotation of the input polarization. With these devices both a very high contrast ratio and good gray scale performance are possible.[4.11] In the optimization of binary phase elements it is necessary to modulate the phase of the impinging beam without modifying its polarization. This is possible if the LCTV is operated at a lower

voltage range.[4.12] Originally the displays are driven by an AC signal with an amplitude varying between 2.9 - 4.2 Volts. By reducing the amplitude of the drive voltage down to the range 1.8 - 2.6 Volts, we are able to achieve 2π phase modulation with only very slight rotation of the polarization. The phase of each pixel may be independently modulated between 0 and 2π, while the amplitude transmittance is only slightly modulated (p-p variation of 8.3%), as shown in Fig. 4.4. Furthermore, since we are simulating binary phase elements, we are modulating the phase between 0 and π. The intensity transmission of the LCTV is essentially the same for these two phase values.

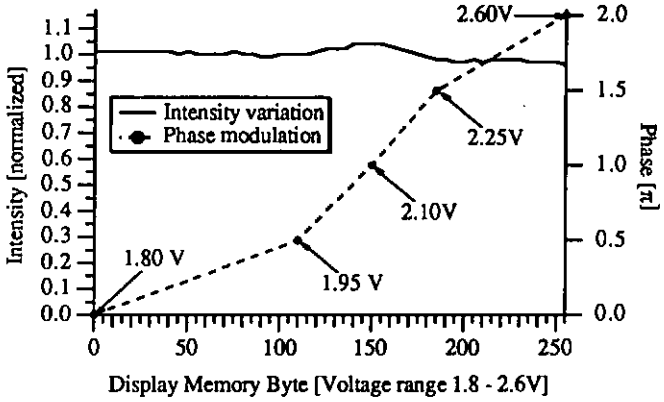


Fig. 4.4 LCTV intensity variation over 2π phase shift.

The ability of the LCTV to modulate phase is shown using a Mach-Zehnder interferometer. The LCTV is inserted in one of the arms of the interferometer and a rectangular region of the LCTV is shifted in phase by π, Fig. 4.5. For the simulation of the phase element to be sufficiently precise, the uniformity of the LCTV must be of high quality. We have measured a phase variation of approximately one fringe (λ = 488 nm) across the entire LCTV surface. To make this measurement, the reference beam of the Mach-Zehnder interferometer is adjusted to compensate the beam transmitted by the LCTV, Fig. 4.6a. When the LCTV is removed, one fringe of phase difference is observed, Fig. 4.6b. Since we are using only a portion of the screen, the phase variation across our unit cell is even less than 2π. The fringes of Figure 4.6b, illustrate a mostly linear phase variation across the display, which results in a slight translation of the Fourier transform in the detector plane. Since in our optical processor the detector is aligned with respect to the beam, this translation effect is insignificant. The magnitude of higher order phase

variations are negligible. Thus, the LCTV is perfectly suited to function as an electrically-addressable phase modulator.

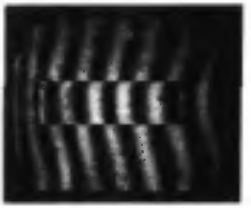


Fig. 4.5 LCTV in Mach-Zehnder interferometer displaying π phase modulation.

(a) (b)

Fig. 4.6 Phase quality of LCTV. Mach-Zehnder interferometer with LCTV (a) and without LCTV (b).

Optical noise associated with one unit cell

The optimization procedure is always applied to the base unit cell. When using FFT techniques the intensity of the diffraction orders is sampled at discrete points and any variation of the envelope function or interference between the orders is not considered. When the number of unit cells that are illuminated is large, the orders are represented by δ functions. However, when only a few unit cells are illuminated the orders are represented by two-dimensional *sinc* functions centered about the δ functions. Neighboring orders have considerable overlap of their *sinc* functions and thus interference effects are observed. In the limiting case, where only one unit cell is illuminated, there are no diffraction orders, only the envelope function which governs the amplitude of the various diffraction orders. The resulting intensity distribution in the Fourier plane displays a random variation of bright and dark spots due to the variation of the envelope function, Fig. 4.7. The intensity variations in the Fourier plane severely limit the performance of the optical parallel processor.

Fig. 4.7 Random intensity variation in Fourier plane due to the variation of the envelope function.

The variation of the envelope function and the interference of the overlapping orders can be easily analyzed using Fourier optics. The calculations are performed for the case of a one-dimensional grating for illustration purposes only. First, the grating can be represented as the convolution of the unit cell with an array of δ functions, called a *comb* function, i.e.

$$\text{grating} = \text{cell} * \text{comb}\left(\frac{x}{X}\right), \tag{4.10}$$

where X is the length of the unit cell. The Fourier transform of the grating is then given by

$$\text{F.T.}\{\text{grating}\} = \text{F.T.}\{\text{cell}\}\ \text{comb}(Xf_x), \tag{4.11}$$

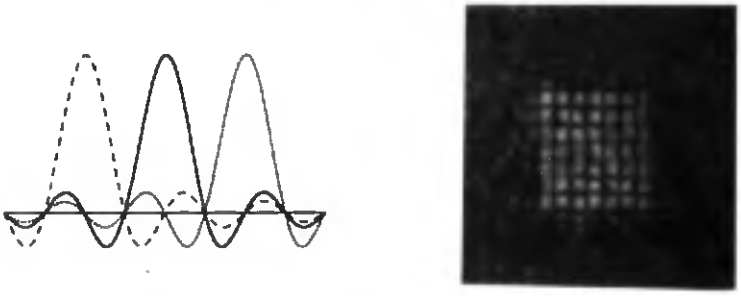
where f_x is the spatial frequency. Figure 4.8a illustrates the amplitude of the ideal Fourier transform of a grating designed for three orders of equal intensity. The unit cell can be defined as

$$\text{cell} = \text{rect}\left(\frac{x}{X}\right)\ \text{grating}. \tag{4.12}$$

The Fourier transform of one unit cell is then represented by

$$\text{F.T.}\{\text{cell}\} = \text{sinc}(Xf_x) * \text{F.T.}\{\text{grating}\}. \tag{4.13}$$

As a result, the δ functions of Fig. 4.8a are replaced by *sinc* functions whose first zeros are located at the neighboring orders, Fig. 4.8b. Thus the neighboring diffraction orders have considerable overlap and produce inhibiting interference fringes. The envelope function described in Fig. 4.8a, is determined by the interference of the overlapping *sinc* functions and depends on the phase of each diffraction order. Since the phases of the orders are not controlled during the optimization, the result is a random variation of the intensity, as witnessed in Fig. 4.7.



(a) (b)

Fig. 4.9 Amplitude of Fourier transform for the case of two unit cells (a) and an experimental example of the separation of the diffraction orders (b).

The result is a decrease in the possible resolution for the unit cell as the number of cells is increased. The subsequent limitations will be discussed in more detail in Sect. 5.

