OPTICAL NEURAL NETS

One of the important applications of analog optical computing is in information processing systems. Processing of information with optics offers many advantages and capabilities including high-speed parallel processing, large volume data handling, compactness, low power consumption, and ruggedness (1–11). Compact custom-made optical hardware can process two-dimensional arrays of data of up to half a million pixels per array at 300 kHz frame rates. Most applications of optical processing systems have been developed for military hardware because of high cost and performance demands. Recent advances in optical material devices and components such as optical memory and optical display devices have made optical information processing systems more attractive for commercial applications. In addition, many of the innovative algorithms developed in the context of optical information processing are also implementable on a digital computer and perform well compared to various algorithms developed by the digital signal processing community.

This article briefly reviews the fundamentals of optical information processing for neural computing. We discuss optical processing techniques, materials, and devices for neural computing, image processing, and pattern recognition. Optical image processing algorithms and architectures as well as basic hardware concepts such as the fundamentals of optical spatial light modulators (SLMs) are reviewed for development and implementation of optical neural networks.

FUNDAMENTALS OF OPTICAL IMAGE PROCESSING

In this section, we briefly discuss the fundamentals of image processing using optical systems. The Fourier transform of an image can be generated in the space domain optically by using a lens (1–5). As a result, various types of image processing algorithms can be implemented optically. A thin lens introduces a phase shift (delay) to an incident wavefront by an amount proportional to both the thickness of the lens at each point and the index of refraction of the lens material. The light distribution $E(\alpha, \beta)$ in the back focal plane of a convex thin lens is the two-dimensional Fourier transform of the light transmittance $e(x, y)$, in the front focal plane [see Fig. 1(a)]. In this figure, (x, y) are the coordinates in the input plane, which is the front focal plane of the lens, and (α, β) are the coordinates in the Fourier plane, which is the back focal plane of the lens. This Fourier transform relationship is the fundamental relationship in the analysis of optical processors. If we place a second lens L_2 behind the Fourier plane as shown in Fig. 1(b), the light distribution at the back focal plane of L_2 is the double Fourier transform of the input field. Here, we have assumed that lenses L_1 and L_2 have the same focal length. Different image processing operations can be achieved by placing a spatial filter at the Fourier plane. For example, by placing an opaque spot at the origin of the Fourier plane, we can block the low spatial frequencies of the input signal, thus, generating a high-pass filtered version of the input field. Similarly, for more sophisticated types of image processing, a complex spatial filter $F(\alpha, \beta)$, where $F(\alpha, \beta)$ is the Fourier transform of a function $f(x, y)$, can be inserted at the Fourier plane, resulting in a light field of $F(\alpha, \beta) \times E(\alpha, \beta)$ leaving the filter plane. Therefore, at the output plane P_3, we obtain the Fourier transform of $F(\alpha, \beta) \times E(\alpha, \beta)$, which is equivalent to the convolution of the input signal $e(x, y)$ with the filter function $f(x, y)$ in the spatial domain. By properly choosing the spatial filter, numerous signal and image processing operations can be performed. There are a number of ways to synthesize the optical spatial filter.

The spatial filter can be generated by a computer and then written onto an optical display device (such as SLM) in the Fourier plane. An alternative way to generate the spatial filter is by holographic techniques (8). Figure 2 shows the procedure to synthesize an optical matched spatial filter by a holographic technique (1,2, and (8). The matched filter is designed for detecting a specific image or target in the presence of noise (12). For white noise, the matched filter is just the target itself in the spatial domain. In the Fourier domain, it is equal to the complex conjugate of the Fourier transform of the target. The matched spatial filter of a reference signal $s(x, y)$ is produced at the filter plane P_2. Referring to Fig. 2, the reference signal $s(x, y)$ is inserted at the input plane P_1. The light distribution $S(\alpha, \beta)$ at the filter plane P_2 is the Fourier transform of the input reference signal $s(x, y)$. A plane wave reference beam of uniform amplitude is incident on the plane P_2 at an angle θ with respect to the optical axis. To produce the matched filter at the Fourier plane for detecting the reference

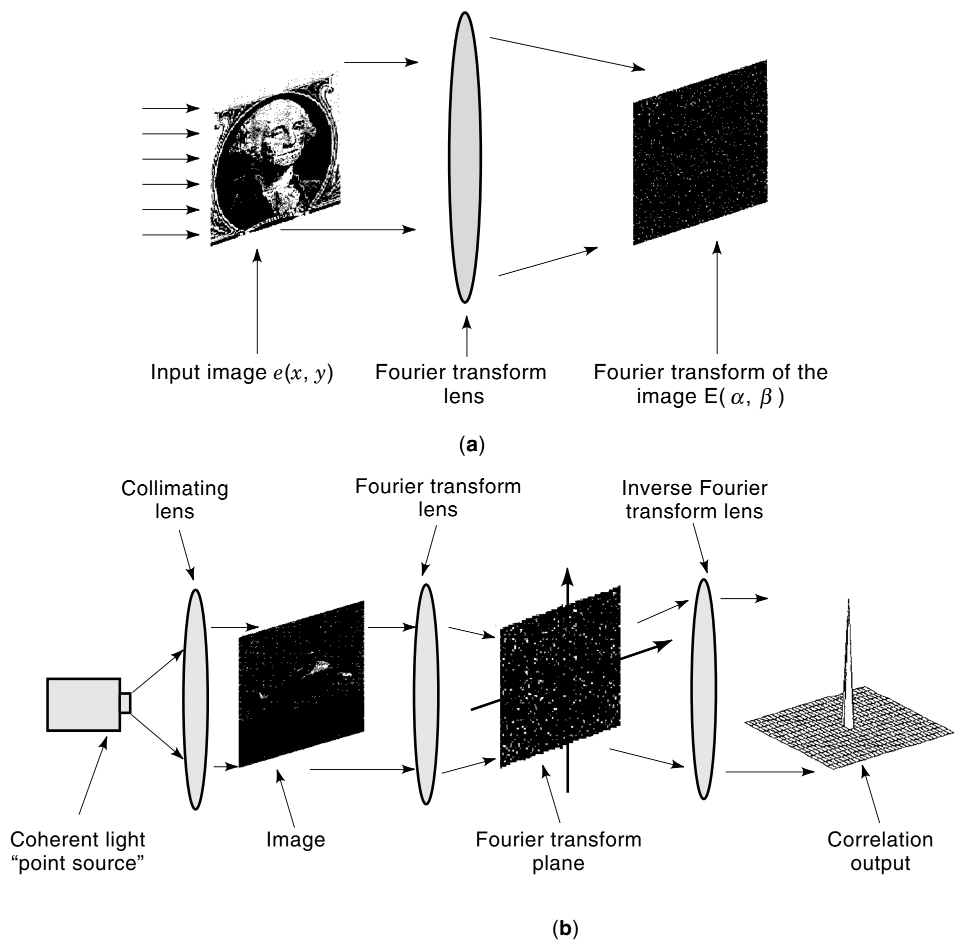


Figure 1. (a) The Fourier transform property of a lens. (b) An optical information processing system. Plane (a, b) is the Fourier plane where a filter function can be inserted to execute different image processing operations.

signal $s(x, y)$, a conventional holographic technique is used to record the interference patterns of the reference signal Fourier transform $S(\alpha, \beta)$ with the reference beam. This can be done by placing a detector such as high-resolution photographic film or a photosensitive recording material at plane P_2. The intensity distribution at the filter plane is obtained when the film is developed to produce a filter transmittance function. Under this condition, the filter transmittance function includes the desired matched spatial filter for $s(x, y)$, which is proportional to the complex conjugate of the Fourier transform of the reference signal $S^*(\alpha, \beta)$. Referring to Fig. 1(b), if the matched spatial filter described previously is placed at the Fourier plane, and an arbitrary signal $g(x, y)$ is inserted at the input plane, then the complex amplitude of the light leaving the filter plane is the product of the filter's transmittance function and the input signal's Fourier transform. Plane P_2 is located at the front focal plane of lens L_2 as shown in Fig. 1(b) which processes the light leaving the plane P_2 and produces its Fourier transform in plane P_3. Therefore, the light pattern in the output plane P_3 is proportional to cross-correlation between the input signal $g(x, y)$ and the reference signal $s(x, y)$. If the input signal is equivalent to the reference signal $s(x, y)$, then the autocorrelation of the reference signal is obtained at the output plane.

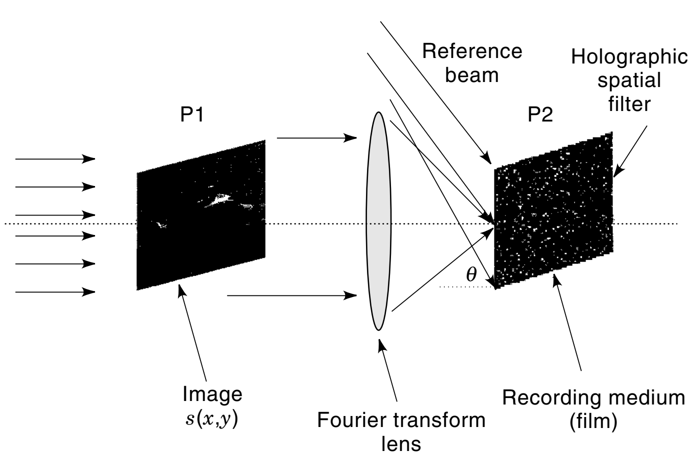


Figure 2. An interferometric holographic technique for synthesizing an optical spatial filter.

