1993-03-10

Analysis of Weighted Fan-out/Fan-in Volume Holographic Interconnections

Gregory P. Nordin
nordin@byu.edu

P. Asthana

B. Keith Jenkins

A. R. Tanguay

Original Publication Citation

P. Asthana, G. P. Nordin, A. R. Tanguay, Jr., and B. Keith Jenkins, "Analysis of Weighted Fan-out/Fan-in Volume Holographic Interconnections," Applied Optics Special Issue on Neural Networks 32(8), pp. 1441-1469 (1993)

Analysis of weighted fan-out/fan-in volume holographic optical interconnections

Praveen Asthana, Gregory P. Nordin, Armand R. Tanguay, Jr., and B. Keith Jenkins

The feasibility of employing volume holographic techniques for the implementation of highly multiplexed weighted fan-out/fan-in interconnections is analyzed on the basis of interconnection fidelity, optical throughput, and complexity of recording schedule or implementation hardware. These feasibility criteria were quantitatively evaluated using the optical beam propagation method to numerically simulate the diffraction characteristics of volume holographic interconnections recorded in a linear holographic material. We find that conventional interconnection architectures (that are based on a single coherent optical source) exhibit a direct trade-off between interconnection fidelity and optical throughput on the one hand, and recording schedule or hardware complexity on the other. In order to circumvent this trade-off we describe and analyze in detail an incoherent/coherent double angularly multiplexed interconnection architecture that is based on the use of multiple-source array of individually coherent but mutually incoherent sources. This architecture either minimizes or avoids several key sources of cross talk, permits simultaneous recording of interconnection weights or weight updates, and provides enhanced fidelity, interchannel isolation, and throughput performance.

1. Introduction

Volume holographic optical elements (VHOE's) have often been suggested as the principal components of an optical interconnection technology for applications such as optical computing and telecommunications that require a large number of interconnections.[1–6] Depending on the application, such interconnection systems may also require varying degrees of interconnection weighting, fan-out, fan-in, and channel independence. In particular, interconnection systems for artificial neural networks provide a specific application for which all of these issues are important.

Artificial neural networks are composed of many highly interconnected nonlinear computational elements (neuron units) that operate in parallel and are arranged in architectural patterns that are motivated to a certain extent by investigations of biological neural networks. The computational elements are in most cases densely interconnected with weighted connection pathways that can be reconfigured and updated to permit either supervised or unsupervised learning. Implementations of adaptive neural networks should optimally permit such pathway reconfiguration and weight updates without excessively compromising either hardware complexity or computational efficiency.

Typical artificial neural-network interconnection topologies require a high degree of both fan-out and fan-in at each neuron unit. A fully interconnected topology for the case of a single-layer network is illustrated schematically in Fig. 1, in which two planes of neuron units are shown. In such a fully connected network, the output of each neuron unit in the input plane is fanned out to all of the neuron units in the output plane. Similarly, each neuron unit in the output plane receives the weighted fan-in of all of the neuron-unit outputs from the input plane. The fan-out/fan-in requirements of neural networks act as a multiplier on the total number of interconnected neuron units, resulting in an interconnection system that must prove capable of supporting very large numbers of independent pathways in a relatively compact topology. For example, a fully connected neural network having 10^5 neuron units in both the input and output planes requires 10^{10} interconnections, as does a partially connected neural network

When this work was performed, the authors were with the Optical Materials and Devices Laboratory, Center for Photonic Technology, Signal and Image Processing Institute, and the Center for Neural Engineering, University of Southern California, University Park, MC-0483, Los Angeles, California 90089-0483. P. Asthana is currently with the Storage Systems Products Division, IBM Corporation, Tucson, Arizona 85744; G. P. Nordin is currently with the Electrical and Computer Engineering Department, University of Alabama in Huntsville, Huntsville, Alabama 35899.

Received 16 July 1992.
0003-6935/93/081441-29$05.00/0.

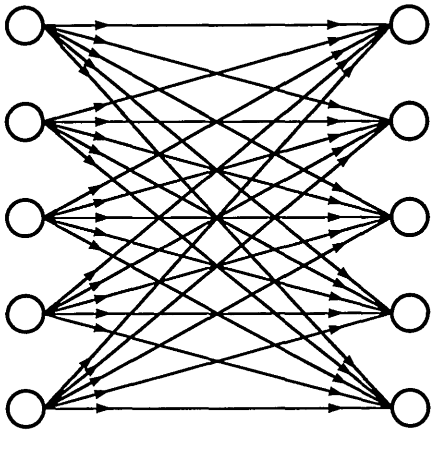

Fig. 1. Schematic representation of fan-out/fan-in interconnections between input and output planes of neuron units.

having 10^6 neuron units in both the input and output planes with a fan-out and fan-in of 10^4.

The use of volume holographic optical elements has been proposed as the basis of an interconnection technology for neural networks precisely because it has the potential to meet the critical requirements of providing both large numbers of interconnections and a weighted fan-out/fan-in topology.[7] Even so, the feasibility of using VHOE's for large-scale weighted fan-out/fan-in interconnection applications depends further on the fidelity with which the interconnection weights can be implemented and on the optical throughput that can be achieved in the volume holographic interconnection system. The optical throughput is a measure of the fraction of the total incident optical power that is diffracted into the set of desired outputs; as such, it provides a quantitative assessment of the overall interconnection system efficiency. Important additional implementation issues include the total number of required exposures, the optical power incident on the holographic recording medium per exposure, the complexity of the implementation hardware, the required exposure schedule (if any), and the total recording time for a given interconnection complexity.

In this paper we quantitatively evaluate the performance characteristics of two distinct volume holographic interconnection architectures in accordance with these criteria. The first interconnection architecture is novel in that an array of individually coherent but mutually incoherent optical sources is used to generate a multiplicity of angularly multiplexed recording beams.[8–14] We hereinafter refer to this type of interconnection architecture as being incoherent/coherent double angularly multiplexed. The second interconnection architecture is based on the use of a single coherent optical source during recording, as has been widely discussed in the literature.[3,15–19] This conventional architecture is referred to herein as a single-source architecture.

In our analysis we use the optical beam propagation method (BPM)[14,20–23] to investigate the diffraction characteristics of weighted fan-out/fan-in interconnections in complex systems that involve both large numbers of holographic gratings and multiple readout beams, and that therefore do not lend themselves readily to analytical solutions. For our purposes herein, we consider explicitly the case of linear holographic materials in order to illustrate key differences in interconnection performance that are architecture dependent. Generalization of these results to nonlinear holographic materials (for example, to certain photorefractive media) is beyond the scope of this paper.

Results are reported herein on the numerical simulation of up to 10- (input node) to-10 (output node) weighted fan-out/fan-in volume holographic interconnections that incorporate between 10 and 190 individual holographic gratings multiplexed within the same region of the volume holographic recording medium, depending on the specific architectural configuration considered. Such simulations require significant run times, even when implemented with the highly efficient BPM algorithm on a supercomputer. To the best of our knowledge, the 10-to-10 case with fully independent weights is one of the most complex volume holographic interconnection systems that have been analyzed to date.

Our simulations demonstrate that the novel incoherent/coherent double angularly multiplexed architecture exhibits both high interconnection fidelity and high optical throughput efficiency even in the presence of fully simultaneous recording. This combination of desirable characteristics derives from the elimination (or minimization) of several distinct sources of interchannel cross talk and throughput loss that are unavoidably present in the more widely investigated single-source architecture. In addition, the mutual incoherence of the readout beams in the incoherent/coherent double angularly multiplexed architecture provides naturally for linearity of summation in an intensity representation without an associated fan-in loss.

In order to compare the fidelity and throughput performance of the incoherent/coherent double angularly multiplexed architecture with an appropriate benchmark, we analyzed the single-source architecture (illustrated in Fig. 2) under directly comparable holographic recording conditions. In particular, we examined the cases of simultaneous, pagewise-sequential, and fully sequential recording of the desired weight updates. These three recording methods represent distinct trade-offs between interchannel cross talk resulting from the presence of extraneous gratings, on the one hand, and recording schedule or hardware complexity on the other. The presence of such extraneous gratings can cause significant errors in the diffracted outputs that are fanned in to a given interconnection node. For example, in the 10-to-10 single-source interconnection system that we modeled, errors as large as 100% in the relative diffracted outputs occur at the peak optical throughput of 50% for the simultaneous recording method. To the best of our knowledge, the analysis of the

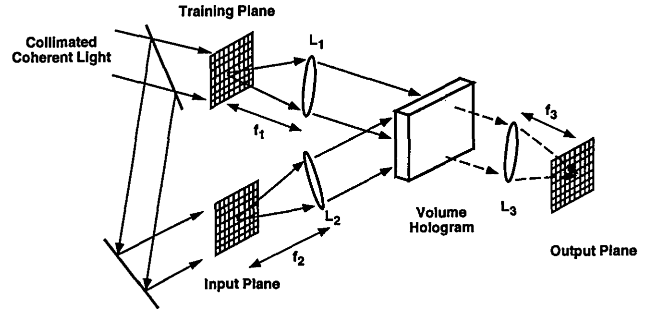

Fig. 2. Schematic diagram of a single-source holographic interconnection architecture in which diffraction gratings in a volume hologram connect pixels in the input plane to pixels in the output plane. Interconnection gratings are formed by recording the interference between light from pixels in the training plane and light from pixels in the input plane. L_1–L_3 are lenses; f_1–f_3 are focal lengths.

simultaneous recording method (as applied to the single-source architecture) presented herein establishes the fidelity errors and throughput losses of this technique quantitatively for the first time. Elimination of the extraneous gratings by using the fully sequential recording method (at a substantial cost in recording schedule or hardware complexity) reduces the largest errors to 13% at an increased peak optical throughput of 85%. These errors for the fully sequential recording case are significantly lower than those described in Ref. 19, at least in part because the relative phase relationships among the recording beams are maintained on readout in our simulations.

In addition, we demonstrate that the severe fan-in loss that unavoidably accompanies the collinear combination of mutually incoherent beams of essentially common wavelength,[24] and that also applies to incoherent readout in the single-source architecture, is one of several effects of beam degeneracy (defined in Section 4.D).[9,12,13,25] This same physical mechanism is also primarily responsible for the residual errors apparent in the fully sequential recording method at peak throughput.

The remainder of this paper is organized as follows. Specific aspects of volume holographic interconnection systems that are essential for the establishment of a valid comparison metric are discussed in Section 2, including the choice of signal representations and weights appropriate for photonic neural networks, the implementation of weighted fan-out/fan-in interconnections in a single-source architecture, and the associated sources of fidelity errors and throughput losses. Section 3 details our modeling methodology and assumptions, anticipating the discussion of the simulation results obtained for the single-source architecture that are provided in Section 4. We present two distinct configurations of the incoherent/coherent double angularly multiplexed architecture and quantitatively analyze their performance characteristics in Section 5. Finally, we compare the relative advantages and disadvantages of each architectural configuration and summarize our results in Section 6.

2. Preliminary Concepts

A. Basic Neural-Network Operation

In many neural networks the computational process of a single network layer can be represented by

$$y_i = f(\rho_i), \tag{1}$$

$$\rho_i = \sum_{j=1}^{N} W_{ij} x_j, \tag{2}$$

in which neuron units i and j are in the output and input planes, respectively, y_i is (proportional to) the output of neuron unit i, x_j is (proportional to) the output of neuron unit j, W_{ij} is the weight of the interconnection between neuron units i and j, N is the number of neurons in the input plane, f is the nonlinear threshold function of each neuron unit, and ρ_i is the activation potential.

In a neural network that incorporates learning, the weights W_{ij} are updated incrementally according to an appropriate learning rule as training patterns are presented to the network. A large class of learning algorithms can be characterized by the following weight-update rule:

$$\Delta W_{ij} = \alpha \delta_i^{(m)} x_j^{(m)}, \tag{3}$$

in which $\Delta W_{ij} = W_{ij}(m) - W_{ij}(m-1)$ is the weight update, α is the learning gain constant, and m is the iteration index. Various learning rules can be formed by suitably choosing the functional form of $\delta_i^{(m)}$; for example, in the case of Hebbian learning, $\delta_i^{(m)}$ is chosen to be y_i.[26]

B. Weighted Fan-Out/Fan-In Interconnections in a Single-Source Architecture

The conventional single-source volume holographic interconnection architecture as configured for photonic neural-network implementation is shown schematically in Fig. 2. This architecture consists of a holographic interconnection medium and three planes of neuron units: a training plane, an input plane, and an output plane. The two-dimensional (2-D) arrays of neuron units or pixels (since in many photonic neural-network implementations each pixel corresponds to a single neuron unit) are placed on these planes. In applications other than neural networks these pixels can represent generic connection nodes. The arrays of pixels can be implemented using, for example, 2-D spatial light modulators (SLM's). Both the training- and input-plane SLM's are illuminated by a single coherent source in the configuration shown in Fig. 2.

A given weighted interconnection between a pixel in the input plane and a pixel in the output plane is physically realized as a single diffraction grating in the holographic medium. Each grating is formed by recording the interference pattern generated by coherent superposition of light from a pixel in the input plane and light from a pixel in the training plane. For simplicity the pixels in both planes are typically assumed to act as point sources[3,18] and lenses L_1 and L_2 function as collimating lenses. Light from each pixel is thus incident as an angularly distinct collimated beam (typically approximated as a plane wave[3,18]) on the holographic medium. The light from each pixel is assumed to fully illuminate the aperture of the holographic medium to ensure the potential for full connectivity.[3,18]

C. Signal Representation in a Single-Source Architecture

Depending on the computational algorithm, the optical signal from each input pixel (after collimation) can be represented by a number of different physical quantities, including intensity, amplitude, polarization, and wavelength. In the two most commonly investigated approaches[3] the signal is represented either by the complex electric field or by the intensity of the light transmitted or reflected by the pixel (depending on the type of SLM).

In the amplitude representation the electric field of the light from pixel j that is incident at the front face of the hologram can be expressed (neglecting the vector nature of the electric field) as $x_j^{\mathrm{amp}} E_0 \exp(i\mathbf{k}_j \cdot \mathbf{r} + i\phi_j - i\omega t)$, in which x_j^{amp} is the amplitude transmissivity (or reflectivity) of the jth pixel, E_0 is the magnitude of the electric field amplitude of the readout beam, ϕ_j is the phase of the readout beam (relative to an arbitrarily chosen coordinate system), and ω is the angular frequency. In this representation a given interconnection weight is proportional to the amplitude diffraction efficiency of the corresponding interconnection grating.

In the intensity representation the output of pixel j is expressed as $x_j I_0$, in which x_j is the intensity transmittance (or reflectance) of the SLM pixel and $I_0 = E_0^2$. In this case each weight is proportional to the intensity diffraction efficiency of its corresponding interconnection grating.

When readout is performed with beams from more than one input pixel, the intensity detected within each output pixel consists of a weighted sum of diffracted signals. If an optical system is constructed such that the light beams from the input pixels are mutually incoherent during readout, the relative detected intensity (using the intensity representation) within the ith output pixel ρ_i is

$$\rho_i = \sum_{j=1}^{N} W_{ij} x_j, \tag{4}$$

in which W_{ij} is an intensity weight (which is proportional to the diffraction efficiency of grating ij).[17] The relationship expressed in Eq. (4) is the same as that in Eq. (2), which defines how the activation potentials are related to the inputs by the interconnection weights of the neural network. Thus readout with mutually incoherent beams in conjunction with an intensity representation leads to an optical system that performs the desired neural interconnection function of Eq. (2).

If the optical system is constructed such that the light beams from the input-plane pixels are mutually coherent (as is typically the case for a single-source architecture), the diffracted signals detected in the output plane can be written (assuming an intensity representation) as

$$\rho_i = \left| \sum_{j=1}^{N} (W_{ij} x_j)^{1/2} \exp[i(\phi_j + \Phi_{ij})] \right|^2, \tag{5}$$

in which Φ_{ij} is the phase of the ijth grating. Each grating phase is set by the relative phases of the recording beams used to form the grating and the properties of the holographic recording medium. If the argument of the exponential is constant for all i and j, the diffracted outputs for a single-source architecture using the intensity representation reduce to

$$\rho_i = \left| \sum_{j=1}^{N} (W_{ij} x_j)^{1/2} \right|^2. \tag{6}$$

This equation is valid only if two specific conditions are met: (1) the readout beams must have the same relative phases as were used during recording of the interconnection gratings, and (2) the phase shift induced by the holographic recording medium itself must be constant for all recorded gratings.

Even with these assumptions, the square-root relationship embodied in Eq. (6) deviates substantially from the desired interconnection function of Eq. (2). The general effect of this deviation on neural-network operation has not yet been established.

As a result, several authors have chosen an amplitude representation for use in the single-source architecture, such that the diffracted outputs are given by[3,27]

$$\rho_i = \left| \sum_{j=1}^{N} W_{ij}^{\mathrm{amp}} x_i^{\mathrm{amp}} \exp[i(\phi_j + \Phi_{ij})] \right|^2, \tag{7}$$

in which W_{ij}^{amp} is the amplitude weight (which is proportional to the amplitude diffraction efficiency). If the argument of the exponential in Eq. (7) is constant for all i and j, the system yields the square of the desired interconnection function, which may be corrected either electronically or by adjusting the nonlinear threshold function of each neuron unit. This simplification is dependent, as is Eq. (6), on the two conditions specified above. The first of these conditions is contingent on the maintenance of rigid optical phase stability in the system, which may be difficult to realize in practice.

A primary advantage of readout with mutually incoherent beams is the avoidance of the rigid phase-stability requirements that are necessary for coherently read out systems.[3,24,28] However, a serious disadvantage associated with the use of mutually incoherent readout beams is the significant throughput loss that is characteristic of all holographic interconnection architectures in which the fan-in is performed collinearly.[24] Since single-source interconnection systems rely on collinear fan-in, they inherently suffer from this throughput loss for readout with mutually incoherent beams. In Section 5 we discuss a method for using readout with mutually incoherent beams without suffering an incoherent fan-in loss.

A further consideration in the implementation of photonic neural networks is that many neural algorithms require the use of bipolar weights and bipolar neuron-unit outputs. One method of achieving this bipolarity for readout with mutually coherent beams is to use an amplitude representation in which a 180° phase shift in the phase of both the electric fields and the diffraction gratings is used to represent negative numbers.[3,27] While attractive in principle (requiring only a single data channel per neuron unit), this approach is difficult to implement in practice, in part because the phases of the resultant diffracted outputs must be detected as well as their amplitudes. An alternative method of achieving bipolarity is to use an intensity representation in conjunction with a dual-rail concept in which each neuron has separate positive and negative channels for both input and output signals with associated weighted interconnections for each pair of channels.[13,27,29,30] Although this method requires two data channels per neuron unit, it is compatible with simple square-law detectors and does not require mutually coherent readout beams.

