Optical network for real-time face recognition

Hsin-Yu Sidney Li, Yong Qiao, and Demetri Psaltis

An optical network is described that is capable of recognizing at standard video rates the identity of faces for which it has been trained. The faces are presented under a wide variety of conditions to the system and the classification performance is measured. The system is trained by gradually adapting photorefractive holograms.

Key words: Optical pattern recognition, neural networks, photorefractives.

1. Introduction

We report the experimental demonstration of a two-layer optical network that accepts input images of faces at standard video rates and classifies them in real time. The adaptable interconnections of the network are implemented with holograms stored in a photorefractive crystal. The optical system that we use in this work is the standard holographic multilayer architecture.[1–6] The second layer has fixed weights, and a simple *ad hoc* procedure is used to train the network. Choosing a training algorithm that is well suited to the optical implementation is the most crucial step in carrying out a successful experiment. The error backpropagation algorithm[7] and its variants are the most popular procedures for training multilayer optical networks.[1,3,4,5] Backpropagation is an example of a learning algorithm that yields distributed representations in the hidden layers of a network. In a distributed representation a large portion (typically half) of the hidden units respond when the input is one of the training samples. In contrast, in a local learning algorithm each hidden unit is trained to respond to only a small number of training examples. The radial basis function classifier is an example of a commonly used local learning algorithm. An optical radial basis function system has been recently demonstrated.[8] The advantage of local algorithms is the fact that the training process is relatively easy. If an input training sample does not cause any of the existing hidden units to respond sufficiently, a new hidden unit is added and devoted to the new sample. The disadvantage of local algorithms is the large network size that is typically obtained. The disadvantage of distributed-representation learning algorithms is the fact that the training is difficult, typically requiring a large number of training cycles.

In selecting an algorithm for training an optical neural network, we can argue that distributed algorithms are well suited for optics because the computational speed of optics can be effectively used to speed up the training. However, the optical implementation of algorithms such as error backpropagation require a dynamic holographic medium that can be accurately controlled. In the experiment described in here we use photorefractive crystals to implement the adaptive interconnections. When a new hologram is recorded in a photorefractive crystal, the previously recorded signal is partially erased. This weight decay in effect limits the number of cycles a training algorithm can run on an optical system, because earlier exposures are erased as the training progresses. Dynamic copying[9–12] can overcome this problem by restoring the strength of the hologram through feedback. However, dynamic copying is still at the early stages of development, and it is premature to construct a large-scale network that uses this approach. Another way of bypassing the weight decay problem is to use local algorithms, since they do not require long training sequences. In this case the large storage capacity of three-dimensional holograms can be used to synthesize the large networks that are required.

The algorithm that we use is a hybrid. It has features of local algorithms in that each hidden unit is trained separately and the training method is not iterative. In contrast, the representations that result are distributed. We found that distributed representations were crucial for two reasons. First, when the optical network was trained with purely local representations, we found that it became ex-

The authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125.

Received 9 November 1992.
0003-6935/93/265026-10$06.00/0.

tremely susceptible to noise and the performance deteriorated very rapidly as the number of hidden units increased. This phenomenon occurs because in a purely local representation, only one hidden unit is on at a time. Because the output is formed as a linear combination of all the hidden units, a small amount of noise from each hidden unit will ultimately overwhelm the signal term as more hidden units are added. Poor generalization performance is the second reason to avoid purely local representations. We found that by switching to distributed representations, the system performed much better when presented with images it had never seen before.

In Section 2 we describe the optical architecture and the overall experimental setup. In Section 3 we describe the training algorithm and the details of the training procedure. In section 4 we describe the performance obtained with the network.

2. Experimental Apparatus

The optical setup is shown in Fig. 1. It is a two-layer network with an optical preprocessing stage that performs edge enhancement. The input device to the network is a liquid-crystal TV (LCTV) that has a resolution of 320 pixels × 240 pixels and 2 cm × 2.5 cm clear aperture. This device was extracted from an Epson television projector. The LCTV is illuminated with collimated light from an argon laser ($\lambda = 488$ nm). Lens L1 produces the Fourier transform of the input image at plane P2. A spatial filter is placed at P2 to accomplish two goals. It blocks the higher diffracted orders that result from the pixelation of the LCTV. The removal of the higher orders gives a smoother, less noisy image but it reduces the light efficiency of the LCTV. The second function of the spatial filter in plane P2 is to block the low-frequency components of the input image that enhance the edges of the input image and dramatically improve the ability of the system to discriminate between inputs from different classes. A photograph of the spatial filter is shown in Fig. 2. It consists of a cross hair and a dc block for high-pass filtering. The purpose of the cross hair is to remove the diffraction pattern at P2 caused by the sharp edges formed at the boundary of the actual area of the LCTV. This boundary, when edge enhanced, yields a strong rectangle that is common to all inputs and makes discrimination difficult. The diameter of the dc block is 260 μm. Given the wavelength of light and the focal length of L1 ($F_{L1} = 50$ cm), we can find the cutoff frequency to be 0.533 lines/mm. Roughly speaking, features in the input plane that are smaller than 1.9 mm are highlighted in the edge-enhanced image. A iris (not shown in Fig. 2) is used to block