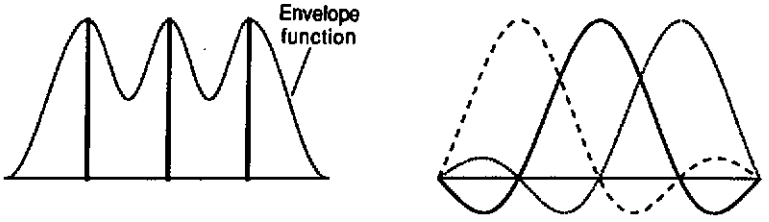
System resolution based on spot size

The resolution of the system is defined as the number of LCTV pixels that make up the unit cell. The more pixels used, the greater the resolution of the unit cell, thus allowing for the definition of smaller features within the cell pattern. Since the LCTV used to encode the unit cells contains pixels of fixed dimensions Δx, the resolution of the system is defined in terms of the number of pixels $N \times N$ which make up the unit cell. Clearly, it is desired to maximize the number of pixels in the unit cell so as to be able to resolve fine features within the design. By increasing the number of unit cells illuminated it is possible to reduce and even eliminate the overlap of two neighboring diffraction orders, while their separation in the focal plane of the Fourier lens remains fixed as

$$\text{separation} = \frac{\lambda f}{N \Delta x}, \tag{4.16}$$

where λ is the laser wavelength, f is the focal length of the Fourier lens, N is the number of LCTV pixels in the length of the unit cell, and Δx is the spacing of the LCTV pixels. The separation of the orders is inversely related to the number of pixels in the unit cell, Fig. 4.10.

The resolution of the unit cell is directly dependent upon the ability to adequately separate the diffraction orders so that they do not overlap. It is assumed that more than one unit cell is illuminated, so as to ensure that the *sinc* functions are sufficiently narrow that their overlap can be neglected, Fig. 4.9a.



(a) (b)

Fig. 4.8 Amplitude of Fourier transform of grating (a) and of unit cell (b).

To improve the performance of the optical processor the effects of the optical noise need to be reduced. It is not possible to destroy the coherence of the neighboring orders, since the coherence of the beam is precisely what produces the diffraction effects of the unit cell being observed. The only way to reduce the noise resulting from the overlapping orders is to reduce the amount of their overlap. As the number of unit cells illuminated is increased, the width of the *sinc* functions decreases towards the ideal limit of a δ function for a grating of infinite extent. By illuminating two unit cells it is already possible to greatly reduce the amount of overlap of the *sinc* functions and thus the resulting optical noise. The case of two unit cells can be expressed as

$$\text{two cells} = \text{rect}\left(\frac{x}{2X}\right)\text{grating}\ , \tag{4.14}$$

and its Fourier transform now becomes

$$\text{F.T.}\{\text{two cells}\} = \text{sinc}(2Xf_x) * \text{F.T.}\{\text{grating}\}\ . \tag{4.15}$$

As a result, the width of the *sinc* functions is reduced to one half of their value for one unit cell, Fig. 4.9a, which significantly reduces the amount of optical noise in the Fourier transform, Fig. 4.9b. The individual diffraction orders within the Fourier transform are now distinguishable and clearly separated from their neighbors.

By further increasing of the number of unit cells illuminated, the width of the *sinc* functions is correspondingly reduced. Thus further reducing the optical noise and resulting in an even better definition of each individual diffraction order. Clearly it is possible to greatly reduce the optical noise due to the interference of neighboring orders. The price to be paid is an increase in the number of LCTV pixels needed to represent the unit cells. Since all of our structures are two-dimensional, the number of unit cells is squared, i.e. 2 x 2 or 3 x 3, requiring four or nine times more LCTV pixels, respectively.

confirmed in the experimental results. For the examples illustrated in Figs. 4.7 and 4.9b, each unit cell contained 32 x 32 LCTV pixels. The results of the two figures verify that as long as the number of unit cells illuminated is greater than one, then the diffraction orders can be sufficiently separated. If the number of unit cells is increased to 6 x 6 for the unit cell with 32 x 32 pixels, the separation of the orders is slightly improved over the case for 2 x 2 unit cells, Fig. 4.12a. However, the separation of the orders is not as good as expected, due to the limitations imposed by the finite spot size. The effects of reducing the width of the *sinc* functions of Eqs. (4.14) and (4.15) are valid only as long as the width of the *sinc* is greater than the beam spot size σ. Experimental results for 2 x 2 unit cells containing 64 x 64 LCTV pixels illustrates the interference of the overlapping orders, Fig. 4.12b. Even by increasing the number of unit cells to 3 x 3, the interference of the overlapping orders can not be avoided since the diameter of the beam spot size is greater than the separation of the orders, Fig. 4.12c. Since the design of the diffractive element will be further improved using an FFT algorithm on a computer, the dimensions of the unit cells are limited to powers of two, i.e. 2^n. Therefore, as illustrated in Fig. 4.12, the unit cells containing 64 x 64 pixels will have overlapping diffraction orders which will inhibit the performance of the system due to the interference noise present in the intensity distribution at the photodetector array. Unit cells with 32 x 32 pixels, however, will produce a large enough separation of the diffraction orders and allow the optical processor to distinguish the individual orders without any interference noise.

(a) (b) (c)

Fig. 4.12 The separated orders achieved with 6 x 6 unit cells of 32 x 32 resolution (a), and the interference noise due to the overlapping orders achieved with 2 x 2 unit cells (b) and 3 x 3 unit cells (c) of 64 x 64 resolution.

4. Experimental results and analysis

The experimental results of the simulated annealing optimization with the optical processor are presented for the two resolutions discussed above, 64 x 64 and 32 x 32 LCTV pixels. While the results of the 64 x 64 resolution exhibit the interference noise due to the overlapping orders, the optical

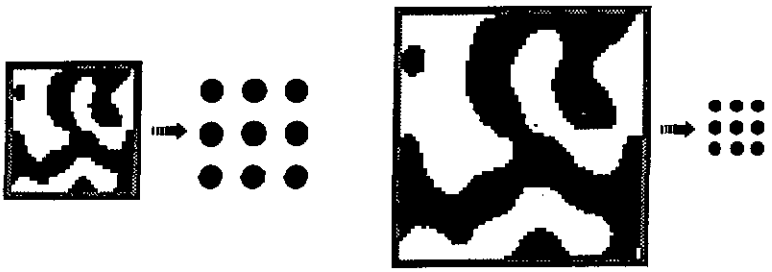


Fig. 4.10 Separation of diffraction orders versus size of unit cell.