For our purposes herein, we consider only unipolar weights and unipolar neural outputs since the dual-rail method can be used to generalize to the fully bipolar case. In addition, we choose to adopt the intensity representation throughout in order to facilitate direct comparisons between mutually coherent and mutually incoherent readout systems. In particular, the intensity representation is perhaps the most natural representation for the incoherent/coherent double angularly multiplexed interconnection architecture, yielding linear sum rules in the diffracted outputs. In any case, this choice does not affect the general conclusions drawn regarding interconnection fidelity and optical throughput. For the various cases of mutually coherent readout treated herein, we assume that both conditions (1) and (2) specified above [following Eq. (6)] are met.

D. Recording Methods in a Single-Source Architecture

For the single-source architecture shown in Fig. 2, in which the light from each pixel is mutually coherent with the light from all such pixels in a given plane, there are several methods by which the desired interconnection gratings may be recorded. All such methods involve a sequence of exposures, and they differ in the nature of each individual exposure. The first method is simultaneous recording, in which mutually coherent light beams from all the input and training pixels are incident simultaneously on the holographic medium during each exposure. By recording the resultant interference pattern, the volume holographic medium forms gratings that connect each neuron unit in the input plane with each neuron unit in the output plane. These constitute the desired *interplanar* connections that perform the weighted fan-out/fan-in function. However, *intraplanar* gratings that connect pairs of pixels within the input plane (and within the training plane as well) are formed as a result of the mutual coherence of the sources.[16,17] These extraneous gratings (termed coherent-recording cross-talk gratings, or cross gratings) introduce a serious source of cross talk into the interconnection system that in turn leads to a loss of both throughput and reconstruction fidelity.

In the pagewise-sequential recording method, a single pixel in the input plane is connected simultaneously with all of its associated training pixels during each exposure, which eliminates the coherent cross-talk gratings within the input plane only. In the fully sequential recording method, the remaining coherent cross-talk gratings in the training plane are eliminated by recording connections between a single pixel in the input plane and a single pixel in the training plane during each exposure. In order to achieve the reduction in coherent-recording cross talk offered by these methods, however, an increase in the recording schedule or hardware complexity must be accommodated. In certain photorefractive media the situation may be complicated further by the need for a complex recording schedule to compensate for the partial erasure of previously recorded interconnection gratings during the recording of later gratings.[3,31]

In a photonic neural-network implementation, the interconnection weights can be either precomputed (in a photonic or electronic computing system) and stored in a permanent holographic medium for later use or obtained adaptively using a suitable learning algorithm and a dynamic holographic medium. For the recording of precomputed weights it is clearly advantageous to minimize the total number of exposures, the total recording time, and the total exposure energy. For a system in which N input nodes are connected to N output nodes (referred to herein as an N-to-N interconnection system) the minimum number of exposures required to record a set of fully independent weights is equal to N, as can be derived from degrees-of-freedom considerations.[32] The assumption of fully independent weights is equivalent to the assumption that the weight matrix is of rank N. If the full set of precomputed weights is recorded pagewise sequentially, therefore, the minimum number of exposures can be achieved.

For adaptive computation of weights a set of training pairs is presented sequentially (one training pair at a time) to the network. Each training pair consists of an input image and its corresponding training image, which are presented on the respective input and training planes of Fig. 2. The mth input image can be represented by the vector $\mathbf{x}^{(m)}$, the components $x_i^{(m)}$ of which are shown in the first term on the right-hand side of Eq. (3). Similarly, the mth training image is represented by the vector $\boldsymbol{\delta}^{(m)}$. Ideally, the full input and training images for the mth training pair $[\mathbf{x}^{(m)}, \boldsymbol{\delta}^{(m)}]$ are presented simultaneously on the input and training planes, respectively, such that only one exposure is required for each training pair. The outer product of the vectors $\mathbf{x}^{(m)}$ and $\boldsymbol{\delta}^{(m)}$ is recorded in each such exposure, which thus corresponds to the simultaneous recording case discussed above. However, if a pagewise-sequential or fully sequential recording strategy is necessary to avoid the deleterious effects of coherent-recording cross talk, then the number of exposures per training pair becomes N and N^2, respectively, for an N-to-N interconnection system. If M training pairs are required to fully train the network, the total number of exposures for the simultaneous, pagewise-sequential, and fully sequential recording methods are M, NM, and N^2M, respectively. If real-time adaptation is not required and precomputation of weights is permitted, then the minimum numbers of exposures required for the three recording methods are N, N, and N^2, respectively.

The choice of recording method has significant practical consequences for photonic neural-network implementations (using the single-source interconnection architecture) in which adaptive computation of the weights is desired. As an example, if the SLM frame time is the temporal bottleneck of the system (which represents perhaps a worst-case estimate, in that the single-pixel access time may be considerably shorter than the full SLM frame time for certain SLM's), then the total amount of time required to train the network scales linearly with the number of required exposures. For large numbers of interconnections (exactly the situation for which holographic interconnections are presumably attractive) this could result in impractically long training sessions. For example, if $N = 10^4$, $M = 10^3$, the SLM's support a 10-ms update rate (such as for nonferroelectric liquid-crystal-based SLM's), and the holographic-material response time at the available power level is fast enough to not provide an even stricter bound, then the simultaneous, pagewise-sequential, and fully sequential recording methods would require 10 s, 10^5 s (28 h), and 10^9 s (32 years), respectively, to record the desired interconnections.

In the pagewise-sequential recording method, one technique for avoiding prolonged training sessions (and consequently inefficient use of available optical power) is to focus the full incident beam on a given pixel of the input SLM and to scan the beam pixel-by-pixel during the recording of a given training pair. If scanning all of the input-plane pixels can be accomplished within the frame time of the SLM (which corresponds to a 1-μs dwell time for the above example), then pagewise-sequential recording can be accomplished in the same total time as simultaneous recording. The concept can also be applied to the fully sequential recording method (which requires 100-ps dwell times for the example above). Decreases in the total training time (with concomitant increases in the efficiency of power and energy use) for both recording methods are achieved at the cost of additional system complexity, particularly since in many cases provision must also be made for simultaneous readout of all input-plane pixels during posttraining computation.

From the joint perspectives of recording schedule and hardware complexity, simultaneous recording may prove to be the most desirable recording method (in the adaptive neural-network paradigm) because the full parallelism of the optical architecture is used, thereby achieving the greatest computational throughput during training without resorting to additional components that increase the system complexity. However, simultaneous recording within the single-source architecture is perhaps the least desirable recording method from the perspective of interconnection fidelity resulting from the deleterious effects of coherent-recording cross talk.

E. Sources of Fidelity Errors and Throughput Losses

In addition to coherent-recording cross talk, there are at least two other sources of fidelity errors and throughput losses that can be present in multiplexed fan-out/collinear fan-in interconnections implemented within the single-source architecture. The first of these is grating-degeneracy cross talk.[16–18,33,34] This form of cross talk arises because of the particular geometric placement of the recording pixels used. If the pixels are placed on regular grids, gratings with degenerate wave vectors may be recorded. Desired interconnections that have degenerate grating wave vectors can have severely distorted weights during reconstruction. This form of cross talk may be alleviated by placing the pixels on fractal sampling grids.[34] The cost, however, is the need to subsample the input, training, and output planes, which decreases the interconnection density of the system.

The second form of cross talk is beam-degeneracy cross talk, which arises from degeneracies in the wave vectors of beams diffracted from different gratings.[9,12,13,25] This form of cross talk is inherent in fan-out/fan-in volume holographic interconnection systems in which the diffracted beams that constitute a given fan-in exit collinearly from the holographic medium; it is present regardless of the sampling grids used for the input and training planes. In such interconnection systems, beam-degeneracy cross talk is also present even in the absence of cross gratings. For readout with mutually coherent beams in the single-source architecture, our simulation results indicate that beam degeneracy is a significant source of fidelity error only when the gratings are overmodulated. Further simulation results indicate that beam degeneracy is the primary physical mechanism responsible for the incoherent fan-in loss observed in single-source architectures when readout is performed with mutually incoherent beams.[12,13,25]

Related effects that can be attributed to beam degeneracy have been observed for 2-to-1 beam combining using a coupled-wave analysis.[35,36] In addition, Lee *et al.* have estimated the magnitude of beam-degeneracy cross talk (identified therein as third-order cross talk) insofar as it affects reconstruction fidelity in a limiting case.[17] Similarly, Slinger mentions that beam degeneracy (identified therein as one of several forms of multiple-grating interactions) is a potential source of cross talk.[19]

3. Modeling Methodology and Assumptions

In this section we describe our modeling methodology and assumptions in detail. The motivation behind our choice of modeling technique (the optical beam-propagation method, or BPM) is discussed in Section 3.A, followed by a brief outline of the pertinent features of the BPM algorithm in Section 3.B. In Section 3.C we discuss the geometric dimensions of the single-source interconnection architecture that we modeled, the recording characteristics of the holographic medium, and the method used to obtain the weight matrices for the various cases.

A. Choice of Simulation Method

The analysis of the diffraction properties of a large number of weighted gratings multiplexed in a volume holographic material has been a difficult problem. Several techniques have been used in the past to analyze multiple-grating diffraction.[19,36–42] Of these techniques, coupled-wave theory has been the most extensively used.[19,36–40] Application of this method often involves making a number of simplifying assumptions so that analytical or numerical solutions are more easily obtainable. Such assumptions include the use of a 2-D model (in which the gratings and the incident and diffracted beams all lie in a plane) and TE polarization of the beams.[19,36–39] With these assumptions and the additional assumption of no undesired cross gratings, analytical solutions have been obtained for the cases of 1-to-N weighted fan-outs,[38] N-to-1 weighted fan-ins,[39] and N-to-N weighted fan-out/fan-ins.[19] In Ref. 19 the assumption was further made that the weight matrix had a rank of one, which is achieved in practice in a linear holographic medium by recording each x_j with the same training vector, $\boldsymbol{\delta}$ (i.e., only a single training pair is recorded). For weight matrices having a higher rank (which is the relevant situation for photonic neural-network implementations) the system of coupled differential equations obtained using coupled-wave analysis has not yielded an analytical solution and thus must be solved numerically.[19] For example, Slinger has numerically modeled a 5-to-5 interconnection system for pagewise-sequential and fully sequential recording of the desired interconnection gratings.[19] However, numerical solutions for the simultaneous recording case have not been presented in the literature to our knowledge.

Rather than obtaining numerical solutions based on the results of a coupled-wave analysis, we used the optical beam propagation method[20] to numerically model readout of a volume hologram in which multiple interconnection gratings are stored. An advantage of this approach is that fewer simplifying assumptions are necessary to make the problem computationally tractable. In coupled-wave analysis, for example, only Bragg-matched interactions between the readout (and diffracted) beams and the recorded gratings are typically retained. In Ref. 19 this led to the assumption that all undesired cross gratings for pagewise-sequential recording operate in the Bragg regime. However, in many physical geometries (depending on the incidence angles of the writing beams and the thickness of the holographic recording material) some or all of these cross gratings may actually operate in the Raman–Nath diffraction regime or have properties that are in the transition regime between Raman–Nath and Bragg diffraction (hereinafter referred to as the transition regime). For example, the interconnection geometry that we simulate (the dimensions of which are discussed in Section 3.C.1) has cross gratings that operate in all three of the possible diffraction regimes.

Coupled-wave theory could be used to model cross gratings that operate in the Raman–Nath or transition diffraction regimes if multiple diffraction orders for each cross grating are retained in the calculations. However, BPM has proved to be a significantly faster numerical technique for solving diffraction grating problems involving many spectral orders when the same number of orders are considered and the same level of accuracy is required.[23]

Restricting the analysis to only Bragg-matched interactions using the coupled-wave approach also neglects a further possible source of cross talk, namely, diffraction from non-Bragg-matched interconnection gratings. By way of contrast, diffraction from non-Bragg-matched gratings is incorporated directly in the BPM algorithm. Although the effects of such non-Bragg-matched interconnection gratings are small, we show in Section 4.E that diffraction from such gratings is the limiting source of fidelity error as the strength of the interconnections in a single-source architecture approaches zero.

B. Optical Beam Propagation Method

The motivation for our choice of computational method is discussed above; herein we give a brief qualitative overview of the BPM algorithm. References 20–23 include discussions of the derivation and/or the validity of BPM for various diffraction problems. The BPM analysis discussed herein considers two physical dimensions [the nominal propagation (z axis) and transverse (x axis) dimensions shown in Fig. 3] and assumes TE polarization for convenience.

The optical beam propagation method simplifies the analysis of volume grating diffraction by replacing the physically distributed modulation/diffraction problem with a sequence of infinitesimally thin modulation layers separated by homogeneous regions in which only diffraction occurs. If a sufficient number of modulation layers are incorporated in the calculation, the system can closely approximate the characteristics of a volume hologram. The calculation

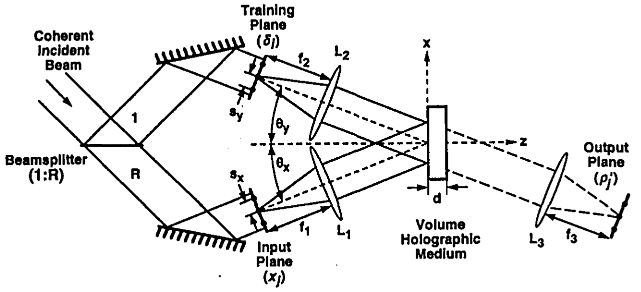

Fig. 3. Layout of the 2-D single-source architecture used in the modeling studies: R is the beam splitter ratio.

methods of the separate modulation and propagation steps are described briefly below.

In order to calculate the effect of the first modulation layer on an incident beam of monochromatic light, the optical (electric) field is multiplied by the phase and/or the amplitude modulation incorporated in that layer. Propagation to the next modulation layer (which implements the process of diffraction) proceeds by first Fourier transforming the optical field to obtain its spectral components. Each spectral component is then propagated to the next layer using the appropriate transfer function (as discussed below), at which point an inverse Fourier transform is performed to obtain the modified electric-field distribution. This series of steps (modulation, Fourier transformation, propagation, and inverse Fourier transformation) is repeated until propagation through the multilayer structure is complete. The use of a fast Fourier transform (FFT) algorithm to perform the forward and inverse Fourier transforms is responsible in large part for the efficiency of the BPM algorithm.

Each of the spectral components obtained after Fourier transformation of the modulated optical field corresponds to a plane wave propagating at a distinct angle θ_m (in which m is the index of the spectral component) with respect to the z axis. Each plane-wave component acquires a phase shift of $\psi(\theta_m)$ after propagation to the next modulation layer. This phase shift can be expressed as

$$\psi(\theta_m) = 2\pi n d_B \cos \theta_m/\lambda, \tag{8}$$

in which n is the refractive index, d_B is the thickness of the homogeneous buffer layer, and λ is the free-space wavelength of the readout beam. For single-grating problems the BPM algorithm is often implemented using a small-angle approximation, in which the cosine term in Eq. (8) is approximated by $1 - {\theta_m}^2/2$ (see, for example, Ref. 23). In our calculations we used Eq. (8) directly, with no small-angle approximation, to compute the phase of each spectral order for propagation from layer to layer.[43] This ensures the accuracy of the relative phases of the orders when many spectral orders are considered, which proves to be critical when modeling diffraction from large numbers of gratings.

Using the optical beam propagation method, we numerically modeled 2-to-2, 3-to-3, 4-to-4, and 10-to-10 interconnection systems to quantitatively investigate the reconstruction fidelity and optical throughput of both single-source and incoherent/coherent double angularly multiplexed architectures. Our main limitation in running arbitrarily larger cases is computational run time. For example, simulation of a 10-to-10 single-source architecture (which for the simultaneous-recording method involves 190 multiplexed gratings) requires > 2.5 h to run on an Alliant FX/2800 minisupercomputer (on a single processing node).

C. Modeling Assumptions

1. *Single-Source Architecture: Geometry and Dimensions*

A schematic diagram of the single-source architecture used in our simulations is shown in Fig. 3. Both the input and training pixel planes are illuminated by beams from the same laser source. Thus light beams from any two pixels in either or both planes are mutually coherent. For simplicity each pixel is assumed to act as a point source such that, after collimation by lens L_1 or L_2, light from each pixel can be approximated as a plane wave. The directions of the incident plane waves in relation to the holographic medium are shown schematically in Fig. 4.

To minimize cross talk resulting from direct overlap of the angular response peaks of the Bragg-regime interconnection gratings, sufficient angular separation must exist between the diffraction peaks of the various gratings.[44] In the system that we model, the angular sensitivities are separated by approximately four to five angular Bragg peak widths (defined by the full width at half-maximum [FWHM]) in order to achieve good angular isolation of the interconnection gratings. The effects of varying this design goal are discussed in Section 4.E.2.

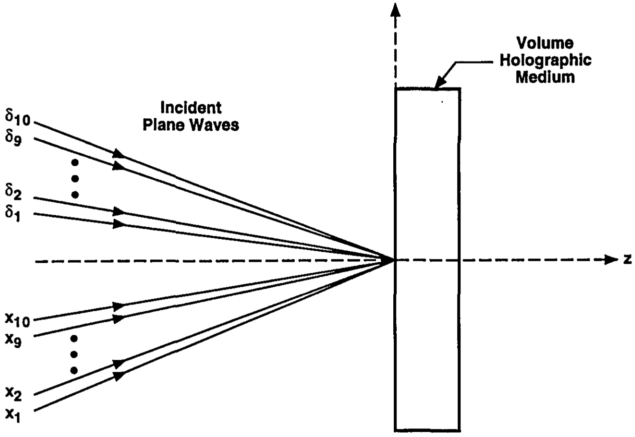

Fig. 4. Schematic representation of the plane waves generated by the pixels in the input and training planes for a 10-to-10 single-source architecture.

To meet the angular-isolation design goal while maintaining reasonable values for the various geometric parameters of the system, we made the following design choices. The pixel spacings s_x and s_y in the input and training planes, respectively, are both 257 μm. The focal length of each lens shown in Fig. 3 is $f_1 = f_2 = f_3 = 50$ mm. The angular separation of plane waves from adjacent pixels is therefore 0.29°. The angles at which the centers of the pixel planes are offset from the normal direction of the volume holographic medium's front surface (the $-z$ axis in Fig. 3) are $\theta_y = 5.9°$ and $\theta_x = 8.8°$ for the training and input planes, respectively. (For generality, θ_y and θ_x are not symmetric about normal incidence; other simulations that we performed in which these angles are symmetric yield the same results as discussed below.) Unless otherwise noted, all angles are specified in air. The wavelength λ of the recording and readout light is 0.514 μm.

For our simulations the thickness of the holographic medium d is chosen as 4.5 mm; its refractive index is assumed to be 2.52 (characteristic of single-crystal bismuth silicon oxide, $Bi_{12}SiO_{20}$). Given this thickness and the above recording geometry, the angular Bragg widths of the interconnection gratings range from 0.055° to 0.080° (FWHM). These values correspond to the smallest (connecting x_1 to ρ_{10}) and largest (connecting x_{10} to ρ_1) interconnection gratings, which have grating periods Λ of 1.7 and 2.5 μm, respectively.

