Optical neural networks using liquid crystal devices

N. Collings
Institute of Microtechnology, University of Neuchâtel,
CH-2000 NEUCHATEL,Switzerland.

ABSTRACT

Work on neural networks which use liquid crystal projection screens to code the input and weight matrices and a liquid crystal light valve to perform integration and thresholding is reviewed.
Key words: optical neural network; liquid crystal light valve.

1. INTRODUCTION

Liquid crystal devices have enabled the construction of high spatial complexity programmable matrix-vector multipliers and thresholding elements which are the building blocks of optical neural networks. The optical matrix-vector multiplier approach was the first method of implementing optical neural networks [1]. An N x N input plane is replicated N x N times onto an N^2 x N^2 analogue weight plane and the product between the replicated input and the weight plane is integrated on a N x N thresholding plane. The resulting N x N thresholded output is either used in a second matrix-vector multiplier (feedforward network) or recirculated through the same matrix-vector multiplier (recurrent network). The thresholding of the N^2 channels is performed by a liquid crystal light valve (LCLV), which is controlled by a single electrical waveform. The novelty of the optical approach in the case of the recurrent network is that the dynamics of the computation depends on the dynamical behaviour of the LCLV.
The following aspects will be discussed in this paper. Firstly, the basic building blocks of these networks, which are the spatially multiplexed optical matrix-vector multiplier and the LCLV. Secondly, an experimental and theoretical model of the dynamical behaviour of the LCLV will be presented, followed by related system experiments.

2. OPTICAL MATRIX-VECTOR MULTIPLIERS

The classic 2D matrix-vector multiplier (MVM) architecture[2] has been explored in neural networks[2], optical interconnects[3], and numerical processing[4]. In the area of numerical processing the limitations to the precision of the computation resulting from the devices used has been assessed[5,6]. In general, the precision will be limited by thermal noise at the detector at low light intensities and by signal-dependent noise in the weight plane at higher intensities. 6-7 bit precision is a best case, where crosstalk due to the optical system is not accounted for. In the area of optical interconnects there has been much emphasis on the signal attenuation in such networks. A per channel loss of 19.5 dB when 4 inputs are connected to 4 outputs through a 4 x 4 shutter array was measured[3]. The low precision and high losses associated with the MVM have hindered the development of MVMs in the areas of numeric processing and interconnects. However, they are of less importance for optical neural networks.

1999 Euro-American Workshop Optoelectronic Information Processing: A Critical Review,
edited by Philippe Refregier, Bahram Javidi, Proc. of SPIE Vol. 10296 (Vol. CR74), 102960F

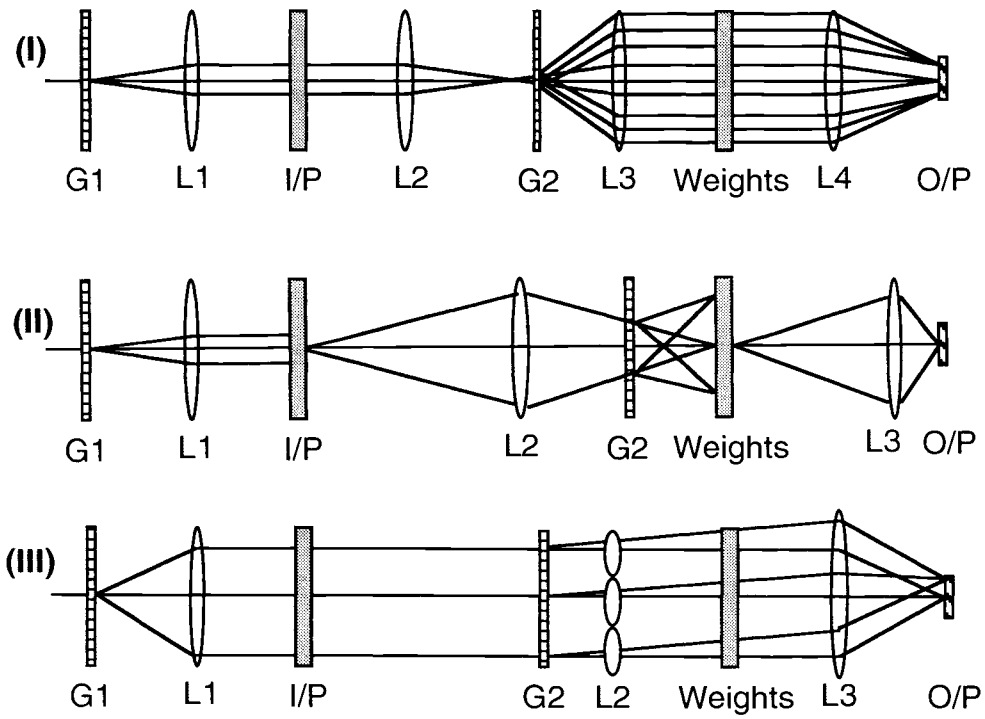


Fig. 1 Optical layouts of three space-invariant fan-out/fan-in systems, comprising: Dammann gratings (G1&2); input (I/P), weight (W) and output (O/P) planes[7]

At the IMT, a number of MVM systems which use diffraction gratings for input array replication have been studied, and three systems were analysed[7] (Fig. 1). In system I, the input plane is replicated in a Fourier optical arrangement using a Dammann grating (DG) displaced from the Fourier plane[8,9]. The displacement of the grating allows the fan-in to be performed by a simple lens. In system II, the replication is performed by the DG in a converging beam in order to tune the spot spacing[10]. In system III, a lenslet array in conjunction with the DG performs a local replication of each input channel (the interlaced fan-out[11]). The following conclusions were drawn. System I requires a high resolution grating for adequate separation of the replications in the weight plane. Since non-uniformities in the fan-out increase with the resolution, the channel capacity of the first system will be limited. Moreover, the spot size in the output plane is comparable to the input beam diameter. System II is not telecentric due to the use of a grating in a converging beam arrangement. However, this allows the spot spacing in the weight plane to be precisely controlled by axially displacing the grating. The consequences are that a large aperture is required for the fan-in optics and lenslets cannot be used behind the weight plane. Because lenslets cannot be used, there must be adequate separation of the replicas in the weight plane, so that the information capacity of this plane is not fully used. System III had no apparent disadvantages other than requiring high quality lenslet arrays for maintaining beamlet collimation between input and weight plane. The main issues in these systems are the quality of the optical elements (lens, gratings, and lenslet arrays) and the complexity of the fan-in optics. For example, in order to process a 16 x 16 neuron input, system I requires a 16 x 16 fan-out grating of period 60 μm and a system length of over 1 metre[12]. The uniformity of the fan-out of such a

high resolution grating would not be very high, with current technological limitations.

The two systems II and III were further developed at IMT. In system II the converging beam fan-out was replaced by a conventional fan-out arrangement, with the grating placed in a Fourier plane, system B shown in Fig. 2 [13,14]. This system is now a telecentric arrangement which allows the use of lenslet arrays after the weight plane to reduce the spot size at the output. The spot spacing was tuned by using a two-lens arrangement for the Fourier transform lenses, whose focal length can be fine tuned by varying the distance between the lenses. This involves more degrees of freedom than the one axial movement of the grating in system II, but has significant advantages for the system.

The spot size and spacing at the three planes of the system (input, weight, and output) are calculated in the diffraction limit. In critical designs, this calculation was supplemented by ray-tracing. The quotient of the area of the elementary cell of the spot intensity distribution to the area of the bright spot has been defined as the 2D compression ratio[15]. It is convenient here to use a 1D compression ratio, C, which is the ratio of the spot spacing to the spot diameter. In the diffraction limit, C is spatially invariant across the three planes. However, when ray aberrations are included, it decreases with the radial dimension in these planes. When C falls below 1, then the space bandwidth product (SBWP) of the system has been reached. When C^2 falls below the reciprocal of the pixel real estate of the device placed in the corresponding plane, then excess loss will be incurred in the system throughput.

