Dynamic digital photorefractive memory for optoelectronic neural network learning modules

Hironori Sasaki, Nicolas Mauduit, Jian Ma, Yeshiahu Fainman, Sing H. Lee, and Michael S. Gray

Neural network modules based on page-oriented dynamic digital photorefractive memory are described. The modules can implement two different interconnection organizations, fan-out and fan-in, depending on their target network applications. Neural network learning is realized by the real-time memory update of dynamic digital photorefractive memory. Physical separation of subvolumes in the page-oriented photorefractive memory architecture contributes to the low cross talk and high diffraction efficiency of the stored interconnection weights. Digitally encoded interconnection weights ensure high accuracy, providing superior neural network system scalability. Module scalability and feedforward throughput have been investigated based on photorefractive memory geometry and the photodetector power requirements. The following four approaches to extend module scalability are discussed: partial optical summation, semiparallel feedforward operation, time partitioning, and interconnection matrix partitioning. Learning capabilities of the system are investigated in terms of required interconnection primitives for implementing learning processes and three memory-update schemes. The experimental results of Perceptron learning network implementation with 900 input neurons with digital 6-bit accuracy are reported. © 1996 Optical Society of America

1. Introduction

Because of the dense and massively parallel interconnections of neural networks, it is generally recognized that optics may play a major role in the implementation of neural networks.1,2 The efficient implementation of neural networks requires realization of the following functionalities: storage of interconnection weights with fast random-access times, calculations of inner-product terms between the interconnection weight vectors and input vectors, and nonlinear thresholding at the output of each neuron. In addition, a large fan-out capability is essential at each neuron to permit system scalability and cascadability. The large capacity and parallel random-access capability of dynamic photorefractive memory3–5 are well suited for the storage, fast access, and fan-out of interconnection weights. Optics are also efficient in performing point-by-point multiplications of large arrays of data. However, accurate summation and thresholding functionalities are more suitable for electronic implementation.

From these considerations, we have investigated a hybrid optoelectronic neural network system that combines the strength of optics and electronics.6 In our proposed system, a series of two-dimensional (2-D) digitally encoded interconnection weight images are stored in the array of small holographic subvolumes. The interconnection weights are decomposed into multiple digital images multiplexed at each subvolume, taking advantage of the photorefractive memory, where the number of multiplexed images corresponds to the number of digital bits. Here we analyze the angular multiplexing scheme,7 but other multiplexing techniques such as wavelength multiplexing7 or orthogonal phase coding8,9 can also be used. The interconnections are implemented as bit serial multiplications between the digitally represented 2-D weight images and the digitally represented 2-D input data planes. The summation and thresholding operations are performed electronically to maintain high accuracy. The use of dynamic digital 3-D photorefractive memory realizes the large capacity for weight storage and...
the interconnection adaptability required by learning. Digital computation achieves less dependence on the nonuniformity of optical components, leading to scalable neural network systems with high throughput.

In Section 2, a neural network architecture based on a digital photorefractive page-oriented optical memory is described and characterized in terms of system scalability and feedforward throughput. Various techniques to increase the system scalability are suggested in Section 3. The learning capability of the proposed system is discussed in Section 4 in terms of various interconnection primitives that are required for learning, and three memory-update schemes are compared for higher learning rates. Section 5 shows an experimental demonstration of Perceptron learning, and Section 6 concludes this paper.

2. Digital Neural Network Module Architecture

Most of the previous optoelectronic neural network implementations\textsuperscript{10–18} are based on the use of analog optical computing. However, because optical components have certain nonuniformities, noise, and limited dynamic range, these factors tend to limit the scalability of these types of optically implemented system.\textsuperscript{19} For these limitations to be overcome, it is justified to implement digital optoelectronic neural networks in which the thresholding and restoring of binary logic levels lead to robust system operation, superior scalability, and superior throughput.\textsuperscript{6} Both the interconnection weight images and input data images are decomposed into the sequences of bit planes. Therefore the number of multiplexed bit planes is determined by the required accuracy of the input data and the interconnection weights. The number of pixels in each interconnection weight plane represents the number of output neurons or input neurons corresponding to two different architectures, which are explained in detail in Subsection 2.A. The feedforward operation of the neural network is performed in parallel as the multiplication of 2-D input data bit planes and 2-D interconnection weight bit planes. The digital multiplication is implemented in a pixel-by-pixel bit-serial manner, i.e., a digital multiplication process is realized in $n^2$-step binary multiplications in which both input data and interconnection weights are represented digitally in $n$ bits. Note that this multiplication occurs in parallel for all the input neurons, i.e., input data and interconnection weight pages. This permits the system to reach a lower bit error rate (BER), which is essential for the reliable operation of large-scale computing systems.

The neural network module implements a two-layer network consisting of an input and an output layer. Such a module can be cascaded to construct a multilayer network.

A. Two Complementary Interconnection Architectures

The optical interconnections are based on page-oriented memory architecture.\textsuperscript{11,16,20–24} Hereafter, we assume the total number of input neurons is $N$ and the total number of output neurons is $M$. In the following discussion, the input array and the output array are represented in vector forms, $\mathbf{x}$ and $\mathbf{y}$, respectively, and the interconnection weight is represented as an element $w_{mn}$ of the interconnection matrix $W$, as shown in Eq. (1):

$$
\mathbf{y}_m = \sum_{n=1}^{N} w_{mn} \mathbf{x}_n.
$$

Although the number of total interconnections between $N$ input neurons and $M$ output neurons is $NM$, the interconnections can be equivalently organized as $N$ sets of fan-out interconnections at $N$ input neurons, or $M$ sets of fan-in interconnections at $M$ output neurons.\textsuperscript{6,25} Using this fact we propose two different architectures, the fan-out and the fan-in interconnection architectures as shown in Figs. 1(a) and 2(a), respectively. Note that both fan-out and fan-in architectures can independently implement any two-layer network. However, there are certain preferences in choosing such two architectures, depending on the number of input and output neurons, as we discuss below.