As noted in Section 3.A, cross gratings exist in each possible diffraction regime for the simultaneous and pagewise-sequential recording methods. One rule of thumb in determining whether a particular grating operates in the Raman–Nath, Bragg, or transition regimes is based on the value of the normalized thickness Q calculated for that grating.[45] The normalized thickness is often defined as $Q = 2\pi\lambda d/n\Lambda^2$.[46] If $Q \leq 1$, Raman–Nath diffraction behavior is typically observed, whereas the Bragg diffraction regime corresonds to $Q \geq 10$. The transition regime occurs for values of Q between 1 and 10. (For a more detailed analysis, see Ref. 46.) Cross gratings formed between adjacent pixels have Q values of ~ 0.6 and thus clearly operate in the Raman–Nath diffraction regime. For pixel separations of four or more the corresponding cross gratings have $Q \geq 10$ (with the largest having $Q = 46$). For comparison, the desired interconnection gratings have typical Q values of ~ 1500.

The geometry of the single-source architecture is assumed to be completely rigid such that optical phase stability (see Section 2.C) is maintained during recording and readout. The SLM pixels of the input and training planes are assumed to be pure amplitude modulators with no residual phase modulation.

2. *Holographic-Medium Recording Characteristics*

Next we define the recording characteristics of the holographic medium used in our modeling and the associated terminology. The medium itself is assumed to be lossless, and the recorded diffraction gratings are assumed to be sinusoidal phase gratings. The material is taken to be linear in the sense that the induced changes in the local refractive index distribution during recording are directly proportional to the associated recording intensity distribution. The available refractive index modulation range is assumed to be unlimited; therefore effects that are caused by a limited available modulation range in real materials are not included. The effects of self-diffraction among the recording beams, erasure, and exposure scheduling are also neglected. Investigation of these effects not only on the optical throughput but also on the reconstruction fidelity is an important area for continuing research.

To illustrate the relationships among the recording intensity, the induced refractive index modulation, and the resultant grating strength for each grating, we consider the simultaneous recording method, using a single training pair $(\mathbf{x}, \boldsymbol{\delta})$. Pixels in the input plane are identified with the indices j and j', while pixels in the training plane are referred to with indices i and i'. The intensity distribution in the holographic medium can be written as

$$\begin{aligned} I(\mathbf{r}) &= I_0 \left| \sum_{j=1}^{N} (x_j)^{1/2} \exp(i\mathbf{k}_j \cdot \mathbf{r}) + \sum_{i=1}^{N} \left(\frac{\delta_i}{R}\right)^{1/2} \exp(i\mathbf{k}_i \cdot \mathbf{r}) \right|^2 \\ &= \frac{I_0}{\sqrt{R}} \Bigg[\sqrt{R} \sum_{j=1}^{N} x_j + \frac{1}{\sqrt{R}} \sum_{i=1}^{N} \delta_i \\ &\quad + 2 \sum_{i=1}^{N} \sum_{j=1}^{N} (x_j \delta_i)^{1/2} \cos(\mathbf{K}_{ij} \cdot \mathbf{r}) \\ &\quad + 2\sqrt{R} \sum_{j=1}^{N} \sum_{j'=j+1}^{N} (x_j x_{j'})^{1/2} \cos(\mathbf{K}_{jj'} \cdot \mathbf{r}) \\ &\quad + 2 \frac{1}{\sqrt{R}} \sum_{i=1}^{N} \sum_{i'=i+1}^{N} (\delta_i \delta_{i'})^{1/2} \cos(\mathbf{K}_{ii'} \cdot \mathbf{r}) \Bigg], \end{aligned} \tag{9}$$

in which $\mathbf{r} = x\hat{x} + z\hat{z}$, R is the (intensity) beam splitter ratio (see Fig. 3), and x_j and δ_i are the transmittances of the jth and ith pixels of the input and training planes, respectively. The wave vectors of the corresponding plane waves are $\mathbf{k}_j$ and $\mathbf{k}_i$, while $\mathbf{K}_{\mu\nu}$ (the grating wave vector that characterizes the intensity distribution recorded as a variation in the refractive index) is defined as $\mathbf{k}_\mu - \mathbf{k}_\nu$, in which μ and ν denote the appropriate pixel index (i, i', j, or j'). For simplicity, the phase of each beam is assumed to be zero at the origin of the x–z coordinate system shown in Figs. 3 and 4.

For the purposes of our analysis we neglect effects that result from any change in the average refractive index and consider only the induced variation in the index. Thus the refractive index modulation $\Delta n(\mathbf{r})$, caused by exposure to the intensity distribution described in Eq. (9), can be written as

$$\Delta n(\mathbf{r}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta n_{ij}\cos(\mathbf{K}_{ij}\cdot\mathbf{r}) + \sum_{j=1}^{N}\sum_{j'=j+1}^{N} \Delta n_{jj'}\cos(\mathbf{K}_{jj'}\cdot\mathbf{r}) + \sum_{i=1}^{N}\sum_{i'=i+1}^{N} \Delta n_{ii'}\cos(\mathbf{K}_{ii'}\cdot\mathbf{r}), \qquad (10)$$

in which the amplitude of the refractive-index modulation for each desired interconnection grating is given by

$$\Delta n_{ij} = C_1(x_j\delta_i)^{1/2}, \qquad (11)$$

while the amplitudes of the intraplanar cross gratings of the input and training planes are described by

$$\Delta n_{jj'} = C_1(Rx_jx_{j'})^{1/2}, \qquad (12)$$

$$\Delta n_{ii'} = C_1(\delta_i\delta_{i'}/R)^{1/2}, \qquad (13)$$

respectively. The proportionality constant C_1 is determined by $2I_0C_0/\sqrt{R}$, in which C_0 is related to the sensitivity of the material and the length of exposure. If we vary the beam splitter ratio R, the relative magnitudes of the desired and cross gratings can be varied. The importance of this ratio is discussed in more detail in Section 4.B.

It is convenient to discuss the relative magnitudes of the various gratings in terms of their grating strength ν rather than their refractive-index amplitudes. The grating strength ν for a given grating is typically defined as

$$\nu = 2\pi\Delta nD/\lambda, \qquad (14)$$

in which Δn is the amplitude of the grating's refractive-index modulation and D is the optical path length in the holographic medium. The strengths of the various types of gratings described by Eq. (10) can be written as

$$\nu_{ij} = (2\pi C_1D/\lambda)(x_j\delta_i)^{1/2}, \qquad (15)$$

$$\nu_{jj'} = (2\pi C_1D/\lambda)(Rx_jx_{j'})^{1/2}, \qquad (16)$$

$$\nu_{ii'} = (2\pi C_1D/\lambda)(\delta_i\delta_{i'}/R)^{1/2}, \qquad (17)$$

in which ν_{ij} represents the grating strength of the ijth interconnection grating and $\nu_{jj'}$ and $\nu_{ii'}$ are the grating strengths of input-plane and training-plane cross gratings, respectively. The strength of each grating is proportional to the square root of the product of the writing intensities. Equations (15)–(17) are valid only for recording with a single training pair $(\mathbf{x}, \boldsymbol{\delta})$. The relevant expressions for multiple training pairs are discussed in Section 3.C.3.

3. *Holographic Weight Formation with Multiple Training Pairs*

We discuss herein the method used to generate the weight matrices employed in our simulations. We are interested in examining the achievable fidelity and throughput of weighted fan-out/fan-in holographic interconnection systems that incorporate *nonsingular* weight matrices, since typical neural-network systems presumably have relatively independent weights.

One method of generating such a nonsingular weight matrix is simply to assign a random number to each weight. However, the desired interconnection weights and associated grating strengths must be consistent with the relative magnitudes of any cross gratings that are present (i.e., for the simultaneous or pagewise-sequential recording methods). This self-consistency requirement derives from the fact that the weights and grating strengths of the cross gratings that are generated in an actual holographic interconnection system are not independent of each other; they are traceable instead to the set of training pairs that were used to record the desired interconnection weights.

Rather than attempting to deduce a self-consistent set of cross-grating strengths given a particular random weight matrix, we use an alternative approach for generating the weight matrix. In this approach we first specify a set of M training pairs and then calculate the corresponding grating strengths for both the desired and cross gratings. This procedure ensures that a set of mutually consistent magnitudes result for all of the gratings applicable to each recording method. For example, to obtain relatively independent weights for the 10-to-10 interconnection case, we used ten random training pairs $[\mathbf{x}^{(m)}, \boldsymbol{\delta}^{(m)}]$ to generate the interconnections. The components of each vector, $\mathbf{x}^{(m)}$ and $\boldsymbol{\delta}^{(m)}$, were taken to be random variables uniformly distributed on the interval $[0, 1]$.

In the remainder of this section we illustrate the multiple-training-pair approach to weight-matrix formation. The primary goal of this effort is to specify the dependence of each weight on the corresponding vector components of the training pairs used to form the weight matrix. In order to achieve this goal we must also determine the appropriate relationship between the grating strength of a given grating recorded in the single-source architecture and its corresponding weight. At the outset, we note that analytical expressions for the relationship between a given grating strength and its associated weight have been derived only for special cases. In particular, no analytical expression has been derived for the case of multiple training pairs (which corresponds to the case of independent weights). In fact, this situation provides a key impetus for complex system simulations using the optical beam propagation method. However, in order to adequately assess the fidelity of a given holographic interconnection system that incorporates independent weights, we must first determine a trial weight matrix in order to compute expected interconnection system outputs analytically for comparison with the simulation results.

As in Section 3.C.2, we base the following example on an analysis of the simultaneous recording method. We first compute the refractive index distribution when multiple training pairs are used. The refractive index distribution is found by summing Eq. (10) over the number of training pairs M with the assumption of holographic medium linearity:

$$\begin{aligned}\Delta n(\mathbf{r}) = \sum_{m=1}^{M} \Bigg[&\sum_{i=1}^{N}\sum_{j=1}^{N} \Delta n_{ij}^{(m)} \cos(\mathbf{K}_{ij}\cdot\mathbf{r}) \\ &+ \sum_{j=1}^{N}\sum_{j'=j+1}^{N} \Delta n_{jj'}^{(m)} \cos(\mathbf{K}_{jj'}\cdot\mathbf{r}) \\ &+ \sum_{i=1}^{N}\sum_{i'=i+1}^{N} \Delta n_{ii'}^{(m)} \cos(\mathbf{K}_{ii'}\cdot\mathbf{r})\Bigg], \end{aligned} \tag{18}$$

in which

$$\Delta n_{ij}^{(m)} = C_1[x_j^{(m)}\delta_i^{(m)}]^{1/2}, \tag{19}$$

$$\Delta n_{jj'}^{(m)} = C_1[Rx_j^{(m)}x_{j'}^{(m)}]^{1/2}, \tag{20}$$

$$\Delta n_{ii'}^{(m)} = C_1[\delta_i^{(m)}\delta_{i'}^{(m)}/R]^{1/2}. \tag{21}$$

Equation (18) can be rewritten as

$$\begin{aligned}\Delta n(\mathbf{r}) = &\sum_{i=1}^{N}\sum_{j=1}^{N} \Delta n_{ij}^{(M)} \cos(\mathbf{K}_{ij}\cdot\mathbf{r}) \\ &+ \sum_{j=1}^{N}\sum_{j'=j+1}^{N} \Delta n_{jj'}^{(M)} \cos(\mathbf{K}_{jj'}\cdot\mathbf{r}) \\ &+ \sum_{i=1}^{N}\sum_{i'=i+1}^{N} \Delta n_{ii}^{(M)} \cos(\mathbf{K}_{ii'}\cdot\mathbf{r}), \end{aligned} \tag{22}$$

in which each refractive-index amplitude is the sum of the contributions from each training pair:

$$\Delta n_{\mu\nu}^{(M)} = \sum_{m=1}^{M} \Delta n_{\mu\nu}^{(m)}, \tag{23}$$

in which μ and ν denote the appropriate pixel indices. Given the preceding relationships, we can immediately determine the dependence of the grating strengths of the desired interconnection gratings and also the input-plane and training-plane cross gratings on the writing intensities of a given set of training pairs, namely,

$$\nu_{ij}^{(M)} = (2\pi C_1 D/\lambda) \sum_{m=1}^{M} [x_j^{(m)}\delta_i^{(m)}]^{1/2}, \tag{24}$$

$$\nu_{jj'}^{(M)} = \sqrt{R}(2\pi C_1 D/\lambda) \sum_{m=1}^{M} [x_j^{(m)}x_{j'}^{(m)}]^{1/2}, \tag{25}$$

$$\nu_{ii'}^{(M)} = (1/\sqrt{R})(2\pi C_1 D/\lambda) \sum_{m=1}^{M} [\delta_i^{(m)}\delta_{i'}^{(m)}]^{1/2}, \tag{26}$$

respectively.

The final step in determining the trial weight matrix formed by the set of training pairs is to specify the relationship between the grating strength of an individual grating and its corresponding weight. Recall that a weight is proportional to the intensity diffraction efficiency of its associated grating (in the intensity representation); i.e., it is a measure of the fraction of a new input x_j that is diffracted to and summed at a pixel i in the output plane. As a starting point we first examine an analytical expression for the relationship between the grating strength of a given grating and its associated weight derived under certain restrictive assumptions. As shown in Ref. 19, by means of coupled-wave analysis each weight W_{ij} is proportional to (in our notation)

$$W_{ij} \propto x_j\delta_i \tag{27}$$

in a single-source architecture when (1) no cross gratings are present, (2) readout is performed with mutually coherent beams, and (3) only a single training pair is recorded (which results in a set of dependent weights). Using Eq. (15), one can re-express each weight in terms of the grating strength of its corresponding grating for this restricted case:

$$W_{ij} \propto \nu_{ij}^{2}. \tag{28}$$

Since a corresponding analytical result is not available for the case of multiple training pairs (that result in independent weights) in either the presence or absence of cross gratings, it is not clear how to generalize from the ideal single-training-pair case. In order to generate an appropriate trial weight matrix for the purposes of our analysis, we therefore assume that the grating strengths deriving from separate training pairs add linearly in so far as they affect the corresponding diffraction efficiency. Thus the elements of a particular trial weight matrix are assumed to be related to the training pairs used to form the weight matrix by

$$W_{ij} \propto [\nu_{ij}^{(M)}]^2 \propto \left\{\sum_{m=1}^{M} [x_j^{(m)}\delta_i^{(m)}]^{1/2}\right\}^2. \tag{29}$$

Whether Relation (Rel.) (29) is an appropriate choice for the relationship between a weight and its associated grating strength depends on the accuracy with which it predicts the diffracted outputs of an actual system [given a set of inputs, as described in Eq. (4)]. As shown in Section 4, Rel. (29) appears to represent a remarkably good choice of a metric for single-source architectures. In fact, Rel. (29) can be taken as an approximate analytical expression of the multiple-training-pair case on the basis of the relatively small errors that result from the use of this metric in simulations of the single-source architecture without the effects of cross gratings. For different architectures other relationships can be selected that provide a better match to the physics of the underlying diffraction process. One such choice is discussed in Section 5.B for a particular configuration of the incoherent/coherent double angularly multiplexed architecture.

To facilitate direct comparisons, we use the same trial weight matrix **W** for all of the 10-to-10 simulations discussed below, including both the single-source and incoherent/coherent double angularly multiplexed architectures. In addition, the same readout vector $\mathbf{x}^{\text{read}}$ is used for all cases. Each component of $\mathbf{x}^{\text{read}}$ is a random variable uniformly distributed over [0, 1].

4. Single-Source Interconnection Architecture: Simulation Results

On the basis of the modeling technique and assumptions discussed in Section 3, we present herein the results of our numerical simulations of the single-source architecture. These results provide a benchmark against which to compare the performance of the incoherent/coherent double angularly multiplexed architecture discussed in Section 5. Simulation results for readout of the single-source architecture with mutually coherent beams for simultaneous, pagewise-sequential, and fully sequential recording methods are discussed in Sections 4.A, 4.B, and 4.C, respectively. For comparison, readout of the fully sequential recording case with mutually incoherent beams is addressed in Section 4.D, in which the effect of beam degeneracy on optical throughput is examined. In Section 4.E the rms errors of the diffracted outputs are directly compared for each case, and we show comparable modeling results for a 4-to-4 interconnection system so that scaling trends can be identified. In Section 4.F we summarize our findings.

A. *Simultaneous Recording and Readout with Mutually Coherent Beams*

As discussed in Section 3.C.3, the refractive-index distribution for simultaneous recording in a single-source architecture that uses multiple training pairs is given by Eq. (22). The first term in Eq. (22) represents the desired interconnection gratings, while the second and third terms describe undesired cross gratings connecting pixels within the input and training planes, respectively. As discussed in Section 2.E, these extraneous gratings introduce a serious source of cross talk into the interconnection system that compromises the independence and isolation of each interconnection pathway.

As a consequence of the extraneous interconnection paths provided by the cross gratings, the actual output of the interconnection system $\boldsymbol{\rho}'$ differs from the ideal output $\boldsymbol{\rho}$ [described by Eq. (6)] when the system is read out with $\mathbf{x}^{\text{read}}$. This lack of reconstruction fidelity is illustrated in Fig. 5 for readout with mutually coherent beams. The diffracted outputs are shown in Fig. 5(a) for the $R = 1$ (unity beam ratio) case as a function of the grating strength of the largest desired interconnection grating recorded in the holographic medium. The magnitude of each diffracted output is defined as the fraction of the total input power that is diffracted into that output.

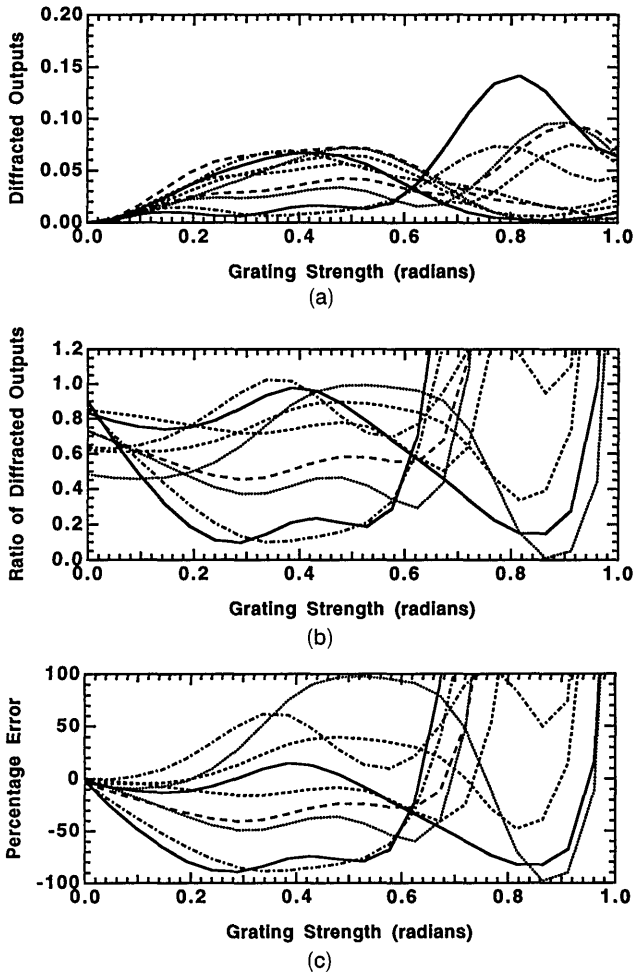

Fig. 5. Single-source architecture simulation results for the simultaneous recording method: (a) diffracted outputs, (b) ratio of each diffracted output to an arbitrarily chosen output, and (c) percentage error of each ratio. Readout is performed with mutually coherent beams. The horizontal axis represents the grating strength of the largest interconnection grating recorded in the holographic medium.