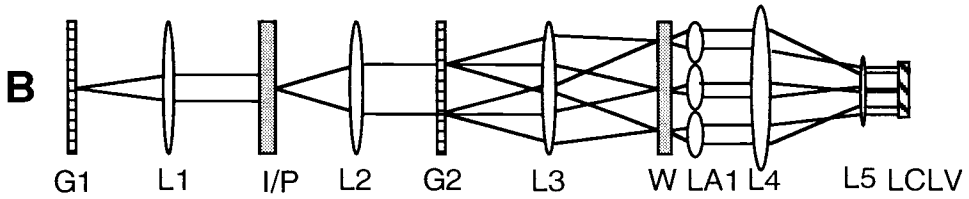


Fig. 2 Layout of optical neural network (system B), including liquid crystal light valve (LCLV) plane.

The 1D compression ratio at the input plane of Fig. 2 is equal to

$$C_{1P} = 1.6\, a/\Lambda_1, \qquad (1)$$

where a is the radius of the input beam, and Λ_1 is the period of 18 x 18 spot grating G1. The numerical value in the system constructed (Λ_1 = 641.7 mm) was 2.5. This is suitable for an input plane where the pixel spacing divided by the pixel aperture is equal to or less than 2.5. This latter ratio is the square root of the reciprocal of the pixel real estate, and we call it the pixel-mark-space (PMS). In the Seiko Epson 640 x 480 VGA screens (P13VM115/125) used in these experiments, the PMS in the horizontal direction is 1.95, and in the vertical direction is 1.33, although the pixel repeat spacing is isometric.

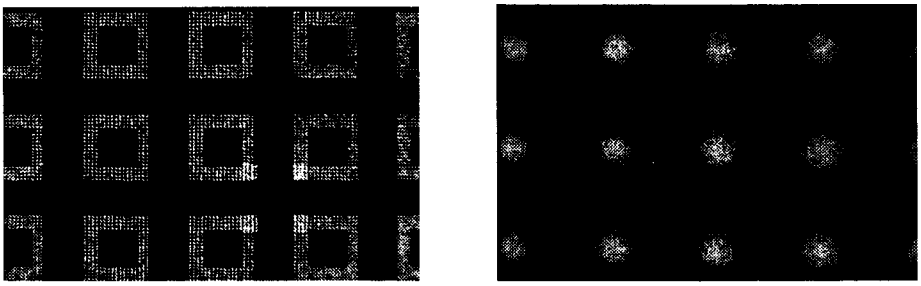
The 18 x 18 spot array on the input plane is relayed with an 18 x 18 fan-out to the weight plane. In order to match the 10^5 spots to the pixel apertures of the VGA screen at W, an accuracy of better that 0.1% is required for the spot repeat spacing. Therefore, the focal lengths of L2 and L3 have to be long (200 mm) in order to avoid excessive field curvature at W, and they have to be tunable. Therefore, a doublet arrangement of two 300 mm lenses separated by an adjustable distance was employed. The tuning required a lot of time, but worked. The 1D compression ratio of the spots in the weight plane was large (about 70) in the diffraction limit, due to the eightfold beam expansion produced by the ratio of the focal lengths of L2/L1.

However, a practical compression ratio should incorporate ray aberrations which are significant over the large image field (20 mm). A significant number of ray tracing simulations have been performed on the problem of minimising field curvature for relaying large image fields between two planes. 200 mm focal length is a reasonable compromise when using stock lenses. However, the ray aberrations remain the principle limitation to the SBWP of this type of system. Larger SBWPs require longer focal length lenses.

The spacing of the sub-arrays is determined by G2 (Λ_2 = 188.6 mm). A 1D compression ratio for the sub-arrays (spacing divided by size) of 1.5 was used. In an ideal system (maximum SBWP), this would be one. In order to fan-in the sub-arrays onto the liquid crystal light valve (LCLV), a lenslet array focusses the sub-arrays to a spot array, followed by image demagnification using a telescope. The telescope does not improve the compression ratio. Therefore, the 1D compression ratio at the LCLV is given by the spacing of the sub-arrays divided by the spot size in the focal plane of the lenslet array

$$C_{LCLV} = p(1.5N-1)f_1/2f_{LA}a, \qquad (2)$$

where p is the pixel spacing (42 μm), N is the fan-out order (18), and f_1,f_{LA} are the focal lengths of L1 and LA1, respectively 25 mm and 2 mm. The calculated value of C_{LCLV} (diffraction limit) is 6.8. The experimental value can be obtained from photos of the weight plane and output plane, where a camera replaces the LCLV (Fig. 3). It is 3, which shows that the lenslet array position is not optimally adjusted.



(a) (b)

Fig. 3 The weight plane (a) and output plane (b) of the system shown in Fig. 2[13].

System III evolved into system C (Fig. 4)[16,17]. A dilute 16 x 16 input array is generated by a high resolution grating G1 (Λ_1 = 268.2 mm) and an expanded beam (a = 3 mm), giving a C_{IP} of 18 (Eq. 1). In order to interlace a 16-way fan-out and separate the fan-out sub-arrays by 8 pixels (giving a sub-array C of 1.5), then C_{IP} should be ≥ 26. This could not be fulfilled with the given G1 because, if the beam was expanded further, the ray aberrations increased. Therefore, the system was operated with spot sizes which overfilled one pixel on the Seiko Epson VGA screen. It is anticipated that the VGA screen will be replaced by dilute source arrays which fulfill the compression ratio criterion.The dilute input array is relayed to the weight plane using a telescopic arrangement of lenslet arrays, and each channel is replicated 16 x 16 times by means of a Fourier plane Dammann grating.

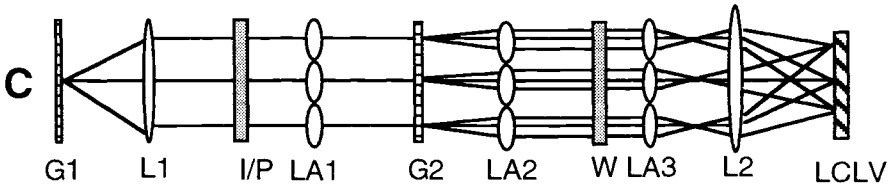
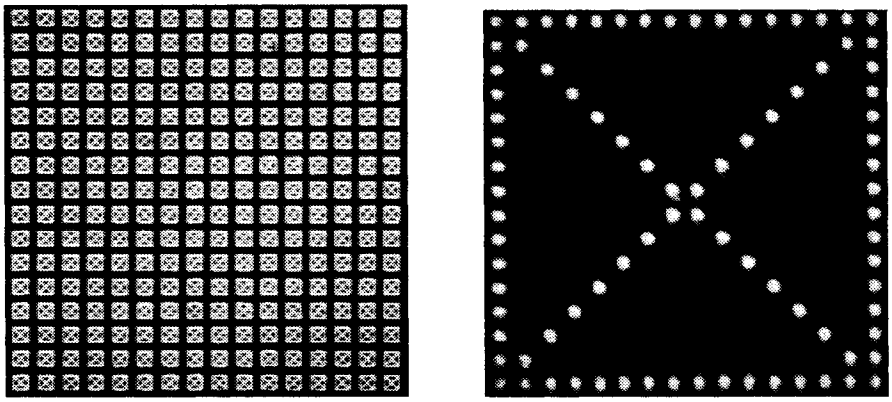


Fig. 4 Layout of optical neural network (system C).

The 1D spot compression ratio in the weight plane is

$$C_{WSP} = 2f_{LA2}\lambda/\Lambda_2\Delta x, \qquad (3)$$

where f_{LA2} is the focal length of LA2, 11.5 mm, Δx is the pixel aperture of LCTV1, and $\Lambda_2 = \Lambda_1$. The resulting C_{WSP} for the smaller aperture dimension is 1.9, which is acceptable for the VGA screen in the weight plane. The fan-in is accomplished by a lenslet array in conjunction with a large aperture lens in a 4f arrangement. Each sub-array is magnified by the ratio of the focal lengths of the large lens and the lenslet array, and all sub-arrays are superposed at the output (Fig. 5).