Figure 1(a) shows the fan-out optical system, in which partitioning into holographic subvolumes is realized by imaging and overlapping two 2-D arrays of mutually coherent light sources onto the photorefractive crystal (PRC). Each fan-out interconnection pattern $w_{mn}$ is displayed on the spatial light modulator (SLM), and its optical Fourier transform (FT) is recorded in the corresponding $n$th photorefractive subvolume. A set of fan-out interconnections, $w_{mn}$, connect the $n$th input neuron to $M$ output neurons. Therefore the total number of holographic subvolumes is equal to the number of input neurons
N. For angular multiplexing, we need to use tilt, i.e., rotation, of the light sources in the input data plane as shown in Fig. 1(a). For wavelength multiplexing, the multiplexed weight images can be selectively accessed by changing the wavelength of the light source. The feedforward operation is obtained by activating the input light source array with corresponding input data. First each bit at the input plane is turned on, depending on the bit representation of the input values at each input neuron. Then all the multiplexed interconnection bit patterns are sequentially read out by changing the tilt of the input data plane, or by changing the light source wavelength for angular multiplexing and wavelength multiplexing, respectively. Next the input plane is changed to the next bit of the input pattern and the same procedure is repeated. These digital bit-by-bit multiplication results are stored at each output neuron to calculate the interconnection between the input data and the interconnection weights: \( \sum_{n=1}^{N} w_{mn} x_n \). The output plane consists of a photodetector array and the associated electronic circuitry to store feedforward results.

To secure digital accuracy at the output neuron plane, we need to modify the optical system shown in Fig. 1(a), as illustrated in Fig. 1(b), where only the portion of the optical system after the photorefractive crystal is described. The rest of the system remains the same, as shown in Fig. 1(a). Here \( f_1 \) denotes the focal length of the FT lens and \( f_2 \) denotes the focal length of the lenslet. In order to realize digital summation at each output neuron, \( N \) incident beams must be detected. This is realized by a lenslet array placed at the back focal plane of the Fourier transform lens after the photorefractive crystal. Each slanted plane-wave incident angle is caused by different subvolume locations on the photorefractive crystal and results in a corresponding slanted angle at the lenslet array plane. This slanted angle difference is again converted to a lateral shift by the lenslet array as shown in Fig. 1(b). By the allocation of \( N \) binary photodetectors at each output neuron, \( N \) incident beams can be detected, and the summation results are obtained. Figure 1(b) shows a case of four photorefractive subvolumes (\( N = 4 \)) and nine output neurons (\( M = 9 \)).

For the proposed fan-in architecture, the fan-in interconnection weight image, \( w_{mn} \), is displayed on SLM1 and its Fourier transform is stored in the corresponding \( m \)th holographic subvolume as shown in Fig. 2(a). A set of fan-in interconnections, \( w_{mn} \), connects \( N \) input neurons to a single output neuron. Therefore the number of subvolumes is equal to the number of output neurons, \( M \). In the fan-in architecture, feedforward operation is obtained by activating the entire output selector light source array and by loading the input neuron pattern, \( x \), on the input SLM2. The inner product between interconnection patterns \( W \) and input pattern \( x \) is performed on a bit-by-bit basis by the transmission through SLM2, and then it is Fourier transformed to yield the output neuron pattern on the output photodetector array.

To secure digital accuracy, we need to add a lenslet array in the back of the Fourier transform lens behind the input SLM2, as shown in Fig. 2(b). Here \( f_1 \) denotes the focal length of the two FT lenses and \( f_2 \) denotes the focal length of the lenslet. The output of SLM2 is demagnified and imaged at each output neuron location, detected, and summed electronically. Figure 2(b) shows an example for \( N = 4 \) input neurons and \( M = 9 \) output neurons.

Because the module system is based on the page-oriented digital holographic memory, the scalability and the throughput of the system are determined by the total memory capacity and space-bandwidth product (SBWP) of 2-D images recorded in the memory. We first estimate the module system scalability by following the procedure described in Ref. 6.

B. Scalability Considerations

A photorefractive crystal used in transmission geometry is assumed for analysis. The number of weight images, \( G_{\text{sub}} \), multiplexed at each photorefractive subvolume is calculated from the required optical power, \( P_{\text{det},b} \), to drive a photodetector and the diffraction efficiency of each weight image. Assuming the use of Kogelnik’s volume diffraction efficiency equation, \( \eta \), we see that the diffraction efficiency, \( \eta \), of each grating is given by

\[
\eta = \eta_s \sin^2 \left( \frac{T \pi n_{\text{max}} / G_{\text{sub}}}{\lambda \cos \theta} \right),
\]

where \( \eta_s \) is the overall optical system efficiency, including component reflection loss and absorption by the photorefractive crystal, \( T \) is the thickness of the crystal, \( n_{\text{max}} \) is the maximum allowable refractive index modulation, \( \lambda \) is the wavelength in vacuum, and \( \theta \) is half the intersecting angle of the object and reference beams inside the crystal. The
required optical power to drive photodetector $P_{\text{det}}$ is given by\textsuperscript{26,27}

$$P_{\text{det}} = \frac{\mu hc}{t_\alpha \eta_q}, \quad (3)$$

where $\mu$ is the number of photons required by a photodetector in order to yield a desired BER rate: typically, $\mu = 1000$ for BER = $10^{-9}$, $h$ is Planck’s constant, $c$ is the speed of light, $t_\alpha$ is the memory access time, and $\eta_q$ is photodetector quantum efficiency. For a given $P_{\text{det}}$, the SBWP of the SLM, $B_{\text{SLM}}$, and laser power illuminating each subvolume, $P_L$, we can find from $P_{\text{det}} = P_{\text{det}}B_{\text{SLM}}$ [see Eqs. (2) and (3)] that the number of multiplexed images, $G_{\text{sub}}$, is given by

$$G_{\text{sub}} = \frac{T \pi \Delta n_{\text{max}}}{\lambda \cos \theta \sin^{-1}(P_{\text{det}}B_{\text{SLM}}/P_L \eta_q)^{1/2}}. \quad (4)$$