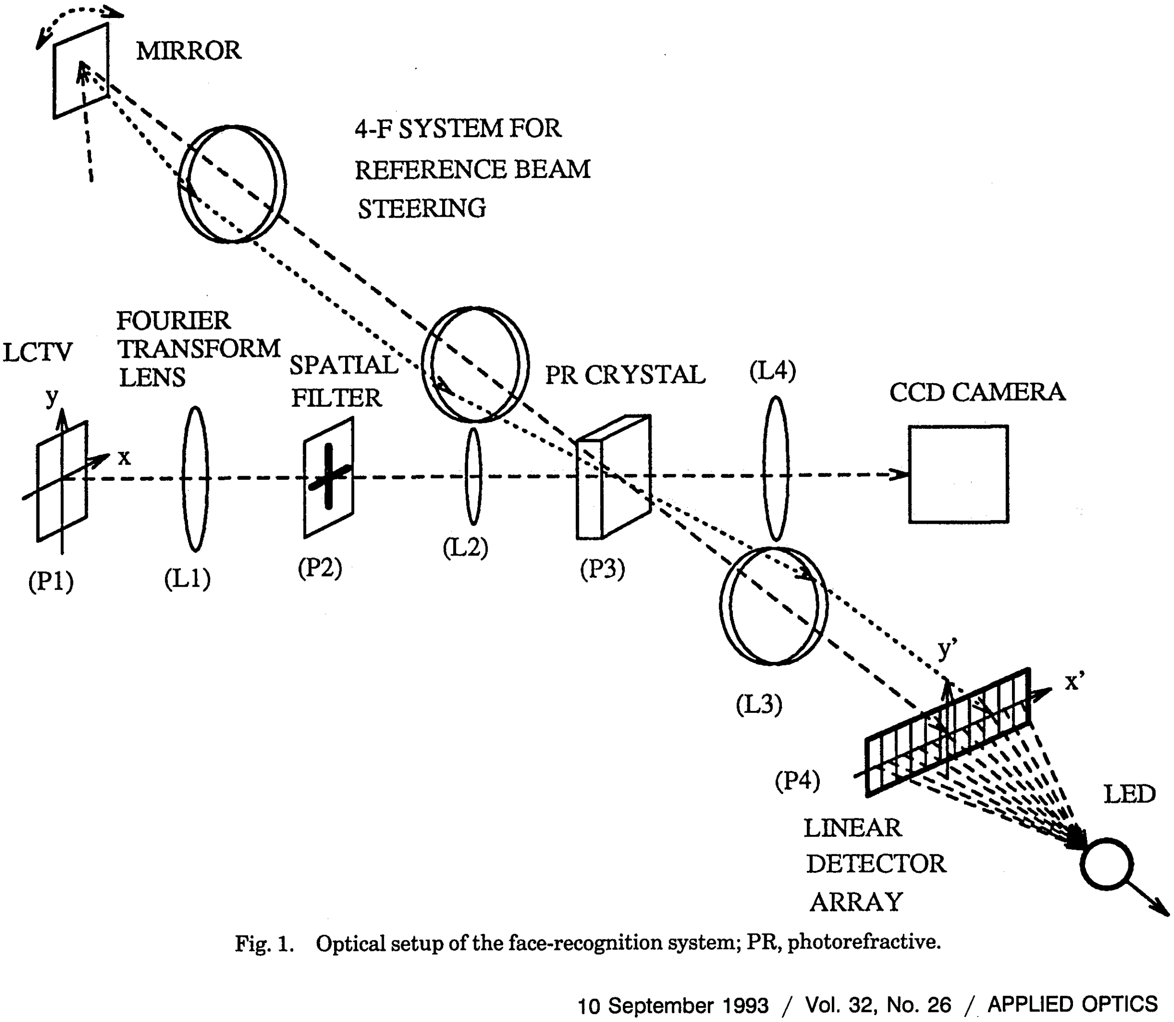

Fig. 1. Optical setup of the face-recognition system; PR, photorefractive.

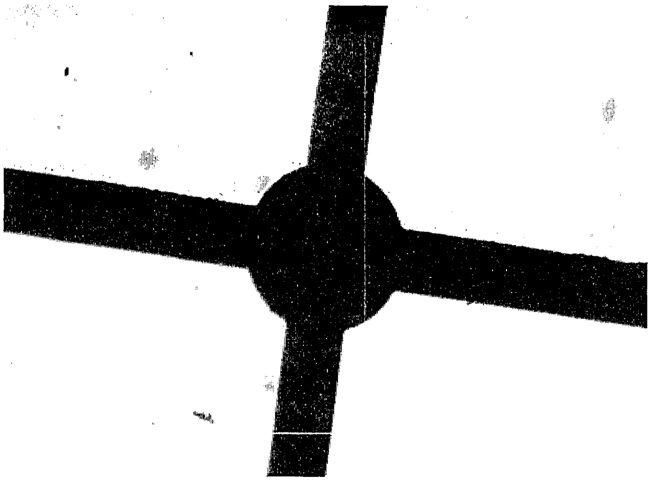

Fig. 2. Spatial filter used in plane P1 of Fig. 1.

the higher orders not blocked by the cross hair. An example of an image of a face and the edge-enhanced version of it that was produced by the optical system is shown in Fig. 3.

Lens L2 images with magnification 1 plane P2 onto plane P3, the plane of the hologram. The size of the spectrum on the hologram is approximately 5 mm in diameter. The hologram is formed by introducing a plane-wave reference. The angle between the signal and reference beam varies from 29° to 31° outside the crystal. The reference beam is reflected off a mirror mounted on a computer-controlled rotation stage. The plane of the rotating mirror is imaged onto the crystal with a unit magnification 4-f system that permits the angle of the reference beam to be scanned without moving the position of the reference beam on the crystal. The crystal is an iron-doped $LiNbO_3$, with a doping level of 0.01%. The c axis of the crystal is in the horizontal direction in Fig. 1. The crystal dimensions are 20 mm × 20 mm × 8 mm.

Lens L4 is a Fourier transform lens that produces an image of the edge-enhanced input image on a CCD for visual assessment. Lens L3 is also a Fourier transforming lens that produces at the output plane P4 the response of the first layer where it is sensed by a linear detector array. A beam splitter placed in front of the array diverts a portion of the light to a CCD camera so that the output of the first layer can be visually monitored. Functionally, the system from the input plane P1 to P4 is an array of image correlators with one-dimensional shift invariance. To understand this, consider the case in which a single hologram is recorded in the crystal at a partici lar angle of the reference beam. In this case th system is a classic VanderLugt[13] correlator excep that a volume hologram is used and the input ha been high-pass filtered. The effect of the volum hologram is to eliminate shift invariance in th horizontal direction in Fig. 1. This happens becaus a horizontal shift at plane P1 will change the angle c incidence at plane P3 and cause the hologram to b Bragg mismatched.[14–16] Specifically, the light distri bution at plane P4 is given by[14]