SPATIAL LIGHT MODULATORS

Spatial light modulators are very important building blocks of optical information processing systems (1–7). SLMs are input–output devices for real-time optical information processing. The SLM can be considered as an addressable photographic film. They are used in optical computing systems, programmable optical interconnects, optical neural networks, and optical pattern recognition systems. The input to an SLM is either a time-dependent electrical signal or a light distribution such as an image. SLMs with electrical or optical input are called electrically addressed and optically addressed SLMs, respectively. The SLM modulates the amplitude and/or phase or polarization of the read-out light beam as a function of the input signal, which can be an optical image or an electrical signal. For optically addressed SLM, the writing light $A_i(x, y)$ is incident on the input of the SLM. In general, $A_i(x, y)$ is a two-dimensional spatially varying amplitude distribution, imaged onto the input of SLM. The output light distribution is a function of the input light amplitude $A_i(x, y)$. In the following sections, we will explain how SLMs are used in optical information processing.

Various SLMs differ in addressing methods and the modulating materials used. The input light may be converted to an electric field distribution by a photoconductor. The electric field is related to the input light intensity for an optically addressed SLM. The electric field can also be directly applied by using transparent conductive electrodes for an electrically addressed SLM. This electric field modifies the properties of the electrooptic or modulating material. For example, it may change the optical refractive index of the modulating material. The read-out light beam is modulated by the modulating element and reflected back to create the read-out image. Some modulating properties are: (1) the electrooptic effect, (2) molecular alignment by the electric field that exists in liquid crystals, (3) the photorefractive effect, (4) electrostatic deformation, and (5) the acoustooptic effect. The electrooptic effect, which is the change in index of refraction of the medium as a function of the applied electric field, is used in a number of SLMs such as Pockels read-out optical modulator. The SLM is used to perform different functions in optical systems. It is used for incoherent to coherent light conversion for converting real-scenes illuminated under natural or other incoherent light into a coherent image. The real scene is imaged onto an SLM that is read-out by coherent light. The coherent image can then be processed by optical signal processing techniques [see Fig. 3(a,b)]. For real-time Fourier plane spatial filtering, spatial filters can be displayed on SLMs in the Fourier plane. In this case, the input image Fourier transformed by a lens is multiplied by the filter on the Fourier plane SLM. An additional Fourier transformation using a lens will produce the convolution between the input image and the impulse response of the filter [see Fig. 3(a,b)]. This can be used in optical spatial filtering, pattern recognition, and neural networks. If the filter is generated electronically by a computer, an SLM with electrical to optical conversion is used to display the filter at the filter plane.

The SLM can also be used for real-time holography. The interference generated between the object beam and the reference beam can be positioned on the SLM, and the holographic pattern, therefore, can be displayed on the SLM. The SLM can store data or images as well. This is useful for optical memory, optical data base/knowledge base processors, optical pattern recognition, and neural networks.

For nonlinear transformation, SLMs can be used to transform an image nonlinearly, such as binarizing an image. This property is also useful for logic operations and switching in digital optical computing (2,3). In information processing, nonlinear characteristics of the SLM can be used in nonlinear filtering and nonlinear signal processing; see Ref. 10 and Chap. 4 of Ref. 6 for the advantages of nonlinear techniques.

There are many considerations in designing and using an SLM for optical processing. Frame rate determines how fast an image written on the SLM can be updated. Spatial resolution is a measure of how finely an image can be displayed on the SLM. Space bandwidth product is a measure of the number of pixels available (the data throughput per frame). Dynamic range is the number of gray levels that can be represented by a pixel. Contrast ratio, a measure of the ability of an SLM to block the light, is the ratio of the maximum and minimum output light levels. Flatness of the mirrors or windows of the SLM to a fraction of a wavelength of the light is important for optical processing where preservation of the precise phase information is critical. Nonlinear input–output characteristics of the SLM are often considered for the specific image processing applications. The exposure sensitivity and read-out light efficiency define the light budget of the system. SLMs also have write-in and read-out wavelength range and electrical driving signal power requirements.

An example of an electrically addressed SLM is the liquid crystal display (LCD). Liquid crystal devices (2) are widely used in small television sets, television projectors, and lap-top computers. These displays have been used in the optical signal processing community for the last several years because of their low cost and commercial availability (9). The liquid crystal displays used in liquid crystal TVs were not originally designed for coherent optical systems. Their optical quality is not ideal for a coherent system mainly because of the phase variation of materials and the surface of devices, which are less critical in an incoherent optical system than a coherent one. However, recent experiments show that the liquid crystal TV is still a good device in the applications where cost is an important factor and an electrically addressed device is needed.

The liquid crystal display in a liquid crystal TV consists of a 90° twisted liquid crystal layer sandwiched between two polarizers that have parallel polarization direction. The image may be converted to an electric field using a detector array. The electric field is applied to the liquid crystal by using two transparent conductive electrodes on the two sides of the liquid crystal layer. The transparent electrodes are pixelated and can be electrically addressed. When there is no electric field applied, the orientation of the input light is rotated by 90° from one side of the liquid crystal layer to the other side and results in no light passing through because the two polarizers are parallel. When an electric field is applied, the twist and the tilt of the liquid crystal molecules are altered depending on the voltage across the liquid crystal layer. As a result, a fraction of the light passing through the liquid crystal layer will retain the same polarization as the input light and, therefore, passes through the second polarizer. The fraction of the light that passes through the display is proportional to the voltage applied to the liquid crystal layer. Liquid crystal displays used in projector-type liquid crystal TVs usually have about 1000×1000 pixels.

A variety of other optical materials can be used for information processing. Photorefractive materials can store optical images using variations in the index of refraction through the electrooptic effect (1,2,6, and 10). A photorefractive material, upon exposure to a light beam or an image, produces a spatially dependent electric field that changes the index of refraction of the material through the electrooptic effect. The variations in the index of refraction result in the refraction or

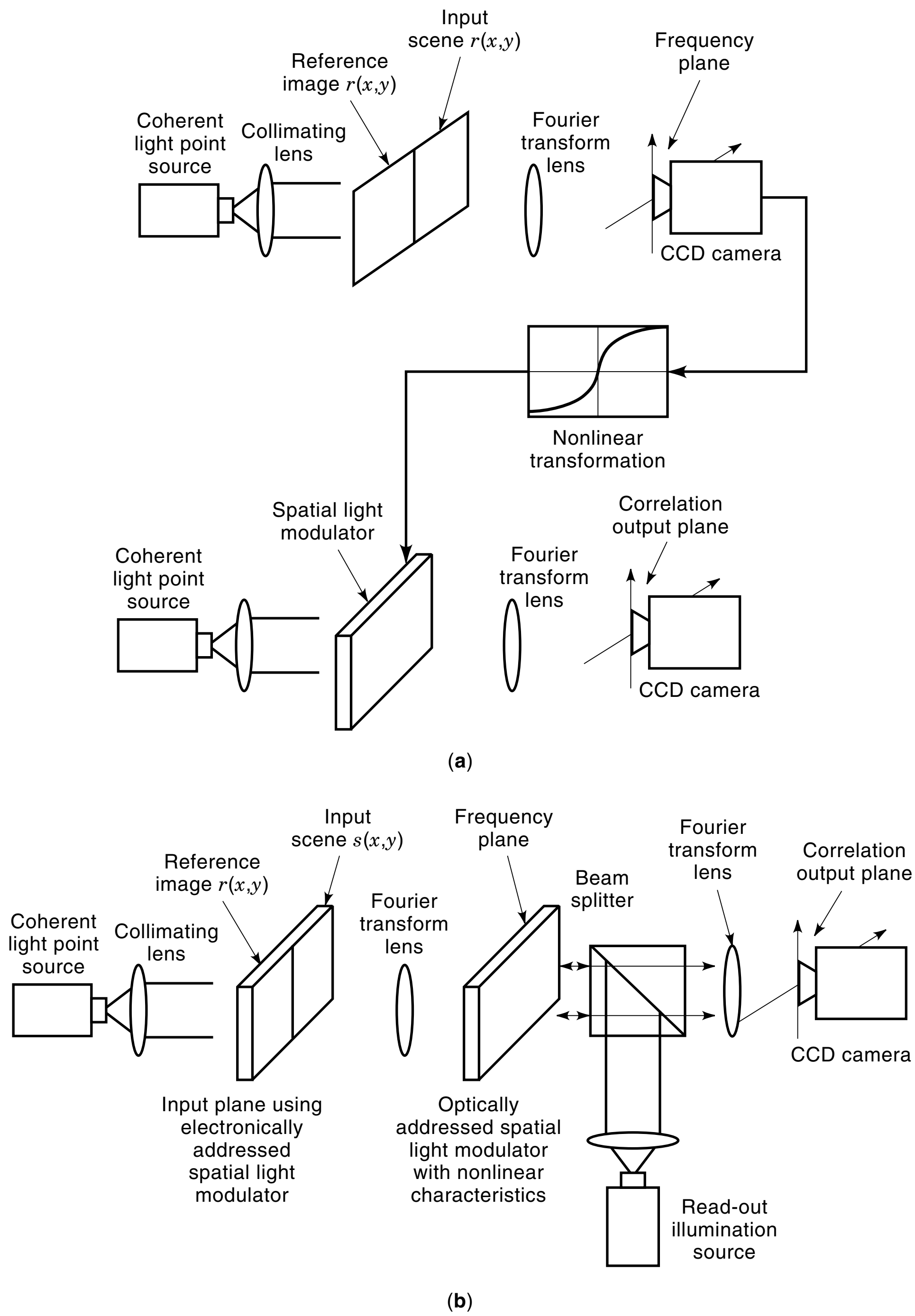


Figure 3. Optical architecture for implementing the nonlinear joint transform correlator for image recognition. (a) An electrically addressed SLM is used in the Fourier domain. (b) An optically addressed SLM is used in the Fourier domain.

diffraction of light. The image stored in the photorefractive material can be read out by an optical beam. For a one-dimensional signal with no applied field, the change in the index of refraction $\Delta n(x)$ as a function of the input intensity $I(x)$ is $\Delta n(x) = -K\ \Delta I(x)/I(x)$, where K is a constant dependent on the electrooptic coefficient, refractive index of the material, temperature, and electron mobility. Photorefractive materials are used in optical storage and memory, real-time optical information processing, neural networks, holography, distortion compensation, and phase conjugation.