The number of holographic subvolumes in the given photorefractive crystal, denoted by $Q$, that corresponds to the number of input and output neurons for fan-out and fan-in architectures, respectively, is defined by using the geometrical constraint of the photorefractive crystal [see Figs. 3(a) and 3(b)]:

$$Q = \left( \frac{S}{\Delta S} \right)^2, \quad (5)$$

where $S$ is the linear dimension of the photorefractive crystal and $\Delta S$ is the linear dimension of each subvolume. The linear dimension of subvolume $\Delta S$ is determined by using the trapezoid defined by the focal length of the FT lens, $f_1$; the linear dimension of the SLM, $\Delta a \sqrt{B_{\text{SLM}}}$, where $\Delta a$ is the pixel size of the SLM; and the distance between the first lobes of the FT of a square SLM pixel, $\Delta x$, as shown in Fig. 3(a):

$$\Delta S = \frac{T}{2f_1} (\Delta a \sqrt{B_{\text{SLM}}} - \Delta x) + \Delta x, \quad (6)$$

$$\Delta x = \frac{2f_1}{\Delta a}. \quad (7)$$

The focal length of the FT lens, $f_1$, is determined as a function of the lens $f$-number, $F/#$, which is shown in Fig. 3(b) to be

$$f_1 = \sqrt{2}(S + \Delta a \sqrt{B_{\text{SLM}}})F/#. \quad (8)$$

Figure 4 shows the number of possible multiplexed weight images $G_{\text{sub}}$ at each subvolume of LiNbO$_3$ and the number of subvolumes $Q$ as functions of the SBWP of the SLM, $B_{\text{SLM}}$. The SLM SBWP corresponds to the number of output and input neurons for fan-out and fan-in architectures, respectively. The number of subvolumes corresponds to the number of input and output neurons for fan-out and fan-in architectures, respectively. The following parameters are used in the calculations for the rest of this paper: $T = 2$ mm, $\Delta n_{\text{max}} = 1 \times 10^{-3}$, $\lambda = 0.514$ $\mu$m, $\theta = 10^\circ$, $\mu = 1000$ for BER = $10^{-9}$, $h = 6.63 \times 10^{-34}$, $c = 3 \times 10^8$, $t_\alpha = 1$ $\mu$s, $P_L = 10$ mW, $\eta_q = 0.5$, $\eta_\beta = 0.1$, $F/# = 3$, $\Delta a = 100$ $\mu$m, and $S = 5$ cm. Figure 4 shows that a SLM SBWP as great as $1000 \times 1000$ can be implemented in this system, whereas for a smaller SBWP the system is underutilizing the potential memory capacity of the photorefractive memory. For example, for a SLM SBWP of $100 \times 100$, more than 100 weight images can be multiplexed with enough diffraction efficiency. However, 100-bit digital interconnection accuracy may be unnecessary for any neural network applications. Below we discuss
how to increase system scalability by using this potential memory capacity gap. The number of the subvolumes decreases as the SBWP of the SLM increases, because the larger numerical aperture of the object beam leads to the larger size of the subvolume.

C. Feedforward Throughput

Because all the holographic subvolumes are accessed fully in parallel, a large value for feedforward throughput is expected. The feedforward throughput, $T_p$, is given by

$$ T_p = \frac{P_{SLM} Q}{t_w G_{bit}^2}, $$ \hfill (9)

where $G_{bit}$ is the required bit accuracy to calculate feedforward operations. Here we assume that both the input data and the interconnection weights have the digital accuracy of $G_{bit}$, and that the multiplication of the values require $G_{bit}^2$ steps of memory access. Figure 5 shows the feedforward throughput (interconnections per second) as a function of the SBWP of the SLM. For calculations, $G_{bit} = 6$ was assumed. The throughput gradually increases as the SLM SBWP increases. For a system with a SLM SBWP larger than $3 \times 10^3$, a throughput of more than $10^{12}$ interconnections/s can be achieved.

For fan-out architecture, the number of input neurons, $N$, identical to the number of holographic subvolumes, decreases as the SLM SBWP increases (see Fig. 4). This is because the numerical aperture increase of the incident beam requires a large subvolume area for a given photorefractive crystal thickness. However, feedforward throughput increases as the SLM SBWP increases. For fan-in architecture, the number of output neurons, $M$, shows the same behavior as the number of input neurons, $N$, for fan-out architecture, because both systems are complementary. Although it may be possible to make the number of input neurons much larger than the number of output neurons for fan-out architecture, the large number of holographic subvolumes is not preferable, especially if we consider the total recording time required for learning. Therefore, in general, fan-out architecture is suitable for cases requiring large numbers of output neurons, whereas fan-in architecture is more suitable for networks with large numbers of input neurons.

3. Techniques to Construct More Scalable Systems

As we discussed in Section 2, the optical system is capable of handling up to a $1000 \times 1000$ SBWP for SLMs. However, the area required by electronic circuitry at the output neuron plane may be a scalability limiting factor. Given the photodetector array size, $S_{det}$, in each dimension, we see that the corresponding photodetector size, $\Delta s_{det}$, in each dimension, including the necessary electronic circuitry, is given by

$$ \Delta s_{det} = \frac{S_{det}}{(QB_{SLM})^{1/2}}. $$ \hfill (10)

The total number of photodetectors is a product of the SLM SBWP and the number of photorefractive subvolumes, $QB_{SLM}$. Because the decrease in the number of subvolumes is moderate, compared with the increase of the SLM SBWP as shown in Fig. 4, the photodetector area decreases as the SLM SBWP increases, as shown in Fig. 6. For calculations, the photodetector array plane was assumed to be a square of 10 cm in each dimension. Therefore, it is likely that the proposed neural network’s system scalability might be limited by the photodetector and the corresponding electronic circuit area requirements.

Below we will discuss possible ways to increase the scalability of both fan-out and fan-in architectures without significantly increasing the SLM and output neuron plane sizes.

A. Partial Optical Summation

The area limitation caused by electronic circuitry will be reduced if the required number of photodetectors is reduced. For example, each output neuron has $N$ photodetectors to sum all the $N$ input beams. Therefore, if the photodetector can detect $N$ different incident beam levels, $N$ photodetectors can be re-
duced to a single photodetector, leading to a reduced area requirement.

The photodetector area requirement may be alleviated by assigning more than one input beam or bit to each photodetector, leading to partial optical summation of the incident beams. For example, if each photodetector can detect four input beams, the scalability limitation caused by the photodetector area limitation might be improved by a factor of 4. Although every optical component has spatial nonuniformity in its characteristics, it might be possible to realize partial optical summation if the number and the physical position of the incident beams of interest are limited. For example, in fan-out architecture, N incident beams pass through the same pixels of SLM1 during the recording process. Therefore, by limiting the number of partially optically summed beams at the output neuron originating from the adjacent holographic subvolumes, we can greatly reduce the spatial nonuniformity of the incident beams [see Figs. 1(a) and 1(b)]. For fan-in architecture, N incident beams at the same output neuron are recorded in the same subvolumes and pass through the adjacent pixels of both SLM1 and SLM2 [see Figs. 2(a) and 2(b)]. Therefore, by locally summing the limited number of incident beams, we may expect that the spatial nonuniformity of the device will be greatly reduced.