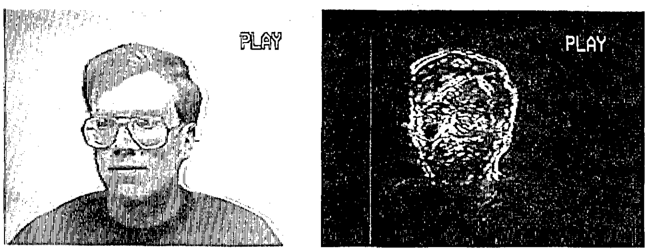

Fig. 3. Edge-enhanced image and original face.

$$g(x', y') = \iint f(x, y)h(x - x', y - y')\mathrm{d}x\mathrm{d}y \ \mathrm{sinc}(\alpha x'), \qquad (1$$

where $f(x, y)$ and $h(x, y)$ are the input and filte functions, respectively. The input coordinates ar (x, y) and the output coordinates are (x', y'). Th thickness of the crystal is L, θ is the angle of th reference beam, and

$$\alpha = \frac{L \sin \theta}{2\lambda F} \cdot \qquad (2$$

We see from Eq. (1) that the effect of the thic hologram is to mask off the two-dimensional correla tion pattern except for one vertical strip whose pos tion depends on the angle of the reference bean The amount of shift invariance that can be tolerate in the horizontal direction is approximately equal t $1/\alpha$ plus the width of the correlation peak in th horizontal direction. The system retains its shi invariance in the vertical direction. If we change th angle of the reference beam and record a differer hologram at each angle, then the one-dimensiona strip of the two-dimensional correlation function wi be produced at a different horizontal location. In th experiment that we will describe, holograms ar recorded at 40 separate angles separated by 0.05 yielding a system that has 40 correlators with one dimensional shift invariance.

The experiment in Fig. 4 demonstrates the opera tion of this part of the system. In this case eacl filter was a recording of the face of the same person a different scales. What is shown in Fig. 4 is the inpu to the network for four different size images, alon with the corresponding response at the right-han side of each picture. We see that as the size in creases, the strongest response of the system is a different vertical positions. In the optical setup, th correlation responses shown in the right-hand side o each picture is actually horizontal, and the displa was created by simply rotating the CCD camera b 90°.

The role of the second layer is twofold. The firs task is to take advantage of the vertical shift invari ance of the first layer, and the second task is t combine the outputs of the 40 correlators and mak the final classification. We first discuss the shif invariance. Suppose that an image at a particula

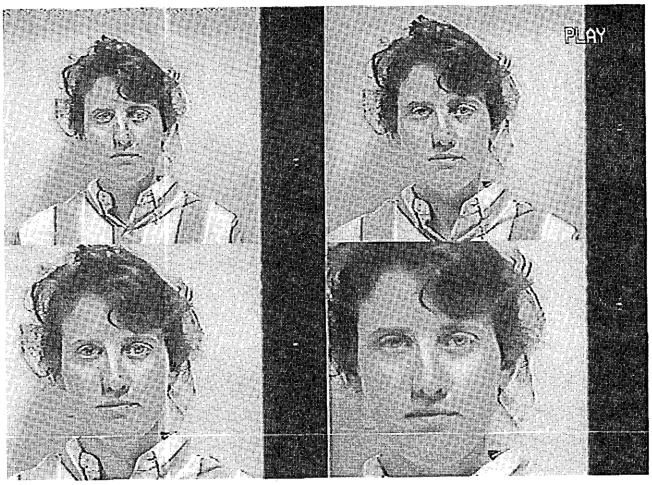

Fig. 4. Experiment showing the position of the correlation peak to be proportional to the size of the input face.

location at the input produces a strong correlation peak somewhere at the output. If the input is horizontally translated by approximately 0.4 mm then the correlation peak disappears. If the input is translated vertically then the correlation peak moves vertically also. What we really need for shift-invariant recognition is a system whose output does not change as the input shifts. To accomplish this we use long-detector elements in the vertical direction as shown in Fig. 1. These long detectors collect the correlation peak and continue to produce a strong output signal as the input image shifts vertically. Unfortunately, we cannot use an arbitrarily long-detector element to obtain full shift invariance vertically, because then the detector would simply collect all the diffracted energy from the corresponding filter stored in the hologram. Roughly speaking, all input signals with the same total energy would yield the same response. A shorter detector responds more selectively to the correlation peak, and hence the degree of match between the input and the reference, but it sacrifices shift invariance. Thus there is a basic trade-off between shift invariance and discrimination capability. In our network we made this compromise by trial and error. By repeating the experiment with a horizontal slit of varying width placed in front of the detector array, we find that the amount of shift invariance in the vertical direction is roughly 3 mm or equivalently 12% of the size of the input image. As we see below, this choice yields good discrimination capability.

The second layer also puts together all the vertically integrated responses from the first layer and produces the final output. Because the output of the detector array in plane P4 is electronically available, we can implement the second layer either electronically or optically. We have done both with comparable performance. The optical implementation of the second layer is realized by thresholding the output of the detector array and then feeding it to a second LCTV. The inner product between the signal recorded on the LCTV and a weight vector stored in the form of a transparency is then optically formed. This inner product is electronically thresholded to produce the final output. In the current system we describe in this paper, the operations of the second layer are so simple that it was easier to do them electronically. Specifically, all the weights of the second layer have the same value. In other words, the second layer simply integrates the output of the first layer. The electric signal from each detector is the square of the light amplitude of the total signal incident at each element. The signal from the detector can be thresholded electronically. However, we get the best performance by simply using the square-law nonlinearity. In this case, the system becomes similar to a quadradic associative memory.[17,18] Notice that the nonlinearity performed at P4 is crucial in this system. If the outputs of all the correlators from the first layer were somehow coherently added without the inclusions of the nonlinearity, then the overall system would simply be equivalent to a single correlator.

A schematic diagram of the overall system is shown in Fig. 5. The input images are detected by a standard television camera. The video signal is either stored on a VCR to form a training set or directly to the LCTV during real-time operation. The two-layer optical network is the system we described above. A personal computer controls the experiment during the training phase by instructing the VCR to advance the video by one frame and pause so that the training algorithm can be executed in the optical system. The output of the hidden layer determines whether the hologram should be modified by the current input image. If a holographic exposure is needed, the computer opens two shutters (one for the signal and one for the reference beam) for a specified time and the hologram is recorded. During the execution of the algorithm the computer also controls the angle of the reference beam so that different hidden units can be trained. After the training is completed, the computer is no longer involved in the operation of the system except to record the output data if desired.