Photorefractive devices can be used to generate spatial filters or holograms in real time. An image $I(x, y)$ is spatially mixed with a reference beam, and their interference intensity is recorded in a photorefractive device. The interference intensity changes the refractive index $\Delta n(x, y)$, which is stored in the form of a volume phase hologram. When the device is illuminated by the reference wave, the object beam $I(x, y)$ is reconstructed. For spatial filtering, the Fourier transform of $I(x, y)$ is stored in the device as a filter function.

So far we have been discussing two-dimensional spatial optical processors that use spatial light modulators to modulate the information of the light beam. Another class of optical processing spatial systems exist. They are basically one-dimensional and use ultrasound or acoustooptical principles to perform signal processing of temporal data (11,13). Acoustooptic cells can be used for data processing as well as light deflection and scanning by setting up the grating structure inside the acoustooptic cell.

OPTICAL PATTERN RECOGNITION

In this section, we briefly discuss some algorithms and systems for optical pattern recognition. Much research has been done to develop optical neural network systems based on optical correlators. We refer the reader to the references cited for more details on optical pattern recognition systems (1,5–7,11–14). The matched filter (15) has extensively been used for optical correlators. It was originally used for extracting radar returns from a noisy background. It provides the optimum theoretical response of a filter when the input signal is corrupted by additive overlapping Gaussian noise. In the derivation of the matched filter, optimum is defined to maximize the signal-to-noise ratio at the sample point, which is defined as the ratio of the output signal peak to the root mean square of the output noise. The definition of "optimum" and the fact that the noise overlaps or blankets the target or the signal is very important (see Ref. 16 and Chap. 1 of Ref. 6). If different criteria are used, the matched filter is no longer optimum. In many pattern recognition applications, however, the input scene noise does not overlap the target (sometimes called disjoint noise). It means that the target is in the foreground and blocks the scene noise. For this class of problems, the matched filter and the optimum filter derived under the overlapping input target and scene noise assumption may not perform well (see Chap. 1 of Ref. 6). Recently, algorithms have been developed for target tracking, which contains a target in the presence of noise and includes nonoverlapping scene noise as well as additive noise on the target. The location of the target in the input scene is unknown, and it varies randomly. One solution for this problem is to use multiple hypothesis testing to design an optimum receiver for the disjoint input target and scene noise (see Chap. 1 of Ref. 6). It is shown that for a noise-free target, the optimum receiver is similar to a correlator normalized by the input scene energy within the target window. In addition, given that the target is noise-free and the scene noise probability density function is bounded, then the actual scene noise statistics becomes irrelevant to the detection process.

Another solution is the optimum filter approach (Chap. 1 of Ref. 6) for detecting targets in spatially disjoint scene noise. The filter is designed by maximizing a performance metric, peak-to-output energy, which is defined as the ratio of the square of the expected value of the output signal at the target location to the expected value of the average output signal energy. The filter produces a sharp output signal at the target location with a low output noise floor. We provide test results of the optimum filter to show its performance. Three target tanks and two objects (a car and a vehicle) are embedded in white Gaussian-distributed background noise with mean of $m_b = 0.4$ and standard deviation of $\sigma_b = 0.3$ [Fig. 4(a)]. Target Tank 1 is identical to the reference tank used in the filter design. Target Tank 2 is rotated by 4°. Target Tank 3 is scaled up by 10%. The additive noise parameters are chosen with mean $m_r = 0$ and standard deviation $\sigma_r = 0.2$ in the filter design. The optimum filter output is plotted in Fig. 4(b) and

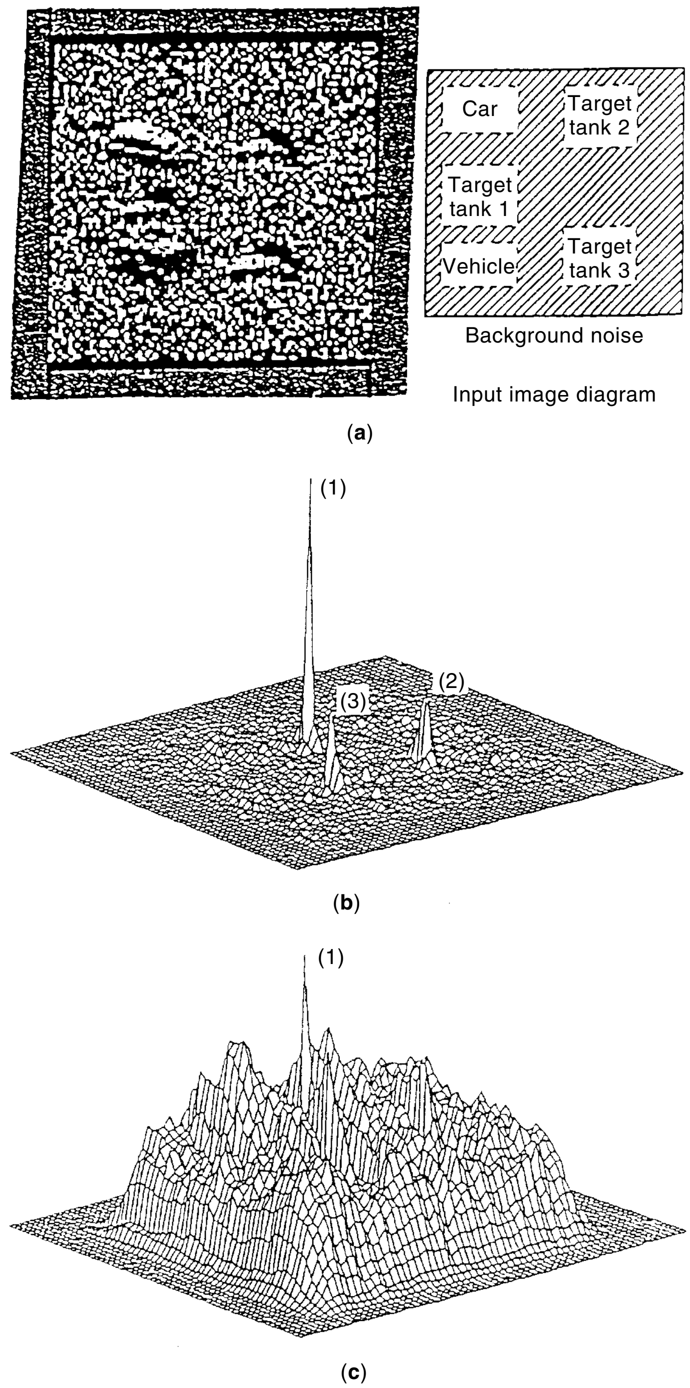


Figure 4. The performance of the optimum filter for detecting noisy targets in background noise. (a) The input scene where three target tanks and two objects (a car and a vehicle) are embedded in white Gaussian-distributed background noise with mean of $m_b = 0.4$ and standard deviation of $\sigma_b = 0.3$. Target Tank 1 is identical to the reference tank used in the filter design. Target Tank 2 is rotated by 4°. Target Tank 3 is scaled up by 10%. The noise added to the targets has mean $m_r = 0$ and standard deviation $\sigma_r = 0.2$. (b) Correlation output of the optimum filter. (c) Output of the matched filter, which fails to detect the target.

compared with the output of the conventional matched filter in Fig. 4(c).

Another architecture for correlation of objects is the joint transform correlator (JTC) (1) as shown in Fig. 3. The reference function $r(x, y)$ and an unknown input object $s(x, y)$ are presented together in the input plane, and their combined or joint Fourier transform is produced in the focal plane behind the first lens. If the joint Fourier transform is recorded on a photosensitive detector such as photographic film and a second Fourier transform is taken, a correlation of the two objects can be realized. The main advantage of the joint transform correlator is that both the input signal and the reference signal are Fourier transformed simultaneously, and the interference between the transforms is achieved in one single step. The input images can be displayed on an SLM for a real-time operation. The JTC is less sensitive to alignments than the standard correlator described earlier.

The implementation of the joint transform correlator using a spatial light modulator (17) is shown in Fig. 3 (see Chap. 4 of Ref. 6). Plane P1 is the input plane that contains the reference signal $r(x + x_0, y)$ and the input signal $s(x - x_0, y)$. The amplitude of the light distribution at the back focal plane of the transform lens FTL_1 is the sum of the Fourier transforms of the two input image functions. We denote $S(\alpha, \beta) \exp[j\Phi_S(\alpha, \beta)]$ and $R(\alpha, \beta) \exp[j\Phi_R(\alpha, \beta)]$ as the Fourier transforms of the input and reference signals $s(x, y)$ and $r(x, y)$, respectively. The Fourier transforms interference intensity distribution at plane P2 is obtained using an optical sensor such as a detector array [see Fig. 3(a)], or an SLM [see Fig. 3(b)], and it includes the cross-power spectrum of the input signals. For the linear or classical joint transform correlator, the inverse Fourier transform (or the Fourier transform with coordinates reversed) of the Fourier transform interference intensities will produce the correlation signals at the output plane. More recently, nonlinearities were introduced into the joint transform correlator. The binary joint transform correlator is obtained by binarizing the joint power spectrum into two values (see Chap. 4 of Ref. 6). It has been shown that, in terms of discrimination, a binary joint transform correlator has superior performance compared with that of the conventional linear joint transform correlator. The binary joint transform correlator was generalized to form a family of correlators called the kth law nonlinear joint transform correlators, which includes the conventional joint transform correlator where $k = 1$ and the binary joint transform correlator where $k = 0$. Here, k represents the severity of the nonlinearity of the transformation of $\mathrm{sgn}(E_m)|E_m|^k$, where E_m is the modified joint power spectrum ($E_m = E - S^2 - R^2$), E is the joint power spectrum, and $\mathrm{sgn}(\cdot)$ is the signum function. Theoretical and experimental studies have shown that the nonlinear joint transform correlator can produce very good correlation performance. The nonlinear joint transform correlator can use the nonlinearity of a nonlinear device such as an SLM at the Fourier plane to alter the Fourier transform interference intensity. It has been shown that when compared with the classical correlator, the compression type of nonlinear joint transform correlator ($k < 1$) provides higher peak intensity, larger peak-to-sidelobe ratio, narrower correlation width, and better correlation sensitivity.