(a) (b)

Fig. 5 The weight plane (a) and output plane (b) of system C[16].

Since, the compression ratio is unaltered by the magnification, it is the same as C_{WSP}, namely 1.9. The advantage of this system is that the SBWP is not limited by the ray aberrations of the lenses used to relay the input to the weight plane. The size of both planes can be increased when the number of lenslets is increased, and the fan-out can be increased by increasing the diameter of the lenslets.

3. OPTICAL THRESHOLDING/INTEGRATION DEVICE

The first stage of integrating the matrix vector product arrays is performed when the fan-in optics creates a summed intensity array (Figs. 3b and 5b). The summed intensity of each spot of the array must be sensed and thresholded. This function is performed by a liquid crystal light valve (LCLV)[18]. Four types of nematic LCLV have been used in the neural network systems at IMT; three reflective and one transmissive. The transmissive LCLV (Microoptics SPT-25) and one of the reflective LCLVs (SOI, St. Petersburg) used a chalcogenide glass semiconductor photoconductor (CGS), which absorbs in the blue/green and transmits red wavelengths. Both valves had a high sensitivity ($< 10\ \mu W/cm^2$) due to the long

response times (several hundred msec). The liquid crystal layer was a 90° twisted nematic in the case of the transmissive LCLV. The contrast ratio of the HeNe laser read-out beam exceeded 100:1 but the activation of the photoconductor limited the read-out intensity to less than 50 $\mu W/cm^2$. Therefore a reflective LCLV was developed using a 45° twisted nematic [19], which reduces the operating voltage and increases on-state reflectance [20]. The read light was reflected from aluminium pixels of sizes 200 x 200 μm with 100 μm gaps or 800 x 800 μm with 200 μm gaps (Fig. 6). When the read light (I^R) is incident on the 45° twisted nematic reflective cell using a polarizing beam splitter (PBS) (Figs. 9 and 10), then the reflected beam I^O has a low intensity when the write intensity (I^W) is low, and a high intensity when I^W is high (excitatory characteristic).

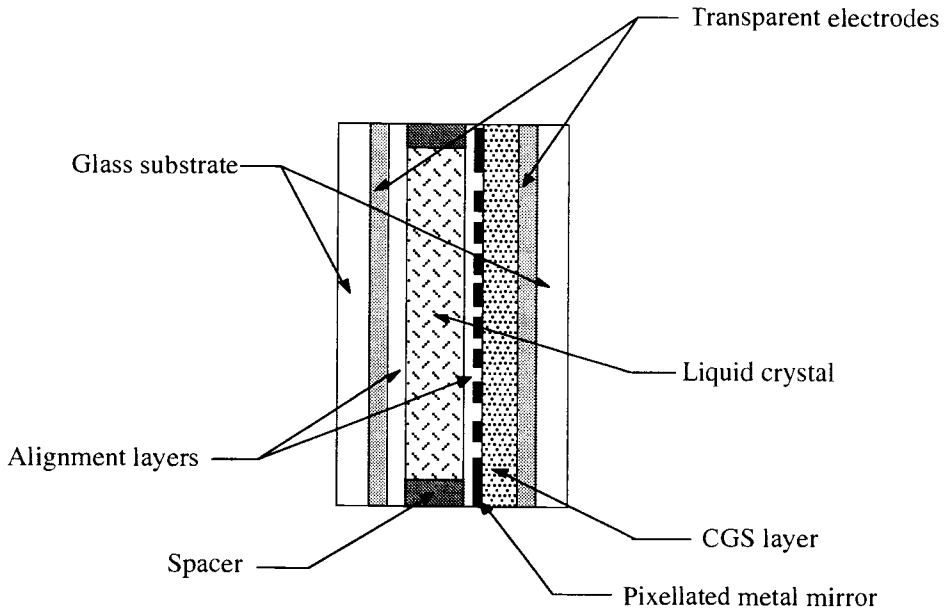


Fig. 6 Cross section of the chalcogenide glass semiconductor photoconductor reflective liquid crystal valve with pixellated metallic mirror[13].

Two further reflective valves used a parallel aligned nematic liquid crystal [21,22]. The dark state (low intensity reflected light) for the parallel aligned LCLV is set by the drive voltage. The dark state can be set for low write intensity (excitatory characteristic) or high light intensity (inhibitory characteristic). For the inhibitory characteristic the reflected light is high intensity when the write light is low intensity.

The characteristics of the reflective LCLVs are tabulated in Table 1.

LCLV	SOI	Asulab	Lebedev
Photoconductor	As_xSe_{1-x}	CdS	GaAs
Mirror	Al-pixels	Dielectric	Dielectric
Reflectivity	70%	70%	75%
Spectral sensitivity	Blue/green	Blue/green	Visible
Write sensitivity	5 $\mu W/cm^2$	45 $\mu W/cm^2$	15 $\mu W/cm^2$
Contrast ratio	100:1	40:1	10:1
Gain	60 dB	12 dB	60 dB
Rise time (ms)	40	120	900
Decay time (ms)	140	540	200

Table 1 Characteristics of reflective LCLVs used at the IMT [19,24,16].

The main issues for LCLVs in this application are the shape of the transfer characteristic (plot of reflected I^R versus I^W), the spatial uniformity, and the gain. The ideal curve for the feedforward network (Fig. 9) is a sigmoid characteristic (Fig. 7). Typically, the bottom left-hand asymptote is curtailed in LCLV, but by computer modelling the LCLV the learning algorithm of the neural network can be appropriately adapted [23].

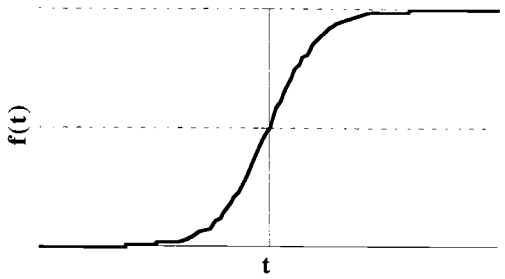


Fig. 7 Shape of sigmoid characteristic.

Both the 45° twist and parallel aligned nematics modulate the phase of the read light beam, and this phase modulation is converted to amplitude modulation by polarizing elements. Therefore, phase uniformity is important for spatial uniformity. The LCLV should have less than half a fringe over the aperture used. In system B the aperture is reduced in comparison with the weight plane aperture by the demagnification factor of the telescope, whereas in system C the aperture is comparable to that of the weight plane. Therefore, the spatial uniformity requirements will be satisfied more readily by system B.

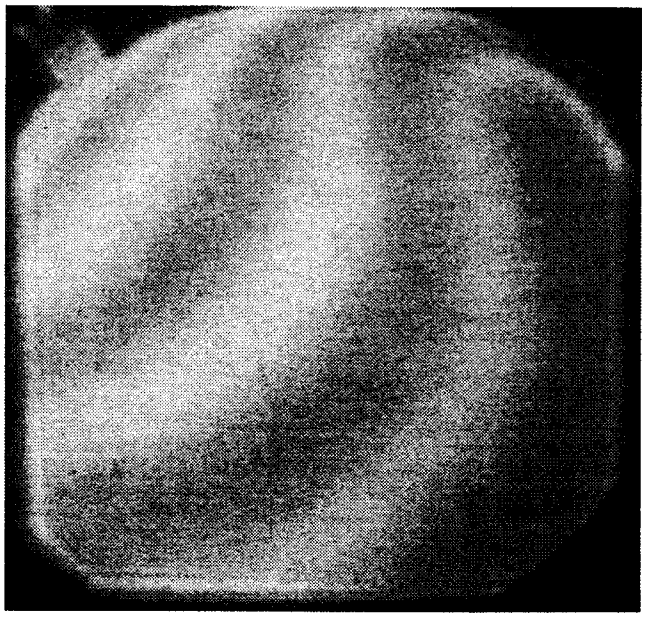


Fig. 8 Photograph of the read-out side of the Lebedev LCLV when it is uniformly illuminated with a coherent read-out beam, using a PBS, and the write intensity and applied voltage are zero[25].