3. Training Procedure

The training algorithm that we use is partially motivated by the tiling algorithm.[19] In the tiling algorithm, individual units are trained separately for a fixed number of iterations. Once a unit is trained the algorithm moves on to a new unit and trains it to make up for the deficiencies in the performance obtained with the previous units. In this way networks with multiple layers and many neurons per layer can be built up and trained. In the standard tiling algorithm, each unit is trained with the perceptron algorithm with the entire training set. In our algorithm each unit is trained by a subset of the training set that consists of similar images. This similarity measure is enforced by training each unit to respond to a contiguous short segment of the

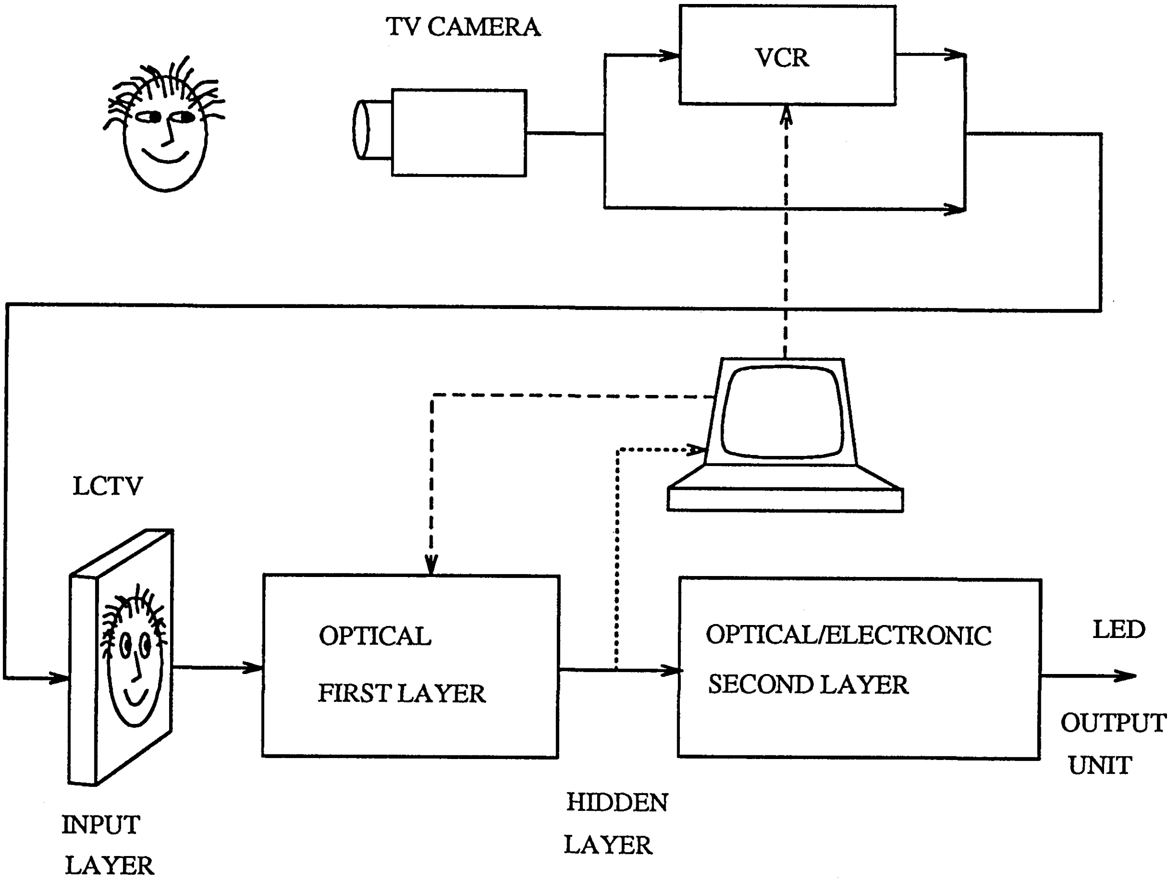

Fig. 5. Schematic diagram of overall system.

training video. In this way, each unit is trained to respond to a specific aspect of the input face. This simplifies the training of individual units, and the overall training procedure results in networks of predictable size.

The flow chart for the algorithm we use is shown in Fig. 6. Let us more specifically describe the algorithm. Let $\mathbf{f}^k$ denote the kth image in the training sequence stored in the VCR and let w_{ij} denote the weight of the first layer connecting the ith input pixel to the jth hidden unit. The training algorithm is as follows.

```
set $e = 0$   ($e$ is the number of exposures per hidden unit)
set $j = 1$   ($j$ enumerates the hidden units)
while ("there are more training examples")
do {   (go through the training set one frame at a time)
    $h = 0$   ($h$ is the number of hidden units turned on)
    for $j' = 1$ to $j$, if $\Sigma_{i'=-I/2}^{I/2} |\Sigma_i f_i^k w_{i\ -i',j}|^2 > \theta$ then
    $h = h + 1$   (count the number of hidden units that are on)
    if ($h < H$ and $\Sigma_{i'=-I/2}^{I/2} |\Sigma_i f_i^k w_{i-i',j}|^2 < \theta$)
        (less than $H$ hidden units are on, and the current unit is off)
        then $w_{ij} = w_{ij} + f_i^k$ and $e = e + 1$
                (make an exposure
    if ($e > E$)   (more than $E$ exposures on curren unit)
        then $j = j + 1$ and $e = 0$   (create new hidde unit)
    "go to next frame"
}
```

The user must select the parameters θ, H, and E before the algorithm begins. In what follows we wil explain the role of each parameter and how it affect the performance of the trained network. The variable j counts the hidden units. We begin trainin the first unit ($j = 1$) by presenting frames to th system in sequence (incrementing k). The kth inpu is added to the weights of the first unit if the respons of the first hidden unit is below a threshold θ. Notic that in the optical system the response of the hidde unit is not simply the inner product between th input and the weight vector, but an integration over I pixels of the center of the correlation function, as w described earlier. If θ is set too high then the unit become very highly tuned to respond to the particula images they are trained for. If the threshold is to low then too much cross talk with unfamiliar face results, leading to erroneous classifications. Ideally