To examine the fidelity of the actual output vector $\boldsymbol{\rho}'$ as compared with the desired output vector $\boldsymbol{\rho}$, we first show the ratios of the diffracted outputs (with respect to an arbitrarily chosen output, in this case ρ_4) in Fig. 5(b). In an ideal interconnection network each diffracted output ratio should be independent of grating strength. The actual ratios vary dramatically as a function of grating strength, indicating that the reconstruction fidelity is grating-strength dependent. As one possible measure of the error in each component of the diffracted output $\boldsymbol{\rho}'$, we show in Fig. 5(c) the deviation of the ratios of Fig. 5(b) (including the effects of cross coupling) from the corresponding ratios calculated from the components of the ideal output vector $\boldsymbol{\rho}$ normalized by the appropriate ideal output-vector ratio. The percentage error $\Gamma_{ii'}$ is defined as

$$\Gamma_{ii'} = 100\left(\frac{\rho_i'/\rho_{i'}' - \rho_i/\rho_{i'}}{\rho_i/\rho_{i'}}\right), \tag{30}$$

in which ρ_i' is a component of $\boldsymbol{\rho}'$. As can be seen from Fig. 5(c), the percentage errors again vary dramatically as a function of grating strength. Since there is no apparent correlation among the errors in the individual outputs, the errors cannot be corrected in any systematic fashion. Further simulations show that the errors are both input-signal and weight-matrix dependent.

The magnitudes of the output errors must be evaluated in conjunction with the optical throughput, which is shown as a function of grating strength in Fig. 6. The throughput is defined, as above, as the fraction of the total incident power that is diffracted into all of the desired outputs. At the peak throughput of nearly 50% (in this case) the percentage errors are as large as 100%. Backing off in throughput to 10% still results in up to 50% errors in the ratios. To reduce the errors to less than 10%, the strength of the interconnections must also be reduced so that the throughput is less than 1.5%. As demonstrated by these results, the simultaneous recording case is characterized by significant inherent fidelity errors, particularly for relatively high throughput. Examination of the pagewise-sequential and fully sequential results in Sections 4.B and 4.C indicates that nearly all of this error is attributable to the cross gratings.

B. Pagewise-Sequential Recording and Readout with Mutually Coherent Beams

The primary motivation for using pagewise-sequential instead of simultaneous recording is to reduce the deleterious effects of the undesired coherent-recording cross-talk gratings. The validity of this argument for a single-source architecture can be examined by comparing the grating strengths of the desired and cross gratings for the two recording methods.

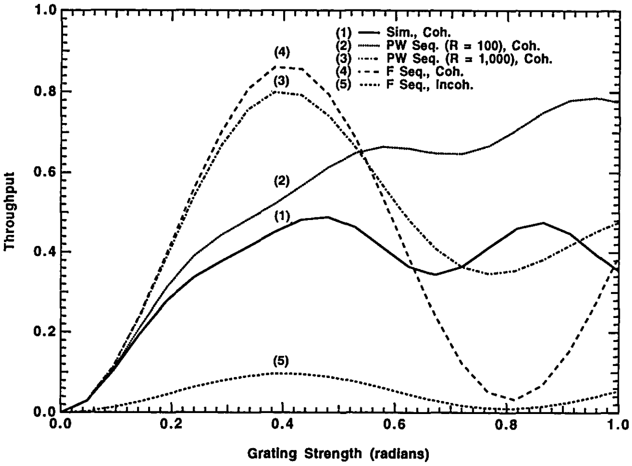

Fig. 6. Simulation results showing the optical throughput (i.e., the amount of incident power diffracted into the desired outputs) for several combinations of recording and readout methods for the single-source architecture. Clarification of the legend is as follows: (1) simultaneous recording method with coherent readout (see Fig. 5), (2) pagewise-sequential recording method with a beam splitter ratio of 100 and coherent readout, (3) pagewise-sequential recording method with a beam splitter ratio of 1000 and coherent readout (see Fig. 7), (4) fully sequential recording with coherent readout (see Fig. 8), and (5) fully sequential recording with incoherent readout (see Fig. 9). The horizontal scale is the same for all single-source-architecture 10-to-10 simulation results, shown in Figs. 5, 7–9, and 11.

Assuming M training pairs, the refractive-index distribution for pagewise-sequential recording is

$$\Delta n(\mathbf{r}) = \sum_{m=1}^{M} \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{N} \Delta n_{ij}^{(m)} \cos(\mathbf{K}_{ij} \cdot \mathbf{r}) + \sum_{i'=1}^{N} \sum_{i''=i'+1}^{N} \Delta n_{i'i''}^{(m)} \cos(\mathbf{K}_{i'i''} \cdot \mathbf{r}) \right] \right\} \tag{31}$$

for a linear recording medium. Using Eq. (23), we can rewrite this as

$$\Delta n(\mathbf{r}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta n_{ij}^{(M)} \cos(\mathbf{K}_{ij} \cdot \mathbf{r}) + N \sum_{i=1}^{N} \sum_{i'=i+1}^{N} \Delta n_{ii'}^{(M)} \cos(\mathbf{K}_{ii'} \cdot \mathbf{r}). \tag{32}$$

A comparison of Eq. (32) with Eq. (22) shows that the input-plane cross-grating terms are, of course, absent. For pagewise-sequential recording, however, an extra factor of N multiplies the training-plane cross-grating terms [the second summation on the right-hand side of Eq. (32)], because for each training pair the training-plane cross gratings are exposed N times, while each desired grating is exposed only once. Thus over the full set of M training pairs, each cross grating is exposed MN times, while each desired interconnection grating receives only M exposures. The grating strength of the desired interconnection gratings is given by Eq. (24), while that of the training-plane cross gratings is

$$\nu_{ii'}^{(M)} = (N/\sqrt{R})(2\pi C_1 D/\lambda) \sum_{m=1}^{M} [\delta_i^{(m)} \delta_{i'}^{(m)}]^{1/2}. \tag{33}$$

This result calls into question the common assertion that the training-plane cross gratings can be made arbitrarily small relative to the interconnection gratings simply by using a sufficiently large beam splitter ratio R.[19,38,47] While this is true in principle for each exposure, the beam splitter ratio must be further increased to overcome the additional effects of exposing the cross gratings $N - 1$ more times than the desired gratings. According to Eq. (33), the beam splitter ratio must be $R = N^2$ just to achieve parity in the grating strengths of the interconnection and cross gratings. For even relatively small numbers of pixels, the beam splitter ratio therefore becomes very large; for example, if $N = 100$, R must be at least 10,000. Although not directly addressed herein, the use of large beam splitter ratios can result in significant throughput trade-offs for implementation in certain types of holographic materials (such as photorefractive single crystals), which are in turn caused by reductions in the grating modulation depth during recording.[12]

Regardless of the practicality of pagewise-sequential recording, our modeling of a 10-to-10 single-source architecture shows that if the beam splitter ratio is made large enough to reduce the magnitudes of the cross gratings relative to the interconnection gratings, then the fidelity and throughput of the interconnection system can be improved as compared with the simultaneous-recording method. For example, Fig. 7 shows the diffracted outputs, ratios, and ratio percentage errors for readout with mutually coherent beams. A beam splitter ratio of 1000 is assumed, which adjusts the average strength of the interconnection gratings to approximately three times the strength of the cross-talk gratings. As seen in Fig. 7(c), the percentage errors of the ratios show a marked improvement over the simultaneous-recording case. However, at the peak throughput of nearly 80% (see Fig. 6), the largest percentage errors in the ratios are still 50%. At 10% throughput the largest error is ~15%. As shown in Section 4.C, the fidelity can be further improved when the cross gratings are eliminated entirely by using fully sequential recording.

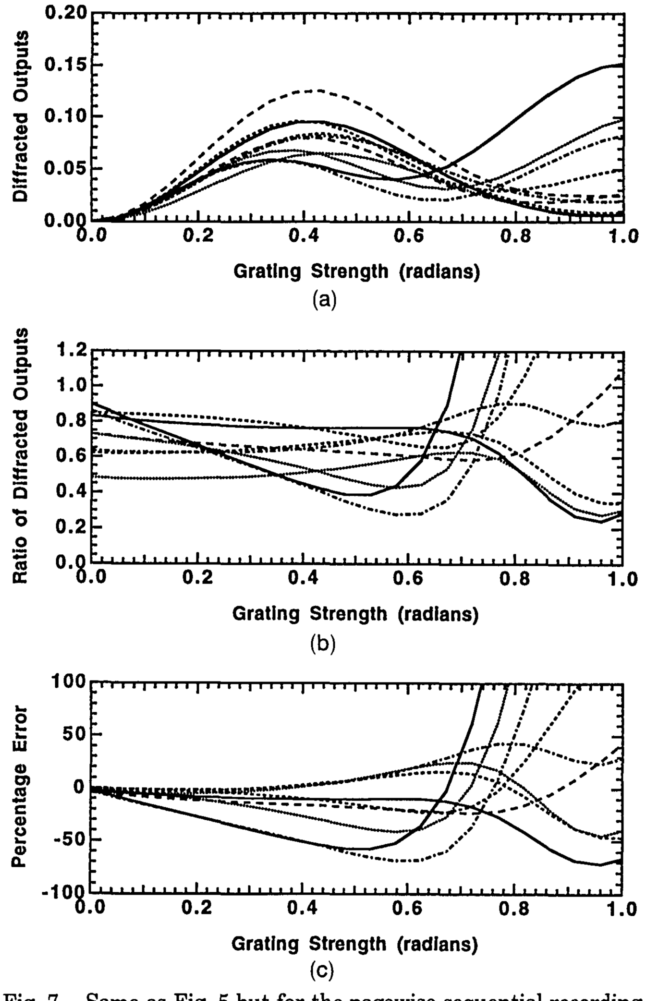

Fig. 7. Same as Fig. 5 but for the pagewise-sequential recording method with a beam splitter ratio R of 1000.

Because increasing the beam splitter ratio can reduce the effects of the cross gratings for the pagewise-sequential recording method, one might ask whether increasing the beam splitter ratio for the simultaneous-recording method in the single-source architecture offers any benefits. As can be seen by comparing Eqs. (24)–(26) for the grating strengths of the interconnection cross gratings for the simultaneous-recording method, increasing the beam splitter ratio reduces the relative magnitudes of the training-plane cross gratings by a factor of $\sqrt{R}$ [Eq. (26)] as compared with the desired interconnection gratings [Eq. (24)]. However, the relative magnitudes of the input-plane cross gratings increase by a factor of $\sqrt{R}$ [Eq. (25)]. The net result is that varying the beam splitter ratio from unity always enhances one set of cross gratings relative to the desired set of interconnection gratings. Our simulations show that this results in significantly reduced fidelity and throughput performance. Thus a unity beam splitter ratio is optimal for the simultaneous recording method.

C. Fully Sequential Recording and Readout with Mutually Coherent Beams

Fully sequential recording with M training pairs results in a refractive-index distribution in the holographic medium of

$$\Delta n(\mathbf{r}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta n_{ij}^{(M)} \cos(\mathbf{K}_{ij} \cdot \mathbf{r}), \tag{34}$$

in which there are no undesired cross gratings. The corresponding grating strengths are given by Eq. (24).

Simulation results for readout with mutually coherent beams are shown in Fig. 8. The beam splitter ratio R is assumed to be unity during recording since there are no cross gratings to minimize. Each diffracted output [Fig. 8(a)] appears more sinusoidal in nature as a function of grating strength than that for either the pagewise-sequential or simultaneous recording cases discussed above. However, the first peak (and the following minimum) of each diffracted output occurs at a different grating strength, with the result that the reconstruction fidelity is somewhat grating-strength dependent [as indicated in Fig. 8(b)]. This result is similar to the behavior noted in Fig. 5 of Ref. 19 for a 5-to-5 interconnection, which was modeled numerically using a coupled-wave approach. The percentage error of the ratios, shown in Fig. 8(c), shows a significant improvement over the pagewise-sequential and simultaneous recording cases that is attributable to elimination of the cross gratings. At the peak throughput of >85% (see Fig. 6) the largest ratio percentage error is 15%. For 10% throughput the largest error is only 3%.

Although the fidelity and throughput performance are greatly improved when the cross gratings are eliminated, the fully sequential recording method has several serious difficulties for practical systems, including large numbers of recording steps, inefficient

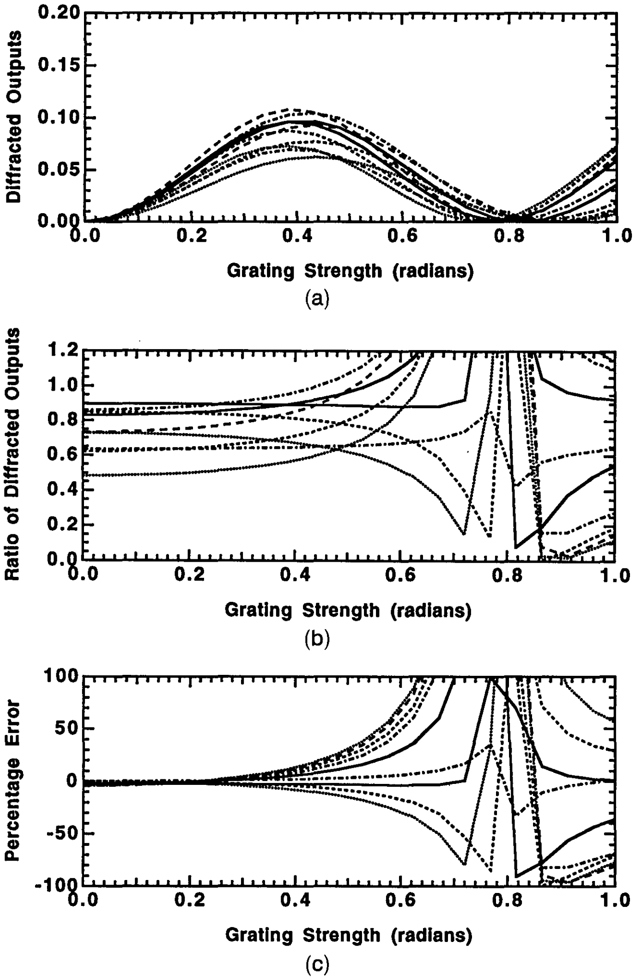

Fig. 8. Single-source architecture simulation results for the pagewise-sequential recording method for a beam splitter ratio, R, of 1000: (a) diffracted outputs, (b) ratio of each diffracted output to an arbitrarily chosen output, and (c) percentage error of each ratio. Readout is performed with mutually coherent beams.

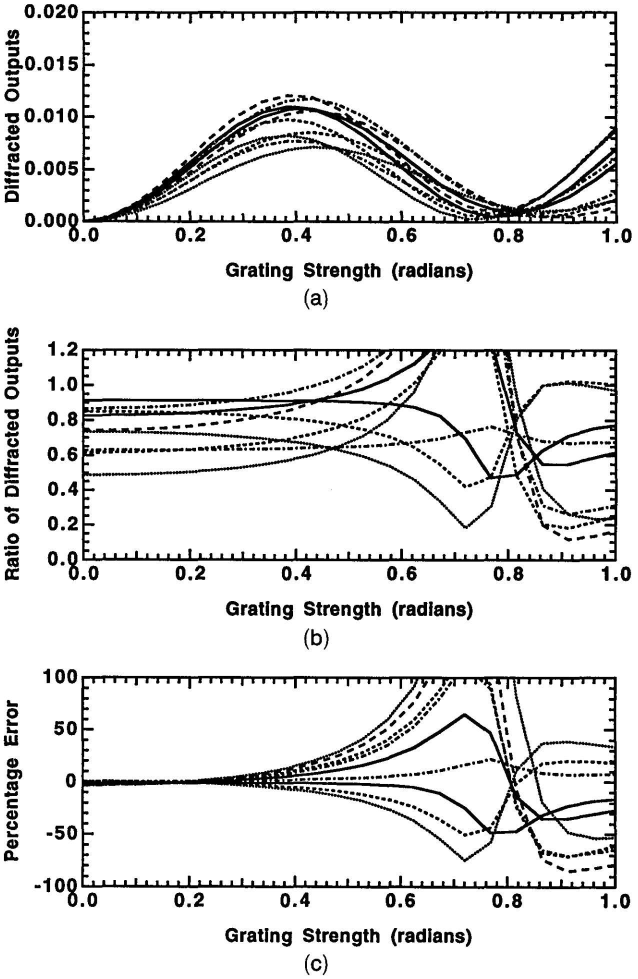

Fig. 9. Single-source architecture simulation results for the fully sequential recording method: (a) diffracted outputs, (b) ratio of each diffracted output to an arbitrarily chosen output, and (c) percentage error of each ratio. Readout is performed with mutually coherent beams.

use of the illuminating beams, and complicated recording schedules (see Section 2.D).

D. Readout with Mutually Incoherent Beams

To this point in the discussion we have considered readout of single-source architectures with mutually coherent beams only. In order to better understand the innovations incorporated in the incoherent/coherent double angularly multiplexed architecture discussed in Section 5, we briefly discuss readout of the single-source architecture with mutually *incoherent* beams. For the purposes of the section we consider the fully sequential recording method, such that only the desired interconnection gratings are present in the holographic medium.