Unfortunately, all the devices we worked with had a several fringes across the active aperture. The fringe pattern on the Lebedev LCLV is shown in Fig. 8. This limited the progress in the system experiments. The maximum gain of the LCLV is the ratio of the maximum intensity of reflected read light to the intensity of write light required to switch the valve ON to 90% of its saturation level. The maximum I^R is set by the level at which the read beam begins to activate the photoconductor[26]. The need for high gain is illustrated in the two system configurations that were investigated at IMT (Figs. 9 and 10).

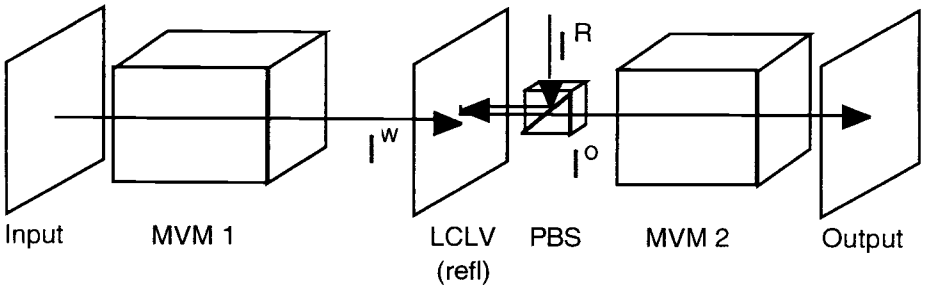


Fig. 9 Two-layer feedforward network formed from two matrix-vector multipliers (MVM1&2) and a reflective LCLV for integration/thresholding of the hidden layer, and a PC for integration/thresholding of the output layer. I^W, I^R and I^O are the write, input read, and output read beams for the LCLV. PBS is the polarizing beamsplitter.

The input plane is illuminated with a high power blue or green laser and passes through a high complexity MVM before arriving as the write beam on the LCLV. The losses incurred in the MVM can be up to 35 dB [27]. I^W must be sufficient to

saturate the response of the LCLV when all pixels in the weight plane of MVM1 are fully transmitting. Equally, when all pixels are switched OFF, I^W must be insufficient to activate the LCLV. It is convenient to set the operating regime of the LCLV such that it is 50% activated when a weight plane of random weights is used.

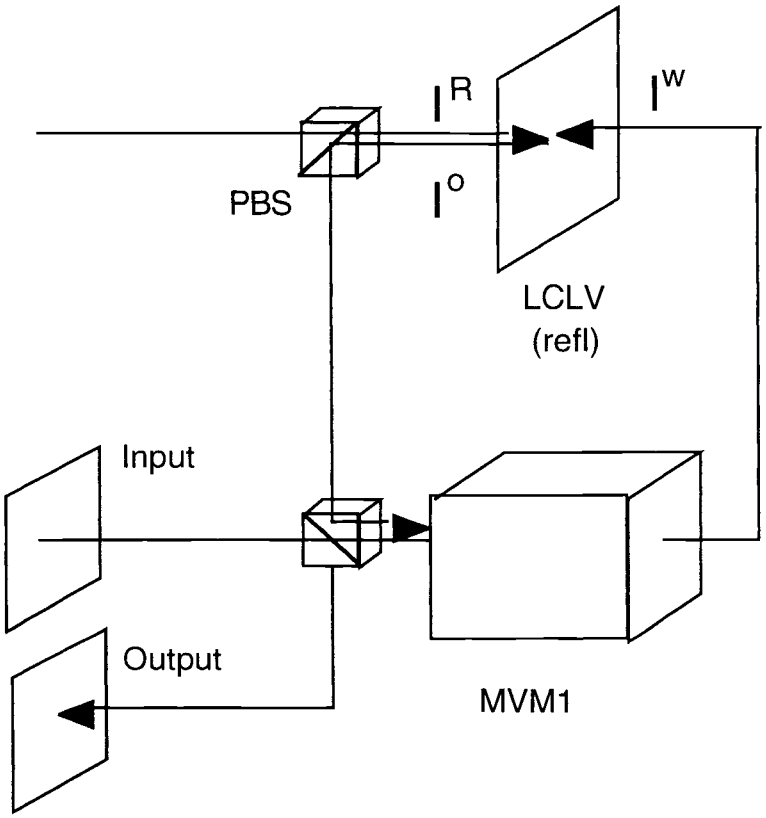


Fig. 10 Recurrent network formed from one matrix-vector multiplier (MVM1) and a reflective LCLV for integration/thresholding of the hidden layer, and a PC for integration/thresholding of the output layer.

Another liquid crystal device which has been tested for this application is a smart pixel array [28]. The idea here is that a custom circuit for integration/thresholding can be made in VLSI silicon and the liquid crystal built on top of it. In fact, the results were disappointing: large non-uniformity, poor gain, and inconvenience due to the arrangement of photodetectors and light modulators on the same side of the device.

4. DYNAMIC CHARACTERISTICS OF THE LCLV

At low write light intensities, there is a relationship between the speed of response of the LCLV and the write light intensity. This is commonly approximated as an inverse proportionality relationship and the sensitivity of a light valve is given in units of the product of the two, or energy density. In order to analyse the behaviour of the LCLV in a dynamic network such as Fig. 10, the exact temporal behaviour must be measured and modelled. The measurements give a limited information which guides the modelling. At first the theoretical results will be presented, followed by the measurements, and then a generalised model.

The Asulab LCLV was modelled on the Hughes LCLV (Model # H4010) in which a light blocking layer of CdTe is interposed between the dielectric mirror and the CdS photoconductor. The temporal response of the photoconductor plus light blocking layer used in the Hughes valve has been analysed experimentally and theoretically for the case of an excitatory characteristic [29,30]. It was observed that the rise time, τ_r, is inversely proportional to the write intensity when I^W is lower than 100 $\mu W/cm^2$, but above this intensity it is inversely proportional to the square root of I^W. When the write light is switched off, the decay time of the photoconductor, τ_d, is independent of the initial light intensity. These results were explained theoretically on the basis of a two-trap model. It was also noted that, in the case of thermal equilibrium during decay, τ_d will be inversely proportional to the number of free carriers which have been excited, ie I^W.

We have made corresponding measurements on the Asulab LCLV. Since it is used in the inhibitory characteristic in the system experiment (see later), switch-on corresponds to a decay of the output from a constant high value to a low value which depends on the write intensity. Conversely, switch-off corresponds to a rise of the output from a low value which depends on the write intensity to a constant high value. The decay curves have been plotted for a number of write light intensities (Fig. 11).

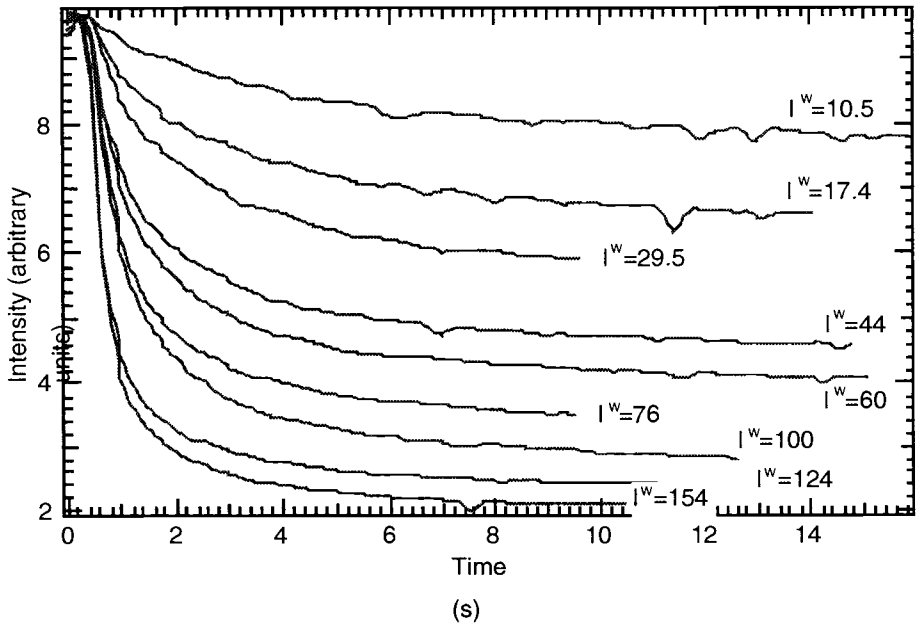


Fig. 11 Decay of read light intensity reflected from the Asulab LCLV with the write light in $\mu W/cm^2$ as the variable parameter[24].