B. Semiparallel Feedforward

Another way to reduce the photodetector area requirement at the output neuron plane is to access holographic subvolumes in a semiparallel fashion. With this approach, Qs out of Q subvolumes are accessed in parallel. The feedforward throughput, \( T_p \), given by Eq. (9), will be reduced to \( T_p' \):

\[
T_p' = \frac{B_{SLM} Q_s}{t_e G_{bit}}. \tag{11}
\]

The values of \( Q_s \) can be changed between \( Q \) and unity, corresponding to the fully parallel access and the fully sequential access, respectively. The total number of photodetectors at the output plane is reduced from \( B_{SLM} Q \) to \( B_{SLM} Q_s \). The physical locations of subvolumes are preserved at the photodetector plane for both the fan-out and fan-in architectures, as shown in Figs. 1(b) and 2(b). Therefore it is necessary to adjust the subvolume location such that each beam spot falls on the same photodetector for every reconstruction step in the process of semiparallel feedforward operations. This can be realized by placing the photorefractive crystal on a translation stage or by simply rotating the photorefractive crystal.

C. Time Partitioning

Because the number of multiplexed images in each holographic subvolume is small and the system’s scalability is not limited by the storage capacity of the page-oriented digital photorefractive memory, it is possible to store more than \( G_{bit} \) weight images in each photorefractive subvolume, where \( G_{bit} \) is the number of bit accuracy required for the interconnection weights. If \( 2G_{bit} \) weight images can be stored in each holographic subvolume, then an extra set of \( G_{bit} \) weight images can be used to double the number of input neurons or output neurons of the network.

For fan-out architecture, the number of either input neurons or output neurons can be doubled by multiplexing twice the number of interconnection weight images at each photorefractive subvolume. The input vector is divided into two portions. On one hand, each divided portion sequentially drives a corresponding 2-D light source array as shown in Fig. 1(a). On the other hand, in order for the number of output neurons to be doubled, the interconnection matrix is divided into two portions, and two different portions of interconnection weight images are multiplexed at each subvolume.

For fan-in architecture, either the number of input neurons or the number of output neurons can also be doubled by multiplexing twice the number of interconnection weights at each photorefractive subvolume. In order for the input neurons to be doubled, fan-in interconnection weights are divided into two portions. Two sets of interconnection weights are displayed sequentially on SLM1 in Fig. 2(a), and they are multiplexed at the same photorefractive subvolume. Note that the input image is also divided into two portions and loaded sequentially into SLM2 in Fig. 2(a), during the feedforward process. The number of output neurons can also be doubled by multiplexing two sets of interconnection weights at the same subvolume.

In this time-partitioning approach, the increase in neural network scalability is realized by increasing the feedforward execution time. Time-partitioning factor \( D_t \) is obtained from Eq. (4) and the binary-represented interconnection weight accuracy \( G_{bit} \), and it is given by

\[
D_t = G_{sub}/G_{bit}. \tag{12}
\]

By utilizing the potential memory capacity of the digital photorefractive memory, we can increase the number of input neurons, \( N \), to \( ND_t \), or we can increase the number of output neurons, \( M \), to \( MD_t \). Figure 7 shows numerical calculation results of
time-partitioning factor \( D_t \) as a function of the SBWP of the SLM, \( B_{\text{SLM}} \). For the calculations the interconnection weight accuracy was assumed to be \( G_{\text{bit}} = 6 \). The effect of time partitioning is apparent in the lower value regions of the SBWP of the SLM’s. This is because the possible number of multiplexed gratings decreases as the SLM SBWP decreases (see Fig. 4). Therefore time partitioning may be useful in improving scalability if the fully parallel scalability is limited by either the SBWP of the SLM’s or by the output neuron plane’s photodetector area. Although system scalability is increased by time partitioning, the system throughput remains the same because the scalability increase is attained by the time-sequential feedforward operations.

D. Interconnection Matrix Partitioning

The interconnection matrix, \( W \), always can be divided into smaller submatrices:

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n \\
  y_{m+1} \\
  \vdots \\
  y_M
\end{bmatrix}
= \begin{bmatrix}
  w_{11} & \cdots & w_{1n} & w_{1n+1} & \cdots & w_{1N} \\
  \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  w_{m1} & \cdots & w_{mn} & w_{mn+1} & \cdots & w_{mN} \\
  \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  w_{M1} & \cdots & w_{Mn} & w_{Mn+1} & \cdots & w_{MN}
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  \vdots \\
  x_n \\
  x_{n+1} \\
  \vdots \\
  x_N
\end{bmatrix}
\]

(13)

In Eq. (13) the interconnection matrix is divided into two-by-two submatrices, and both input and output are divided into two subvectors, respectively. In this case, Eq. (13) is realized by putting four optoelectronic neural network modules together. The modular design of the neural network enables us to increase the scalability of the system by partitioning the interconnection matrix. Because the number of both input and output neurons is decreased, the scalability limitation imposed by the photodetector area requirement is alleviated. However, at the same time, the interconnection matrix partitioning leads to the decrease in the utilization of the storage capacity of the photorefractive memories.

4. Learning with Neural Network Modules

The adaptability of neural networks to a given problem by learning is an important feature. As examples of interconnection primitives, we employ Widrow–Hoff learning.\(^{28}\) Widrow–Hoff learning is expressed by

\[
W^{(\text{new})} = W^{(\text{old})} + \alpha \delta^T \cdot x, \quad (14)
\]

where \( W^{(\text{old})} \) are the old weights, \( W^{(\text{new})} \) are the updated weights, \( w_{mn} \) is the interconnection weight connecting input neuron \( x_n \) and output neuron \( y_m \), \( \alpha \) is the learning rate, and \( \delta_m \) is the element of error vector \( \delta \), corresponding to the error between the network output and the desired output at the \( m \)th output neuron, respectively. In both fan-out and fan-in architectures, new interconnection weights are read out, calculated, and stored for every subvolume in a sequential fashion. More specifically, weight images corresponding to a row vector in Eq. (15) are multiplexed in the same subvolume in the fan-out architecture. In contrast, a column vector is represented by a set of multiplexed weight images at each subvolume for the fan-in architecture.