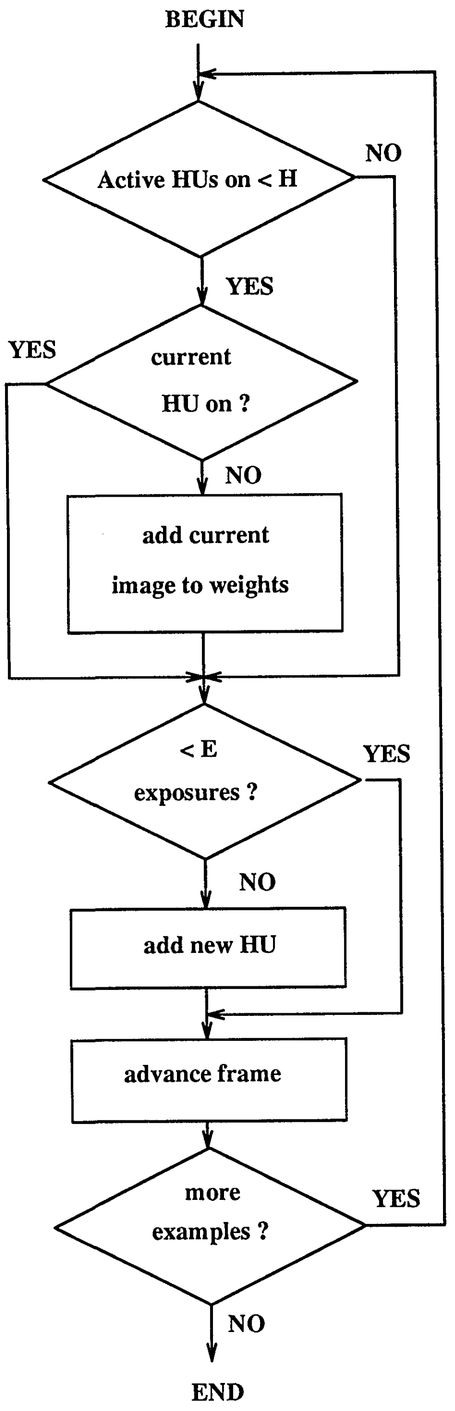

Fig. 6. Flow chart for the algorithm used to train the network; HU, hidden unit.

θ should be lowered as the training proceeds and hidden units are added, because this weakens all the stored holograms. In the experiment we describe we used a constant θ. The first unit continues to accumulate training examples in this way until a total of E exposures have been made to it. At that point a new hidden unit is created (j is incremented) by rotating the mirror that controls the angle of the reference beam. We would like to have E large in order to have each unit be responsive to as many training examples as possible. However, because we are only presenting positive examples to the system (i.e., we never subtract anything from the weights but always add to them), if too many examples are accumulated the weight is simply the average of the subject's faces, which is similar to the average of anybody's face, and loses its discrimination capability. The first H hidden units are trained in exactly the same manner as the first. When j exceeds H, the current input frame is added into the weights of the jth hidden unit only if fewer than H units are above threshold. If H is set to 1, then the training of the early units is identical to the rest. However, this results in a hidden layer response that has only one unit on at a time. We have already commented that we found that this results in poor performance on the training set because of susceptibility to noise and poor generalization. By requiring that at least H hidden units are on at any one time for the training set, we improve the robustness of the system and improve generalization. If H becomes too large, we would need too many hidden units to enforce this requirement, and the encoding becomes inefficient.

The discussion in the above paragraph describes the basic trends that we predict and experimentally observe as the parameters E, H, and θ are adjusted. The experiment that we will describe in this paper was carried out with $H = 3$, $E = 6$, and θ set equal to three times above the noise background level. These values were arrived at empirically by running the experiment several times and measuring the generalization performance. The system performance is sensitive to the setting of θ (it should be set relatively low), but not as sensitive to changes in H and E. These settings worked best for all the face-recognition experiments we tried. Unfortunately, there is no guarantee that these settings are the best for other problems.

The most attractive feature of this algorithm is that it can be easily implemented with the optical system described in Section 2 while yielding remarkably good classification performance, as we will see in Section 4. The algorithm requires two basic operations from the optical system: evaluation of the response of the hidden units to an input image so that the computer can compare it with a threshold, and addition of the current image into the hologram that specifies the weights of the unit. We have already described how the system evaluates the response of the hidden units. We will discuss here how the weight updates are performed. When a hologram is exposed to light, the strength of an individual holographic grating (or connection w_{ij}) is modified according to the following equation[3]:

$$\tau \frac{dw_{ij}}{dt} + w_{ij} = \beta m_{ij}, \tag{3}$$

where τ is the time constant of the holographic recording in the photorefractive crystal, β is a constant that depends on the crystal properties, and m_{ij} is the modulation depth of the frequency component of the illuminating light that matches the grating w_{ij}. For a short light exposure of duration Δt, we can approximate the change in the hologram by

$$\Delta w_{ij} \approx -\frac{\Delta t}{\tau} w_{ij} + \frac{\Delta t}{\tau} \beta m_{ij}. \tag{4}$$

In other words, each exposure reinforces each weight in proportion to the strength of the corresponding frequency component of the illuminating light. However, each exposure also erases all the weights in proportion to their current strength. This is the well-known weight-decay problem that plagues photorefractive memories[20] and photorefractive neural networks.[3] Several solutions to this problem have been proposed.[9,10,21] We use a simple exposure schedule in our experiment, in which the later exposures are linearly shortened to compensate for the decay of the earlier holograms, resulting in an approximately uniform final recording. Specifically the mth exposure, t_m, is set equal to $t_m = 3 - m/240$ s. Thus the exposure varied from 3 s at the beginning of the exposure sequence to 2 s at the end, with a total light intensity equal to 10 mW/cm^2 and a modulation depth of approximately 0.1.