In Fig. 3(b), the joint transform correlator is implemented using an optically addressed SLM, at the Fourier plane. The Fourier transform interference pattern is displayed at the input SLM to obtain the intensity of the Fourier transform interference. For the nonlinear joint transform correlator, the SLM nonlinearly transforms the joint power spectrum according to the nonlinear characteristics of the device.

For a kth law nonlinearity, the Fourier transform of the signal $g(E)$ is

$$g(E) = [R(\alpha, \beta) \times S(\alpha, \beta)]^k \exp\{j[\Phi_S(\alpha, \beta) - \Phi_R(\alpha, \beta)]\} \quad (1)$$

and generates the correlation. In Eq. (1), $k = 1$ corresponds to a linear correlator, and $k = 0$ corresponds to a binary nonlinearity. Varying the severity of the nonlinearity k will produce correlation signals with different characteristics. For highly nonlinear transformations (small k), the high spatial frequencies are emphasized and the correlation becomes more sensitive in discrimination.

To allow for target distortion such as rotation and scale variations, a composite reference is synthesized by using a training set of target images. For rotation invariant pattern recognition, the training set includes a number of rotated images of the target. Much research has been conducted on how to perform distortion invariant pattern recognition (1,4,6,13,14). Figure 5 shows the performance of a nonlinear composite filter implemented by the nonlinear joint transform correlator. Figures 5(a,b) show two versions of target, a Mig-29 rotated by 0° and 45°, respectively. Figure 5(c) is a composite filter that was synthesized from 19 training images of a Mig-29 rotated from 0° to 90° in increments of 5°. This composite filter is used as a reference image in the nonlinear joint transform correlator. The input scene shown in Fig. 5(d) contains two rotated versions of a Mig-29 rotated by 60° and 75° and buried in disjoint background noise as well as additive noise. The reference composite filter and the input scene are put alongside in the input plane of the nonlinear joint transform correlator. The nonlinearities are applied in the Fourier domain. The mesh plots of the correlation outputs of the nonlinear joint transform correlator for $k = 0.2$, and $k = 1$ are shown in Fig. 5(e,f), respectively. The nonlinear joint transform correlator detects the two versions of target successfully, whereas the linear joint transform correlator ($k = 1$) fails to detect the targets.

A large number of reference images can be stored on optical memory. The main advantage of optical memory is its parallel access capability, which may prove advantageous over magnetic storage (1,2,18,19). By illuminating an optical memory disk with a broad optical beam, many stored bits can be accessed in parallel. The illuminated bits on the disk are imaged onto the optical processor for processing or detector array and converted into electronic signals to be used by the computer. Given the commercially available optical components and devices, one million channels can be accessed in parallel. The access time to a block of data is a few tens of milliseconds.

The optical disk is very useful to store data as two-dimensional blocks for neural networks and image processing. The parallel access optical memory has numerous applications in image processing, database management, and neural networks. It provides the capability to access large volume of data rapidly. Parallel access optical memory is attractive in neural networks, pattern recognition, and associative memory by recording a large number of reference patterns. The pattern to be inspected/searched is displayed on a spatial light

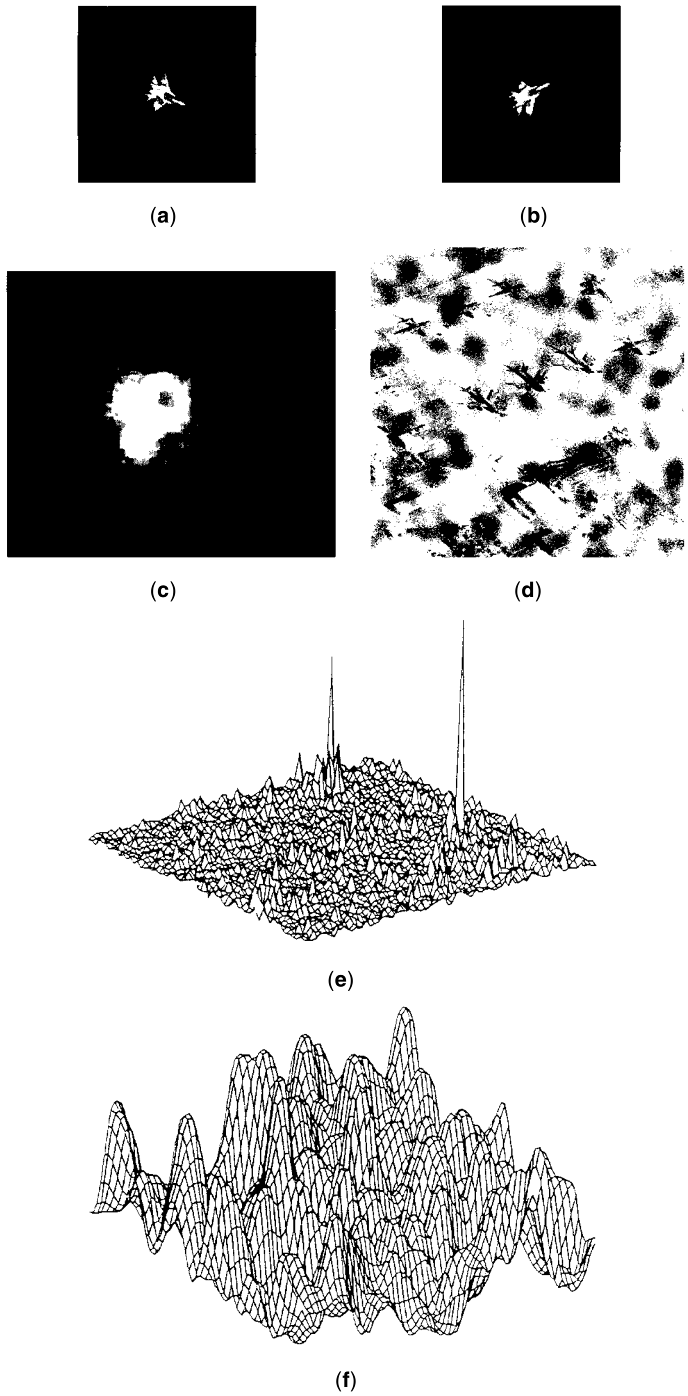


Figure 5. Performance of the nonlinear joint transform correlator for image recognition: (a) Original target, a Mig-29 rotated at 0°, (b) original target, a Mig-29 rotated at 45°, (c) composite reference image, (d) input scene (e) correlation output of the nonlinear joint transform correlator, and (f) output of the matched filter correlator.

modulator to obtain the product between the input pattern and the reference patterns. The product is imaged on a detector, normalized according to the intensity of the input pattern, and is maximized when the input pattern matches the data illuminated on the disk. This process is iterated by rotating the disk and illuminating various portions of the disk to search the entire data to obtain the maximum output. This system is attractive because it provides processing for a large volume of stored data in one disk revolution.

Additional improvements in the storage capacity of optical memory can be accomplished by using thick medium optical disks such as photorefractive materials described in the section entitled "Spartial Light Modulators." The data are stored holographically by interfering the data (such as images) and a plane wave reference beam as described in the section entitled "fundamentals of image processing." The image is retrieved by re-illuminating the hologram on the optical memory with a similar reference beam. It is possible to store multiple images in the same hologram using angular multiplexing, that is changing the angle of the reference beam during hologram recording. Retrieving or accessing the images is achieved by using the proper angle of illumination. The three-dimensional disk is mechanically rotated to scan the entire volume disk. The photoreflective recording materials can provide up to 10^{12} pixels for a 5 cm radius disk. Using acoustooptic deflectors, the access time to each stored data in hologram is 10 μs, which makes the total access time 10 s. Rapid advances have taken place in photorefractive holographic materials such as the commercially available photo polymers and optoelectronic devices, and the need for large storage capacity memory has stimulated much interest in research and development of optical memory. Parallel access and fast data transfer rates seem to be the key to successful applications of optical memory.

OPTICAL NEURAL COMPUTING

Artificial neural networks, which are also referred to as neuromorphic systems, parallel distributed processing models, and connectionist machines, are intended to provide humanlike performance by mimicking biological neural systems (17,20). They are used in image processing, signal processing, and pattern recognition. Neural networks are characterized by massive interconnection of simple computational elements, or nodes, called neurons. Neurons are nonlinear and typically analog and can have a slow response, typically several hundred hertz. A neuron produces an output by nonlinearly transforming a sum of N inputs shown in Fig. 6(a), where $f(\cdot)$ represents the nonlinear characteristics of the neuron and w_i is the weight of the interconnection. Three types of neuron nonlinearity are shown in Fig. 6(b). Neural networks provide many computational benefits. The information is stored in the interconnections. Training or learning changes the interconnection weights w_i.

Because their large degree of parallelism and massive interconnection capability, neural networks provide fault tolerance. Losing a few nodes will not affect the overall performance significantly. Neural networks do not require complete knowledge of the statistical models of the signals to be processed and instead use available training data. Neural systems are best for problems with no clear algorithmic solutions.