In this situation, the minimum separation of the diffraction orders will be determined by the spot size σ of the beam in the focal plane of the Fourier lens, Fig. 4.11. Therefore, the maximum number of LCTV pixels N_{max} that a unit cell can contain is limited by the spot size according to

$$N_{max} = \frac{\lambda f}{\sigma \Delta x} . \tag{4.17}$$

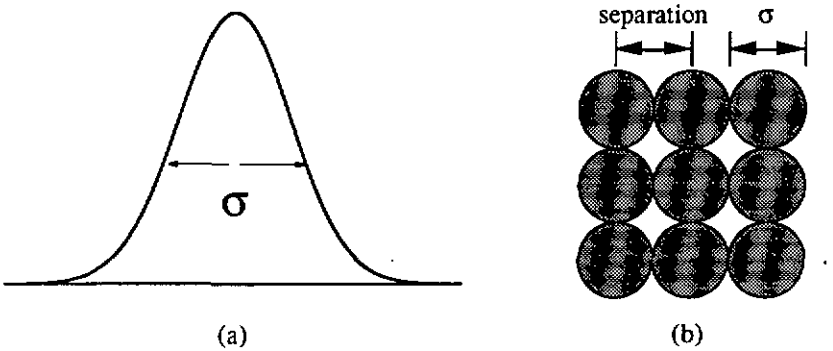
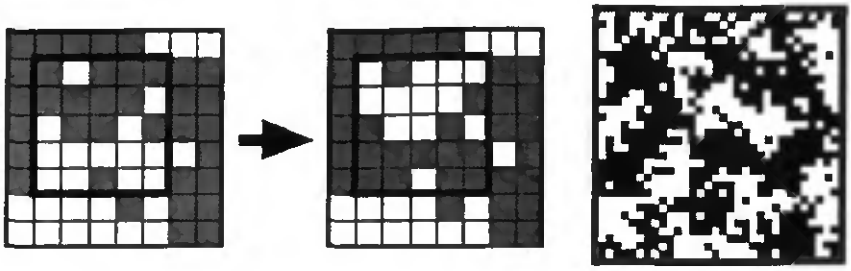


Fig. 4.11 Influence of the beam spot size (a) on the minimum separation of the diffraction orders (b).

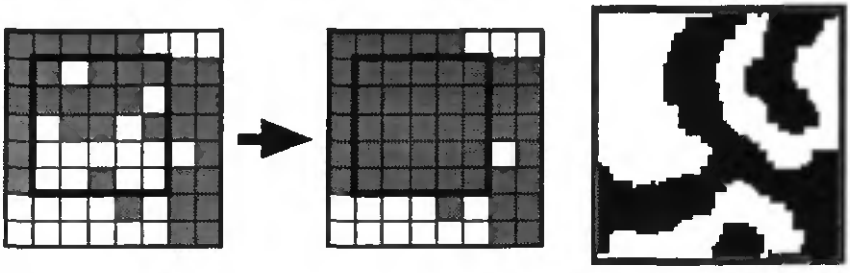
The Fourier lens has a listed resolution of 32 lines/mm, corresponding to a spot size slightly greater than 30 μm. Using a knife-edge test to measure the diameter of the spot size in the focal plane, a diameter of approximately 40 μm was found. Aberrations in the optical wavefront due to the imaging system and the optical quality of the LCTV account for the deviations of the measured spot size from the limiting resolution of the Fourier lens. The limitations imposed by the spot size are easily observed in the intensity distribution at the Fourier plane. For the experimental realization of the optical processor, the maximum resolution of the unit cell N_{max} must be less than 50 LCTV pixels, where $\lambda = 488\ \mu m, f = 380\ mm$, $\sigma = 40\ \mu m$, and $\Delta x = 90\ \mu m$. This limit is



(a) (b)

Fig. 4.13 Inversion of phase state of all pixels in a 5 x 5 block (a) and an example of the random pixels resulting in the optimized design (b).

A large improvement in the optimized designs is achieved by altering the pixel modification algorithm. The algorithm was modified to set all of the surrounding pixels to the same phase state as the center pixel of the group, Fig. 4.14a. This approach is much better adapted for the suppression of the noisy pixels. Essentially, the structures that exist are eroded or dilated, thus suppressing any isolated pixels. Consequently, the definition of the optimized designs is greatly improved as witnessed in Fig. 4.14b.



(a) (b)

Fig. 4.14 Adaptation of the phase states of the surrounding pixels to that of center pixel in a 5 x 5 block (a) and the resulting improvements in the optimized design (b).

A final adjustment of the pixel modification algorithm was introduced so that the algorithm could produce an island of pixels of one phase state in a sea of pixels of the other phase state. As the algorithm was described above, it is incapable of producing an island of phase inversion in a region of uniform phase distribution. Also, as the algorithm was described there would be times when the randomly selected center pixel was surrounded by pixels of the same phase, and thus no modification was made to the unit cell design and time was wasted during the optimization process. To allow the algorithm to produce an

optimization is still capable of iterating to a good design provided the intensity distribution in the diffraction orders desired is not too complex. The results of the 32 x 32 resolution illustrates the capabilities of the processor when the diffraction orders are well separated in the Fourier plane.

The optimized unit cell designs are further optimized using a downhill algorithm on a computer and employing a Fast Fourier Transform (FFT) routine. The performance of the optimized designs before and after the downhill optimization are compared to give an idea to what extent the optical processor is capable of iterating to a good solution. The performance of the designs are judged by their diffraction efficiency and their uniformity. The diffraction efficiency η is defined to be the percentage of the total transmitted intensity located in the optimized diffraction orders, i.e.

$$\eta = \frac{\sum I_{order}}{I_{trans}}, \qquad (4.18)$$

where I_{order} is the intensity in the optimized order and I_{trans} is the total transmitted intensity. The uniformity δ of the design is expressed in terms of the standard deviation of the intensities σ_I from the mean value $<I>$. The standard deviation is normalized with respect to the mean value, thus its magnitude is a percentage of the mean value. The uniformity δ is given by

$$\delta = \frac{\sigma_I}{<I>}, \qquad (4.19)$$

where σ_I is the standard deviation of the intensity distribution and $<I>$ is the mean value of the intensity.

Pixel modification algorithm

Initially, one pixel of the unit cell was randomly selected and its phase modified from 0 to π or from π to 0. However, the resulting changes in the value of the cost function were smaller than the amount of noise present. The change in intensity in the Fourier plane was smaller than the electronic noise associated with the reading of the array of photodetectors. Therefore, due to the limited dynamic range of the photodetectors, it was necessary to modulate a group of pixels within the unit cell, so that the effective change in the cost function would be larger than the noise level. In the initial results, the phase state of every pixel in the block of pixels was inverted as illustrated in Fig. 4.13a. However, with this approach randomly distributed noisy pixels result in the optically optimized design, as shown in Fig. 4.13b, which have to be corrected during the downhill optimization on the computer.

island structure in a region of uniform phase and to ensure that at each trial some modification was made to the existing cell design, one additional condition was joined to the pixel modification algorithm. In the case that the randomly selected center pixel is surrounded by pixels of the same phase, then all the pixels of the block are inverted. With this adjustment, the algorithm now is capable of producing an island of phase if necessary, and the design of the unit cell is modified at every trial.