A. Learning Architecture

For fan-out architecture, the process of learning is as follows. First, error \( \delta_m \) is calculated at \( M \) output neurons as the difference of the network output and the desired output. Then each error \( \delta_m \) (\( m = 1, \ldots, M \)) is multiplied by an input value \( x_n \) and summed with old weight \( w_{mn}^{(\text{old})} \) as shown in Eq. (15). In our proposed system, we need to read the old weight images out of the holographic memory. In other analog approaches based on photorefractive crystals,\(^{12,14,17,18}\) the stored interconnection weights are directly modified by using the multiplication of error and input signal, based on the dynamics of photorefractive materials. Because the interconnection is expressed in a digital fashion in our proposed system, we need to read out weight images and calculate new weights by using the old weights.

Figure 8 shows the modified architecture for fan-out learning architecture. Each output neuron is equipped with one modulator, and the newly calcu-
lated weights are fed back to SLM1 by imaging. The newly calculated weights are detected at SLM1, and the weight images are recorded at each subvolume. The number of output neurons, $M$, is different from the SBWP for SLM1, $N$ [see Figs. 2(a) and 2(b)] for the fan-in architecture. In order to overcome this SBWP mismatch, the new weights must be calculated in SLM2, which also has a detector array to receive old weights from the photorefractive subvolume, as shown in Fig. 9. At SLM2, the input data are multiplied by error signal $\delta_m$ that is fed back from the output layer, and summed with old weights $w_m^{\text{old}}$. The newly calculated interconnection weights are then fed back to SLM1 by imaging.

**B. Interconnection Primitives for Learning**

Input data $x_n$ have to be broadcast to update new weights. For fan-out architecture, the inputs $x_n$ at each input neuron have to be broadcast to all $M$ output neurons in order to modify the $n$th set of interconnection weights according to

$$
\begin{bmatrix}
w_{1n}^{\text{new}} \\
\vdots \\
w_{Mn}^{\text{new}}
\end{bmatrix} =
\begin{bmatrix}
w_{1n}^{\text{old}} \\
\vdots \\
w_{Mn}^{\text{old}}
\end{bmatrix} + \alpha \begin{bmatrix}
\delta_1 x_n \\
\vdots \\
\delta_M x_n
\end{bmatrix}.
$$

(16)

This can be implemented by adding one more weight image at every subvolume. This weight image is used to connect one input to all the output neurons with interconnection weights of unity. Similarly, for fan-in architecture, we add one more weight image whose interconnection weights are unity for each output neuron. This modification enables us to broadcast all the input $x$ to SLM2 and update the $m$th set of interconnection weights given by

$$
[w_{m1} \cdots w_{mN}]^{\text{new}} = [w_{m1} \cdots w_{mN}]^{\text{old}} + \delta_m x_n.
$$

(17)

For both fan-out and fan-in architectures, the old weights have to be broadcast from the photorefractive crystal to the processing plane as shown in Eqs. (16) and (17), respectively. This is realized by simply reconstructing $G_{\text{sub}}$ multiplexed weight images for both fan-out and fan-in architectures by setting all the inputs unity.

Backward error propagation\(^{29}\) is used to train a three-layer network consisting of input, hidden, and output layers. The calculated error signals are backpropagated from the output layer to the hidden layer in order to train the hidden layer. The error is multiplied by interconnection weights and backpropagated to the hidden layer. For fan-out architecture, the summation of the backpropagated error is calculated at the output plane, by multiplying the error by the reconstructed interconnection pattern and by locally summing such multiplication results (see Fig. 8). Each backpropagated error will be fed back to the hidden layer either electronically or optically in a semiparallel manner. In Fig. 9 the output error, $\delta_m$, is fed back to the output selector array in the fan-in architecture, and the fan-in interconnection weights are reconstructed and multiplied by error $\delta_m$. The feed-back error is obtained at SLM2 fully in parallel and sent back to the previous layer.

C. Comparison of Memory-Update Strategies

Because photorefractive crystals are real-time holographic recording media, learning can be realized by modifying weight images stored in the photorefractive memory. In this section several different memory-updating strategies are investigated and compared. In the following discussions the dynamics of both writing and uniform erasure of a weight image are assumed to be exponential with the constant material response time, $\tau$.

Suppose one out of $G_{\text{sub}}$ multiplexed weight images is to be erased and a new weight image is overwritten at the same multiplexing order. It has been shown that selective erasure, a superposition of a $\pi$-phase-shifted image onto the original one, is much faster than uniform erasure.4 In this approach all the $G_{\text{sub}}$ weight images are recorded for a short time increment, $\Delta t$, based on incremental recording.8 Only the weight image that is to be updated is selectively erased for $\Delta t$ by adding a $\pi$-phase shift to one of either an object or a reference beam. The other $G_{\text{sub}} - 1$ weight images simply repeat the incremental recording cycles. After the desired image is completely erased, a new image is overwritten by the incremental recording technique. The total update time, $T_1$ (see Appendix A), is given by

$$
T_1 = \frac{\tau G_{\text{sub}}}{2(G_{\text{sub}} - 1) \ln \frac{\gamma}{\gamma - (1 - \xi)/G_{\text{sub}}}} - \tau \ln(1 - \sigma),
$$

(18)

where

$$
\xi = (1 + \varepsilon)^{-1/2(G_{\text{sub}}/(G_{\text{sub}} - 1))},
\gamma = (1 + \varepsilon)^{-1/2(G_{\text{sub}}/(G_{\text{sub}} - 1))} - (1 + \varepsilon)^{-1/2}.
$$

(19)

In Eqs. (18) and (19), $\tau$ is the material response time, $\sigma$ is a saturation parameter in incremental recording by which the incremental recording process cycle is stopped when the refractive index modulation is equal to $\sigma \Delta n_{\text{max}}/G_{\text{sub}}$, and $\varepsilon$ is the allowable variation.
of diffraction efficiencies of multiplexed weights. Total update time $T_1$ is almost independent of the number of the multiplexed gratings, $G_{\text{sub}}$, as shown in Fig. 10.