Neural networks are characterized by the network topology, neuron input–output characteristics, and learning rules. Learning plays an important role in the performance of the neural networks. The ability to adapt the weights is essential in applications such as pattern recognition where the underlying statistics are not available, and the new inputs are con-

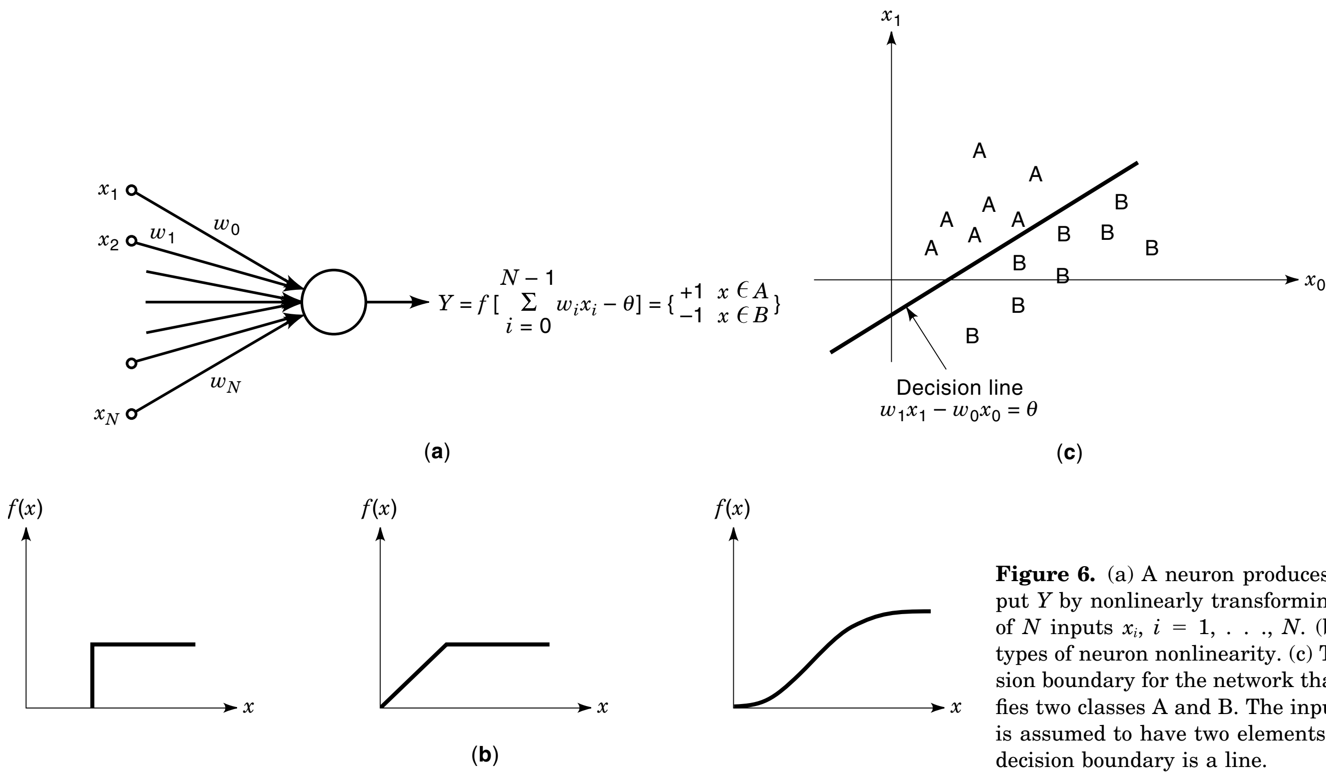


Figure 6. (a) A neuron produces an output Y by nonlinearly transforming a sum of N inputs x_i, $i = 1, \ldots, N$. (b) Three types of neuron nonlinearity. (c) The decision boundary for the network that classifies two classes A and B. The input vector is assumed to have two elements and the decision boundary is a line.

tinuously changing. Conventional statistical techniques are not adaptive and tend to perform poorly when the input changes. For classification, an algorithm is used to compute matching values between the input and the stored data and then to select the class that generates the minimum value. A probabilistic model is used to compute the likelihood or probability that the input belongs to a certain class. If Gaussian distribution is used, tractable solutions can be obtained.

An important application of neural networks is to identify/classify the class of an input pattern when the input is partially obscured or distorted. This has applications in pattern recognition and classification. Neural network classifiers may outperform conventional statistical techniques when the underlying distribution of data is generated by nonlinear processes and is strongly non-Gaussian. Neural network classifiers contain more than one stage. The output of the first stage exhibits the degree of matching between the input and the weights stored in the network. The maximum of these values is enhanced, and the outputs are forwarded to the second stage. This provides a strong output corresponding to the most likely class. If supervision is provided, this information can be used to adapt the weights of the network using a learning rule that will improve the performance of the system by reducing the probability of error.

The perceptron learning rule can be implemented in both single-layer and multilayer networks. Figure 6(a) is a single-layer perceptron with a single output that classifies an input into two classes: A and B. The decision boundary for this network is a hyperplane that divides the space representation of the input. For example, if the input vector has only two elements, the decision boundary is a line [Fig. 6(c)]. The output is the inner product of the inputs and the weights adjusted by a threshold and binarized. An output of $+1$ represents class A, and -1 represents class B. The connection weights can be adapted using the following learning rule: $w_i(t+1) = w_i(t) + \alpha[d(t) - y(t)]x(t)$. Here $d(t)$ is the desired output response that provides maximum separation between class A and class B, that is, $d(t) = +1$ when x belongs to class A, and $d(t) = -1$ when x belongs to class B. $0 < \alpha < 1$ is a positive gain, $x(t)$ is the input, and $w_i(t)$ is the weight. If the inputs from the two classes are separable such that they are on opposite sides of a hyperplane, then the perceptron classifier works successfully by placing the decision boundary between the two classes.

When the classes cannot be represented by hyperplane decision boundaries and are separated by complex decision surfaces, multilayer perceptron is needed. A multilayer perceptron is a feed-forward network that consists of an input layer, an output layer, and as many hidden layers as needed. A two- layer perceptron is shown in Fig. 7. The nonlinearities used within the nodes of the multilayer perceptron provide the capability to generate the complex decision boundaries. In

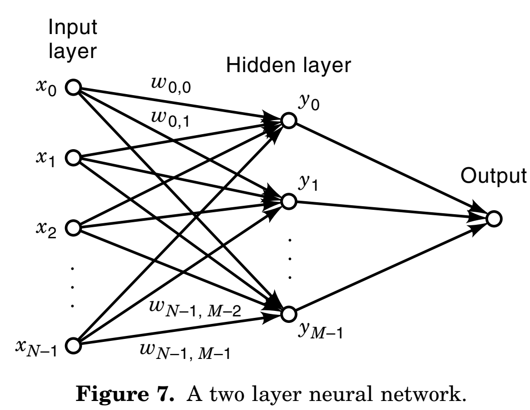


Figure 7. A two layer neural network.

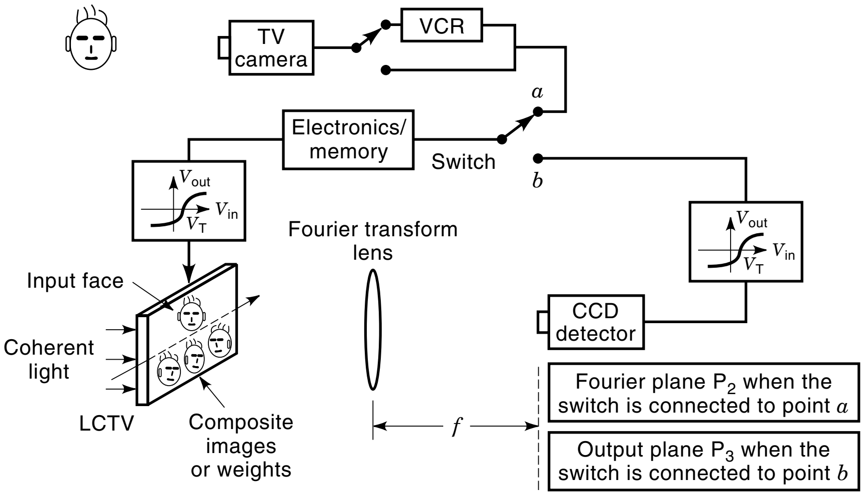


Figure 8. A single SLM nonlinear JTC-based two-layer neural network for pattern recognition. The liquid crystal TV is used to display the input image and the composite images as well as the joint power spectrum by using time multiplexing.

this algorithm, the procedure focuses on the error between the ideal output and the actual output, which represents the overlap between the different classes.

The back-propagation algorithm is used to train the multilayer perceptron. It is an iterative algorithm designed to minimize the mean square error between the desired output and the actual output of a multilayer network. Thus, for each training input, a desired output is specified, and continuous differentiable nonlinearities are used in the network. The actual outputs y_i are calculated using the input, weights, and nonlinearities. The weights are adapted to minimize the mean square difference between the desired output and the actual output. Also the number of the nodes, the number of the hidden layers, and the thresholds need to be set.

Fourier Nonlinear Filter-Based Neural Networks

Figure 8 presents a nonlinear JTC-based optoelectronics neural network associated with a supervised learning algorithm for pattern recognition (21). The system is a two-layer neural network as shown in Fig. 7. The first layer is implemented using a joint transform correlator (please see the section entitled "Optical Pattern Recognition") and the second layer is implemented electronically because of the small number of the hidden layer neurons. The system is trained with a sequence of input images, is able to classify an input in real time, and is easy to implement optically. The system is trained by updating the reference images (weights) in the input that can be stored in electronic or optical memories. The processor uses commercially available optoelectronics devices and can be built as a low-cost compact system. The output of the first layer of a perceptron is the nonlinear correlation between the input pattern and the weights followed by thresholding. The correlation signals are detected by a CCD detector interfaced with electronics to implement the second layer. The nonlinear thresholding is performed electronically to obtain the output neuron. Updating of the network weights can be carried out electronically, and the results are displayed on the input device.

Based on the characteristics of the nonlinear JTC, the proposed system has the following features.