Results of 2 x 2 unit cells with 64 x 64 resolution

The initial result of the optimizations using 2 x 2 unit cells of 64 x 64 resolution was realized before the final pixel modification algorithm was employed. The algorithm set all the surrounding pixels to the value of the center pixel, but did not provide the inversion of the whole block as discussed above. Therefore, the algorithm was not yet adapted to provide for the realization of islands of phase in regions of uniform phase. It was the results of this initial result that provoked the final adaptation of the algorithm.

To begin the optimization procedure the unit cell was initially set to the ambiguous pattern of Fig. 4.15a. A group of 7 x 7 pixels, centered around a randomly selected pixel in the unit cell, were modified in each of the 2 x 2 unit cells. At the completion of the simulated annealing optimization, the pattern had evolved to its final configuration, Fig. 4.15b. The results of the optical

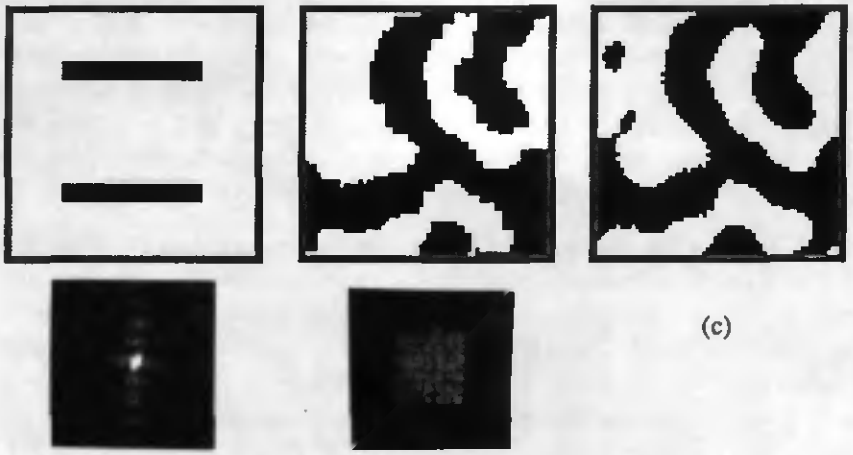


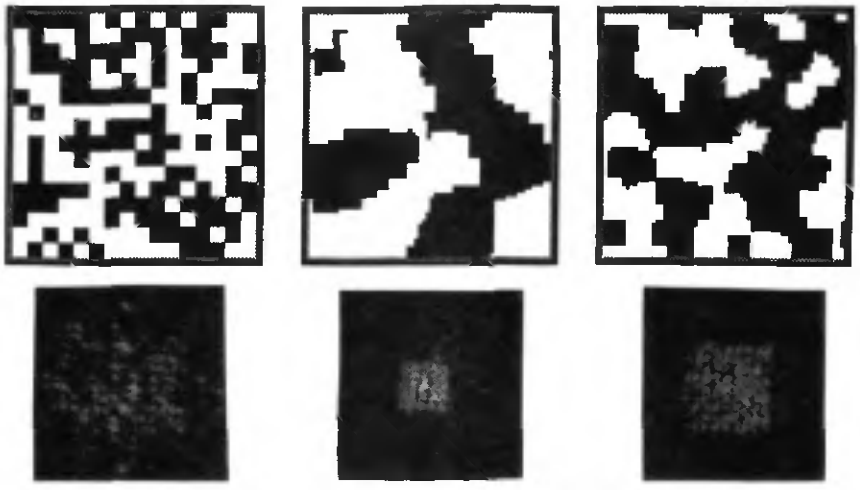
Fig. 4.15 The 64 x 64 pixel design and Fourier distribution for the initial starting point (a), the result of the optical optimization for 7 x 7 orders (b), and final result after the downhill optimization on the computer (c).

optimization provide a good starting point for the downhill optimization realized with FFT techniques on a computer. It is necessary to continue the optimization procedure on a computer to improve the uniformity of the design while also compensating for any adverse effects the optical system may have introduced into the optimized design. When analyzed using an FFT, the optically optimized result exhibited an efficiency of $\eta = 77.7\%$, and a uniformity of $\delta = 40.9\%$. After completing the downhill optimization of the design on the mainframe computer, the final design, Fig. 4.15c, was characterized by an efficiency of $\eta = 77.0\%$ and a uniformity of $\delta = 1.6\%$.

The final adaptation of the pixel modification algorithm was introduced as a direct consequence of the results for 7 x 7 orders seen in Fig. 4.15. The downhill optimization on the computer modified the optical design only slightly. The most notable change being the addition of a dark circle in the upper left comer. Using only the dilation technique, without the ability to inverse an entire block of pixels, the optical processor was unable to produce the missing dark circle. The final version of the pixel modification algorithm was applied in all of the other optimization results presented.

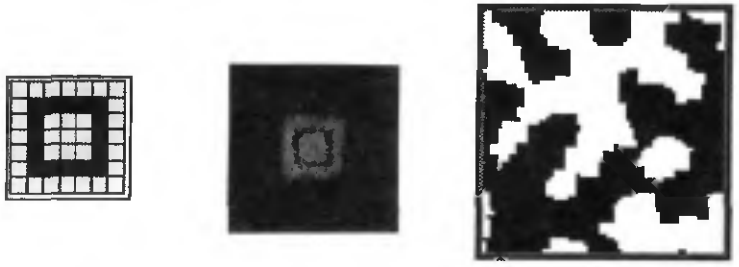
Beginning from a randomly initialized unit cell, Fig. 4.16a, the optical processor has been used to optimize the unit cell for 5 x 5 orders of uniform intensity. The optimized design achieved with the optical processor, Fig. 4.16b, is characterized by an efficiency of $\eta = 74.0\%$ and a uniformity of $\delta = 41.4\%$. The downhill optimization on the computer improves the uniformity to $\delta = 2.0\%$ while the efficiency remains relatively unchanged at $\eta = 75.7\%$. Applying the same technique to the optimization of 9 x 9 orders results in the design shown in Fig. 4.16c, which yields an efficiency of $\eta = 72.8\%$ and a uniformity of $\delta = 57.0\%$. After the downhill optimization the efficiency of the final design is $\eta = 74.9\%$ and the contrast is improved to $\delta = 2.7\%$.

For a final illustration of the capability of the optical processor using 2 x 2 unit cells with a resolution of 64 x 64, a varying distribution of the intensity was sought in the optimized orders. The desired intensity distribution is illustrated in Fig. 4.17a, where the bright pixels correspond to maximum intensity and the dark pixels to zero intensity. Even with the interference noise due to the overlapping orders, the optical processor was able to realize the desired intensity distribution, Fig. 4.17b. The optimized design shown in Fig. 4.17c, has an efficiency of $\eta = 70.6\%$ and a uniformity of $\delta = 33.9\%$. The downhill optimization improves the design to an efficiency of $\eta = 75.4\%$ and a uniformity of $\delta = 1.6\%$.