The training set for the experiment was a video recording of one of the authors (Yong Qiao) moving his head in front of the camera, turning, nodding, tilting his head, smiling, etc. The total number of images in the training set is 5400 frames. The execution of the algorithm modified the hologram with only 240 of these images. The rest produced an acceptable hidden layer response. Because each hidden unit receives six exposure, a total of 40 hidden units were created. The maximum number of hidden units that the system can support is limited by two factors. One is the dynamic range of the photorefractive hologram. In this case a total of 240 holograms are superimposed. If we assume that all these exposures are statistically uncorrelated (i.e, each exposure simply erases all the previously recorded holograms and does not ever reinforce them), then the diffraction efficiency of each hologram would fall by a factor equal to $(240)^2$ (Ref. 9) compared with the efficiency with which a single hologram is stored. Because as many as 5,000 holograms[22] have been superimposed in LiNbO$_3$, the dynamic range was not a problem in our experiment. The second limitation is the numerical aperture of the optical system that permitted all the reference beams to enter the crystal. The system we used in the experimental had the capability to implement in excess of 100 units, and it is possible to build systems with more than 1000 units. Therefore, this particular training set did not stretch the limits of the system's capabilities. The entire training cycle lasted ~40 min, which includes the time for hologram exposure and controlling the system by computer.

Figure 7 is a composite photograph showing a short sequence of the training session. Each picture in the composite shows the current input frame; at the right, vertically displayed, is the optical response of the hidden units. The first event in the sequence is at the top left in Fig. 7, and it shows the frame shortly after the hologram is exposed. As time progresses the hidden layer response changes (upper right-hand corner) and gradually dims (lower right-hand corner). Ultimately there are fewer than three units on, and the system is triggered to make another exposure (lower right-hand corner). The white ribbon on the left of the input image where the hidden layer normally appears indicates that the hologram is being exposed to light and the camera that monitors the hidden layer response is flooded with light.

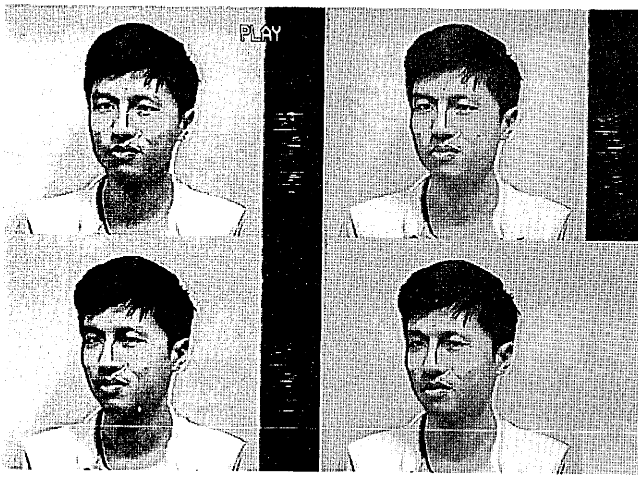

Fig. 7. Photographs showing part of the training session.

4. Classification Performance

In this section we describe the performance of the trained network. Once the network is trained it operates in real time, processing 30 frames/s directly from the input TV camera. The outputs from the detector array are simply added together electronically, and this sum is then thresholded to produce the final output. The holograms will decay when exposed to light during the testing phase. We can overcome this decay by either thermally fixing the hologram[23] or by using dynamic copying.[10–12] In this experiment we adopted a simpler route that temporary overcomes this problem. By reducing the readout light intensity by a factor of 20, compared with the total writing intensity, we can calculate that the holograms will decay after several hours of constant illumination. The holograms were sufficiently strong that the reduction in the readout intensity yielded a sufficient signal at the detector. The system was tested with the training set and with a wide variety of test sets, including Yong presented to the system under various conditions and other people in an attempt to confuse the system. Shown in Fig. 8 is the signal at the output of the system before final thresholding. The entire recorded presentation shown in Fig. 8 lasts for ~10 min. The first minute is a portion of the training set. The next 2 min are a real-time input of Yong, who looks into the TV camera and moves around in a manner similar to that in the training set. While he does this, he does not

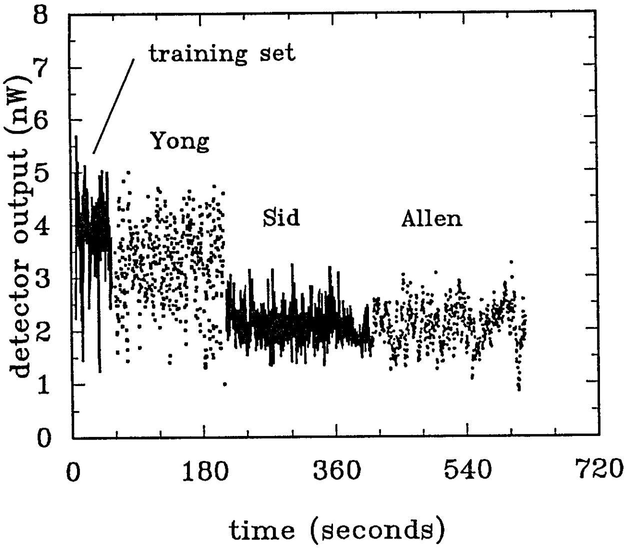

Fig. 8. System response before thresholding.

have access to any information from the network. The rest of the sequence is the response of the system to two other persons (Sid and Allen). We can see that the average response is highest for the training set, and it remains almost as high for the rest of the time when Yong is the input. The average response for the other two subjects is markedly lower. The variance of the response is higher for Yong, because he was exhibiting a wider range of head perspectives, compared with Sid and Allen, to test the limits of the system. Similar behaviors were observed for all 14 members of our group.

To make the final classification, we need to threshold the signal shown in Fig. 8. In the actual system, this is done electronically in real time. The optimum threshold was determined from the data shown in Fig. 8. Shown in Fig. 9 is a plot of probability of error as a function of the output threshold level. The three curves correspond to the probability of error for Yong, Sid, and Allen, estimated by classifying the data in Fig. 8 with different thresholds. If we want to minimize the overall probability of error, the optimum threshold level is approximately 2.5 nW, giving a probability of error of ~ 12%. If we set the threshold slightly above 3 nW, then we almost never make a false recognition while correctly identifying Yong approximately 70% of the time.