1. It is easy to implement optically and is robust in terms of system alignment.
2. The system can be integrated into a low-cost compact prototype.
3. The system is trained by updating the reference images (weights) in the input which can be stored in electronic or optical memories and no filters or holograms need to be produced.
4. Because nonlinear JTCs use nonlinear transformation in the Fourier plane, the system is robust to illumination variations of the input image and has a good discrimination sensitivity.
5. The system is shift-invariant.

The first layer can be implemented by using a nonlinear JTC. The nonlinear thresholding is performed electronically to obtain the output neuron. In the training of the first layer, the reference images (weight images) are formed by using perceptron learning. The input image is correlated with the stored weight images displayed at the nonlinear JTC input. If the degree of similarity exceeds a threshold and the input image belongs to, say, class C_1, a match is declared, and the input image is discarded. If the degree of similarity is below the threshold and the input image belongs to the correct class C_1, the input image is added into the weights to create a new weight image. For the input images that do not belong to C_1, they are either subtracted from the weights or discarded. Thus, each weight image (or composite image) is formed by the superimposition of a number of images that are selected from the training set. The updated weights are used to test a new input, and the process is repeated until M composite images are formed. When a new input needs to be added to or subtracted from the weights, only the corresponding composite image needs to be retrained. The input is compared with weight images either sequentially or in parallel.

Parallel implementation reduces the training time and requires that all the weights be displayed at the input simultaneously. However, the composite images can be simultaneously displayed at the input plane if they are spatially separated. The number of images that can be spatially separated at the input is determined by the available input space-bandwidth product of the display device and the size of the images. If an SLM such as a liquid crystal television (LCTV) is used to display the weights, it can handle only a limited number of images in each composite image. The number of superimposed images in each composite image is limited by the grayscale LCTV. If the LCTV with 4 bits of grayscale is used, we find that at most, eight images can be superimposed for each composite image. The weights of the second layer can be determined according to the a priori probabilities of the input images.

We should point out that the proposed system discussed earlier classifies the input data to the system as either an image to be accepted or rejected. The network itself is trained only by a class of images which are to be accepted. The training is done in a such a way that the input, which do not belong to the class of images to be accepted, will produce a degree of similarity that will, in general, be lower than a preset threshold.

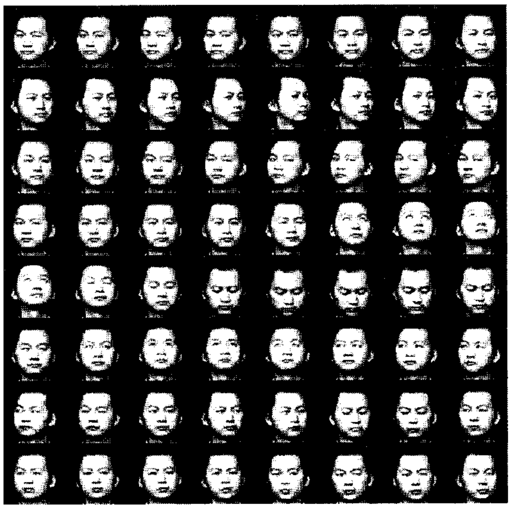
In the tests presented here, we have assumed that the probability of an input image belonging to each composite image is the same. That is, the various distorted input images are equally likely. Thus the weights of the second layer are equal. This is intuitively satisfying in terms of the Bayes' cost, and, as we will show, it produces good experimental results. The outputs of the first layer are nonlinearly transformed and are summed up to produce the second-layer output.

If the input distortions that are used to train the composite images are not equally likely, the weights of the second layer can be adjusted according to the a priori probabilities of the composite images. We present some experiments for facial recognition.

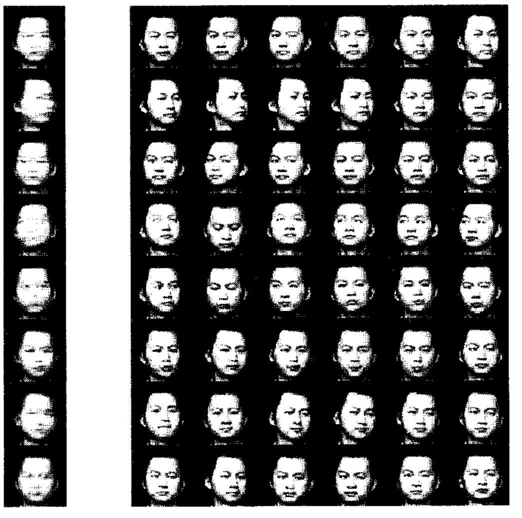
One application of neural network described earlier is in face recognition. Face recognition is an important application of pattern recognition and neural networks. It is, however, a difficult task because the facial appearance is constantly changing as a result of different head perspectives, different illuminations, and different hair styles. Using neural networks is an attractive solution for this problem. For the system described in this section, the facial images are captured by the video camera and stored on the video tape or in the computer. The images used as the testing set of the neural networks are different from those of the training set. The input images are compared with the composite images stored in the database. The comparison is conducted by running the program designed to simulate the nonlinear JTC. Each hidden unit is simulated by the response of the neural networks to the input image when only one composite image is used. The overall output is the response of the system when all the composite images are used when the outputs of the hidden units are added and the sum is passed through a binary thresholding stage. In the training procedure, a total of 128 facial images including different head perspectives and different facial motions are captured. Figure 9(a) shows some examples of various head perspectives used for training. Each image size is 64×64 pixels, which is sufficient to represent the required facial features for pattern recognition. These images are used as a training set. 48 training samples (images) are selected during the training procedure and stored into 3-D composite images with each one having six images to recognize the face of one person. Figure 9(b) shows the selected training samples and the constructed composite images (leftmost column) to be used as weights.

When the training for one person's images is completed, the system is capable of recognizing the distorted input facial image. The same procedure is used to train the system for other facial images. For classification, the input is compared with the composite images corresponding to various people. The output response of the neural network is produced when the output peak intensities (outputs of the first layer) that result from the composite images of a specific person to an input image are summed. The response of the neural network that exceeds the threshold determines whether the person belongs to the class represented by the composite images.

When the system works in conjunction with a badge or password identification system to confirm an identity, it becomes a two-class classification system. When a person wants to pass a security check, he or she enters his or her name or identification number while the camera captures his or her facial image. Based on the name or identification number, the corresponding composite images are compared with the input



(a)



(b)

Figure 9. (a) Examples of various head perspectives used in the training process, (b) selected training samples (right) and composite images (leftmost column) used as the weight functions. Each composite image is constructed by six images shown at right. These composite images are displayed at the input of the nonlinear JTC.

image. If the response of the system exceeds a threshold, the system confirms the identity; otherwise, the input is rejected, and access is denied.

Figure 10 presents computer simulation results for face identification. Figure 10(a) shows the input plane of the system that displays the composite images for image class 1 and an input image of class 1. Class 1 corresponds to the facial images of person 1. The input image shows the person wearing glasses to simulate a distortion not accounted for during the training. The composite images are partially superimposed to make better use of the available input space-bandwidth product. A kth law nonlinear JTC with a nonlinearity $k = 0.3$ is employed. A 256×256 fast Fourier transform subroutine is used in the simulations. The photograph and three-dimensional mesh plot of the first-layer output plane are presented in Fig. 10(b). The first-layer output contains a number of correlation peaks between the composite images and the input image. When the input is person 2, as shown in Fig. 11(a), the first-layer output of the system has a low response, as shown in Fig. 11(b).

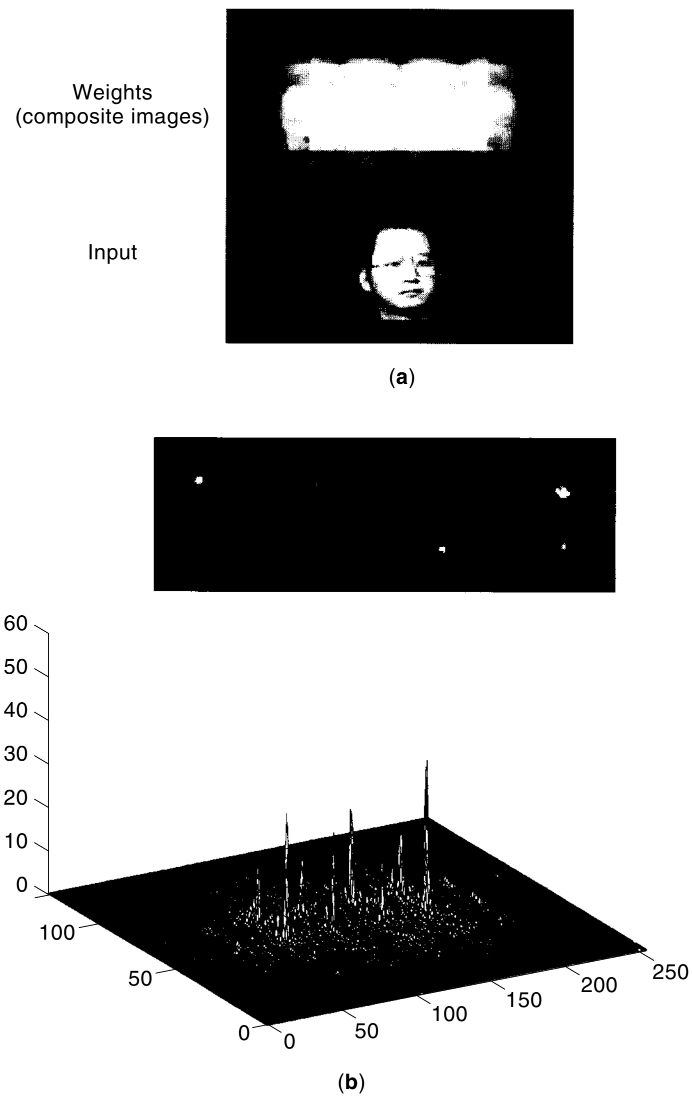


Figure 10. Computer simulations for face recognition. (a) Input plane of the system displaying input images for class 1 and eight composite images for class 1. The composite images are spatially multiplexed when they are displayed next to one another. (b) Output plane of the first layer showing the response to the input image of (a).