(a) (b) (c)

Fig. 4.16 The unit cell and its Fourier distribution for the random initialization (a), a fan-out of 5 x 5 orders (b), and a fan-out of 9 x 9 orders (c).



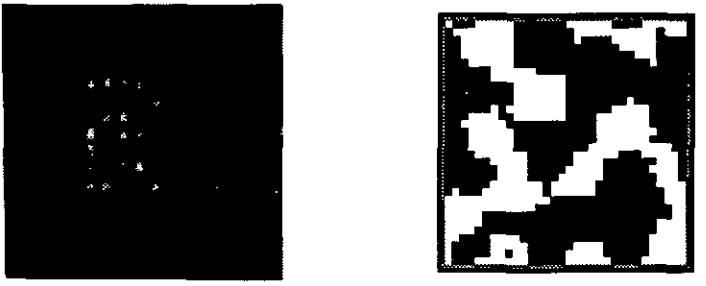
(a) (b) (c)

Fig. 4.17 The desired intensity distribution amongst the 7 x 7 orders (a) and the results achieved by the optical processor (b), with the optimized design (c).

Results of 6 x 6 unit cells with 32 x 32 resolution

The interference noise due to the overlapping orders is no longer present during the optimizations using 6 x 6 unit cells of 32 x 32 resolution. Each of the diffraction orders is well separated from the others and is imaged onto an individual detector of the photodetector array. Consequently, the intensity distribution in the diffraction orders may be optimized by the optical system into more complex arrangements. The limit to the complexity of the desired intensity distribution is determined by the resolution of the unit cell. The orders are well separated, but the minimum feature size is severely limited in the unit cell containing only 32 x 32 pixels.

Beginning from a randomly initialized unit cell, the optical processor has been used to optimize the unit cell for 7 x 7 orders of uniform intensity. The desired intensity distribution in the diffraction orders is achieved, Fig. 4.18a, by the optimized design shown in Fig. 4.18b. The optimized design is characterized by an efficiency of η = *69.1%* and a uniformity of δ = *37.3%*. The downhill optimization on the computer improves the efficiency to η = *75.1%* and the uniformity to δ = *2.4%*.



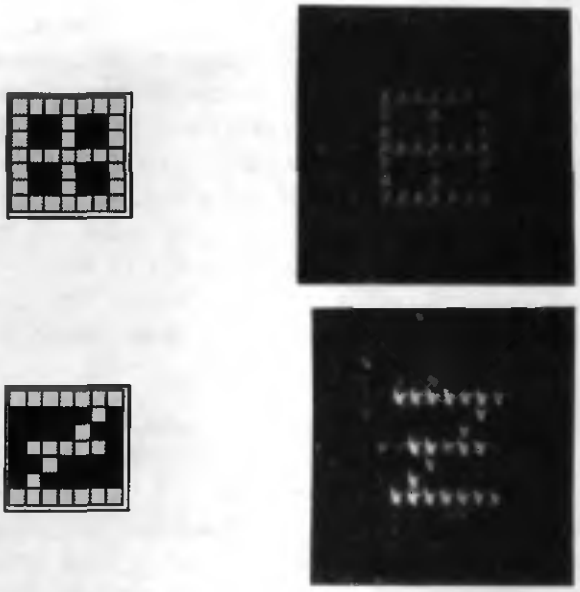
(a) (b)

Fig. 4.18 The intensity distribution in the 7 x 7 orders (a) achieved with the optimized unit cell (b).

To illustrate the capability of the optical processor using 6 x 6 unit cells with a resolution of 32 x 32, a more complex distribution of the intensity was sought in the optimized orders. The two patterns described in Fig. 4.19a were used as the target values for the optimization with the optical processor. In both cases, the processor was able to optimize the design of the unit cell so as to achieve the desired intensity distribution amongst its diffraction orders, Fig. 4.19b. The diffraction orders are well separated and good contrast is observed between the orders optimized for high intensity and those optimized for low intensity. The optimized design for the "⊞" pattern has an efficiency of η = *64.1%* and a uniformity of δ = *29.1%*. After the downhill optimization, the final design has an efficiency of η = *73.5%* and a uniformity of δ = *1.9%*. The optimized design for the "Z" pattern has a similar performance with an efficiency of η = *64.7%* and a uniformity of δ = *29.4%* before the downhill optimization and η = *72.2%* and δ = *1.6%* afterwards.

Summary of experimental results

In all of the designs realized by the optical processor, the uniformity of the optimized orders is extremely poor. The standard deviation was always between 30% - 40% of the mean value. The optical system is unable to achieve a uniform distribution of the intensity in the diffraction orders. The intensity



(a) (b)

Fig. 4.19 The desired intensity distribution amongst the 7 x 7 orders (a) and the results achieved by the optical processor (b).

variation amongst the orders is quite large, even when a design that has been optimized on the computer to have good uniformity is used. Therefore, the theoretical uniformity of the diffraction orders is not observable using the optical system. Consequently, during the optical optimization using the simulated annealing algorithm, the weighting of the contrast in the cost function, see Eqs. (4.7) and (4.8), was always zero. Thus the uniformity of the optimized orders was never emphasized as part of the desired goal. The purpose of the optical optimization was to produce a design that yields a highly efficient distribution of the intensity amongst the desired orders. The uniformity of the distribution was addressed only during the downhill optimization on the computer.

The optical system with a resolution of 64 x 64 pixels was able to perform well despite the presence of the interference noise due to the overlapping orders. The optimized designs achieved with the 64 x 64 resolution were characterized by higher efficiency than the designs achieved with the 32 x 32 resolution. The finer details achieved with the increased resolution allowed for the realization of more efficient unit cell designs. The example illustrated in Fig. 4.17, shows the capability of the optical system to distribute the intensity

into certain orders without influencing the neighboring orders. The complexity of the desired distribution is limited for the unit cells containing 64 x 64 pixels. While the higher resolution of the cells allows for the definition of finer structures, the noise resulting from the overlapping diffraction orders inhibits the realization of complex intensity distributions amongst the orders. The separation of the orders achieved with unit cells containing 32 x 32 pixels allows for the precise definition of the intensity in each order. The absence of the interference noise allows for the realization of more complex intensity distributions amongst the diffraction orders. However, the reduced number of pixels limits the minimum size of the features in the unit cell, thus reducing its efficiency.