Simulation results for the diffracted outputs in this case are shown in Fig. 9. The reconstruction fidelity is essentially the same as for readout with mutually coherent beams (Fig. 8), even though the ideal outputs are expressed in this case by Eq. (4) instead of Eq. (6). Since beam degeneracy is always present in the single-source architecture, the low fidelity error results shown for both mutually incoherent and coherent readout beams indicate that beam degeneracy is not a significant source of fidelity error (at least, not for grating strengths up to ~ 0.4 rad, at which the peak in throughput occurs). However, for readout with mutually incoherent beams the peak throughput drops to 10% (as shown in Fig. 6) because of incoherent fan-in loss, which limits the maximum throughput to $1/N$ for an N-to-N interconnection system with collinear output summation.[24]

An understanding of the physical mechanism responsible for incoherent fan-in loss in a single-source holographic interconnection system permits us to design a system that avoids this loss (as discussed in Section 5). The fan-in loss in a single-source architecture for readout with incoherent beams can be explained by the *beam degeneracy* that is present in the diffracted outputs.[9,10,13,25] Beam degeneracy refers to the **k**-vector degeneracies in beams diffracted from different gratings in holographic interconnection systems that have collinear fan-in. This can be understood by reference to Fig. 10(a), in which a

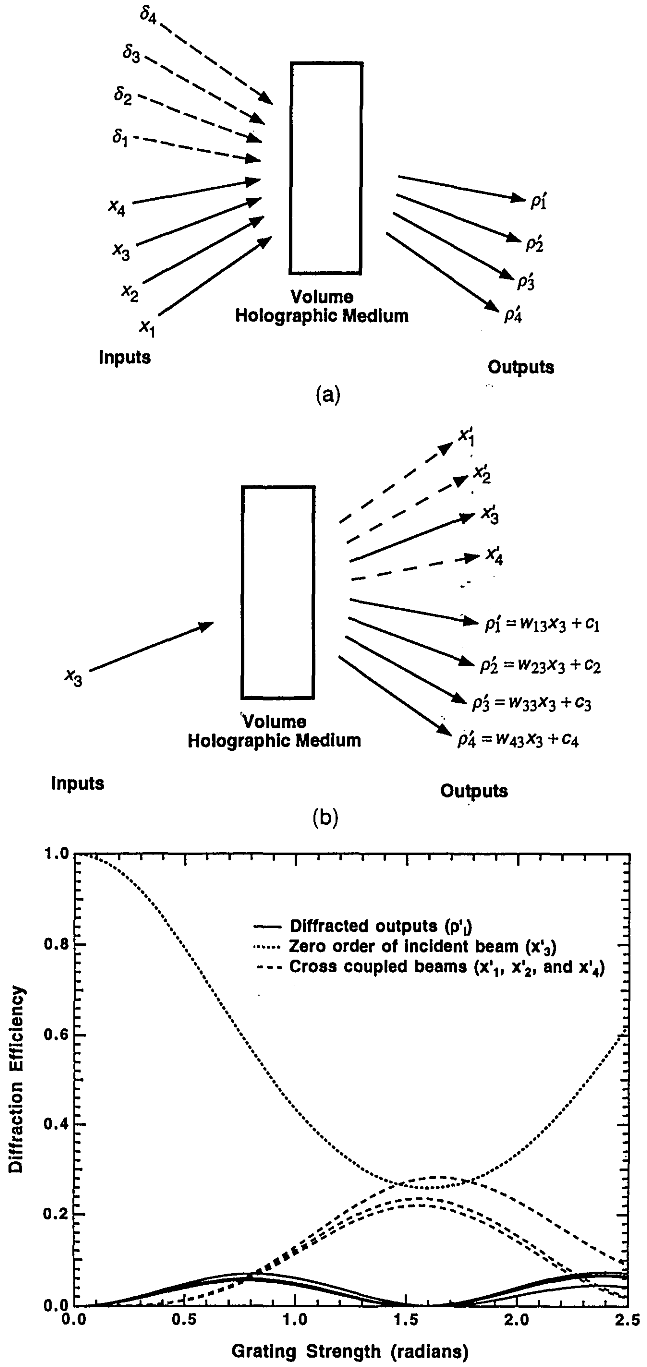

Fig. 10. (a) Schematic diagram of the recording geometry of a 4-to-4 interconnection system; (b) schematic diagram of the readout geometry of the 4-to-4 interconnection system with a single beam (x_3) and the resulting outputs; (c) simulation results for readout of the 4-to-4 interconnection system, in which the power cross coupled by beam degeneracy from the desired outputs to the x_1', x_2', and x_4' beams can be significant.

4-to-4 interconnection system is represented schematically. As in Fig. 4, each collimated beam is represented by its **k** vector. When readout is performed using all four input beams (x_1–x_4), four reconstructed beams emerge collinearly in each of the four ρ_i' directions. Thus the four beams that are diffracted in a given ρ_i' direction have degenerate wave vectors; i.e., they are *degenerate beams.*

To understand how this degeneracy leads to a fan-in loss for readout with mutually incoherent beams, consider readout of the 4-to-4 interconnection system described above with a single input beam, x_3. As shown by the solid arrows in Fig. 10(b), diffracted outputs are generated in each of the four output directions (ρ_1'–ρ_4') as well as the zero-order direction x_3'. However, each diffracted beam ρ_i' is in turn diffracted into the directions x_1', x_2', and x_4' by the gratings recorded among the training beams δ_1–δ_4 and the other input beams x_1, x_2, and x_4. These interactions cause power to be coupled out of the desired outputs ρ_i' and into the zero orders of x_1, x_2, and x_4, generating the cross-coupled beams x_1', x_2', and x_4'.

The magnitude of this effect can be illustrated by modeling a 4-to-4 interconnection system (using the BPM) in which a single beam of unit intensity [corresponding to x_3 in Fig. 10(b)] is used for readout. Uniform weights are assumed ($W_{ij} = 1$ for all i and j), and the geometry of the system used in the modeling is the same as that described in Section 3.C.1. The results of the BPM calculation are shown in Fig. 10(c), in which the diffraction efficiencies of all of the diffracted beams are shown as a function of grating strength. The diffraction efficiencies of the desired outputs ρ_i' are indicated with solid curves, while the zero-order and cross-coupled beams are represented by the dotted and dashed curves, respectively. The peak diffraction efficiency into the desired outputs ρ_i' occurs at ~0.8 rad. The summed power in the four desired outputs is only 24%, which is 1/4 (or $1/N$) of the maximum available power. The rest of the power either remains in the zero order (x_3') or is diffracted into the cross-coupled beams (x_1', x_2', and x_4'), each of which has approximately the same diffraction efficiency as the desired beams (at 0.8 rad).

It is clear that if the other three input beams (x_1, x_2, and x_4) are also incident and that if all of the incident beams are mutually incoherent, then the summed diffraction efficiency into the desired outputs is still only 25% at best, because the light in each output beam adds incoherently. The end result in this case is analogous to that corresponding to the collinear combination of N mutually incoherent quasi-monochromatic beams of essentially identical wavelengths using $N - 1$ beam splitters; only $1/N$ of the power at most can be diverted into the desired direction, while the remaining power exits the system in the direction of the original input beams.[13]

However, as illustrated in the simulation results of Section 4.C, it is possible to achieve high optical throughput in the single-source architecture if the incident beams are mutually coherent and the proper phase conditions are satisfied. Alternatively, if a holographic interconnection system is designed such that each detected output consists of *angularly* instead of *collinearly* fanned-in beams, the throughput loss that is due to the effects of beam degeneracy can be avoided even for readout with mutually incoherent beams. The incoherent/coherent double angularly multiplexed architecture discussed in Section 5 illustrates one method of achieving such angular fan-in at each pixel in the output plane with consequent high throughput for readout with mutually incoherent beams.

E. Discussion

1. *Comparison of Recording Methods Using an rms Fidelity Error*

The percentage-error ratio metric used above to assess the reconstruction fidelity of the single-source architecture under various recording conditions is attractive in that it graphically illustrates the variation in fidelity among the individual diffracted outputs. However, a lumped error measure is more convenient for comparing not only different recording methods for the single-source architecture but also for comparing the reconstruction fidelity of different architectures.

One possible error measure is given by[19]

$$\epsilon = |\hat{u} - \hat{u}'| = \left[\sum_{j=1}^{N}\left(\frac{\rho_j}{|\rho|} - \frac{\rho_j'}{|\rho'|}\right)^2\right]^{1/2}, \qquad (35)$$

in which $\hat{u} = \rho/|\rho|$ is the unit vector in the direction of ρ; $\hat{u}'$ is defined similarly. The error measure ϵ can be interpreted as the rms error of the components of the normalized diffracted output vector $\hat{u}'$. The separate normalizations of the ideal and actual components by $|\rho|$ and $|\rho'|$, respectively, permit a change in the throughput that does not in turn bias the relative fidelity of the components. The maximum value of the rms error ϵ is $\sqrt{2}$ since $\hat{u}$ and $\hat{u}'$ are unit vectors having unipolar components.

The rms errors of the various recording methods discussed in Sections 4.A–4.D for the 10-to-10 simulations of the single-source architecture are shown in Fig. 11 as a function of the grating strength of the largest interconnection grating. For simultaneous recording the rms error starts out relatively small at essentially zero grating strength (and practically no throughput, as shown in Fig. 6), and then becomes large quite rapidly as the grating strength is increased. If just the cross gratings among the input-plane pixels are removed using pagewise-sequential recording with $R = N^2 = 100$ (i.e., only the desired interconnection gratings and the training-plane cross gratings are present, with comparable grating strengths), neither the fidelity nor the throughput improves significantly. Decreasing the magnitudes of the training-plane cross gratings relative to the desired interconnection gratings (using pagewise-sequential recording with $R = 1000$) results in substantial improvement in both the fidelity and the peak throughput.

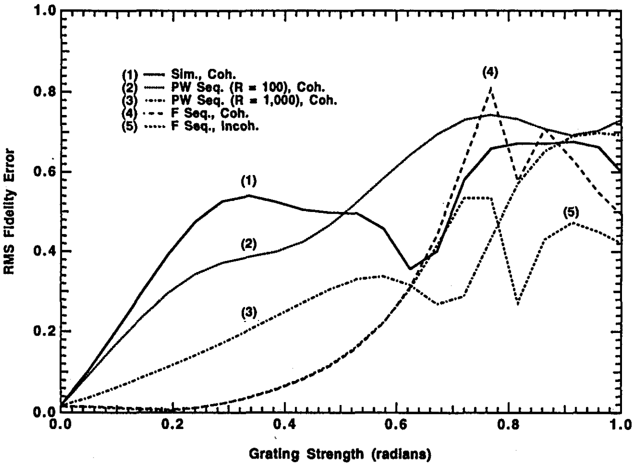

Fig. 11. Simulation results showing the rms fidelity error [as defined in Eq. (35)] for several combinations of recording and readout methods for the single-source architecture. An explanation of the legend is provided in Fig. 6.

Complete elimination of the cross gratings by using sequential recording yields a further significant improvement in fidelity and a marginal increase in the peak throughput, as shown in Fig. 6. The reconstruction fidelity that we obtain in the absence of cross gratings is much higher than that shown by Slinger.[19] This difference can be understood as follows. Slinger's analysis considers both random complex weights (i.e., each interconnection grating has a random phase as well as a random amplitude) and random-amplitude readout beams that have either a 0° or 180° relative phase. The direct implication is that the readout beams do not have the same relative phases as the recording beams that were used to create the weights [i.e., the argument of the exponential in Eq. (7) is not constant for all i and j, such that condition (1) discussed following Eq. (6) in Section 2.C is violated]. The net result is poor reconstruction fidelity.

In our analysis of the single-source interconnection architecture the readout beams have the same relative phases as the recording beams, which permits the appropriate phasing conditions to be satisfied upon reconstruction such that good fidelity results. This same-relative-phase condition is consistent with the implementation of unipolar weights and outputs, as is discussed in Section 2.C (recall that a dual-rail strategy can be generalized to the bipolar case instead of requiring the interconnection system to implement bipolarity directly). A comparison of our results with Slinger's indicates that (in the absence of cross gratings) a single-source architecture can in fact yield high reconstruction fidelity for an appropriate mapping of neural interconnection requirements to the architecture.

2. *Effects of Sidelobe Overlap on Reconstruction Fidelity*

A curious feature of the simulation results shown in Fig. 11 is that the fidelity error for each recording method does not go to zero with decreasing grating strength. Furthermore, the asymptotic value of the fidelity error near zero grating strength is the same for each recording method. This indicates that the limiting value of the error is independent of the presence or absence of cross gratings.

The cause of this somewhat counterintuitive behavior appears to be diffraction from non-Bragg-matched interconnection gratings.[48] The physical mechanism for this effect is illustrated in Fig. 12, in which the +1-order diffraction efficiency of two sinusoidal phase gratings is shown as a function of the incidence angle of a plane-wave readout beam. The gratings are assumed to have the same grating period, but their fringes are oriented at different slant angles relative to the front face of the holographic medium. If a readout beam is incident at θ_1 (Bragg matched to grating G_1), it is diffracted by both G_1 (through G_1's mainlobe response) and grating G_2 (through G_2's sidelobe response at θ_1). Since G_1 and G_2 have the same grating period, the light beams diffracted from both gratings are collinear upon exiting the holographic medium. Hence the overall diffraction response at incidence angle θ_1 is composed of a small contribution from G_2 as well as the main contribution from G_1. Upon coherent addition of the two contributions the net diffraction efficiency in general differs from what it would be in the absence of G_2. If G_1 implements a weighted interconnection, this difference results in a small error in the weight of the grating. We refer to the source of this error as sidelobe overlap.

In the single-source architecture illustrated in Figs. 3 and 4, there are multiple sets of gratings that have the same grating period but different slant angles because of the existence of multiple pairs of writing beams originating from the pixels of the input and training planes that have equal angular separations. For example, the grating written by beams x_1 and δ_1 (Fig. 4) has the same grating period (but a different slant angle) as the grating written by beams x_2 and δ_2. As a result of the presence of sidelobe overlap among various sets of gratings, one would expect the weights of the affected gratings to exhibit a small residual error upon reconstruction. This variation in the recorded weights should be present even as the grating strength goes to zero because both the main Bragg response of a particular grating and the overlapping sidelobes of the other gratings scale in proportion to each other as a function of grating strength.

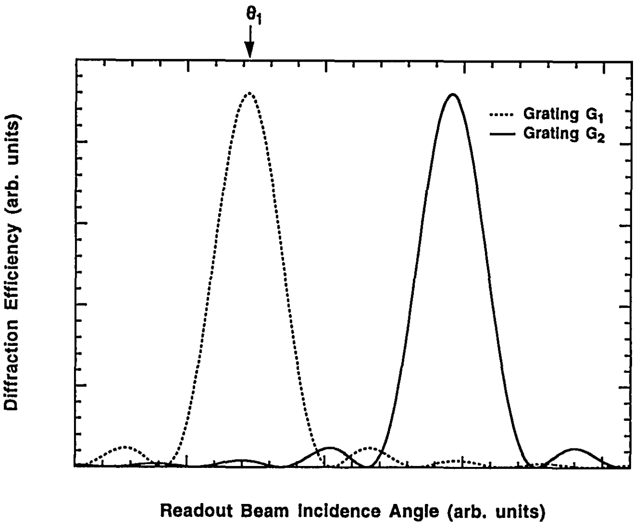

Fig. 12. Angular response characteristics of two Bragg gratings that have the same grating period and slightly different slant angles. Although the main lobes of the angular responses are well separated, the sidelobes and the main lobes overlap.

One method of testing the assertion that sidelobe overlap affects the reconstruction fidelity as described above is to vary the thickness of the holographic medium such that the degree of sidelobe overlap among the various interconnection gratings is changed (since the angular width of the diffraction response of each grating is inversely proportional to the thickness[49]). If sidelobe overlap is present, one would expect the limiting value of the rms fidelity error to generally increase with decreasing thickness, and vice versa. This behavior is exactly what we observe in our simulations.

For the simultaneous and pagewise-sequential cases shown in Fig. 11, the effect of sidelobe overlap is apparent only at small grating strengths because the cross gratings are the main source of fidelity error at larger grating strengths. For the fully sequential recording case with coherent readout, sidelobe overlap appears to be the dominant source of reconstruction error up to ~0.25 rad. As borne out by other simulation results, the reconstruction fidelity in this regime can be further improved simply by increasing the thickness of the holographic medium.

3. *Scaling Trends*

To examine how our modeling results scale with the number of nodes in the interconnection system, we also simulated a 4-to-4 single-source interconnection architecture having a 4 × 4 weight matrix that is a subset of the 10 × 10 weight matrix used above. The corresponding four input beams from $\mathbf{x}^{\mathrm{read}}$ are used to read out the interconnection system. Simulation results for the rms fidelity error of the normalized output vectors and the throughput for the 4-to-4 case are shown in Figs. 13(a) and 13(b), respectively, for the simultaneous, pagewise-sequential, and fully sequential recording methods.

The overall performance characteristics noticed in the 10-to-10 simulations are also present in the 4-to-4 results, in that the rms fidelity errors and throughputs for the various cases show no identifiable trends for scale-up from the 4-to-4 case to the 10-to-10 case. As a specific example, the simulation results for the relative behaviors of the pagewise-sequential and simultaneous recording methods in the 4-to-4 case are similar to the 10-to-10 results (see Fig. 13). The rms fidelity error for the simultaneous recording method in both cases increases rapidly as the grating strength increases from zero. When the input-plane cross gratings are removed (i.e., when pagewise-sequential recording is used with $R = 16$ for the 4-to-4 case and $R = 100$ for the 10-to-10 case), neither the fidelity nor the throughput improves significantly

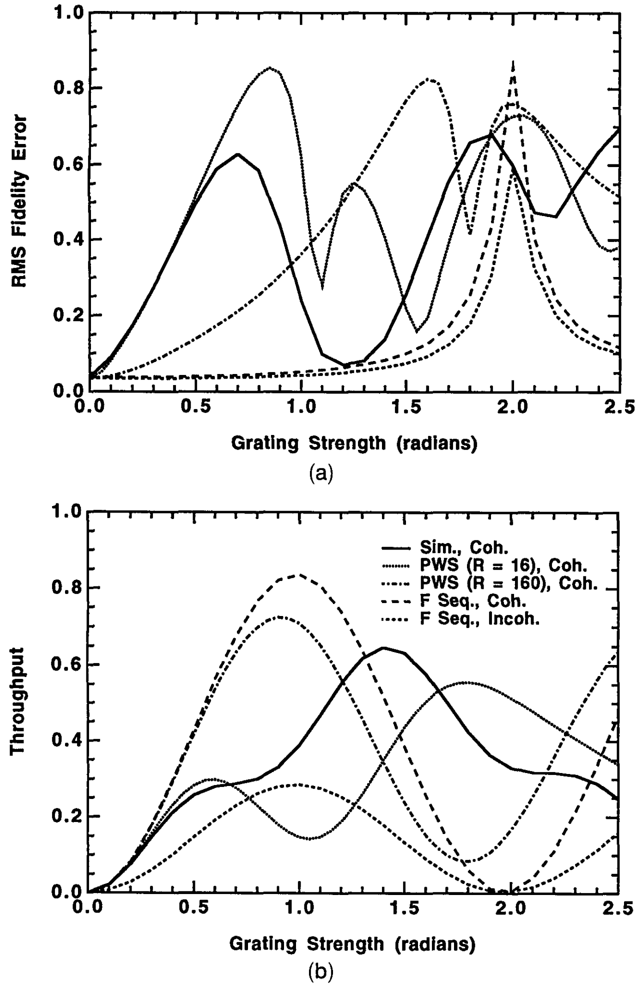

Fig. 13. Simulation results for a 4-to-4 single-source architecture. Shown are (a) the rms fidelity error for various recording and readout combinations and (b) the optical throughput for the various recording and readout combinations. Clarification of the legend [in (b), from top to bottom] is as follows: simultaneous recording method with coherent readout, pagewise-sequential recording method with a beam splitter ratio of 16 and coherent readout, pagewise-sequential recording method with a beam splitter ratio of 160 and coherent readout, fully sequential recording with coherent readout, and fully sequential recording with incoherent readout.

in either case. Increasing the strength of the desired interconnections to approximately three times the training-plane cross gratings (i.e., when pagewise-sequential recording is used with $R = 160$ for the 4-to-4 case) reduces the fidelity error and significantly increases the peak throughput, just as in the 10-to-10 case. In fact, the rms fidelity error is nearly the same for both simulations at a grating strength corresponding to the peak throughput in each case. At the current level of simulation complexity it is not yet clear whether this apparent insensitivity to the dimensions of the interconnection system is generalizable.