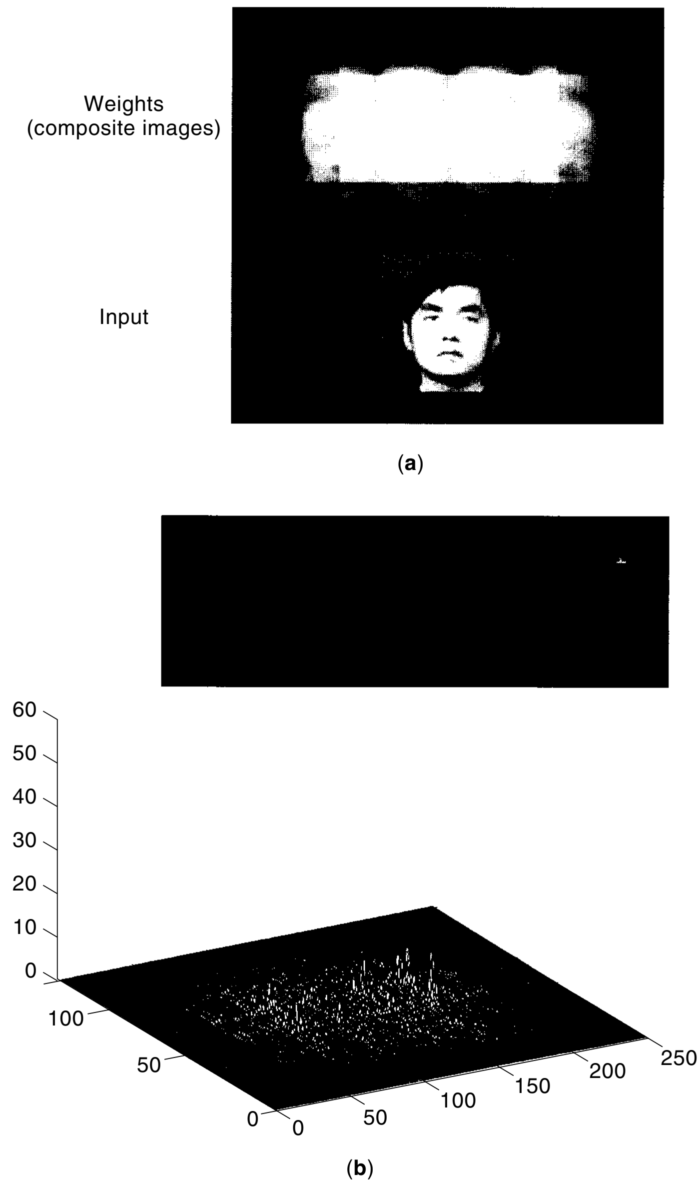
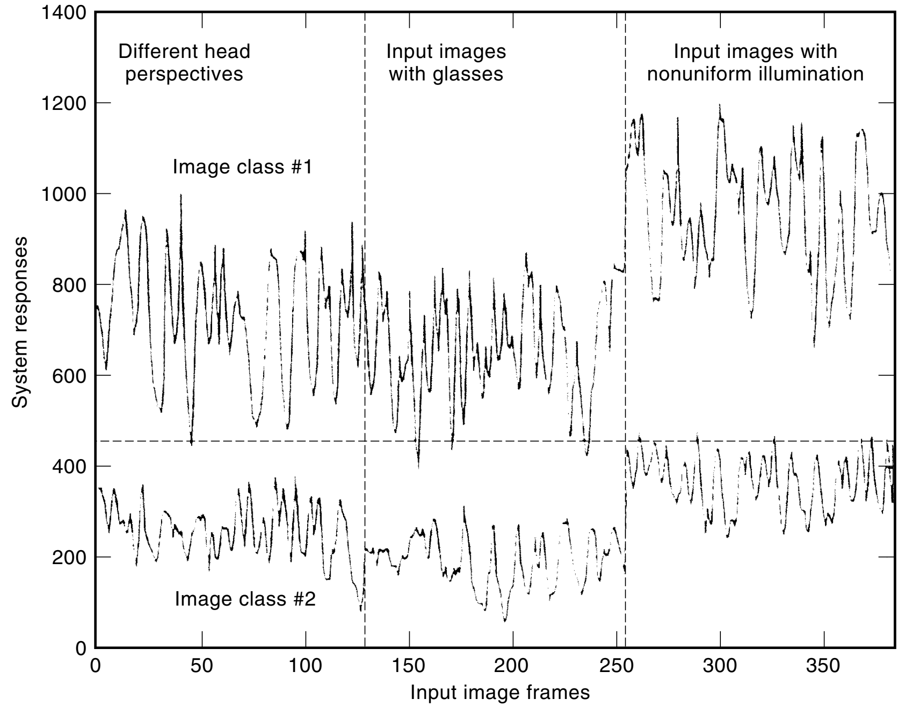
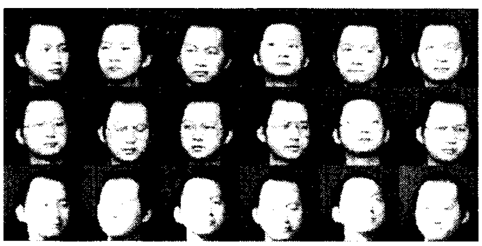


Figure 11. (a) Input plane of the system displaying input image of the class 2 and composite images for class 1; (b) output plane of the first layer showing the response to the input image of (a).

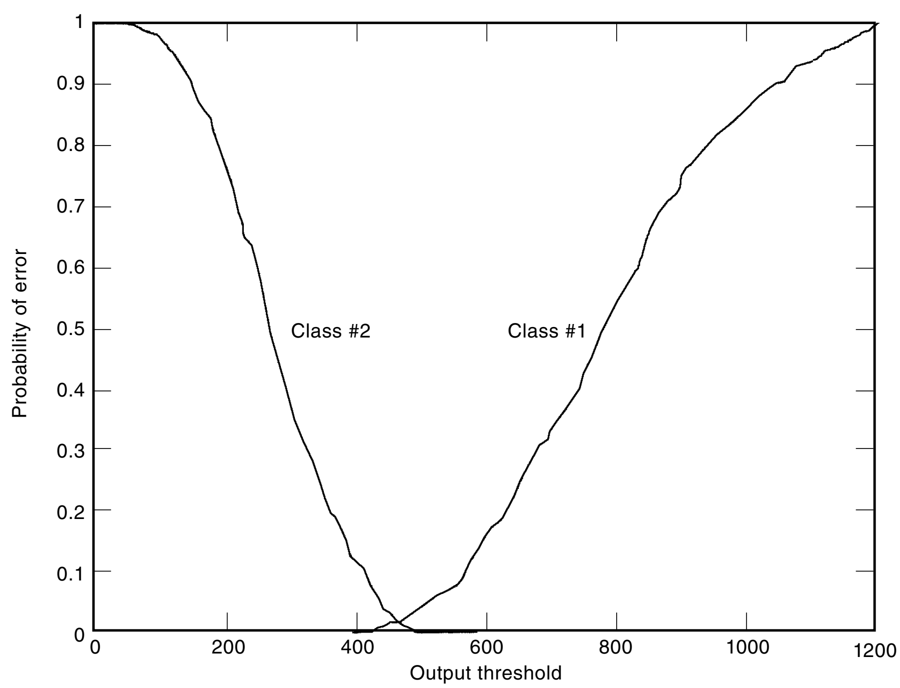
Figure 12(a) illustrates examples of various distorted facial images of class 1 from a testing set used in testing the neural network. Figure 12(b) shows the system response or the output of the second layer in response to the distorted facial images of class 1 and class 2 from the testing set [see Fig. 12(a)] with different head perspectives and various distortions. Class 2 corresponds to the facial images of person 2 in Fig. 11(a). Here the network is trained to recognize class 1, that is, the composite images for class 1 are used as the weights. The first region of the plots is the system responses to input facial images with different head perspectives. The second region of the plots is the system response when the inputs are wearing glasses and have different head perspectives. The



(b)



(a)



(c)

Figure 12. (a) Examples of various distorted facial images of class 1 used as input images in testing the neural network. (b) Neural network response to facial images of class 1 and class 2 with different head perspectives and various distortions. Weights for class 1 are used in the tests. (c) Plot of the error probability versus the output threshold level.

third region corresponds to the case in which the input light is illuminating the face nonuniformly from the top and from the sides. During the training, a uniform input light is illuminating the front of the face. It can be seen that the system is capable of handling the nonuniform illumination and various distortions. The classification of images is dependent on thresholding of the second-layer output. Figure 12(c) presents a plot of the error probability versus the second-layer output threshold level. If we choose a threshold level of 460, the overall error probability of the system is ~2%.