The results obtained with the optical optimization and the subsequent downhill optimization on the computer are compared to the results obtained by applying only a downhill algorithm on the computer. The comparison is to give an appreciation of the quality of the results and of the benefits, if any, of the simulated annealing phase of the optimization. Some of the optimization examples are summarized in Table 4.1 along with the results of the downhill optimization on the computer. In each of the examples, the efficiency is better for the results obtained from the optically implemented simulated annealing followed by the downhill optimization on the computer. The results of the downhill algorithm alone are plagued by the problem of local minima which was the original motivation to the simulated annealing technique. The uniformity of the optimized designs are relatively equivalent. The differences observed are small enough to be neglected. Therefore, it can be concluded that the simulated annealing algorithm produces a good starting point for the downhill algorithm, allowing the algorithm to avoid many of the local minima which would limit the design performance.

	Fig. 4.17 64 x 64		Fig. 4.19 "⊞" 32 x 32		Fig. 4.19 "Z" 32 x 32	
	η	δ	η	δ	η	δ
Optical + Downhill	75.4%	1.6%	73.5%	1.9%	72.2%	1.6%
Downhill only	73.6%	0.9%	70.5%	3.1%	70.6%	2.3%

Table 4.1 Comparison of results of optical optimization followed by downhill on computer to results obtained with downhill algorithm only.

The comparison to the downhill algorithm alone served to illustrate the improved performance obtained by initially optimizing the design with the simulated annealing algorithm on the optical system. The results of the example illustrated in Fig. 4.15 are compared to the same optimization procedure implemented entirely on the computer, i.e. the simulated annealing phase and the downhill phase of the optimization on the computer, to determine the effects, if any, of implementing the simulated annealing algorithm on the optical system. The results are summarized in Table 4.2. The performance of the two optimized designs is basically identical. The uniformity of the implementation on the computer is better, but the difference is insignificant. Therefore, the optical implementation of the simulated annealing phase of the optimization has not degraded the performance of the algorithm.

Simulated Annealing	η	δ	Simulated Annealing	η	δ
Optical	77.0%	1.6%	Computer	77.2%	0.9%

Table 4.2 Comparison of the results obtained with simulated annealing algorithm implemented optically and with a computer.

5. Limitations of system architecture

The performance of the optical parallel processor is greatly affected by the noise sources associated with the system's implementation. There are four main areas in which the system components limit the capabilities of the optical processor: (1) the resolution of the unit cells, (2) the dynamic range of the photodetectors, (3) the uniformity of the intensity in the Fourier plane, and (4) the speed of the optical processor. The limitations (1), (3), and (4) are all related to the LCTV, whereas the limitation (2) is due to the array of photodetectors.

There are two effects governing the number of pixels per unit cell, the number of pixels available in the LCTV and the spot size of the beam in the focal plane of the Fourier lens. The LCTV consists of 320 x 220 pixels. At least 2 x 2 unit cells must be illuminated so that the sinc functions resulting in the Fourier plane will not significantly overlap. Better distinction of the individual orders is achieved with the illumination of 3 x 3 unit cells. Furthermore, the number of pixels in the unit cell is constrained to be a power of two 2^n, because of the FFT operations during the downhill optimization. Therefore, for the LCTV used in this experimental realization, the maximum

resolution of the unit cell is limited to 64 x 64 pixels. However, for the results presented, the resolution was limited by the spot size of the beam in the focal plane of the Fourier lens. The measured spot size of approximately $\sigma = 40\ \mu m$ limited the system to a resolution of 32 x 32 pixels for results free of noise due to the interference of overlapping orders. For the system architecture to separate the orders for a resolution of 64 x 64 pixels, a spot size of $\sigma \leq 32\ \mu m$ would be necessary. Since the spot size is not diffraction limited, but is dominated by aberrations, it can be significantly decreased by improving the optical quality of the beam. In which case, the maximum resolution of 64 x 64 pixels can be realized without any degradation in the measured intensities due to the interference of the neighboring orders.

A further constraint on the system resolution comes from the limited performance of the photodetector array used. The multiplexed readout of the photodetector array has a noise level equal to about 10% of the detector's full dynamic range. Consequently, the minimum detectable intensity change is much larger than desired. This condition requires that a group of pixels be modified together so that the resulting intensity change in the Fourier plane is greater than the detector's noise. Thus, the signal-to-noise ratio is increased at the cost of the resolution of the minimum features within the unit cell. However, while the size of the minimum features is increased, the precision of the position of the features is maintained. The result of the increased feature size is a limitation on the complexity of the optimized orders. The number of orders that can be resolved with the optical processor is directly related to the minimum feature size, since finer structures are required to produce the high frequency diffraction orders.

There exists a variation of the intensity in the Fourier plane following the LCTV. Close observation of the photos of the experimental results, e.g. Fig. 4.9b, illustrates a noticeable variation in intensity along the diagonal across the diffraction orders. Theoretically, the intensity at opposite sides of the diagonal should be the same for binary phase elements. The intensity gradient appears to be due to the refresh rate of the drive voltage of the individual pixels of the LCTV. The state of each pixel is refreshed at the video rate of 60 Hz in the sequence illustrated in Fig. 4.20. Consequently, the pixels in one corner of the display are out of phase with the other corner, which produces a type of gradient effect which is observed in the diffraction orders. The variation of the intensity limits the optical processors ability to optimize the uniformity of the diffraction orders. The uniformity of the optically optimized designs is very poor, between 30% - 40% of the mean intensity value. The intensity

variation in the diffraction orders due to the refresh of the LCTV limits the optical processor to optimizing the designs for high efficiency only. Consequently, the poor uniformity of the design must be compensated during the downhill optimization on the computer.

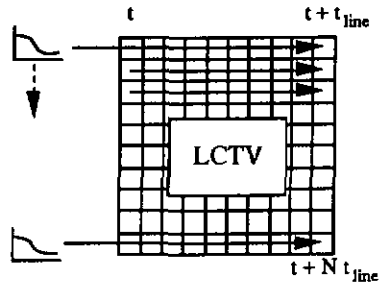


Fig. 4.20 Pixel refresh within LCTV.

The speed of the optical parallel processor is limited by the slow response time of the LCTV and the time necessary to readout the detector array. The LCTV uses twisted-nematic liquid crystal displays which have a response time on the order of tens of milliseconds. The electronics of the video projector drive the LCTV at the video rate of 60 Hz, which corresponds well with the speed of the crystals themselves. Clearly, there is no need to update the signal sent to the displays more quickly than the displays can react. Consequently, each modification of the pixels of the LCTV takes at least 30 - 50 msec.