We can improve the performance of the system further by using the time domain. If the input face is moving and presents different views to the system, we can eliminate many of the errors by using a period of time longer than the duration of a single frame to do the classification. Specifically, we classify the current frame to be Yong if M out of the N previous frames give us a positive response. In implementing such an algorithm, we need to select N, M, and the threshold level. Shown in Fig. 10 is a plot of probability of error on the same three data sets as before as a function of the threshold level for $M = 7$ and $N = 25$. Notice that if the threshold level is selected in the range of 2.75–3 nW, the estimated probability of error is zero. In this example, the decision is made based on observation of the input video for 6 s (the computer sampled the output at four samples/s). In general, there is a trade-off between performance and observation time.

The next sequence of experiments we describe were carried out to evaluate the kind of generalization obtained by the network. In this case, the subjects (Yong and others) were allowed to look at the output of the network, and adjustments were made to test the limits of the system. Examples from this series of experiments are displayed in the composite of Fig. 11. The pictures are arranged in a 4 × 4 matrix. We will assign to each picture a pair of numbers (i, j), with the picture at the upper left-hand corner being

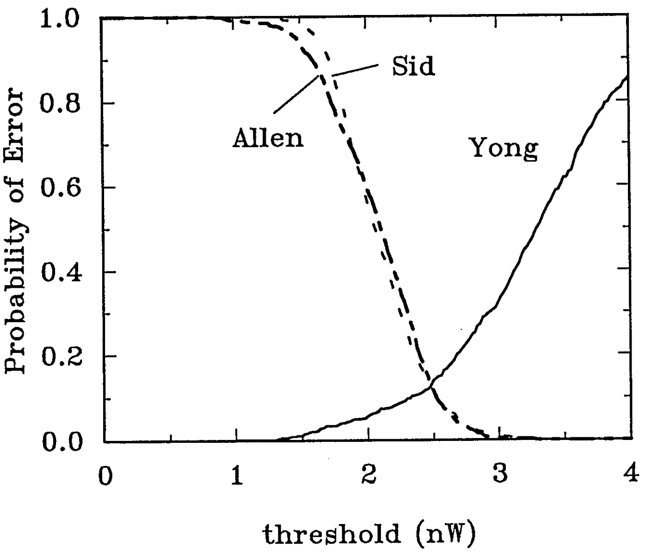

Fig. 9. Probability of error as a function of the output threshold level.

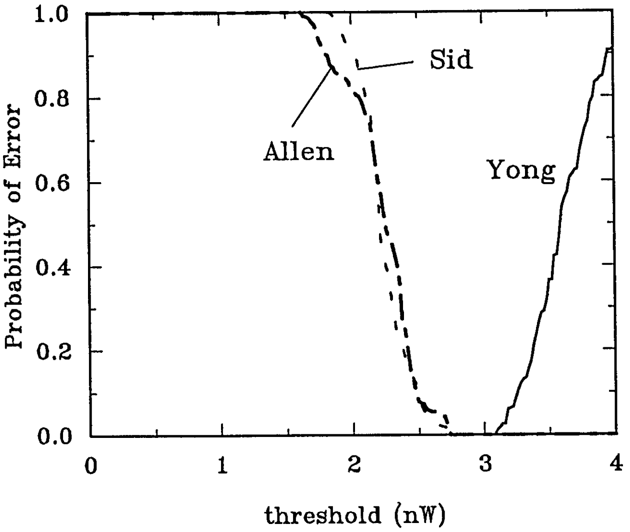

Fig. 10. Probability of error as a function of the output threshold level when the output is observed for 6 s to perform the classification.

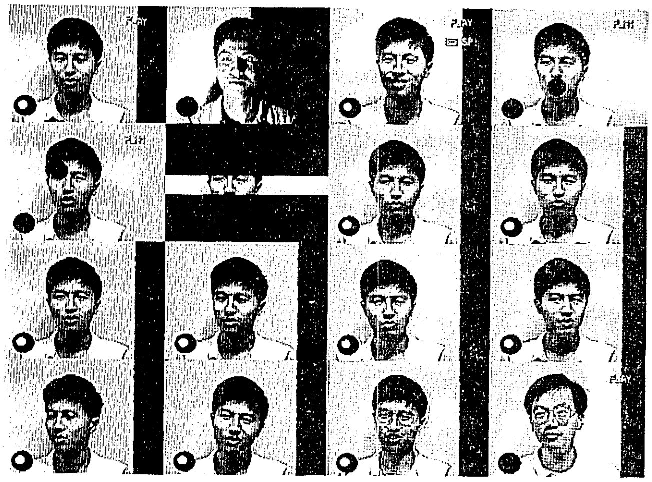

Fig. 11. Examples demonstrating the generalization capabilities of the system. A bright dot in the circle at the lower right-hand corner of each photograph indicates that the system classifies the input image as the person it was trained to recognize.

(1, 1) and the one at the upper right-hand corner being (1, 4). The small black circle within each picture displays the final output of the system after thresholding. If a bright dot appears in the circle, the system makes a positive identification of Yong. Picture (1, 1) is an example of Yong being correctly recognized by the system. Picture (1, 2) shows Yong illuminated from below and the side, whereas during training the illumination was from above. We can see that the system is sensitive to the direction of illumination because of the edge enhancement that is performed by the system. As the direction of illumination changes, the edges move around. To obtain invariance to illumination direction, we need to include in the training set examples of different lighting. Picture (1, 4) and (2, 1) show that key features such as the mouth and the eye are crucial for recognition. However, as picture (2, 2) shows, the eyes alone are not enough for a positive identification. Picture (2, 3) is meant to display the invariance of the system to up and down motion. It is difficult to assess this from the still photo. However, we measured a tolerance to vertical shifts of ~5% of the whole scene. The optical system was arranged such that vertical shifts of the input image become horizontal shifts on the LCTV. We made this arrangement because we need more tolerance to horizontal input shifts (people move side to side much more than up and down), and the optical system provides shift invariance in the vertical direction at the LCTV plane. Prior to the training, the tolerance to vertical input shifts was 2% of the whole scene. Training more than doubled the tolerance of the vertical shift. The tolerance of the system to nodding up and down was recorded by measuring the vertical motion on the screen of a fixed point on Yong's forehead, as he nodded up and down. According to this measure, the spot on his forehead can move by 1 cm without loss of recognition. From this measurement, and by measuring the dimensions of Yong's head, we obtain a crude estimate of 5° for the maximum tolerable angle of forward head tilt Picture (3, 3) shows an example of the tolerance of th system to horizontal shifts of the input image. I this direction the optical correlator provides considerable shift invariance. We measured the maximun horizontal shift to be ~13% of the total horizonta extent of the input frame. Overall, the system ha more than three times better tolerance to shifts in th horizontal than the vertical direction. Pictures (3, 4 and (4, 1) demonstrate the system's ability to tolerat turning of the head, which we measured to be 30° i either direction. The maximum tilt of the head picture (4, 2), was measured to be 12° in eithe direction. We did not seriously test the response o the system to scale changes.