These curves have been fitted with the function

$$I^O(t) = K0 + K1.e^{-K2.t}. \quad (4)$$

This can be written in a form which relates to the physical properties of the LCLV

$$I^O(I^W, t) = I^O(I^W, \infty) + \left[I^\circ(0,0) - I^\circ(I^w, \infty)\right].\exp\left\{-\frac{t}{\tau_d(I^W)}\right\}, \quad (5)$$

where $I^O(I^W, \infty)$ is the static transfer characteristic of the valve, $I^O(0, 0)$ is the output intensity before the light was switched-on, and $\tau_d(I^W)$ is the decay time. In order to simplify the notation for the following development, a separate function $g(I^W)$ will be used for the static transfer function of the valve. Then, equation (5) becomes

$$I^O(t) = g(I^W) + [\, g(0) - g(I^W)] \;.\exp\left\{ - \frac{t}{\tau_d(I^W)} \right\} . \qquad (5')$$

A log-log plot of the decay time against the write light intensity I^W gave a straight line with no change of slope. The gradient of the line was -0.68 and the intercept with the log (τ_d) axis was 1.3. Therefore, the exponent of the write intensity is intermediate between the inverse square root and the inverse quoted previously for the rise time of an excitatory characteristic. The intercept on the write intensity axis is also greater than that measured in [29] (19 s as opposed to about 1 s).

The behaviour of the rise time τ_r with varying write intensity was then measured. In this case, the valve is exposed to a variable write intensity in the initial state, and then the write light is switched off. The equation corresponding to (5') in this case is

$$I^O(t) = g(0) - [\, g(0) - g(I^W)] \;.\exp\left\{ - \frac{t}{\tau_r(I^W)} \right\} , \qquad (6)$$

where $\tau_r(I^W)$ is the rise time. The log-log plot of the rise time against the write light intensity I^W gave a straight line, whose gradient was -0.50 and the intercept with the log (τ_r) axis was 0.89 (or intercept on the τ_r axis of 7.8 s). This is intermediate between the no write light dependence and the inverse proportionality on write light intensity mentioned earlier.

Although there is both a dielectric mirror and a light blocking layer between the input (write) and output (read) sides of the valve, the light leakage from the read side is sufficient to alter the dynamic transfer characteristic. The alteration is a uniform upward translation of the I^O v. I^W curve. At the levels of read light used in the experiment , about 5 mW/cm^2, this translation increases I^O by 20%. However, since the read light is constant in time, this breakthrough has no influence on the dynamic behaviour.

In order to understand the dynamic behaviour of the valve in a system, where the input changes in a continuous manner, equations (5') and (6) must be generalised. If the write light was switched-off before the asymptotic output intensity was reached in Fig. 10 then the output intensity would not reach the asymptote, because the LCLV continues to integrate the write light intensity after switch-on. It is assumed that this integration is that of a linear, time-invariant and causal system. Therefore, the output can be written as the convolution of a (nonlinear) function of $I^W(t)$ with the impulse response of the system, which is an exponential decay (or rise) with the time constant elucidated above,

$$I^O(t) = I^O(-\infty) + \int_{-\infty}^{t} G(I^W(t')).\exp\left\{ - \frac{(t - t')}{\tau} \right\}.dt' \;, \qquad (7)$$

where τ is either τ_d or τ_r.

For the decay time experiment, we had

$$I^W(t') = I^W.H(t'-t_0), \tag{8}$$

where H(t) is the Heaviside function, and t_0 is the time at which the light was switched-on. When Eq. (8) is inserted into Eq. (7), we get

$$I^O(t) = I^O(-\infty) + \tau_d.G(I^W) - \tau_d.G(I^W).\exp\left\{ - \frac{(t - t_0)}{\tau_d} \right\}. \tag{9}$$

This is the same as Eq. (5') with the identities $I^O(-\infty) \equiv g(0)$, $t_0 = 0$, and $\tau_d.G(I^W) = g(I^W) - g(0)$.
For the rise time experiment,

$$I^W(t') = I^W - I^W.H(t'-t_0), \tag{10}$$

where t_0 is now the time at which the light was switched-off. When eq. (10) is inserted into equation (7), then

$$I^O(t) = I^O(-\infty) + \tau_r.G(I^W).\exp\left\{ - \frac{(t - t_0)}{\tau_r} \right\}. \tag{11}$$

This is the same as Eq. (6) with the identities $I^O(-\infty) \equiv g(0)$, $t_0 = 0$, and $\tau_r.G(I^W) = g(I^W) - g(0)$.
Therefore, Eq. (7) is validated to the extent that it predicts the correct form of response when step functions are applied to the write beam intensity.

5. OPTICAL NEURAL NETWORK SYSTEMS

The integrating/thresholding element (LCLV) represents an array of neurons whose activation values o_j are the sum of inputs x_j that arrive via weighted pathways. The input from a particular pathway is an input signal x_j multiplied by the weight W_{ij} of the pathway. A bias term θ_j is included in the sum in order to provide a variable threshold

$$o_j = \sum_{i=0}^{N} x_i W_{ij} + \vartheta_j. \tag{12}$$

The weights are the pixel transmission values of the LCTV in the weight plane. The outgoing signal (reflected read light intensity) is $y_j = f(o_j)$ where y_j is a non-linear function (transfer curve) of the activation value (write light intensity).
The investigation of a feedforward optical nework of the type shown in Fig. 9 was motivated by the desire of our collaborating computer scientists to realise a multilayer perceptron hardware which had the potential of high parallelism. A major difficulty was that only unipolar coding is available in an incoherent optical system which uses intensity as the sole coding modality. The multilayer perceptron algorithm had to be adapted to the optics [31]. In this method the network is trained on the computer with unipolar inputs and bipolar weights. During the recall, bipolar weights W_{ij} are transformed to all positive weights W_{ij}' according to the input x_i

$$W'_{ij} = \max\left\{ \left(W_{ij} - W_{min}\right) \cdot \left(1 - \frac{\vartheta_j - W_{min} \sum_i x_i}{\sum_i \left(W_{ij} - W_{min}\right) x_i} \right), 0 \right\}, \tag{13}$$

where W_{min} is the minimum of all original weight values and local threshold ϑ_j. The neuron activation value becomes then

$$\sum_i W_{ij}x_i - \vartheta_j = \sum_i W'_{ij}x_i . \tag{14}$$

The intermediate matrix-vector product $\sum(W_{ij} - W_{min})x_j$ in Eq. (13) and all positive weights W_{ij}' are computed off-line (in a host computer). The final matrix-vector product $\sum W_{ij}'x_i$ and the thresholding are performed in the optical system. The disadvantage is that it cannot be used on-line. The recall task is shared between computer and optics and hence does not take total advantage of the optics. The advantage is that a multilayer implementation is possible. This is not the case for simple weight constraint, where negative values of the weight matrix are represented by zeroes. Although this works with reduced memory capacity for single layer networks, multilayer networks do not converge. A technique which is extensible to multilayer networks is the weight bias method [32,33]. In this method the network is trained on the computer with bipolar inputs and initial weights. The global threshold θ and the slope β of the thresholding function are set to arbitrary values. Once the network is trained, bipolar weights and inputs must be transformed to unipolar all-positive inputs for optical implementation. Considering inputs and weights normalized to the [-1,1] interval, this transformation can be done either by adding one to the weights and inputs or by subtracting them from 1

$$\text{if } x \in [-1\ldots1] \quad \Leftrightarrow \quad \frac{(1 \pm x)}{2} \in [0\ldots1]. \tag{15}$$

After this modification, the output of the network using complementary inputs and weights can be calculated as follows

$$\begin{aligned} y_j &= f\left[\sum_i \frac{(1+W_{ij})}{2}\frac{(1+x_i)}{2} + \sum_i \frac{(1-W_{ij})}{2}\frac{(1-x_i)}{2}\right] \\ &= f\left[\frac{N}{2} + \frac{1}{2}\sum_i W_{ij}x_i\right] \end{aligned} \tag{16}$$

where N is the number of elements of the input vector. The drawbacks of this method are that the thresholding function of the LCLV should be changed when the input x_i changes, and there is a loss of SBWP due to the use of complementary inputs and weight matrices. However, the complementary outputs of the LCLV are readily available when a polarizing beamsplitter is used. A variant of the weight bias method is the Reversal Input Superposing Technique (RIST)[34]. The final method for the efficient use of unipolar coding in multilayer networks is to use mixed excitatory and inhibitory characteristics in the neural plane thresholding[35].
The implementation of this method would require a LCTV in the read beam of the LCLV (Fig. 12).