The second strategy is to erase the weight image selectively and to overwrite a new one without incrementally recording other $G_{\text{sub}} - 1$ weight images. After the new weight image is updated, the incremental recording resumes. Update time $T_2$ for a single weight image (see Appendix B) is given by

$$T_2 = \tau \ln \left( \frac{(1-\sigma)G_{\text{sub}} + \sigma + 1}{(G_{\text{sub}} - 1)(1-\sigma)} \right).$$

During this update time, other $G_{\text{sub}} - 1$ weight images suffer from refractive index modulation degradation $R$, defined by

$$R = \frac{G_{\text{sub}} - 1}{G_{\text{sub}} + 1}.$$  

Numerical simulation results of the normalized update time for both strategies 1 and 2 are compared in Fig. 10 as a function of the number of multiplexed weight images, $G_{\text{sub}}$. The update time is normalized by material response time $\tau$. For numerical calculations, we assumed $\sigma = 0.95$ and $\epsilon = 0.1$. It is clear that strategy 2 has a shorter update time than strategy 1. However, in order to utilize strategy 2, degradation $R$ during selective erasure and overwriting must be negligible. For example, if we want $R > 95\%$, $G_{\text{sub}}$ must be larger than 40.

The third update strategy is simply to overwrite $G_{\text{sub}}$ weight images in the subvolumes without selective erasure. In digital neural network implementation, interconnection weights are recorded in digital representation. Once the interconnection is to be updated, all the $G_{\text{sub}}$ weight images may need to be updated. The old weight images are uniformly erased during the incremental recording cycles. Total time $T_3$ required for incremental recording is given by

$$T_3 = -\tau \ln(1-\sigma).$$

After the incremental recording, the refractive index modulation of the old weight images is erased by a factor $E$, defined as

$$E = 1 - \sigma.$$  

If $\sigma = 0.95$, $E$ becomes 0.05. This implies that the diffraction efficiency of old weight images decreases to 0.25% of their original value. It is clear from Eq. (22) that the total time to update $G_{\text{sub}}$ gratings is independent of $G_{\text{sub}}$. To evaluate total time $T_3$, we compare $T_3$ with the total time of strategy 2, $T_2$. From Eq. (20), the total time to update $G_M$ weight images out of $G_{\text{sub}}$ gratings based on strategy 2 is

$$T_M = G_M\tau \ln \left( \frac{(1-\sigma)G_{\text{sub}} + \sigma + 1}{(G_{\text{sub}} - 1)(1-\sigma)} \right).$$  

By equating Eq. (22) and Eq. (24), we obtain the break-even number, $G_M$, for strategy 2 at which the total update time for both strategies 2 and 3 are equal:

$$G_M = \frac{\ln(1-\sigma)}{\ln((1-\sigma)G_{\text{sub}} + \sigma + 1) - \ln((G_{\text{sub}} - 1)(1-\sigma))}.$$  

The numerical simulation result of $G_M$ from Eq. (25) is shown in Fig. 11 as a function of $G_{\text{sub}}$. For theoretical calculation, $\sigma = 0.95$ is used. If the number of the weight images that have to be updated in the same subvolume is larger than $G_M$, strategy 3 leads to shorter update times than strategy 2. From Fig. 11, break-even number $G_M$ is approximately 10% of the entire multiplexed gratings. Therefore we may conclude that update strategy 3 is virtually always better than strategies 1 and 2.

D. Learning Rate Consideration

Here we discuss learning capabilities based on the contents of Subsection 4.C. We assume strategy 3 as a memory-update method. Because the entire contents at a subvolume are updated, newly calculated weights as well as the unchanged weights are
temporarily stored in another buffer subvolume and then written back to the desirable subvolume. Here we investigate and compare the learning rates based on incremental recording and scheduled recording. Figure 12(a) shows a diagram of memory transfer from an original subvolume to a buffer subvolume by incremental recording. Since the number of the memory access to the original subvolume is large because of the incremental recording process, the memory degradation of the original subvolume is not negligible. In order to prevent this readout degradation, the contents of the original subvolume is also refreshed during the readout process by incremental recording. Therefore the memory transfer process requires twice the update time, $T_3$, and the corresponding memory access time. Then the number of readout cycles can be obtained by dividing the total recording time by the incremental recording time, $D_{t}$, given in Eq. (A1) in Appendix A, leading to

$$T_{4,\text{incremental}} = 2T_3 \left(1 + \frac{t_a}{\Delta t}\right).$$

As shown below in Fig. 13, the memory access time is not negligible when the number of multiplexed gratings is large and incremental time $\Delta t$ is much shorter than memory access time $t_a$.

As proven in Appendix C below, the times required for both incremental and scheduled recordings are the same. However, memory access cycles to the original subvolume are reduced to the number of multiplexed gratings, $G_{\text{sub}}$, by scheduled recording. Therefore, we can simply read out the interconnection weight from the original subvolume and record it in a buffer subvolume by assuming the readout degradation is negligible, as shown in Fig. 12(b). Thus, the corresponding total update time, $T_{4,\text{schedule}}$, is given by

$$T_{4,\text{schedule}} = 2(T_3 + G_{\text{sub}}t_a).$$

Figure 13 shows learning rate $T_r$ as a function of the SLM SBWP. The interconnection bit accuracy was assumed to be $G_{\text{bit}} = 6$. The memory access time is again assumed to be $t_a = 1 \mu$s. For calculations we assumed the following thickness: 2 mm for LiNbO$_3$, 5 mm for CdTe, and 1 cm for strontium barium niobate (SBN). The values of $\Delta n_{\text{max}}$ are assumed to be $1 \times 10^{-3}$, $5 \times 10^{-4}$, and $1 \times 10^{-4}$ for LiNbO$_3$, CdTe, and SBN, respectively. A wavelength of 0.514 $\mu$m is assumed for LiNbO$_3$ and SBN, and 1.06 $\mu$m for CdTe. The material response time is taken experimentally (SBN) and from the literature (LiNbO$_3$ and CdTe)$^{30-32}$.
assumed the use of a single external light source and laser power $P_L = 1$ W to increase the learning rate. The saturation parameter is assumed to be $\sigma = 0.95$. All other parameters are the same as used before. Because SBN and CdTe have much faster response times than LiNbO$_3$, their learning rates are much higher than that of LiNbO$_3$.