5. Conclusion

The main goal of this experiment was to use a combination of existing optical techniques and algorithmic ideas to build a trainable real-time face-recognition system that works. This system gave us remarkably good performance and yet it greatly underutilizes the full capabilities of the optical network. In contrast, there are many ways we can seek to improve the performance of the system. For instance, to incorporate invariances to scale or illumination, we would need to expand the training set to have all possible *combinations* of scale and illumination conditions of interest as well as all the invariances that the current system incorporates. For example, to accommodate five different scales, we would need to expand the size of the training set by roughly a factor of 5. The number of hidden units that are needed with the approach we use usually scales proportionally to the size of the training set. Expanding the size of the optical system from the current 40 hidden units to approximately 1000 units is within reach. It should therefore be possible to expand the variety and range invariances accordingly. In addition, we can seek ways to build in some of the invariances, in addition to the one-dimensional shift invariance afforded by the Fourier transform holograms. For instance, we can have an adaptive optical system that is trained to recognize eyes independently of the identity of the face. This feature detector can then be used to normalize the input for vertical position or head rotation. These modifications of the system and its extension to multiperson recognition will be the subject of a future paper.

This work is supported by the Defense Advanced Research Projects Agency. We thank A. Grot and Jiafu Luo for helping us fabricate the spatial filters, K. Curtis and R. Denkewalter for many stimulating discussions and their help, and all the members of our group for their patience in participating in this experiment as training or testing sets.

References

1. K. Wagner and D. Psaltis, "Multilayer optical learning networks," Appl. Opt. **26**, 5061–5076 (1987).
2. Y. Owechko, G. J. Dunning, E. Marom, and B. H. Soffer,

"Holographic associative memory with nonlinearities in the correlation domain," Appl. Opt. **26**, 1900–1910 (1987).

3. D. Psaltis, D. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," Appl. Opt. **27**, 1752–1759 (1988).
4. Y. Qiao and D. Psaltis, "Local learning algorithm for optical neural networks," Appl. Opt. **31**, 3285–3288 (1992).
5. D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in artificial neural networks," Nature **343:** 325–330 (1990).
6. H. Yoshinaga, K. Kitayama, and T. Hara, "All-optical error-signal generation for backpropagation learning in optical multilayer neural networks," Opt. Lett. **14**, 202–204 (1989).
7. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing,* D. E. Rumelhart and J. L. McClelland eds. (MIT, Cambridge, Mass., 1986), Vol. 1, Chap. 8.
8. M. A. Neifeld and D. Psaltis, "Optical implementation of radial basis classifiers," Appl. Opt. **32**, 1370–1379 (1993).
9. D. Brady, K. Hsu, and D. Psaltis, "Periodically refreshed multiply exposed photorefractive holograms," Opt. Lett. **15**, 817–819 (1990).
10. Y. Qiao, D. Psaltis, C. Gu, J. Hong, and P. Yeh, "Phase-locked sustainment of photorefractive holograms using phase conjugation," J. Appl. Phys. **70**, 4646–4648 (1991).
11. H. Sasaki, Y. Fainman, J. E. Ford, Y. Taketomi, and S. H. Lee, "Dynamic photorefractive optical memory," Opt. Lett. **16**, 1847–1876 (1991).
12. S. Boj, G. Pauliat, and G. Rossen, "Dynamic holographic memory showing readout, refreshing, and updating capabilities," Opt. Lett. **17**, 438–440 (1992).
13. A.B. VanderLugt, "Signal detection by complex spatial filtering," IEEE Trans. Inf. Theory **IT-10**, 139–145 (1964).
14. J. Yu, F. Mok, and D. Psaltis, "Capacity of optical correlators," in *Spatial Light Modulators and Applications II,* U. Efron, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **825**, 114–120 (1987).
15. C. Gu, J. Hong, and S. Cambell, "2-D shift-invariant volume holographic," Opt. Comm. **88**, 309–314 (1992).
16. M. A. Neifeld and D. Psaltis, "Optical implementations of radial basis classifiers," Appl. Opt. **32**, 1370–1379 (1993).
17. C. L. Giles and T. Maxwell, "Learning, invariance, and generalization in high-order neural networks," Appl. Opt. **26**, 4972–4978 (1987).
18. D. Psaltis, C. H. Park, and J. Hong, "Higher-order associative memories and their optical implementations," Neural Net. **1**, 149–163 (1988).
19. M. Mezard and J. P. Nadal, "Learning in feedforward layered networks—the tiling algorithm," J. Phys. A **22**, 2191–2203 (1989).
20. K. Bløtekjaer, "Limitations on holographic storage capacity of photochromic and photorefractive media," Appl. Opt. **18**, 57–67 (1979).
21. Y. Taketomi, J. E. Ford, H. Sasaki, J. Ma, Y. Fainman, and S. H. Lee, "Incremental recording for photorefractive hologram multiplexing," Opt. Lett. **16**, 1774–1776 (1991).
22. F. Mok, "Applications of holographic storage in lithium niobate," in *Annual Meeting,* Vol. 23 of 1992 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1992), p. 102.
23. D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple storage and erasure of fixed holograms in Fe-doped $\mathrm{LiNbO_3}$," Appl. Phys. Lett. **26**, 182–184 (1975).