In addition to these similarities, there are a few interesting differences among these particular 4-to-4 and 10-to-10 cases. For simultaneous recording the 4-to-4 case shows a significant drop in the rms fidelity error at ~ 1.2 rad (which happens to correspond to 50% throughput). This result is weight-matrix and input-vector dependent and cannot be relied on to occur in general. Similar comments apply to the observed drop in fidelity error for the $R = 16$ pagewise-sequential recording case that occurs at grating strengths in excess of 0.85 rad. Despite these differences, the general trends do not appear to be weight-matrix and input-vector dependent.

Furthermore, as the grating strength goes to zero the fidelity error for the 4-to-4 case [Fig. 13(a)] is the same for all of the recording methods, as in the 10-to-10 cases. However, the actual value of the error is somewhat smaller for the 10-to-10 results than for the 4-to-4 simulations. This difference is probably caused by the increased number of overlapping sidelobes that affect a particular weighted interconnection in the 10-to-10 case. As the number of overlapping sidelobes increases, their effect on the interconnection fidelity may tend to decrease because the sign of each sidelobe's contribution to the weighted interconnection can be either positive or negative, depending on which particular sidelobe is accessed for the off-Bragg grating. The contributions of a large number of sidelobes may therefore tend to average to zero. The implication is that the limiting fidelity error that is due to sidelobe overlap should decrease as the number of nodes in the interconnection architecture increases (at least until some other limiting phenomenon is reached). This should result in better fidelity performance for the sequential recording method for throughputs at which the fidelity is limited by sidelobe overlap.

F. Single-Source Architecture Simulation Results: Conclusions

Our simulation results show that, as expected, the coherent-recording cross-talk gratings for simultaneous recording in a single-source architecture cause a significant degradation in reconstruction fidelity for reasonable throughputs. Pagewise-sequential recording shows better fidelity and peak-throughput performance than simultaneous recording if the beam splitter ratio can be made large enough to overcome the undesired strengthening of the training-plane cross gratings, which is caused by the larger number of exposures that they receive relative to the desired interconnection gratings. Sequential recording yields both high reconstruction fidelity and high optical throughput, but at the cost of $N^2 - 1$ more recording steps than for the case of simultaneous recording.

Based on the considerations discussed in this paper, an attractive method of implementing a single-source architecture is to employ simultaneous recording in a geometry that clearly separates the range of spatial frequencies obtained for the desired interconnection gratings from those obtained for the cross gratings; and to use a holographic material that is sensitive to the former range of spatial frequencies and insensitive to the latter.[17,19] Of course this assumes that high throughput is required in the resultant interconnection system. No special requirements are placed on the spatial-frequency sensitivity of the holographic medium if low throughputs are permissible. In this regime the simultaneous recording method yields the same fidelity performance as the other recording methods, all of which are limited by the amount of angular sidelobe overlap that is present.

Even if a single-source architecture is implemented as described above, it has at least one remaining characteristic that detracts from its potential use in implementations of large-scale weighted fan-out/fan-in interconnection systems. As mentioned in Section 2.E, the density of pixels on the input and training planes is limited to a certain degree by grating degeneracy.[33,34]

5. Incoherent/Coherent Double Angularly Multiplexed Interconnection Architecture

In this section we discuss the operation of the incoherent/coherent double angularly multiplexed architecture that we have proposed and investigated recently[8-14] and compare its relative merits to those of single-source architectures. Two configurations of the incoherent/coherent double angularly multiplexed architecture are presented: first, the full-aperture configuration[8-14] and second, the subhologram configuration.[11,12,14,50] Both configurations permit simultaneous recording of each training pair with significantly reduced coherent-recording cross talk as compared with the single-source architecture. In addition, readout is performed with mutually incoherent beams such that, during operation, each diffracted output is described by the usual neural-network summation of Eq. (4) rather than the modified summation of Eq. (6). This is accomplished without sacrificing the high throughput efficiency that is typically associated only with fully coherent systems, because both configurations avoid the presence of beam degeneracy. Furthermore, while grating degeneracy is present in the full-aperture configuration of the architecture, it can potentially be eliminated by using the subhologram configuration and therefore does not require subsampling of the input and training planes.

The full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture is described in Section 5.A, and simulation results are discussed in Section 5.B. Operation of the subhologram configuration of the architecture is discussed in Section 5.C, and simulation results follow in Section 5.D. Section 5.E provides a comparison of the two configurations.

A. Full-Aperture Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Operation

A schematic diagram for one layer of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture is shown in Fig. 14(a). The architecture has three key components: (1) a 2-D array of individually coherent but mutually incoherent sources; (2) optoelectronic neuron-unit arrays that integrate the functions of light detection, neuron-unit nonlinear response, and optical modulation for each pixel (these arrays are denoted as SLM_1 and SLM_2); and (3) the volume holographic interconnection medium. In this section, we briefly review the interconnection method used in the architecture. Additional details are provided in Refs. 12 and 13.

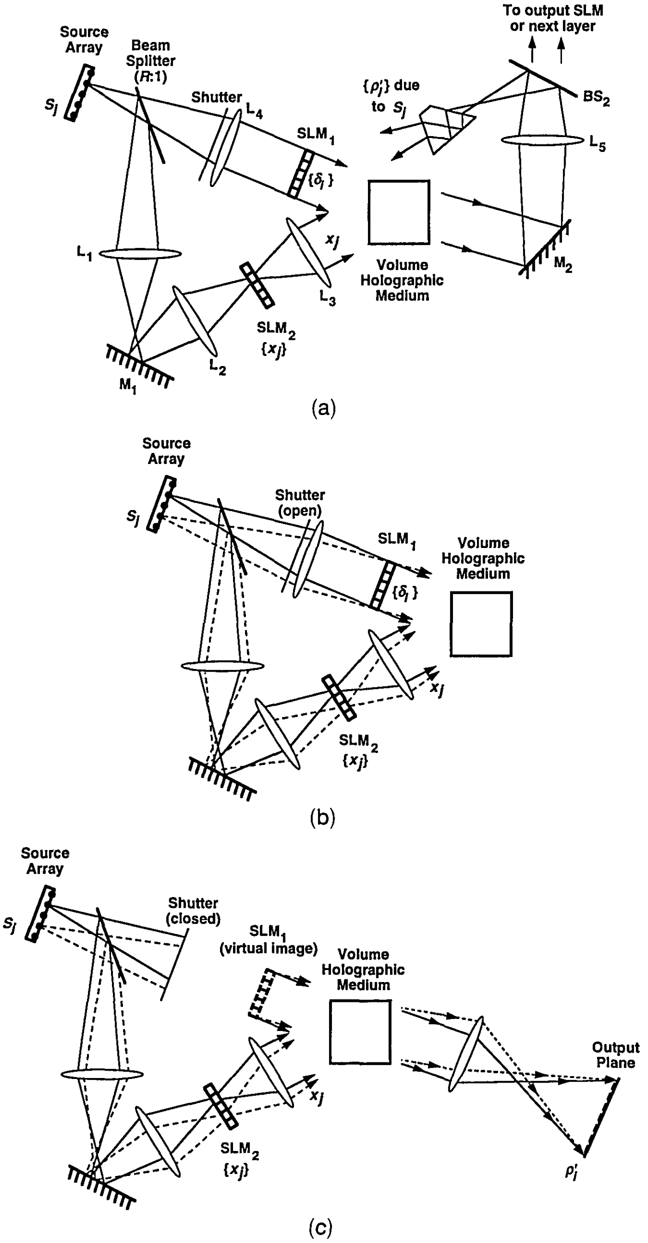

Fig. 14. Schematic diagram of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture showing (a) general layout, (b) recording, and (c) reconstruction. M_1 and M_2 are mirrors; L_1–L_5 are lenses; BS_2 is a second beam splitter.

The process of recording a set of interconnections is illustrated in Fig. 14(b). Assume for the moment that a single pair of training vectors $[\mathbf{x}^{(m)}, \boldsymbol{\delta}^{(m)}]$ is to be recorded. Light from each source S_j is split into two optical paths. In the upper path each beam is collimated and illuminates the full aperture of the training plane (SLM_1). Each beam therefore reads out the training vector $\boldsymbol{\delta}^{(m)}$ that is present on SLM_1 (each of these beams is referred to herein as a $\boldsymbol{\delta}^{(m)}$ beam). The collimated beams deriving from the full set of sources propagate at different angles through the training plane. In the lower path the source array is imaged onto the input plane (SLM_2) such that light from each source illuminates only one corresponding pixel. A collimating lens converts the beams emerging from the pixels of SLM_2 to a set of angularly distinct collimated beams that illuminate the full aperture of the volume holographic medium.

As a result of the individually coherent but mutually incoherent nature of the sources in the source array, interconnection gratings are formed only between each $x_j^{(m)}$ and the corresponding $\boldsymbol{\delta}^{(m)}$ beam with which it is mutually coherent. This permits the simultaneous recording of a set of angularly multiplexed holograms in which each hologram is formed by the interference of an angularly distinct reference beam $x_j^{(m)}$ with a second angularly distinct beam (from source S_j) bearing the image $\boldsymbol{\delta}^{(m)}$. The full set of image-bearing beams $\{[\boldsymbol{\delta}^{(m)}]_j\}$ that derives from all of the sources in the source array $\{S_j\}$ and that encodes the contents of SLM_1 is also angularly multiplexed. Hence we describe this architecture as *double angularly multiplexed.*[13]

Similar to the case of simultaneous recording in a single-source architecture, the incoherent/coherent double angularly multiplexed architecture requires only one exposure to record each training pair $[\mathbf{x}^{(m)}, \boldsymbol{\delta}^{(m)}]$, which is accomplished by turning on all of the sources in the source array simultaneously. Since mutually incoherent beams are used to read out the pixels of SLM_2, no coherent-recording cross-talk gratings among the input-plane pixels are formed. Similarly, cross gratings among the separate beams encoded with $\boldsymbol{\delta}^{(m)}$ do not occur. The only cross gratings that can form in the holographic medium result from overlaps among adjacent diffracted components δ_i within each $\boldsymbol{\delta}^{(m)}$ beam in the Fresnel regime. Depending on the size of the pixels and the distance between SLM_1 and the holographic medium, these cross gratings connect any single pixel only to those in some local neighborhood of the pixel. The effects of such local cross talk can be minimized by adjusting the beam splitter ratio R. This interconnection system therefore permits simultaneous recording of each training pair while minimizing the effects of coherent-recording cross talk.

As illustrated in Fig. 14(c), readout is performed using the lower optical path (with all of the sources turned on simultaneously). The volume holographic optical element, or VHOE, performs the requisite set of weighted fan-outs, while the imaging lens following the VHOE performs an optical fan-in operation by imaging the diffracted beams onto the pixels of the output plane. The angularly distinct set of collimated beams that illuminated SLM_1 during recording is therefore reconstructed by the VHOE; after passing through the lens, the beams form a real image in the output plane, which is conjugate to the SLM_1 plane. The net result is that a fan-in of angularly distinct incoherent beams is performed at each node in the output plane. As long as the angular spread is sufficiently large, an incoherent fan-in can be performed without incurring the usual fan-in loss associated with a collinear incoherent fan-in. If we use appropriate optical elements (depending on the particular neural-network model being implemented), the output plane shown in Fig. 14(c) may be coincident with the input side of SLM_1 itself [as shown in Fig. 14(a)], with SLM_2, with the input SLM of the next layer, or with any combination thereof.

B. Full-Aperture Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Simulation

The optical beam propagation method was used to analyze the fidelity and throughput performance of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture by simulating a 10-to-10 interconnection system in which the same weight matrix was recorded as above for the single-source architecture and in which the same input vector $\mathbf{x}^{\mathrm{read}}$ was used for readout. In all such simulations, readout was performed using mutually incoherent beams.

To facilitate direct comparison with the single-source architecture simulations, we assume that the same holographic medium characteristics (linear material with a thickness of 4.5 mm and a refractive index of 2.52) and the same operating wavelength (0.514 μm) are used. The separation of the sources in the source array is the same as the pixel separation in SLM_1 and SLM_2, which is 257 μm. Instead of using two lenses to image the source array onto SLM_2 [lenses L_1 and L_2 in Fig. 14(a)], we use a single lens only, which is separated by twice its focal length from both the source array and SLM_2. The focal lengths of lenses L_4 and L_3 are assumed to be 50 mm. The separation between SLM_1 and the holographic medium is also assumed to be 50 mm. The beam splitter ratio R is unity.

Given the above parameters, light passing through each pixel of SLM_1 spreads approximately 200 μm laterally in propagating to the holographic medium because of diffraction. Since this is less than one pixel width, we assume for simplicity that geometrical optics adequately describes the propagation of light from SLM_1 to the holographic medium. Our model thus does not consider the effects of any diffraction-induced local-neighborhood cross gratings that may be present. Instead, each beam that reads out SLM_1 [referred to as a $\boldsymbol{\delta}^{(m)}$ beam above] is assumed to propagate essentially unchanged to the holographic medium, forming an exact (rather than an approximate) image of SLM_1 on its front face. The interference of each $\boldsymbol{\delta}^{(m)}$ beam with light from its associated $x_j^{(m)}$ pixel in SLM_2 therefore results in ten (for a 10-to-10 interconnection) distinct grating regions in the holographic medium. In each distinct region the resulting interconnection grating pattern has a different amplitude, proportional to $[x_j^{(m)}\delta_i^{(m)}]^{1/2}$ for a single training pair, while the grating periods and slant angles are the same for all ten regions that derive from source S_j.

For the parameters assumed above for the source and pixel separations, for the focal length of L_4, and for the separation between SLM$_1$ and the holographic medium, each $\delta^{(m)}$ beam is shifted by one pixel from angularly adjacent $\delta^{(m)}$ beams on the face of the hologram. The net result is 19 separate regions in the holographic medium, with between 1 and 10 gratings multiplexed in each region. The 2-D formulation of the BPM discussed in Section 3.3 was used to simulate readout of each distinct region.

Simulation results for the diffracted outputs of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture are shown in Fig. 15 as a function of the grating strength of the largest interconnection grating. The rms fidelity error and throughput are shown in Fig. 16; despite readout by a set of mutually incoherent beams, the peak throughput is nearly 95%.

As in our fidelity analysis of the single-source architecture, the functional dependence of each weight on its corresponding grating strength determines the ideal input/output proportionality factor that must be known for comparison with the actual input/output characteristics of the system. For holographic interconnection systems the ideal input/output relationship is based on the underlying physics of the diffraction process used in the system.

For example, direct application of the weight relationship expressed in Rel. (29) for a single-source architecture to the fidelity analysis of the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture suggests the presence of significant fidelity errors, as illustrated in Figs. 15(c) and 16(a). Although the percentage errors in Fig. 15(c) are small for low throughput, they become quite large (up to 85%) at the peak throughput. Similar behavior is observed for the rms fidelity error in Fig. 16(a) (solid curve).

The reason for the apparent lack of fidelity is that Rel. (29) does not adequately describe the physics behind the diffraction process used in the full-aperture configuration. In this case each grating in a given region of the holographic material is completely independent of the other interconnection gratings in that region, except for effects such as angular sidelobe overlap. We therefore assume that the diffraction efficiency of any particular grating is given by $\sin^2(\nu_{ij}/2)$, in which ν_{ij} is the strength of this grating.[49] On this basis the corresponding weight relationship for the full-aperture configuration of the architecture (for multiple training pairs) is

$$W_{ij} \propto \sin^2[\nu_{ij}^{(M)}/2], \qquad (36)$$

in which $\nu_{ij}^{(M)}$ is given by Eq. (24).

Comparison of the diffracted outputs obtained from the BPM simulations to the ideal outputs calculated using Rel. (36) [as shown in Figs. 15(d) and 16(a)] yields much better measured fidelity performance. The actual diffracted outputs from the holographic medium are of course not changed; rather, the metric against which they are compared is related more closely to the underlying diffraction behavior of the interconnection system. The weight definition of

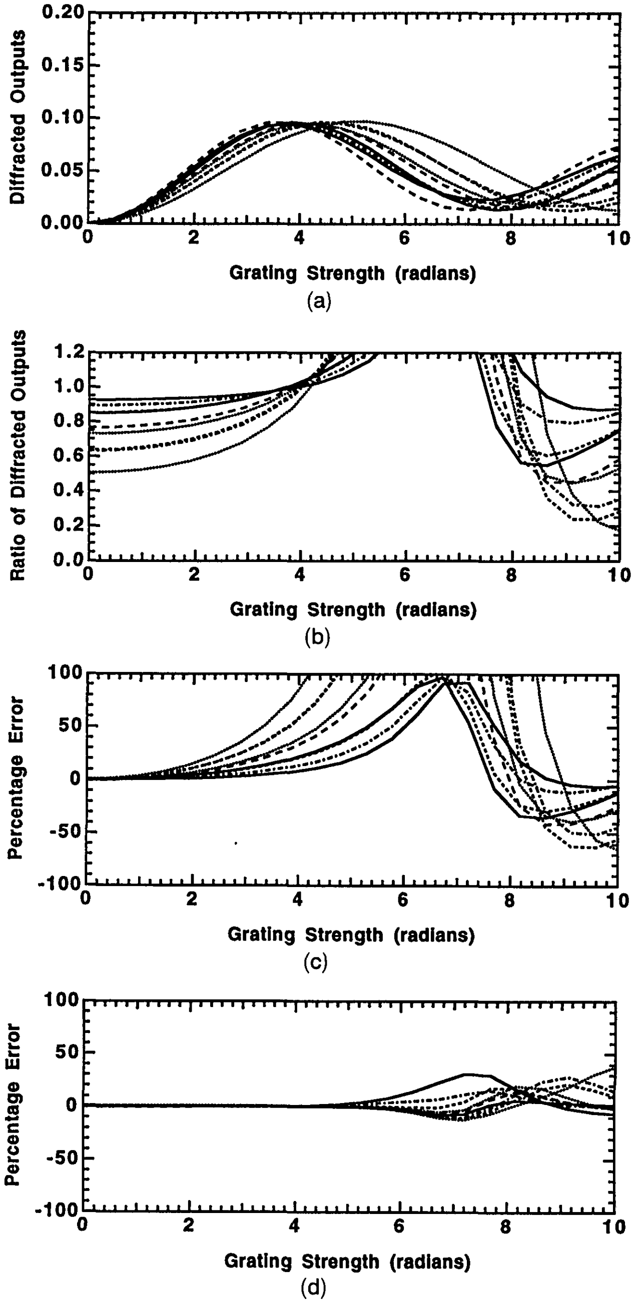

Fig. 15. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (full-aperture configuration) for readout with mutually incoherent beams. Shown as functions of the grating strength of the largest grating are (a) the diffracted outputs, (b) the ratios of the diffracted outputs, (c) the percentage error using Rel. (28) for the dependence of each weight on grating strength, and (d) the percentage error using Rel. (36) for the dependence of each weight on grating strength.