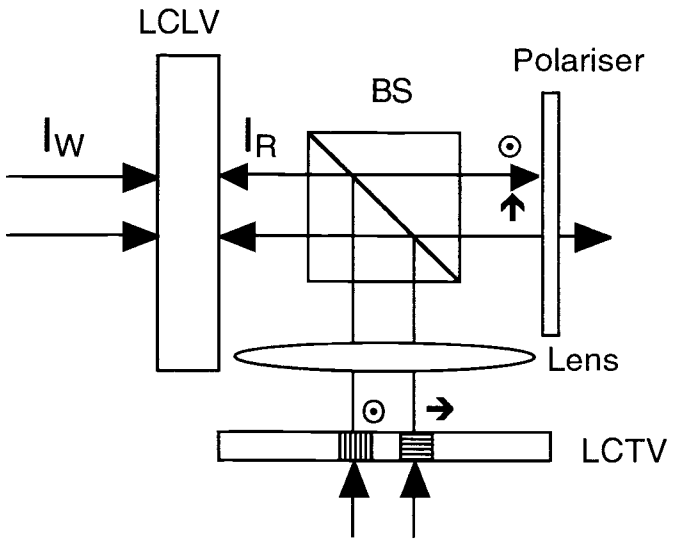


Fig. 12 Implementation of mixed excitatory/inhibitory characteristics for neural plane thresholding.

The investigation of recurrent optical systems of the type shown in Fig. 10 has been motivated by a desire to study the dynamic behaviour of recurrent neural networks rather than fixed state convergence. The dynamic behaviour is a topic of interest in the computer science community[36]. However, it has not been studied in the context of optical implementations because of the overriding interest in static equilibrium states of such networks. It is a relevant topic to study in the optical domain because the necessary analog optical hardware exists and the modelling of the hardware can be readily performed. The dynamic behaviour of the LCLV was mentioned as contributing to the stability of the memory states in ref. 37. Moreover, a dynamic equation for the LCLV different from the one developed here was presented without proof in ref. 38. The algorithm for the recurrent network is the soft-threshold Hopfield algorithm[39]. However, it is formulated in terms of bipolar neural outputs. Therefore, one of the methods discussed above must be used. A combination of weight constraint and the use of inhibitory characteristics which maintains a reasonable memory capacity (number of static equilibrium states) is the Inhibitory Model[40]. The positive components of the weight plane array are set to zero and all the neurons are operated with an inhibitory characteristic. We constructed a recurrent neural network based on the inhibitory model (Fig. 13).

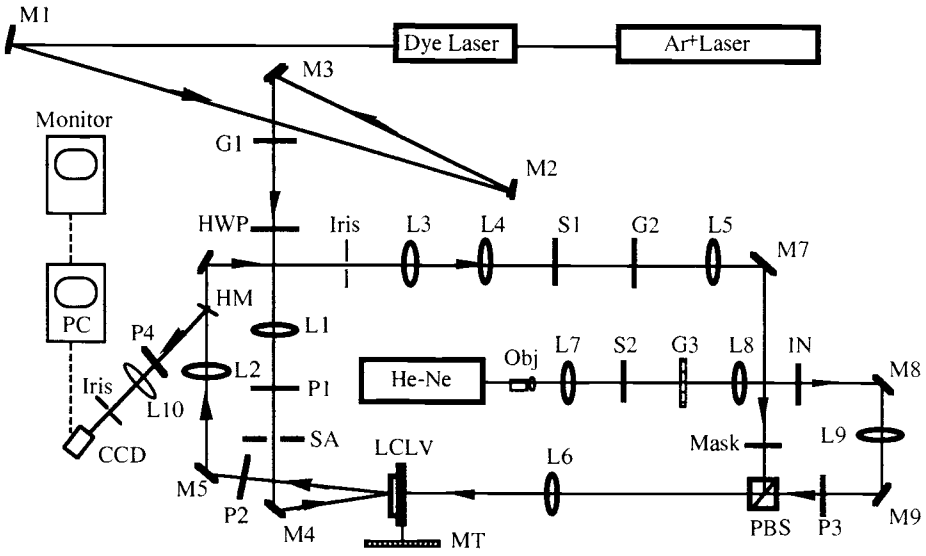


Fig. 13 Recurrent neural network with fixed weight plane (Mask), using the Asulab LCLV and 555 nm write and read wavelengths. Network initiated by a HeNe laser and input mask (IN)[24].

A grating G2 together with lens L5 replicates a 3 x 3 spot array 3 x 3 times on the weight plane (Mask). A lens L6 Fourier transforms the weight plane to a 3 x 3 array on the read side (photoconductor side) of the LCLV. By adjusting the position of G2, the spacing of the write spots can be made to match the spacing of the read spots on the valve. A wedge reflector WR removes 2% of the feedback beam in order that the neural activations can be monitored. A mechanical shutter S1 is used to close the feedback loop when it is desired to stabilise the valve on a static input. The input is effected using a helium neon laser and an input mask. A 3 x 3 spot array is generated at the input mask and imaged onto the read side of the valve by lenses L9 and L6. The input beam also has a shutter S2 which is closed when it is desired to let the feedback loop run freely.

The system operation proceeded as follows. The shutter S1 was closed and S2 opened in order to introduce a initial pattern using the input mask. As soon as the system stabilized on the initial state, S2 was closed and at the same time S1 opened. Then the system ran according to its dynamics. The system at this point operates in continuous time, i.e. the state of each neuron is a continuous function of time and each neuron updates asynchronously with respect to the other neurons.

The main imperfection in this experiment arises from the lack of well-engineered componentry and devices. The standard deviation of the non-uniformity (SDNU) generated by the fan-out G1 was 4%. The SDNU of the OFF-state of the valve (zero write light) was optimised to a minimum of 3% by translating the valve in an xy-sense using a microtranslation stage. The combined SDNU in the reflected spot array was 4%. Due to the larger area and diverging nature of the beam incident on grating G2, the beam fan-out was more non-uniform at this point of the feedback loop. For the beams which will pass through the pixels the weight mask, there was a SDNU of 8%.

Initial attempts to produce grey level weight masks increased this non-uniformity, due to the difficulty of controlling the processing and the critical alignment required for the resulting mask. The mask was designed according to the inhibitory model. Two patterns, a "T" and an "L", were stored in the mask. When the grey level

maskwas fabricated, a SDNU of 11% over the nine fanned-in spots was obtained. In order to improve the uniformity of the fan-in, a binary mask was fabricated to replace the grey level mask. Over each of the 34 transmitting pixels of the binary mask, a piece of polaroid was glued in order to vary manually the grey level transmittance. By this technique a SDNU of 3% over the nine fanned-in spots was achieved. By this means, the non-uniformities of the mask and grating G2 could be corrected. A more satisfactory solution would have been to use an LCTV as the grey level mask as in the MVM1 of Fig. 10. However, the loss in the feedback loop would have been increased and the Asulab LCLV did not have sufficient gain to compensate this loss.