The time required to complete one modification of the unit cell is a combination of the time necessary for the LCTV to receive the modified drive signal, the liquid crystal to react to the change, the readout of the array of detectors, the Metropolis decision process, and if the modification is not accepted, then the time necessary to return the LCTV pixels back to their original state. The total number of modifications during an optimization depends on the resolution of the unit cell and the simulated annealing parameters, such as starting temperature, rate of temperature decrease, and number of iterations at each temperature. The number of iterations can range from several hundreds to several thousands. The typical time needed for the optimization of a unit cell with 64 x 64 pixels was almost 8 hours and included almost 63,000 modifications. Thus, for the system components applied in these experimental results, the optical processor was able to complete one modification on the average every 450 msec. Only 14% of the modifications were accepted in the above example, therefore for 86% of the modifications it was necessary to reset the LCTV pixels back to their original state. Thus, the

determination of the cost function and the Metropolis decision process account for only a small percentage of the computation time. The slow response of the LCTV and the slow readout of the photodetector array are responsible for the majority of the computation time associated with the simulated annealing optimization.

6. Conclusions

The optical parallel processor has been used to realize the simulated annealing optimization of binary phase diffractive elements for fan-out applications. The results obtained with the processor have illustrated the system's ability to produce an optimized design which can serve as a good starting point for the downhill algorithm implemented on a computer. Due to the large variation of the intensity in the Fourier plane, the processor was only able to optimize the design towards a solution with high efficiency. The uniformity of the design could not be corrected during the optical optimization, but was improved during the downhill algorithm on the computer. The experimental results presented and the constraints of the optical implementation which have been addressed, have shown the feasibility of applying an optical parallel processor to such heuristic optimization tasks. However, for the optical processor to compete with the computer techniques normally used, the limitations described in the previous section must be improved. Replacement of the LCTV and improvement of the photodetector array could be all that is necessary to bring the computational capability of the system to a competitive level.

Improvement of the multiplexed readout of the photodetector array is a clear necessity. Unfortunately, for the work presented here, another detector was not available. Consequently, the photodetector array has been the limiting factor for the accuracy of the system's performance. However, with a detector exhibiting a much larger dynamic range, the sensitivity of the system can be increased to a level such that individual pixels can be modulated instead of groups of pixels. At this point, the optical processor can fully exploit the unit cell resolution available.

The LCTV with its twisted-nematic liquid crystal has greatly influenced the performance of the optical system. Referring back to the discussion on the limitations to the system resolution, any further increase in the system resolution beyond the maximum 64 x 64 pixel limit, would require the replacement of the LCTV by another which has higher resolution and good

phase modulation characteristics. The LCTV not only dictated the maximum resolution of the unit cell, but was responsible for the intensity variation observed in the diffraction orders in the plane of the photodetector array. The intensity variation is a consequence of the refresh rate of the drive voltage of the individual pixels of the LCTV. If the intensity variation is static in the Fourier plane, then the possibility exists to calibrate the response of the detectors to compensate for the variation. Another possibility is to replace the nematic LCTV by a ferro-electric device. The ferro-electric liquid crystal contains two bistable states, which remain stable once they have been set. Therefore, a refresh signal is not necessary and the intensity variation in the diffraction orders can be removed. It is possible to adjust the drive voltage so that the ferro-electric crystal yields π phase modulation. The polarization is rotated, but by adjusting a polarizer following the device the intensity of the two phase states can be equalized. Furthermore, the ferro-electric liquid crystal has a much faster response time than the nematic liquid crystal. A factor of 100 could be gained in speed by using a ferro-electric device rather than the nematic LCTV.

The optical parallel processor has been presented for the optimization of binary phase fan-out elements. However, the optical processor can be applied in other areas as well. The extension of the processor for the optimization of multi-level phase elements is straightforward. However, the system limitations discussed earlier are even more critical for the optimization of multi-level phase elements, and more effort is needed with respect to the signal-to-noise ratio. Also, for multi-level elements it would not be possible to replace the LCTV by a ferro-electric device since they have only two phase states. Therefore, for multi-level phase elements it is absolutely necessary to be able to calibrate the intensity variation in the Fourier plane due to the refresh of the LCTV pixels so that it can be removed from the intensity measurements.

Finally, the optical parallel processor has been shown to be effective in the optimization of binary phase fan-out elements using simulated annealing. The phase modulation capability of the LCTV has been exploited to allow the optical system to realize the time consuming calculations associated with the simulated annealing optimization. By compensating for the intensity variation observed in the Fourier plane, it is hoped that in the future the processor can also implement the downhill phase of the optimization process. In which case, the 2^n constraint on the resolution of the unit cells needed for the FFT algorithm can be removed and the full resolution of the LCTV can be exploited. Furthermore, the architecture of the optical processor could be

applied to other types of heuristic optimization tasks, such as the correction of wavefront aberrations or wavefront characterization. As the quality of the LCTV devices increases, the capabilities of the this type of optical processor will similarly increase.

References

4.1 S. Kirkpatrick, C.D. Gelatt, Jr., and M.P. Vecchi, "Optimization by simulated annealing," Science **220**, 671-680 (1983).

4.2 N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. 21, 1087-1092 (1953).

4.3 H.P. Herzig, D. Prongué, and R. Dändliker, "Design and fabrication of highly efficient fan-out elements," Jpn. J. Appl. Phys. 29, L1307-L1309 (1990).

4.4 U. Killat, G. Rabe, and W. Rave, "Binary phase gratings for star couplers with high splitting ratios," Fiber Integr. Opt. 4, 159-167 (1982).

4.5 R.L. Morrison, "Symmetries that simplify the design of spot array phase gratings," J. Opt. Soc. Am. A **9**, 464-471 (1992).

4.6 H. Dammann and K. Görtler, "High-efficiency in-line multiple imaging by means of multiple phase holograms," Opt. Comm. **3**, 312-315 (1971).

4.7 A. Vasara et al., "Binary surface-relief gratings for array illumination in digital optics," Appl. Opt. **31**, 3320-3336 (1992).

4.8 M.R. Feldman and C.C. Guest, "Iterative encoding of high-efficiency holograms for generation of spot arrays," Opt. Lett. **14**, 479-481 (1989).

4.9 M.P. Dames, R.J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," Appl. Opt. **30**, 2685-2691 (1991).

4.10 J. Amako and T. Sonehara, "Kinoform using an electrically controlled birefringent liquid-crystal spatial light modulator," Appl. Opt. **30**, 4622-4628 (1991).

4.11 K.J. Weible, N. Collings, and A.R. Pourzand, "Initial results of a fully interconnected neural network with modifiable interconnects," Optical Memory and Neural Networks **1**, 157-159 (1992).

4.12 N. Konforti and E. Marom, "Phase-only modulation with twisted nematic liquid-crystal spatial light modulators," Opt. Lett. **13**, 251-253 (1988).

V. Conclusion

The experimental realization of the three different types of optical learning systems has provided a much better appreciation of the problems and the promises associated with their optical realization. Each of the implementations is faced with constraints imposed by the system architecture, the devices, and the neural model or optimization algorithm. Some of the limitations are specific to the particular neural model or algorithm, whereas others are more directly related to their optical implementation. The limitations due to the optical implementation can be divided into two categories, those due to the devices, and those due to the system architecture. While improvements in the device technology will remove some of the constraints, the more fundamental restrictions will remain.