Figure 13 shows two types of learning rates corresponding to incremental recording and scheduled recording for three different materials. Because smaller values of the SLM SBWP imply a larger number of multiplexed gratings as shown in Fig. 4, the effect of time lag caused by memory access during the incremental recording is obvious, i.e., the learning rate difference between incremental and scheduled recordings is more apparent for smaller values of the SLM SBWP than for larger values of the SLM SBWP. A faster response time results in a shorter incremental recording time, $\Delta t$. Therefore the learning rate difference between incremental recording and scheduled recording is larger for CdTe than for SBN and LiNbO$_3$. In fact, because of the very large photorefractive response time of LiNbO$_3$, the difference between the learning rate based on incremental and scheduled recordings is independent of the SBWP of SLM. The learning rate gradually increases as the SBWP of the SLM increases. However, the learning rate saturates and starts decreasing at higher SBWP's of the SLM. This is because the total update time is independent of the number of multiplexed gratings and the number of gratings decreases as the SBWP of the SLM increases.

5. Experiments

In order to demonstrate experimentally the implementation of the proposed neural network module system, we optically implemented the learning process of a Sexnet$^{33,34}$ that classifies male and female images by using Perceptron learning.$^{35,36}$ The Sexnet consists of 900 input neurons and a single output neuron. The network output is unity if the input image is classified as male and zero for a female image. The image database used for training has 90 photographs of young adult faces (45 male and 45 female).$^{37}$ Each image is scaled to normalize the distance between the eyes and the distance between the eyes and the mouth. All images are adjusted to the same average brightness. The original faces were sampled at a resolution of $30 \times 30$ pixels. Each pixel represents 64 gray-scale levels corresponding to a 6-bit accuracy. One pixel at the upper left-hand corner of each image is set to be unity working as an offset that is required for Perceptron learning. The network was trained by using the Perceptron learning rule,

$$W^{(\text{new})} = W^{(\text{old})} + (y' - y)x,$$

where $y'$ is the desirable network output and $y$ is the actual network output. At every learning epoch, the network was tested by using 50 test images and the correct output rate was calculated.

The network was originally trained by using floating point numbers of 32-bit accuracy on a digital computer. Then the interconnection weight accuracy was reduced and learning was repeated. We found 6-bit-accuracy interconnection weights converged into the same learning results as 32-bit-accuracy floating numbers, but 5-bit-accuracy interconnection weights failed to converge. The learning results with 6-bit interconnection weights were applied to 50 test images showing a correct rate of 76%, close to the value already reported in Ref. 34. Figure 14 shows the schematic diagram of the optical setup configured as fan-in architecture. The LiNbO$_3$ photorefractive crystal is used as dynamic digital photorefractive memory. Interconnection weights are displayed on the first liquid-crystal light valve, LCLV1, and recorded at each holographic subvolume. Seven weight images, one for the sign bit and six for the 6-bit accuracy interconnection weights, are multiplexed with angular multiplexing. The input image is introduced to LCLV2 and multiplied by the interconnection patterns reconstructed from the digital photorefractive memory. The results of the multiplication are detected by a CCD camera interfaced to a microcomputer. Pixel thresholding is performed over $3 \times 3$ CCD camera pixels. Thresholding and summation processes are performed by the microcomputer. Although the averaged intensity pixels
vary from 120 to 255 on the CCD camera, thresholding and binarization processes demonstrated reliable digital detection during the experiments. Newly calculated weights are simply overwritten based on memory-update strategy 3.

Figure 15 shows the learning process of the optical implementation, identical to the computer simulation results. Figure 16 shows the interconnection pattern images after the learning is complete. Although the learning process requires at least a 6-bit accuracy, the trained network can be operated with a lower accuracy, i.e., 5 bits, because the fifth bit plane does not include any interconnection information as shown in Fig. 16.

6. Conclusions

Optoelectronic neural network modules based on page-oriented digital photorefractive memory have been presented. Physical separation of holographic subvolumes results in low cross talk and high diffraction efficiency, leading to scalable optoelectronic neural network systems. Digitally encoded interconnection weights ensure high accuracy. Two different interconnection architectures, fan-out and fan-in, have been described. Fan-out architecture is suitable for networks that require large numbers of output neurons, whereas fan-in architecture is more suitable for implementing neural networks with large numbers of input neurons. The scalability and feedforward throughput of the systems were investigated by considering transmission geometry, geometrical constraint of the photorefractive memory, and power requirements of the photodetectors. Feedforward throughput of $O(10^{12})$ can be achieved with a SLM SBWP of $10^6$. Learning capability was also investigated, including various interconnection primitives required for learning, as well as memory-update strategies. The memory-updating approach of overwriting entire interconnection gratings has been found to lead to the highest learning rate of $O(10^9)$ and $O(10^7)$ interconnections/s for CdTe and SBN, respectively, with 6-bit computation accuracy. A Sexnet that used the Perceptron learning rule was optically demonstrated with the fan-in architecture for 900 input neurons. Both the input data and the interconnection weights were represented with 6-bit accuracy. The experimental results of the optically trained Sexnet performance were found to be in perfect agreement with digital computer simulations and demonstrated unique capabilities of the system to implement networks that require high computation accuracy.

Appendix A. Memory-Update Strategy 1

In order to estimate the time required to update a single weight image, we have to know how many iteration cycles are needed to selectively erase a single weight image. The incremental recording time, $\Delta t$, is derived from Eq. (4) in Ref. 8 and is given by

$$\Delta t = \frac{\tau}{2(G_{sub} - 1)} \ln(1 + \varepsilon), \quad (A1)$$

Fig. 16. Interconnection weight images multiplexed in a photorefractive subvolume after learning is complete. MSB, most-significant bit; LSB, least-significant bit.
where \( \tau \) is the material response time and \( \varepsilon \) is the allowable variation of the diffraction efficiency of multiplexed weight images in the incremental recording process. Because of the selective erasure process during \( \Delta t \), the initial refractive index modulation of a weight image, \( \Delta n_i \), decreases to \( \Delta n_{i+1} \):

\[
\Delta n_{i+1} = (\Delta n_i + \Delta n_{\text{max}}) \exp(-\Delta t/\tau) - \Delta n_{\text{max}}, \tag{A2}
\]

where \( \Delta n_{\text{max}} \) is the material’s maximum allowable refractive index modulation. Because of incremental recordings of other \( G_{\text{sub}} - 1 \) weight images, uniform erasure occurs during the time of \( \Delta t \). After the uniform erasure of \( (G_{\text{sub}} - 1) \Delta t \), refractive index modulation \( \Delta n_{i+1} \) decreases to \( \Delta n_{i+2} \):

\[
\Delta n_{i+2} = \Delta n_{i+1} \exp(-(G_{\text{sub}} - 1) \Delta t/\tau). \tag{A3}
\]