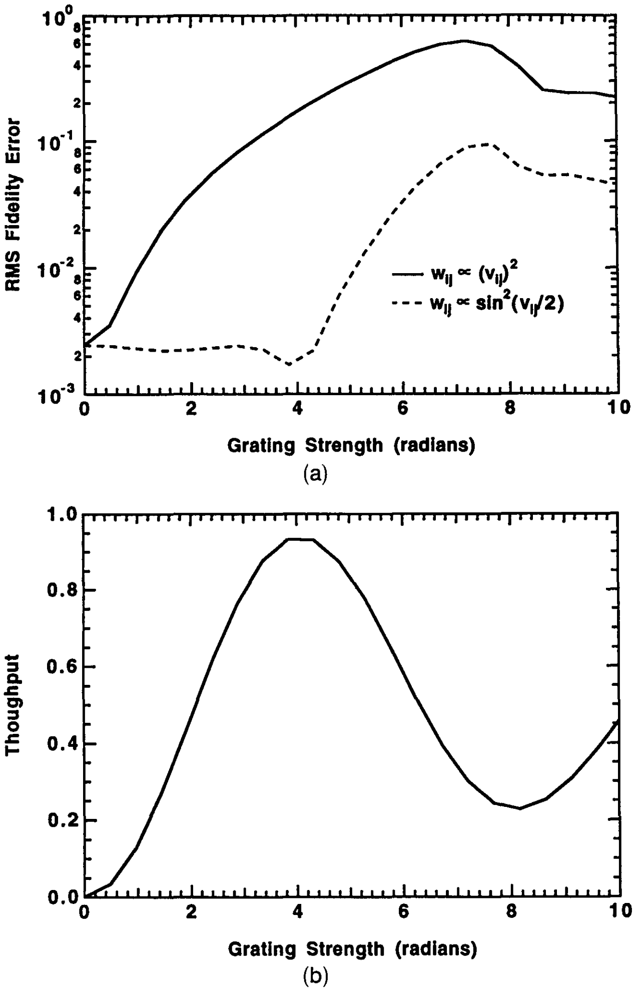

Fig. 16. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (full-aperture configuration). Shown are (a) the rms fidelity error for two different functional dependencies of the weights on the grating strength and (b) the optical throughput.

Rel. (36) as compared with that of Rel. (29) implies a different functional form for the weight-update relationship in an adaptive system. The effect of this altered functional form on the performance of learning algorithms is currently unknown. We conjecture that, given a physical implementation with limited dynamic range for each weight, the soft-limiting characteristic provided by the $\sin^2$ function may prove to be in some respects preferable to a hard-clipping saturation characteristic.

As shown in Fig. 15(d), the percentage errors of the ratios of the diffracted outputs obtained using the $\sin^2$ weight relationship are very close to zero for throughputs up to and including the peak throughput of 95%. This result shows that the principal source of error in this configuration indeed derives from the metric rather than from some source of cross talk. The comparison of the two metrics is illustrated clearly in Fig. 16(a), in which the rms fidelity errors are plotted on a log scale. The rms fidelity error for the $\sin^2$ metric is relatively flat between 0 and 4 rad of grating strength (at which the peak throughput occurs) and does not go to zero as $\nu \rightarrow 0$. The principal reason for the nonzero fidelity error throughout this region appears to be overlap of the angular sidelobes of the gratings multiplexed in each region of the hologram. Increasing the number of interconnection nodes in the incoherent/coherent double angularly multiplexed architecture should lower this level of error by averaging out the contributions of an increased number of sidelobes, as discussed in Section 4.E.3.

For the 10-to-10 interconnection system analyzed, the rms fidelity error shown in Fig. 16(a) with a weight relationship given by Rel. (36) is dramatically smaller than for the sequentially recorded single-source architecture using either mutually coherent or incoherent readout beams. For example, at the grating strengths corresponding to peak throughput, the error for the incoherent/coherent double angularly multiplexed architecture is more than an order of magnitude lower than that for the sequentially recorded case of the single-source architecture.

For completeness we applied the weight relationship of Rel. (36) to the fidelity analysis of the single-source architecture and found that the fidelity errors were essentially the same as shown in Fig. 11. This result is not surprising, because in the single-source architecture the grating strength of each individual grating is small (0.4 rad for the largest interconnection grating at the peak throughput), and Rel. (36) reduces to Rel. (29) for small $\nu_{ij}^{(M)}$.

In certain volume holographic media, such as photorefractive crystals, the presence of many overlapping incoherent beams in the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture results in a small modulation depth for each pair of recording beams, which in turn significantly reduces the achievable optical throughput.[12] This problem may be avoided by using the subhologram configuration of the incoherent–coherent double angularly multiplexed architecture, which is discussed in Section 5.C.

Furthermore, as mentioned in the beginning of Section 5, the full-aperture configuration of the incoherent–coherent double angularly multiplexed architecture is additionally subject to the effects of grating degeneracy. By modifying the architecture slightly to realize the subhologram configuration, we can potentially avoid this source of cross talk without using fractal sampling grids. The trade-off, however, is permitting the presence of additional cross gratings. As discussed below, these can in turn be minimized by adjusting the beam splitter ratio.

C. Subhologram Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Operation

As shown in Fig. 17, the subhologram configuration of the double angularly multiplexed architecture can be created by inserting an additional lens (L_6) between SLM_1 and the holographic medium. The lens

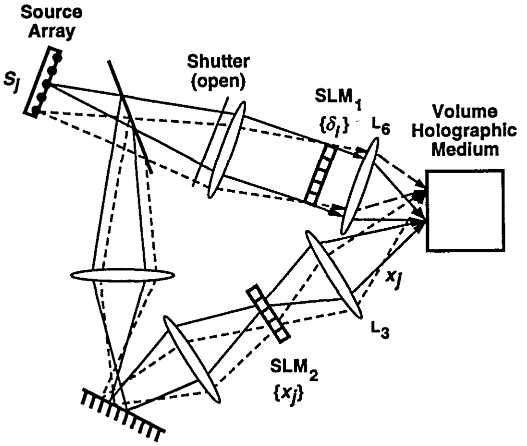

Fig. 17. Schematic diagram of the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture.

is positioned at a distance of one focal length from the medium such that it performs a Fourier transform of the beams emerging from SLM$_1$ (this corresponds to reimaging the source array onto the medium). In addition, lens L$_3$ is adjusted to image SLM$_2$ onto the medium, which also effectively reimages the source array instead of collimating the light from each pixel in SLM$_2$ as in the full-aperture configuration of the architecture.

The optical system is designed such that the images of the source array through the upper and lower paths of the architecture are in registry at the holographic medium so that spatially distinct holograms (i.e., subholograms) are formed across the face of the medium (as shown schematically in Fig. 18). To understand the nature of each subhologram, let us focus on only one source S_j that is imaged onto the hologram through both paths. The image of S_j through the upper path contains the Fourier transform of the image of $\delta^{(m)}$ that is on SLM$_1$, while the image of S_j through the lower path has an intensity proportional to $x_j^{(m)}$. The interference between the two beams creates weighted interconnections between the jth pixel in the input plane and all of the training-plane pixels. However, interference among the components of the Fourier transform of $\delta^{(m)}$ causes cross gratings that form intraplanar connections among the training-plane pixels. By adjusting the beam splitter ratio R, we can decrease the magnitude of these cross gratings relative to the desired interconnection gratings.

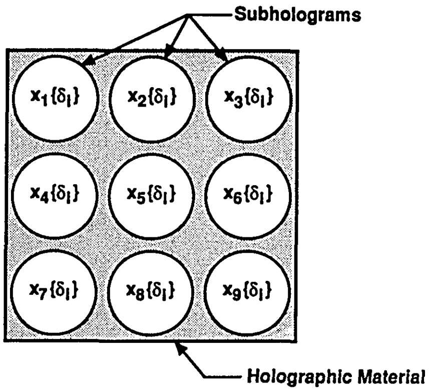

Fig. 18. Schematic diagram of a subhologram array. Each subhologram is shown as spatially separate in this case.

Each of the subholograms connects a single pixel in the input plane to all of the pixels in the training plane and thus performs a 1-to-N weighted fan-out upon reconstruction. As in the full-aperture configuration of the architecture, an imaging lens is used after the holographic medium [shown as L$_5$ in Fig. 14(a)] to perform the fan-in to each node in the output plane. The subholograms in general will at least partially overlap within the volume holographic medium, depending on the focal lengths of the lenses and the spacings of the pixels and of the sources. However, this spatial overlap does not cause additional cross gratings to form during simultaneous exposure of the set of subholograms because the sources are mutually incoherent. During recording, all of the sources are turned on simultaneously such that the recording of M training pairs requires only M exposures, just as in the full-aperture configuration of the architecture. Full illumination of both SLM apertures is accomplished with the entire source array on, which provides efficient power transfer to the holographic medium during each exposure.

In Section 5.D we discuss simulation results and scaling trends for the subhologram configuration of the double angularly multiplexed architecture.

D. Subhologram Configuration of the Incoherent/Coherent Double Angularly Multiplexed Architecture: Simulation

In order to determine the relationship between an individual weight and the strength of its associated grating within the subhologram configuration of the incoherent–coherent double angularly multiplexed architecture, we use the fact that each spatially segregated subhologram implements an independent 1-to-N fan-out. As discussed in Section 3.B, an analytical solution for the diffraction efficiency of a 1-to-N weighted fan-out has been obtained using coupled-wave theory under the assumption that no cross gratings are present.[38] The net result is that the weights and grating strengths for the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture are related by (in our notation)

$$W_{ij} \propto [\nu_{ij}^{(M)}]^2, \qquad (37)$$

which is the same as Rel. (29) for the single-source architecture. The fidelity analysis presented in this section is based on the use of Rel. (37) to compute the ideal input/output characteristics of the subhologram configuration.

For the simulations discussed in this section, the weight matrix and the readout vector used are the same as those described in Section 3.C.3. Also, the parameters of the optical components are chosen to be the same as for the full-aperture configuration of the architecture, except that the focal length of L_4 is 25 mm, that of L_3 is 100 mm, and that of L_6 is 50 mm. For the purposes of our modeling, the resultant subholograms (~ 500 μm in diameter, set on 500-μm centers) are considered to be fully separated, and the pixels in the training plane are treated as point sources. The optical beam propagation method was used to model the readout of each subhologram.

Simulation results are shown in Figs. 19 and 20. In Fig. 19 the individual diffracted outputs are computed for a beam splitter ratio R of 100. The horizontal axis, as above, refers to the grating strength of the largest interconnection grating in the holographic medium. The ratio percentage errors in Fig. 19(c) are not much different than those shown in Fig. 8(c) for sequential recording in the single-source architecture. For example, at the peak throughput of over 95% [as shown in Fig. 20(b)] the largest ratio error for the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture (with R = 100) is 25%; at 10% throughput, the largest ratio error is only 4%.

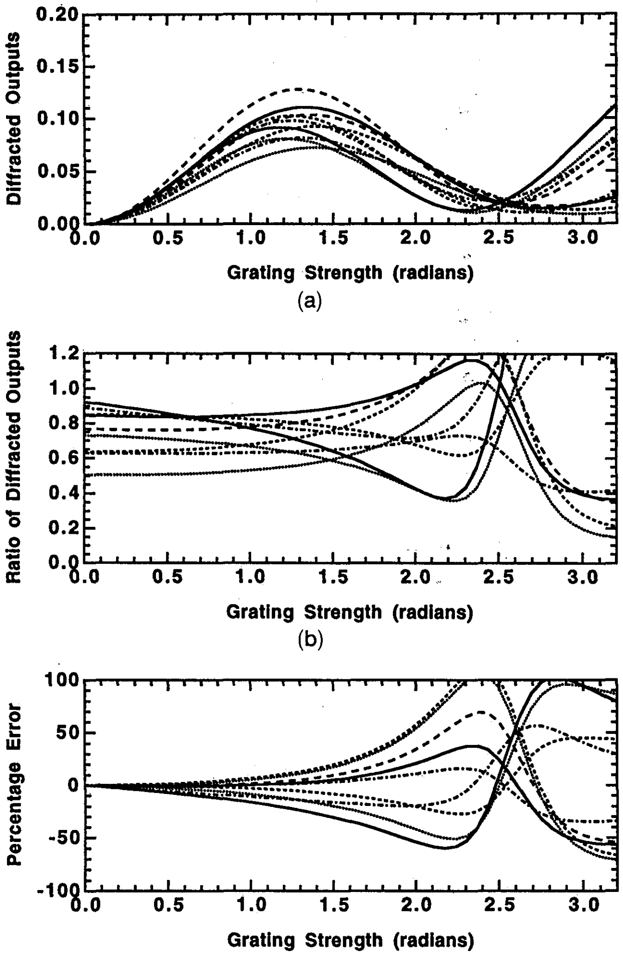

Fig. 19. Simulation results for the 10-to-10 incoherent/coherent double angularly multiplexed architecture (subhologram configuration) for readout with mutually incoherent beams. Shown as functions of the grating strength of the largest grating are (a) the diffracted outputs, (b) the ratios of the diffracted outputs, and (c) the percentage error of each ratio.

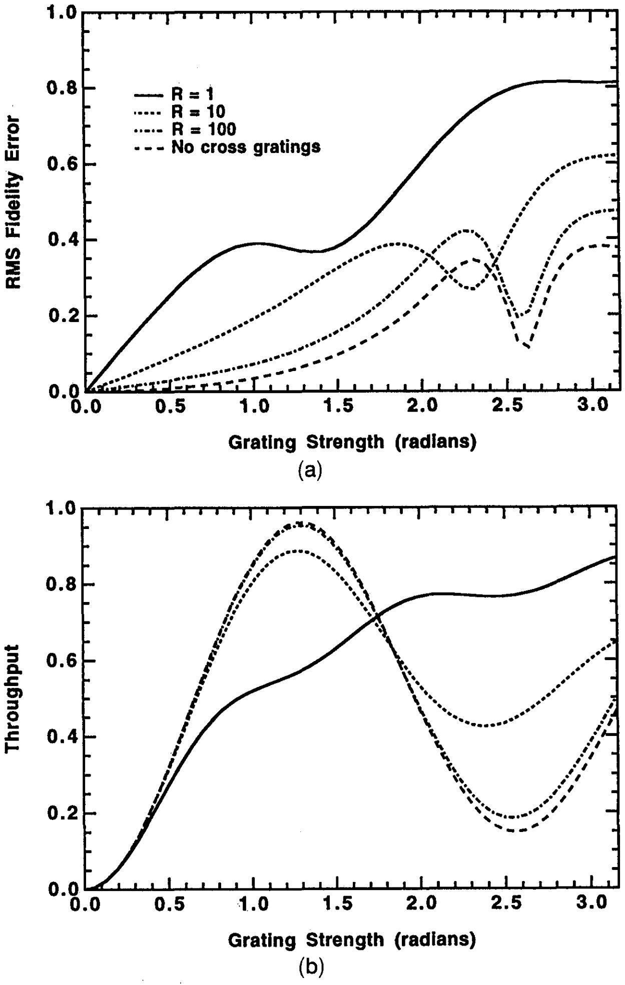

Fig. 20. (a) The rms fidelity error and (b) the throughput for various beam splitter ratios in the 10-to-10 incoherent–coherent double angularly multiplexed architecture (subhologram configuration). When R = 100, the fidelity error and throughput approach the case for which there are no cross gratings. Readout is performed with mutually incoherent beams.

The rms fidelity error for the subhologram configuration is shown in Fig. 20(a) for several values of the beam splitter ratio, and the corresponding throughputs are shown in Fig. 20(b). When the beam splitter ratio is unity (i.e., when the cross gratings have the same relative amplitudes as the desired interconnection gratings), the cross gratings cause large fidelity errors except in the limit of low throughput (similar to the results obtained for the single-source architecture). As the relative magnitudes of the cross gratings decrease with increasing beam splitter ratios, the fidelity and throughput both improve. For comparison, a case is also shown for which no cross gratings are present. In all cases, the fidelity error asymptotically approaches zero with decreasing grating strength because angular sidelobe overlap that might affect the desired interconnections does not occur in the subhologram configuration.

In order to examine how these results scale with the number of interconnection nodes in the training and input planes, we simulated a 4-to-4 interconnection system using the same weight matrix and readout vector as in the single-source architecture 4-to-4 simulations. The fidelity error and throughput results are shown in Fig. 21 for the same beam splitter ratios as shown in Fig. 20. For $R = 100$ neither the fidelity nor the throughput seems to differ significantly between the 10-to-10 and the 4-to-4 results. For the case $R = 1$, however, substantial improvement in both fidelity and throughput is observed in scaling up to the 10-to-10 interconnection system from the 4-to-4 system (except at small grating strengths).

E. Discussion of the Incoherent/Coherent Double Angularly Multiplexed Architecture Configurations

Our simulations demonstrate the ability of the incoherent/coherent double angularly multiplexed architecture to obtain high optical throughput (at least for linear holographic materials) when mutually incoherent beams are used during readout of the holographic interconnections. By avoiding beam degeneracy in both configurations of the architecture, we can circumvent the usual incoherent fan-in loss found in the single-source architecture.

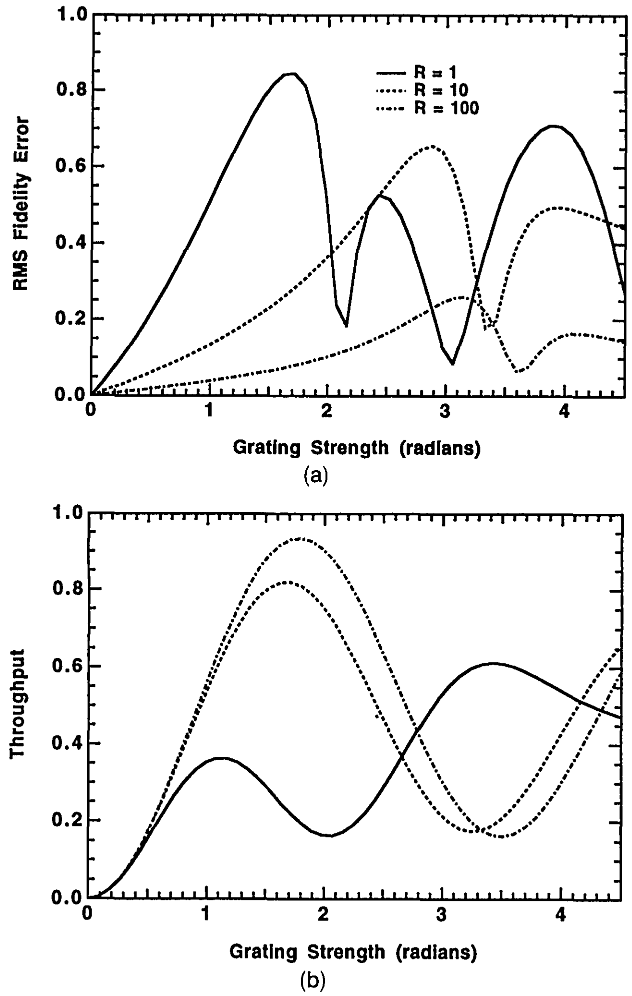

Fig. 21. (a) The rms fidelity error and (b) the throughput for various beam splitter ratios in the 4-to-4 incoherent/coherent double angularly multiplexed architecture (subhologram configuration). Comparison with Fig. 20 indicates how fidelity and throughput variations scale with the number of interconnection nodes.