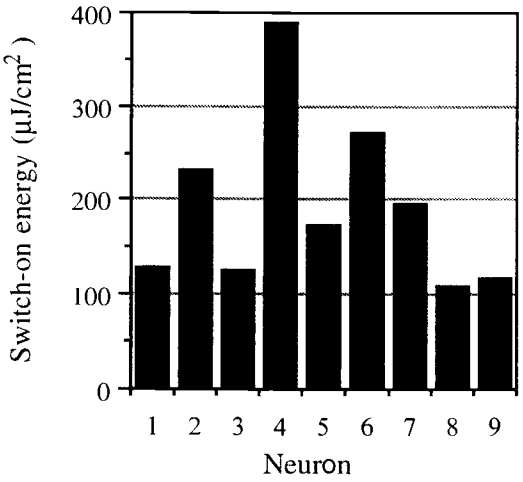


Fig. 14 Non-uniformity of write sensitivity of the Asulab LCLV[24].

The sensitivity of the LCLV, in the area selected for uniform OFF-state, was itself quite non-uniform (Fig. 14). This resulted in the above system always settling down to a steady state equilibrium. Since the facility for changing the interconnection weights was in place, due to the adjustable polaroid on the mask, it was decided to adjust the fan-in to pre-compensate this non-uniformity. This adjustment gave rise to a dynamical equilibrium state as opposed to a static equilibrium state (Fig. 15).

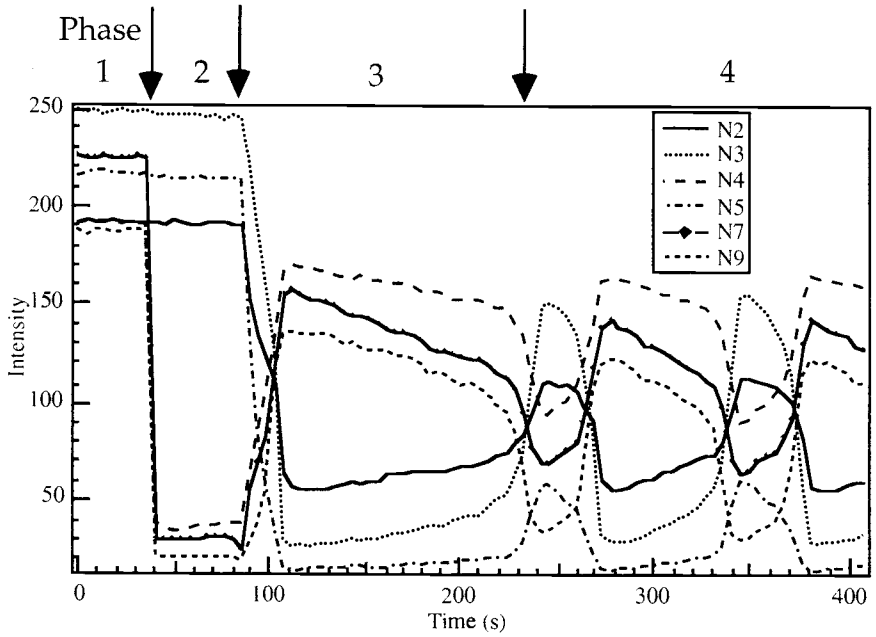


Fig. 15 Intensity of 6 spots of the reflected read beam from the LCLV in Fig. 13 as a function of time.

The temporal evolution of the intensities of each of the nine neurons can be measured using the camera and image analysis software. The interconnection matrix (Mask) is such that the nine neurons can be divided into two independent groups, one of six and the other of three neurons, which act only amongst themselves. The temporal evolution of the six neurons, which form two competing groups of three neurons, for the case when the input image is an "L", is shown in Fig. 15. It can be seen from the figure that the system goes through five distinct phases: In the first phase, the feedback loop was open (S2 closed) and there was no input (S1 closed), therefore all neurons were ON though the intensities were not uniform. In the second phase (S2 closed, S1 open), the initial pattern "L" was introduced in the system. As soon as the feedback loop was opened and the input switched off (S2 open, S1 closed), the system underwent a period of transition, the third phase. The length of the transition takes about 20 secs. In the fourth phase, all the neurons tend to their opposite states (a "T") in a quasi-static equilibrium. Then a fifth phase of regular oscillation between states which are almost "L" and states which are almost "T". There is an asymmetry in this oscillation which favours the "L" state. The cycle time is about 100 seconds. When the initial input image is a "T" the fifth phase is similar with a cycle time of 100 s, but with a different asymmetry, viz the "T" phases of the cycle last longer than the "L" phases. If the intensity of the read-in light was decreased below a value of 700 $\mu W/cm^2$, the oscillation stopped and all neurons remained ON.

The interesting aspects of Fig. 14 are phases 3 and 4. Here, some neurons are abruptly switched from a $I^w = 0$ to a value of I^w determined by the feedback loop. Yet other neurons are switched from a high value of I^w to a low value determined by the feedback loop. What is important to note is that the time constants of the decaying (and rising) neurons are roughly similar. The transition from the third to the fourth phase is interesting because at this point the loop dynamics takes over. The dynamic equation (7) of the LCLV is required to explain the equilibrium dynamic behaviour of the network, ie the oscillations. If it is differentiated, then

$$\frac{dI^{O}(I^{W}, t)}{dt} = \tau G\left[\frac{dI^{W}(t)}{dt}\right] + G(I^{W}). \tag{17}$$

If the first term on the RHS is ignored, then

$$\frac{dI^{\circ}(I^{W}, t)}{dt} = \frac{g(I^{W}) - g(0)}{\tau}. \tag{18}$$

It was mentioned earlier that six neurons form two competing groups. Let these be called A and B, so that $I^{W}_{A}=K_{AB}I^{O}_{B}$, where K is the transmission through the feedback loop from output B to input A. Similarly, $I^{W}_{B}=K_{BA}I^{O}_{A}$. For ease of analysis, it is assumed that $K_{BA} = K_{AB}$. Then Eq. (18) can be written

$$\frac{dI^{\circ}{}_{A}}{dt} = \frac{g(I^{W}{}_{A}) - g(0)}{\tau}. \tag{18'}$$

Differentiating wrt t gives

$$\frac{d^{2}I^{\circ}{}_{A}}{dt^{2}} = \frac{1}{\tau}.g\left(\frac{dI^{W}{}_{A}}{dt}\right). \tag{19}$$

But

$$\frac{dI^{W}{}_{A}}{dt} = K_{AB}.\frac{dI^{O}{}_{B}}{dt} = K_{AB}.\frac{g(I^{W}{}_{B}) - g(0)}{\tau} = K_{AB}.\frac{K_{BA}g(I^{O}{}_{A}) - g(0)}{\tau}. \tag{20}$$

If the nonlinear function g() in the middle of the characteristic (see Fig. 7) is approximated by a straight line of slope M (dynamic gain), then (19) and (20) can be combined to form a harmonic equation of the form

$$\frac{d^{2}I^{O}{}_{A}}{dt^{2}} = K_{AB}.\frac{K_{BA}M^{2}I^{O}{}_{A}}{\tau^{2}}. \tag{21}$$

The period T of the harmonic oscillation is given by

$$\frac{2\pi}{T} = \frac{K_{BA}M}{\tau}. \tag{22}$$

Approximate values for K, M, and t are 0.07, 4, and 1 s (the latter from Fig. 11 when $I^{W} = 20\ \mu W/cm^{2}$). Therefore, T = 21 s, which differs by a factor of 5 from the period observed. This calculation represents a first approach where numerous approximations hinder a closer match with experiment.

6. CONCLUSIONS

This review notes the highlights of one step along the path to understanding and implementing optical neural networks. Very often a lot of work was expended on blind alleys. For instance, a lot of time was spent on the early LCTVs in interfacing, which was later redundant because of the advent of multimedia LCTVs. Of course,

this meant that new optical elements had to be designed for each new device used. In order to extend these studies to UXGA LCTVs, another design/fabrication cycle would have to be performed.