We obtain the expression of refractive index modulation \( \Delta n_k \) after the \( k \)th iteration of selective erasure and uniform erasure as follows:

\[
\Delta n_k = \Delta n_0 \zeta^k + \Delta n_{\text{max}} \gamma \frac{1 - \zeta^k}{1 - \zeta}, \tag{A4}
\]

where

\[
\zeta = (1 + \varepsilon)^{-1/2(G_{\text{sub}}/(G_{\text{sub}} - 1)}),
\]

\[
\gamma = (1 + \varepsilon)^{-1/2(G_{\text{sub}}/(G_{\text{sub}} - 1)} - (1 + \varepsilon)^{-1/2}, \tag{A5}
\]

and \( \Delta n_0 \) is the initial refractive index modulation.

Making \( \Delta n_k = \Delta n_{\text{max}}/G_{\text{sub}} \) and \( \Delta n_k = 0 \), we obtain the solution of \( k \) from Eq. (A4):

\[
k = \frac{1}{\ln \zeta} \ln \frac{\gamma}{\gamma - (1 - \xi)/G_{\text{sub}}}. \tag{A6}
\]

After the weight image is selectively erased, the new one is overwritten based on the incremental recording schedule [see Eq. (7) in Ref. 8]. Total update time \( T_1 \) is the sum of the selective erasure process time and the incremental recording process time, given by

\[
T_1 = k \Delta t G_{\text{sub}} - \tau \ln(1 - \sigma),
\]

\[
= \frac{\tau G_{\text{sub}}}{2(G_{\text{sub}} - 1) \ln \zeta} \ln(1 + \varepsilon) \ln \left[ \frac{\gamma}{\gamma - (1 - \xi)/G_{\text{sub}}} \right] - \tau \ln(1 - \sigma), \tag{A7}
\]

where \( \sigma \) is a saturation parameter. The incremental recording process is stopped when the grating’s refractive index modulation becomes equal to \( \sigma \Delta n_{\text{max}} / G_{\text{sub}} \).

### Appendix B. Memory-Update Strategy 2

The selective erasure time, \( t_{se} \), to erase a single weight image of refractive index modulation \( \Delta n_{\text{max}} / G_{\text{sub}} \) is

\[
t_{se} = \tau \ln(1 + 1/G_{\text{sub}}). \tag{B1}
\]

Overwriting time \( t_w \) to record a weight image up to \( \Delta n_{\text{max}}/G_{\text{sub}} \) is

\[
t_w = -\tau \ln(1 - 1/G_{\text{sub}}). \tag{B2}
\]

The degradation of refractive index modulation \( R \) for other \( G_{\text{sub}} - 1 \) gratings caused by selective erasure and overwriting is

\[
R = \exp(-(t_w + t_{se})/\tau). \tag{B3}
\]

Inserting Eqs. (B1) and (B2) into Eq. (B3) yields

\[
R = \frac{G_{\text{sub}} - 1}{G_{\text{sub}} + 1}. \tag{B4}
\]

Required time \( t_i \) for incremental recording is

\[
t_i = -\tau \ln(1 - \sigma) + \tau \ln \left(1 - \frac{G_{\text{sub}} - 1}{G_{\text{sub}} + 1} \sigma \right). \tag{B5}
\]

From Eqs. (B1), (B2), and (B5), update time \( T_2 \) for a single weight image is

\[
T_2 = \tau \ln \left[ \frac{(1 - \sigma)G_{\text{sub}} + \sigma + 1}{(G_{\text{sub}} - 1)(1 - \sigma)} \right]. \tag{B6}
\]

### Appendix C. Comparison of Total Recording Time for Incremental and Scheduled Recordings

In scheduled recording, the \( n \)th grating’s recording time, \( t_n \), is given as

\[
t_n = \tau \ln \left[ \frac{(n - 1)\beta + 1}{(n - 2)\beta + 1} \right], \tag{C1}
\]

where \( \beta \) is the parameter determined by the first grating’s writing time, \( t_1 \), and is given by

\[
\beta = 1 - \exp(-t_1/\tau). \tag{C2}
\]

From the scheduled recording as given by Eq. (C1), each grating’s refractive index modulation becomes a constant value of

\[
\Delta n = \Delta n_{\text{max}} \frac{\beta}{(G_{\text{sub}} - 1)\beta + 1}, \tag{C3}
\]

where, \( \Delta n_{\text{max}} \) is the maximum value of the refractive index modulation and \( G_{\text{sub}} \) is the number of multiplexed gratings at each subvolume. Using Eq. (C1), we obtain the total recording time for a scheduled recording:

\[
T_{\text{schedule}} = \tau \ln \left[ \frac{(G_{\text{sub}} - 1)\beta + 1}{1 - \beta} \right]. \tag{C4}
\]

The total recording time for an incremental recording is given as a function of saturation parameter \( \sigma \) as

\[
T_{\text{incremental}} = -\tau \ln(1 - \sigma). \tag{C5}
\]
Because saturation parameter $\sigma$ is related to parameter $\beta$ by the following relationship,
\begin{equation}
\sigma = \frac{\beta G_{sub}}{(G_{sub} - 1)\beta + 1}, \tag{C6}
\end{equation}
substituting Eq. (C6) into Eq. (C5) leads to
\begin{equation}
T_{\text{incremental}} = \tau \ln \left[ \frac{(G_{sub} - 1)\beta + 1}{1 - \beta} \right], \tag{C7}
\end{equation}
which is identical to the total recording time of the scheduled recording as given by Eq. (C4).

The authors are grateful to T. J. Sejnowski for valuable suggestions regarding neural network simulations. Various suggestions from reviewers are greatly appreciated. H. Sasaki acknowledges financial support from Oki Electronic Industry Co., Ltd., Japan.

References