Our simulation results (presented in Section 5.B) further demonstrate that high reconstruction fidelity is achievable in the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture for a fidelity metric based on the diffraction properties of its interconnection system. As noted above, grating degeneracy is present in the full-aperture configuration such that fractal sampling grids may be required in certain applications. By contrast, the subhologram configuration of the architecture avoids the presence of grating degeneracy, which may permit an increased interconnection density for a given physical system volume relative to both the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture and the single-source architecture. However, the subhologram configuration involves a fundamental trade-off between reconstruction fidelity and the beam splitter ratio resulting from the presence of cross gratings that are not present in the full-aperture configuration.

An important aspect of the subhologram configuration is the incorporation of both spatial and angular multiplexing in the holographic medium to obtain independence of the interconnection gratings. In the limiting case of complete spatial separation of the subholograms, only spatial multiplexing is used. In this case, a thin holographic material could in principle be used in the interconnection system. However, since large numbers of interconnections (10^8–10^{10}) are anticipated for photonic neural networks, space–bandwidth limitations will in general necessitate some degree of subhologram overlap in order for compact system implementations to be realized. In this case, the independence of the interconnection gratings with nonnegligible subhologram overlap necessitates angular (or wavelength) multiplexing to achieve Bragg isolation, which in turn requires the use of a thick holographic medium. The set of overlapping subhologram configurations spans the continuum between the full-aperture and full-subhologram configurations and as such may yield an optimum compromise between these two extremes. In fact, the optimal degree of subhologram overlap may well prove to be material dependent.

6. Comparison of Holographic Interconnection Techniques

In this section we compare and contrast the full-aperture and subhologram configurations of the incoherent/coherent double angularly multiplexed architecture with the single-source architecture configured using different recording methods.

The fidelity and throughput performance of the single-source architecture and of both configurations of the incoherent/coherent double angularly multiplexed architecture are summarized in Fig. 22, in which the rms fidelity errors for the 10-to-10 simulations are shown as a function of optical throughput instead of grating strength. For each curve in the figure the right-hand end point represents the peak throughput achieved in the simulation for the particular interconnection architecture and recording method to which that curve corresponds.

As a result of the effects of cross gratings, use of the simultaneous recording method in the single-source architecture yields both poor reconstruction fidelity and a peak throughput of 50% (for the case simulated). For a beam splitter ratio of 100 (the same as shown for the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture) the pagewise-sequential recording method does not yield significant performance improvement. As mentioned in Section 4.B, a serious drawback for the use of this recording method is that the beam splitter ratio required for a given level of fidelity error increases quadratically with the number of nodes in the interconnection system. When all of the cross gratings are eliminated in the single-source architecture by using a sequential recording technique, both high throughput and good reconstruction fidelity are achievable.

As the pagewise-sequential and fully sequential recording methods within the single-source architecture require a significantly larger number of exposure steps per training pair and greater hardware complexity than the simultaneous recording method, they are potentially less attractive options for implementation of large-scale adaptive neural-network systems. Alternatively, simultaneous recording in a single-source architecture suffers from a lack of reconstruction fidelity for significant optical throughput. If low throughput is tolerable in a given computational architecture, or if the effects of the cross gratings can be minimized by using the spatial-frequency-sensitive properties of a particular holographic material, a single-source architecture with mutually coherent readout beams and an appropriate fractal sampling grid becomes a viable option.

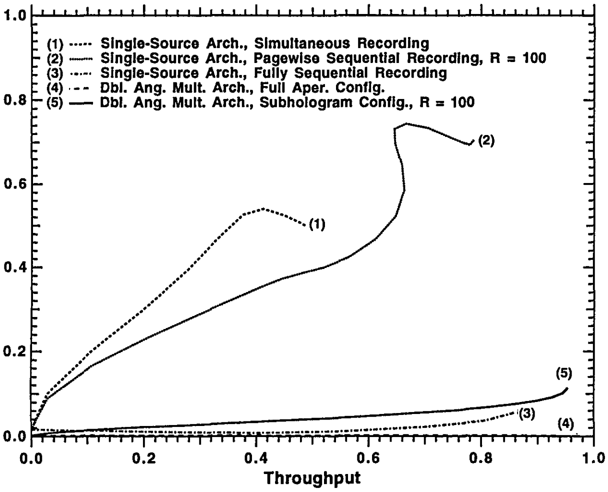

Fig. 22. Rms error as a function of the throughput for the single-source interconnection architecture (parameterized by recording method) and for the two configurations of the incoherent/coherent double angularly multiplexed architecture. In all cases, the single-source architecture is read out with mutually coherent beams and the incoherent/coherent double angularly multiplexed architecture is read out with mutually incoherent beams. The curve for the full-aperture configuration of the incoherent/coherent double angularly multiplexed architecture lies almost directly on the horizontal axis.

As illustrated in Fig. 22, both configurations of the incoherent/coherent double angularly multiplexed architecture can achieve both high fidelity and high optical throughput while using simultaneous recording (with a large enough beam splitter ratio in the case of the subhologram configuration). High optical throughput proves to be obtainable (in a linear holographic medium) despite the use of mutually incoherent readout beams because each configuration avoids the presence of beam degeneracy.

In addition to providing for linear summation of the diffracted output intensities, readout with mutually incoherent beams in the incoherent/coherent double angularly multiplexed architecture avoids (during operation, not training) the rigid optical phase stability requirements needed in a single-source architecture that is read out with mutually coherent beams. This feature reduces the degree of vibration isolation required and hence increases the practicality of operating a trained photonic neural network in an industrial or field environment.

A further advantage of the incoherent/coherent double angularly multiplexed architecture is that the interconnection gratings in the volume holographic medium can be copied into a second volume holographic recording medium in a single recording step.[51] For example, the full set of interconnections that are learned in a primary adaptive system can easily be reproduced in any number of secondary permanent holographic media for operational use. In contrast, direct single-step copying of an interconnection pattern within the single-source architecture is not possible without sacrificing either interconnection fidelity or optical throughput. Instead, it appears that at least N (if not N^2) exposure steps are required for duplication of an N-to-N interconnection system within a single-source architecture.

As mentioned in Section 5.E, greater interconnection densities may be achievable if grating degeneracy (and hence the use of fractal sampling grids) can be avoided. Of the interconnection techniques discussed herein, only the subhologram configuration of the incoherent/coherent double angularly multiplexed architecture offers the potential of avoiding the presence of grating degeneracy.

In this paper we have quantitatively evaluated the performance characteristics of the incoherent/coherent double angularly multiplexed architecture (based on the use of an array of individually coherent but mutually incoherent sources) for highly multiplexed volume holographic interconnection applications. In addition, we have quantitatively evaluated the directly comparable performance characteristics of conventional single-source architectures. As discussed below, there are several clear directions for continuing research.

7. Future Research Directions

In this study we have performed a detailed comparison of a number of holographic interconnection architectures that can be used to implement weighted interconnections with a high degree of fan-out and fan-in. For the most part, this comparison has been made on the basis of interconnection pathway independence (lack of cross talk) and insertion loss (optical throughput efficiency). For the neural-network application in particular, it would be of considerable interest to determine the appropriate levels of interchannel isolation and insertion loss permissible in the context of particular learning models without compromising overall system performance. In other words, to what degree are certain neural-network models sensitive (or insensitive) to these effects, if we rely to a greater or lesser extent on the learning capacity of the network to obviate the necessity for ideal interconnection behavior? Preliminary experimental and theoretical studies suggest the ability of some learning algorithms to overcome a certain degree of cross talk in the interconnection system,[52,53] but a more comprehensive study of this issue is necessary.

The simulation studies presented herein should be expanded to evaluate the additional limitations imposed by the effects of self-diffraction among the recording beams, grating erasure, exposure scheduling, finite pixel size, and finite range of grating-strength modulation on both reconstruction fidelity and throughput, particularly as the number of interconnections is increased. Significantly increasing the number of interconnection nodes considered in this analysis will enable scaling trends for the relative errors of each architecture to be further identified and compared. Inclusion of the grating recording characteristics of photorefractive media in the holographic-recording model will permit the effects of material nonlinearities to be determined and the utility of these materials for the implementation of adaptive photonic neural networks to be evaluated. Furthermore, extension of the BPM simulations to three dimensions will permit verification of the trends observed using a two-dimensional model and will also permit direct investigation of grating-degeneracy effects.

In addition to further modeling studies, previous laboratory work that has confirmed the basic features of the incoherent/coherent double angularly multiplexed architecture[12] can be expanded to include a more detailed study of various implementation issues, particularly as applied to photorefractive media. Such issues include the quantitative comparison of experimentally determined fidelity errors and optical throughput losses with simulation results; the effects of scale-up in the number of nodes on interconnection performance; the effects of subhologram overlap and the beam splitter ratio on reconstruction fidelity and throughput; and continued device development that will permit the eventual integration of mutually compatible source arrays, neuron-unit arrays, and volume holographic media into a practical system.

This research was supported in part by the Defense Advanced Research Projects Agency (both the Program on Neural Networks and the National Center for Integrated Photonic Technology) and the U.S. Air Force Office of Scientific Research (University Research Initiative: Center for the Integration of Optical Computing). Greg Nordin gratefully acknowledges support from a Hughes Doctoral Fellowship and a Center for Photonic Technology Graduate Prize Fellowship at the University of Southern California.

References

1. R. Kostuk, J. Goodman, and L. Hesselink, "Design considerations for holographic optical interconnects," Appl. Opt. **26**, 3947–3953 (1987).
2. D. Z. Anderson and D. M. Lininger, "Dynamic optical interconnects: volume holograms as optical two-port operators," Appl. Opt. **26**, 5031–5038 (1987).
3. D. Psaltis, D. J. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," Appl. Opt. **27**, 1752–1758 (1988).
4. M. Cronin-Golomb, "Dynamically programmable self-aligning optical interconnect with fan-out and fan-in using self-pumped phase conjugation," Appl. Phys. Lett. **54**, 2189–2191 (1989).
5. J. H. Hong, S. Campbell, and P. Yeh, "Optical pattern classifier with perceptron learning," Appl. Opt. **29**, 3019–3025 (1990).
6. E. S. Maniloff and K. M. Johnson, "Dynamic holographic interconnects using static holograms," Opt. Eng. **29**, 225–229 (1990).
7. See, for example, the feature on neural networks, Appl. Opt. **26**, 4909–5111 (1987).
8. B. K. Jenkins, G. C. Petrisor, S. Piazzolla, P. Asthana, and A. R. Tanguay, Jr., "Photonic architecture for neural nets using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest* (ICO-90 Organizing Committee, Kobe, Japan, 1990).
9. P. Asthana, H. Chin, G. P. Nordin, A. R. Tanguay, Jr., S. Piazzolla, and B. K. Jenkins, "Photonic components for neural net implementations using incoherent/coherent holographic interconnections," in *OC'90 Technical Digest* (ICO-90 Organizing Committee, Kobe, Japan, 1990).
10. P. Asthana, H. Chin, G. Nordin, A. R. Tanguay, Jr., G. C. Petrisor, B. K. Jenkins, and A. Madhukar, "Photonic components for neural net implementations using incoherent–coherent holographic interconnections," in *1990 OSA Annual Meeting,* Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 57.
11. B. K. Jenkins, A. R. Tanguay, Jr., S. Piazzolla, G. C. Petrisor, and P. Asthana, "Photonic neural-network architecture based on incoherent–coherent holographic interconnections," in *1990 OSA Annual Meeting,* Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 56.
12. P. Asthana, "Volume holographic techniques for highly multiplexed interconnection applications," Ph.D. dissertation (University of Southern California, Los Angeles, Calif., 1991).
13. B. K. Jenkins and A. R. Tanguay, Jr., "Photonic implementa-

tions of neural networks," in *Neural Networks for Signal Processing,* B. Kosko, ed. (Prentice-Hall, Englewood Cliffs, N.J., 1992), Chap. 9, pp. 287–382.

14. G. P. Nordin, "Volume diffraction phenomena for photonic neural network implementations and stratified volume holographic optical elements," Ph.D. dissertation (University of Southern California, Los Angeles, Calif., 1992).
15. D. Psaltis, A. A. Yamamura, K. Hsu, S. Lin, X.-G. Gu, and G. Brady, "Optoelectronic implementations of neural networks," IEEE Commun. Mag. **27**(11), 37–40 (1989).
16. H. Lee, "Volume holographic global and local interconnecting patterns with maximal capacity and minimal first-order crosstalk," Appl. Opt. **28,** 5312–5316 (1989).
17. H. Lee, X.-G. Gu, and D. Psaltis, "Volume holographic interconnections with maximal capacity and minimal cross talk," J. Appl. Phys. **65,** 2191–2194 (1989).
18. C. X.-G. Gu, "Optical neural networks using volume holograms," Ph.D. dissertation (California Institute of Technology, Pasadena, Calif., 1990).
19. C. Slinger, "Analysis of the *N*-to-*N* volume-holographic neural interconnect," J. Opt. Soc. Am. A **8,** 1074–1081 (1991).
20. J. A. Fleck, J. R. Morris, and M. D. Feit, "Time-dependent propagation of high energy laser beams through the atmosphere," Appl. Phys. **10,** 129–160 (1976).
21. J. Van Roey, J. van der Donk, and P. E. Lagasse, "Beam-propagation method: analysis and assessment," J. Opt. Soc. Am. **71,** 803–810 (1980).
22. D. Yevick and L. Thylen, "Analysis of gratings by the beam-propagation method," J. Opt. Soc. Am. **72,** 1081–1089 (1982).
23. R. V. Johnson and A. R. Tanguay, Jr., "Optical beam propagation method for birefringent phase grating diffraction," Opt. Eng. **25,** 235–249 (1986).
24. J. W. Goodman, "Fan-in and fan-out with optical interconnections," Opt. Acta **32,** 1489–1496 (1985).
25. P. Asthana, G. Nordin, S. Piazzolla, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of interchannel cross talk and throughput efficiency in highly multiplexed fan-out–fan-in holographic interconnections," in *1990 OSA Annual Meeting,* Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 242.
26. D. O. Hebb, *Organization of Behavior* (Wiley, New York, 1949).
27. C. Peterson, S. Redfield, J. D. Keeler, and E. Hartman, "Optoelectronic implementation of multilayer neural networks in a single photorefractive material," Opt. Eng. **29,** 359–368 (1990).
28. S. Kessler and R. Hild, "A new method for simultaneous complex addition and subtraction," Opt. Quantum. Electron. **15,** 65–70 (1983).
29. N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the Hopfield model," Appl. Opt. **24,** 1469–1475 (1985).
30. G. C. Petrisor, B. K. Jenkins, H. Chin, and A. R. Tanguay, Jr., "Dual-function adaptive neural networks for photonic implementation," in *1990 OSA Annual Meeting,* Vol. 15 of 1990 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1990), p. 56.
31. F. H. Mok, M. C. Tackitt, and H. M. Stoll, "Storage of 500 high-resolution holograms in a $\mathrm{LiNbO_3}$ crystal," Opt. Lett. **16,** 605–607 (1991).
32. D. Brady and D. Psaltis, "Control of volume holograms," J. Opt. Soc. Am. A **9,** 1167–1182 (1992).
33. D. Psaltis, D. Brady, X.-G. Gu, and K. Hsu, "Optical implementation of neural computers," in *Optical Processing and Computing,* H. Arsenault, ed. (Academic, New York, 1988), pp. 251–276.
34. D. Psaltis, X.-G. Gu, and D. Brady, "Fractal sampling grids for holographic interconnections," in *Optical Computing '88,* P. Chavel, J. W. Goodman, and G. Roblin, eds., Proc. Soc. Photo-Opt. Instrum. Eng. **963,** 468 (1988).
35. E. N. Glytsis and T. K. Gaylord, "Three-dimensional (vector) rigorous coupled-wave analysis of anisotropic grating diffraction," J. Opt. Soc. Am. A **7,** 1399–1420 (1990).
36. S. K. Case, "Coupled-wave theory for multiply exposed thick holographic gratings," J. Opt. Soc. Am. **65,** 724–729 (1975).
37. L. Solymar, "Two-dimensional *N*-coupled-wave theory for volume holograms," Opt. Commun. **23,** 199–202 (1977).
38. L. Solymar and D. J. Cooke, *Volume Holography and Volume Gratings* (Academic, New York, 1981), p. 212.
39. C. W. Slinger and L. Solymar, "Volume phase holograms reconstructed by the object wave," Opt. Quantum Electron. **16,** 369–372 (1984).
40. E. N. Glytsis and T. K. Gaylord, "Rigorous 3-D coupled wave diffraction analysis of multiple superposed gratings in anisotropic media," Appl. Opt. **28,** 2401–2421 (1989).
41. K.-Y. Tu, T. Tamir, and H. Lee, "Multiple-scattering theory of wave diffraction by superposed volume gratings," J. Opt. Soc. Am. A **7,** 1421–1435 (1990).
42. K.-Y. Tu, H. Lee, and T. Tamir, "Analysis of cross talk in volume holographic interconnections," Appl. Opt. **31,** 1717–1729 (1992).
43. M. D. Feit and J. A. Fleck, Jr., "Light propagation in graded-index optical fibers," Appl. Opt. **17,** 3990–3998 (1978).
44. R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography* (Academic, New York, 1971).
45. W. R. Klein and B. D. Cook, "Unified approach to ultrasonic light diffraction," IEEE Trans. Sonics Ultrason. **SU-14,** 123–134 (1967).
46. T. K. Gaylord and M. G. Moharam, "Thin and thick gratings: terminology clarification," Appl. Opt. **20,** 3271 (1981).
47. B. Benlarbi and L. Solymar, "The effect of the relative intensity of the reference beam on the reconstructing properties of volume phase gratings," Opt. Acta **26,** 271–278 (1979).
48. W. J. Burke and P. Sheng, "Crosstalk noise from multiple thick-phase holograms," J. Appl. Phys. **48,** 681–685 (1976).
49. H. Kogelnik, "Coupled wave theory for thick hologram gratings," Bell Syst. Tech. J. **48,** 2909–2947 (1969).
50. G. P. Nordin, P. Asthana, A. R. Tanguay, Jr., and B. K. Jenkins, "Analysis of weighted fan-out/fan-in volume holographic interconnections," in *Diffractive Optics: Design, Fabrication, and Applications,* Vol. 9 of 1992 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1992), pp. 165–167.
51. S. Piazzolla, B. K. Jenkins, and A. R. Tanguay, Jr., "Single-step copying process for multiplexed volume holograms," Opt. Lett. **17,** 676–678 (1992).
52. E. G. Paek, J. R. Wullert III, and J. S. Patel, "Holographic implementation of a learning machine based on a multicategory perceptron algorithm," Opt. Lett. **14,** 1303–1305 (1989).
53. C. W. Slinger, "Weighted volume interconnects for adaptive networks," Opt. Comput. Process. **1,** 219–232 (1991).