Three components of the network have been described, the MVM module, the LCLV and the system architecture. Work has advanced most rapidly in the first, and we have now an idea on how to make a robust unit of high connectivity. The second area is where more effort is needed in order to make reliable, uniform devices with frame speeds up to 1 kHz. In the third area, we have been able to show where optics may hold unique advantages, but there is a lot to do in order to confirm these advantages and to develop appropriate algorithms and applications. In particular, an efficient means of implementing bipolar connections or neurons would be important, as would a storage prescription which made efficient use of available memory. For example, an XGA LCTV which could be used at 52% capacity is a 1 Mbyte memory, which could be accessed in parallel at 1 kHz rates, using a state of the art LCLV. The area of application would be in tasks which are not efficiently performed on conventional computers, such as optimisation and pattern recognition.

ACKNOWLEDGEMENTS

I would like to thank Wei Xue, Ken Weible, Christoph Berger, and Ali Pourzand and numerous diploma and semester students who have made all the experimental and simulation work on these systems, and all those in the IMT Optics group and at the CSEM Zurich who have contributed to the microoptical element fabrication in these studies. I also wish to thank the Swiss National Science Foundation who have financed the studies on Optical Neural Networks at IMT between 1990 and 1998.

REFERENCES

1. N.H. Farhat, D. Psaltis, A. Prata, and E. Peak, "Optical implementation of the Hopfield model", Appl. Opt., 24, 1469-1475 (1985).

2. J.W. Goodman, A.R. Dias, and L.M. Woody, "Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms", Opt. Lett., 2, 1-3 (1978).

3. A.R. Dias, R.F. Kalman, J.W. Goodman, and A.A. Sawchuk, "Fiber-optic crossbar switch with broadcast capability", Proc SPIE 825, 170-177 (1987).

4. M.A.G. Abushagar and H.J. Caulfield, "Optical matrix computations", in Optical Processing and Computing ed. H.H. Arsenault, T. Szoplik, and B. Macukow (Academic Press; 1989).

5. S.G. Batsell, J.F. Walkup, and T.F. Krile, "Noise issues in optical linear algebra processor design", in Design Issues in Optical Processing ed J.N. Lee (Cambridge; 1995).

6. S.B. Odinokov and A.V. Petrov, "Analysis of addition accuracy in optoelectronic matrix-vector multiplier", Proc. SPIE 2430, 270-278 (1994).

7. N. Collings, "Design considerations for a useful two-layer neural network", Euro-American workshop on Optical Pattern Recognition, eds. B. Javidi and P. Réfrégier pp. 314-333 (SPIE Optical Engineering Press, 1994).

8. K.J. Weible, G. Pedrini, W. Xue, and R. Thalmann, "Optical implementation of a neural network associative memory using diffraction gratings", Jap. J. Appl. Phys. 29, L1301-L1303 (1990).

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

10. N. Collings, A.R. Pourzand, R. Völkel, "The construction of a programmable multilayer analogue neural network using space invariant interconnects" Proc. SPIE 2565, 40-47 (1995).
11. N. Collings and C. Berger, "Demonstration and discussion of an interlaced fan-out interconnect", Inst. Phys. Conf. Ser. 139: Part II, 247-250 (IOP Publishing; 1995).
12. K.J. Weible, "Experimental investigation of optical neural networks and learning systems", Ph.D dissertation (University of Neuchâtel, November, 1993).
13. A.R. Pourzand, "Optimization of 2D liquid crystal devices for use in optical information processing systems", Ph.D dissertation (University of Neuchâtel, June 1998).
14. N. Collings, A.R. Pourzand, F.L.Vladimirov, N.I.Pletneva, A.N.Chaika, "The construction of a multilayer analogue neural network using liquid crystal SLMs", Optical Memory and Neural Networks 6, 187-198 (1997).
15. A.W. Lohmann, et al., "Array illuminators for the optical computer", Proc. SPIE 963, 232-239 (1988).
16. C. Berger, "Compact all-optical recurrent neural network", Ph.D dissertation (University of Neuchâtel, October 1998).
17. C. Berger, N. Collings, R. Völkel, M.T. Gale, and T. Hessler, "A microlens-array-based optical neural network application", JEOS A 6, 683-689 (1997).
18. N. Collings and W. Xue, "Liquid crystal light valves (LCLV) as thresholding elements in neural networks: basic device requirements", Appl. Opt., 33, 2829-2833 (1994).
19. N. Collings, A.R. Pourzand, F.L. Vladimirov, N.I. Pletneva, and A.N. Chaika, "Pixellated reflective light valve for neural network application, " submitted to Applied Optics.
20. J. Grinberg, A. Jacobson, W.P. Bleha, L.Miller, L. Fraas, D. Boswell, G. Myer, "A new real-time non-coherent to coherent light image converter: The hybrid field effect liquid crystal light valve", Opt. Eng. 14, 217-225 (1975).
21. Fabricated for our institute by Asulab S.A., Neuchâtel, Switzerland, in 1984.
22. Purchased from the P.N. Lebedev Institute, Moscow, Russia, in 1996.
23. P.D. Moerland, E. Fiesler, and I. Saxena, "Incorporating LCLV non-linearities in optical neural networks", Appl. Opt. 35, 5301-5307 (1996).
24. W. Xue, "Characterization of liquid crystal light valves for neural network application", Ph.D dissertation (University of Neuchâtel, March 1994).
25. C. Berger, N. Collings, and D. Gehriger, "Recurrent optical neural network for the study of pattern dynamics", Proc. SPIE 3402, 233-244 (1997).
26. N. Collings and W. Xue, "Characterization of optically addressed SLM's for recurrent optical neural networks", Int. J. Optical Computing, 2, 97-107 (1991).
27. N. Collings, A.R. Pourzand, F.L.Vladimirov, N.I.Pletneva, A.N.Chaika, "The construction of a multilayer analogue neural network using liquid crystal SLMs", Optical Memory and Neural Networks 6, 187-198 (1997).
28. T.C.B. Yu, R.J. Mears, A.B. Davey, W.A. Crossland, M.W.G. Snook, N. Collings, and M. Birch, "Smart VLSI/FELC spatial light modulators for neural networks", Proc. SPIE 2430, 243-248 (1994).
29. L.M. Fraas, W.P. Bleha, J. Grinberg, and A.D. Jacobson, "ac photoresponse of a large-area imaging CdS/CdTe heterojunction", J. Appl. Phys., 47, 584-590 (1976).
30. L.M. Fraas, J. Grinberg, W.P. Bleha, and A.D. Jacobson, "Novel charge-storage-diode structure for use with light-activated displays", J. Appl. Phys.,47, 576-583 (1976).
31. P.D. Moerland, E. Fiesler, and I. Saxena, "Discrete all-positive multilayer perceptrons for optical implementation", Opt. Eng. 37, 1305-1315 (1998).

32. H.J. Wright and W.A. Wright, "Holographic implementation of a Hopfield model with discrete weightings", Appl. Opt. 27, 331- 338 (1988).
33. B.K. Jenkins and C.H. Wang, "Model for an incoherent optical neuron that subtracts", Opt. Lett. 13, 892- 894 (1988).
34. Y. Hayasaki et al., "Reversal-input superposing technique for all-optical neural network", Appl. Opt. 33, 1477-1484 (1994).
35. F.M. Dickey and J.M. DeLaurentis, "Optical neural networks with unipolar weights", Opt. Comm. 101, 303-305 (1993).
36. IEEE Trans. Neur. Networks 5, (1994) Special issue on dynamic recurrent neural networks.
37. H.J. White, "Experimental results from an optical implementation of a simple neural network", Proc. SPIE 963, 570-575 (1988).
38. K.-Y. Hsu, H.-Y. Li, and D. Psaltis, "Holographic implementation of a fully connected neural network", Proc.IEEE 78, 1637-1645 (1990).
39. J.J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons", Proc. Nat. Acad. Sci. USA 81, 3088-3092 (1984).
40. I. Shariv and A.A. Friesem, "All-optical neural network with inhibitory neurons", Opt. Lett. 14, 485-487 (1989).