The experimental investigation has allowed a critical analysis of the ability of the optical systems to meet the demands of high parallelism sought for optical neural networks and learning systems. The optical system architecture was based on information fan-out using diffraction gratings to achieve a matrix-matrix multiplier type of processor. The computational capacity of the fully interconnected neural system was constrained to a modest limit of 16 x 16 neurons, while requiring large physical dimensions. The system architecture which has been investigated is inhibited by the fan-in of the weighted interconnections. The severe constraints are a direct consequence of the Fourier relation which exists between the input neurons and the output neurons, which limits the minimum separation of the neurons. To envisage a further decrease of the neural separation, thus augmenting the computational capacity, it would be necessary to modify the fan-in technique which was exploited in this system architecture.

The ability to achieve an all-optical implementation is hindered by the requirements of the thresholding device. In most applications, the required threshold function is more complicated than a simple AND or NOR function achieved with a liquid crystal light valve (LCLV). Also the performance of our LCLV was severely limited due to its poor uniformity and sensitivity. The poor performance of the device dictated the transformation to a hybrid implementation, where the thresholding and feedback were realized using a

detector and a personal computer with feedback to a liquid crystal display. Even with improved performance, the LCLVs are incapable of the more complex threshold functions, such as the winner-take-all selection. The most versatile alternative will be the use of integrated devices, where detectors and modulators receive and encode the optically transmitted information and VLSI circuits provide complex logic functions. Integrated devices of this type are currently developing at the prototype level and will be key devices in optical parallel processing systems.

The experimental realization of the three different types of optical learning systems provided a wide range of applications in which to evaluate the merits of the optical implementation of such learning systems. The results of the fully interconnected network illustrate the restrictions due to the system's large fan-out and fan-in requirements, precise alignment constraints, and the large system dimensions. The system architecture employing diffraction fan-out elements is best adapted for applications requiring limited fan-out. This is confirmed in the realization of the Hamming network and the Kohonen network, where the optical implementations performed as well as the simulations. The optical implementation of the simulated annealing optimization of the binary diffractive fan-out elements illustrates another area in which the optical techniques are well applied. The realization of the Fourier transform of the phase information represents a perfect example of where optical techniques can be employed in information processing systems. Similar techniques can be applied in areas such as wavefront aberration correction or wavefront characterization, where the information to be treated is already in optical format.

The prototype systems which have been investigated provide insight into the promises and the problems associated with the optical implementation of neural networks and learning systems. The experimental studies provide valuable feedback on the performance of certain system architectures and the various devices employed. The liquid crystal televisions have been shown to be very efficient devices for modulating either intensity or phase. Whereas, the performance of the LCLVs has fallen far short of the demands of the experimental realizations. The real benefits of the optical learning systems is their ability to adapt their performance so as to compensate for the imperfections within the experimental realization. By feeding back information to the LCTVs the systems are capable of compensating the nonuniformities of the optical system, but are unable to correct for

misalignments in the interconnection plane unless time consuming algorithms are employed.

The prototype systems have allowed the performance analysis of the various learning and optimization algorithms. The algorithms must be adapted to meet the various constraints imposed by the optical implementation, but have been shown to remain effective in iterating the system to optimized behavior. Furthermore, the prototypes provide a framework in which to test and develop modified learning schemes.

Finally, the experimental realizations have provided an indication as to which devices or components limit the physical dimensions of the optical system. The size of the system and the tolerances placed on the alignment of the components are critical parameters for evaluating system architectures. In today's market of ever shrinking system components, the miniaturization of the prototype optical system must be foreseeable. The fan-in conditions and the rigid alignment constraints seen in the fully interconnected network made the reduction of its physical dimensions unimaginable, thus reducing the attractiveness of such a system architecture for large fan-out systems. Whereas the system with reduced fan-out requirements and the Fourier transform system have much less restrictions on their system dimensions and their alignment constraints are also more lax.

The prototype optical information processing systems that have been investigated are far from challenging the competitive techniques employed with electronics. Yet, they serve the purpose of illustrating the strong points that optical techniques have to offer, while uncovering some problem areas that must be taken into consideration. The day of the fully optical digital computer may be far from reality, but the role of optics in the domain of information processing is ever increasing. Prototype systems such as those discussed will help to illustrate areas of potential application of optical techniques.

Acknowledgements

I would like to take this occasion to express my gratitude and appreciation to Prof. R. Dändliker. Six years ago, I showed up on the doorstep of the Institut de Microtechnique looking for a position as an optical engineer within Prof. Dändliker's group. Offering my services to support his already existing research activities, I was prepared to bring my academic career to an end. I found myself refused a position as an engineer, but offered instead the position of Doctorant. Six years later, it can be said that the realization of this thesis is due to Prof. Dändliker's refusal to hire me and for this I am deeply grateful. Throughout these past years, not only have I benefited greatly from his extensive experience and insight in the domain of optical sciences, but have received his continuous support in handling the inconveniences sometimes associated with governmental bureaucracy. For all of the above, I'm indebted to Prof. Dändliker and express my sincere thanks. And finally, it's taken six years, but the position that I came looking for has finally been made available.

I would also like to thank the members of the jury, Dr. X. Arreguit, Prof. F. Pellandini, and Prof. A. C. Walker, for their comments and their interest during the critical reviewing of this thesis. Furthermore, I would like to recognize the Swiss National Science Fund which supported this research in the domain of optical computing and optical neural networks.

I would like to thank Dr. N. Collings for his competence in the domain of neural networks and his suggestions conceming this manuscript. Also, I'd like to thank Xue Wei for his willingness and flexibility in arranging the use of the available lab space and equipment and for his valuable ideas both relating to the liquid crystal light valves and elsewhere. The work that has been achieved has been supported by many students performing their semester or diploma projects. Their efforts have been much appreciated and have contributed greatly to the projects. The most significant effort, which requires further acknowledgement, is the semester and diploma projects of Urban Schnell. The work presented pertaining to the Kohonen model and the traveling salesman problem were a direct result of Urban's efforts and initiative.

During the past years I have benefited greatly from the discussions and helping hands of my colleagues in the optics group. Not only have they been

there when the laser needs aligning, but they've also been present when an occasion needs celebrating. I would especially like to thank Dr. Hans Peter Herzig. While not directly involved with most of the work presented in this thesis, he has always been available with an endless resource of ideas and suggestions. Also, I would like to thank the other groups of the Institut de Microtechnique as well as the secretarial staff for the friendly and open environment which has been created and nurtured within walls of the IMT, be it Breguet, Jaquet-Droz, or Tivoli.

Finally, the words don't exist to express my gratitude for the support and encouragement I've received at home over the past years. Thanks Fa, for everything that has nothing to do with this thesis, but without which this work would never have been completed.

The End