The performance of the system can be improved if time multiplexing of the input images of the same person is used. The output response of the system is determined by examining more than one input image to classify the corresponding person. In the experiments in which time multiplexing of the input image is used, the output response of the system is the

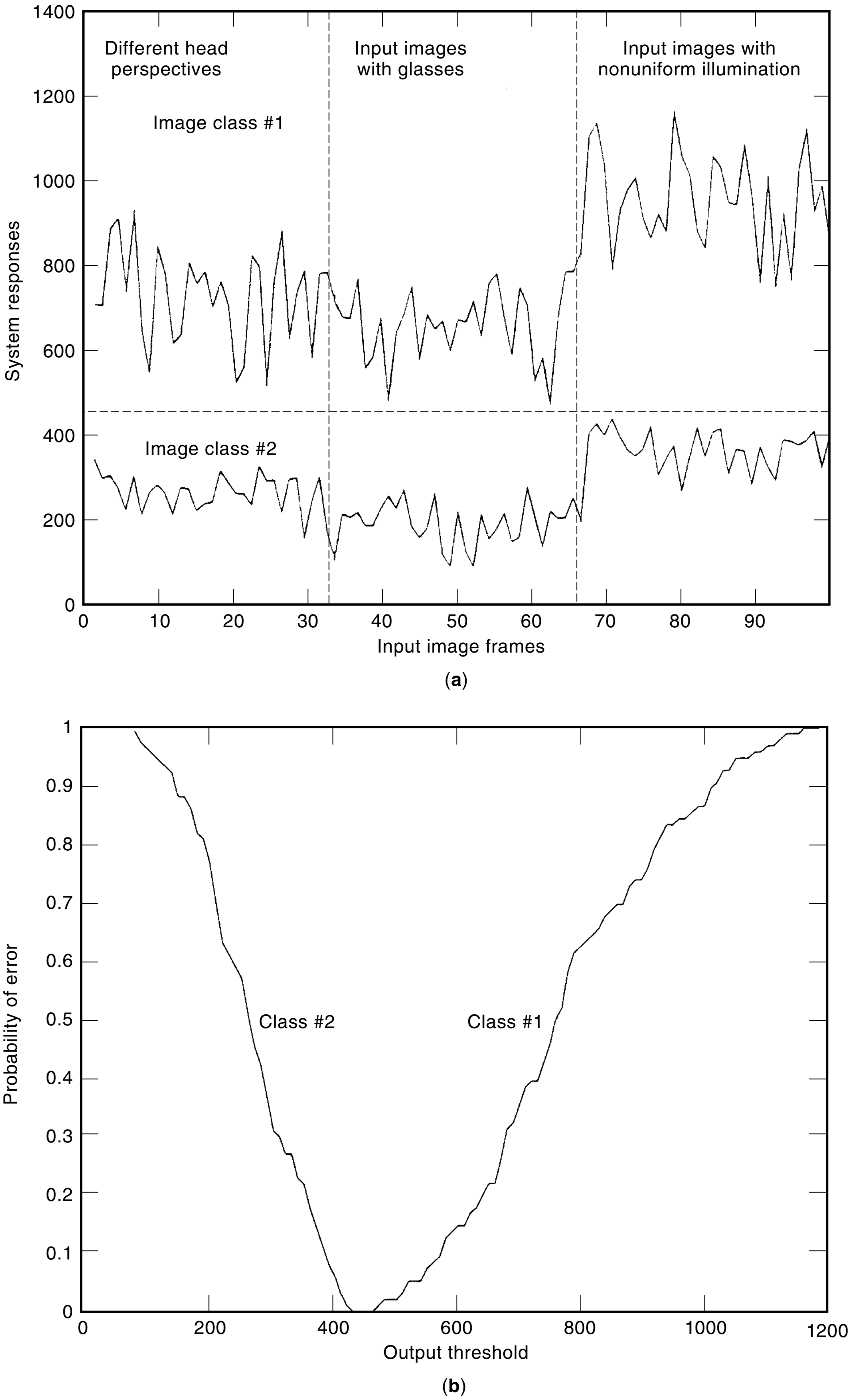


Figure 13. Use of time multiplexing of the input image to reduce the probability of error. Weights for class 1 are used in the tests. (a) System response to the facial image of class 1 and class 2 with different head perspectives and various distortions. (b) Plot of the error probability versus the output threshold level.

average of the system's response to four contiguous distorted input images of the same person. Figure 13(a) shows the system response or output of the second layer in response to facial images of class 1 and class 2 with different head perspectives and various distortions when time multiplexing of the input images is used. Figure 13(b) is the corresponding plot of the probability of error versus the second layer output threshold level. If we choose a threshold level of 460, the overall probability of error of the system when time multiplexing is used is reduced to 0. Once the system is trained with a specific image class, it is capable of recognizing the images in that class and rejecting images from other classes.

Figure 14 shows experimental results for five image classes. In the test, composite images of image class 1 are used as the weights. Figure 14(a) presents examples of input images from five image classes. The leftmost image that is expected to be recognized is from class 1. Figure 14(b) presents the system response to facial images of class 1 and other image classes. Figure 14(c) is the plot of the probability of error versus the second-layer output threshold level when the distorted input images are selected from the five classes. The input distortions are different head perspectives. If we choose a threshold level of 460, even without time multiplexing, the system can classify image class 1 from other image classes with a probability of error of 0.

It would be interesting to compare the performance of the neural network pattern recognition system presented here with a correlator. When only one channel (composite image) is used, the response of the system is a correlation between the input image and the composite image. The composite image is produced by the averaging of the same 48 input images in the training set of class 1 that were used to construct the weight function for the neural network system.

Figure 15(a) is the composite image consisting of 48 reference images. Figure 15(b) presents the correlation between facial images of class 1 and class 2 with different head perspectives and various distortions when the composite image in Fig. 15(a) is used. Figure 15(c) presents a plot of the probability of error versus the output threshold level. The smallest overall probability of error of the system that can be obtained is ~15%, which is significantly larger than the probability of error produced by the neural network system.

The system performance of the face recognition system is assessed by probability of false acceptance and probability of detection. Generally, adjusting the output threshold can vary these probabilities. For example, the threshold can be set such that the probability of false acceptance is very small to prevent impostors, but it will also make it harder to recognize the authorized images. There is no general way to resolve these issues, and the solutions are application dependent. Thresholds can be set using training.

MULTILAYER OPTICAL NEURAL NETS USING HOLOGRAPHIC MEMORY

A handwritten character recognition system was built by Psaltis and Quio (21a). It was realized by a multilayer optical neural network shown in Fig. 16. A rotating mirror was used to change the reference beam in 26 different directions, two LiNbO3 photorefractive crystals were used to represent the input layer and the hidden layer, and a CCD camera was used to represent the output layer. The learning method proposed by Kanerva was used to train the system (22). The weights are initially assigned random values, and are updated with the new inputs. Each input character has 100 pixels, the hidden layer contains 10^5 units, and the output layer has 26 units that represent one of the 26 letters of the alphabet. In tests, 104 patterns were used to train the system and a test set of 520 patterns was used. The error probability was 44%. It is claimed that the system has a processing rate of 10^{12} multiplications per second.

A two-layer network is shown in Fig. 17 (23). The input device to the network is a liquid-crystal TV. The liquid crystal

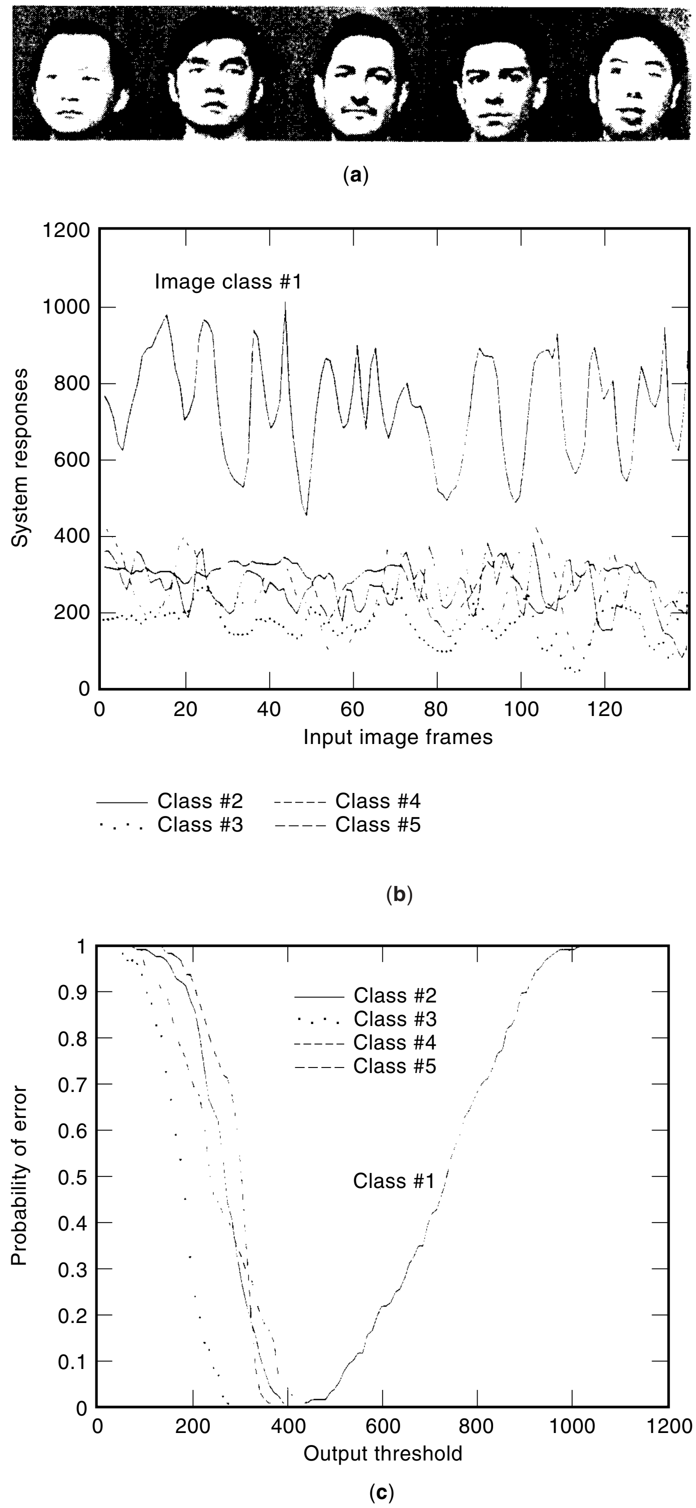


Figure 14. (a) Examples of input images from five images classes. The neural network is programmed to recognize the leftmost image, which is from class 1. (b) System response to facial images of class 1 and other classes. (c) Plot of the error probability versus the output threshold level.

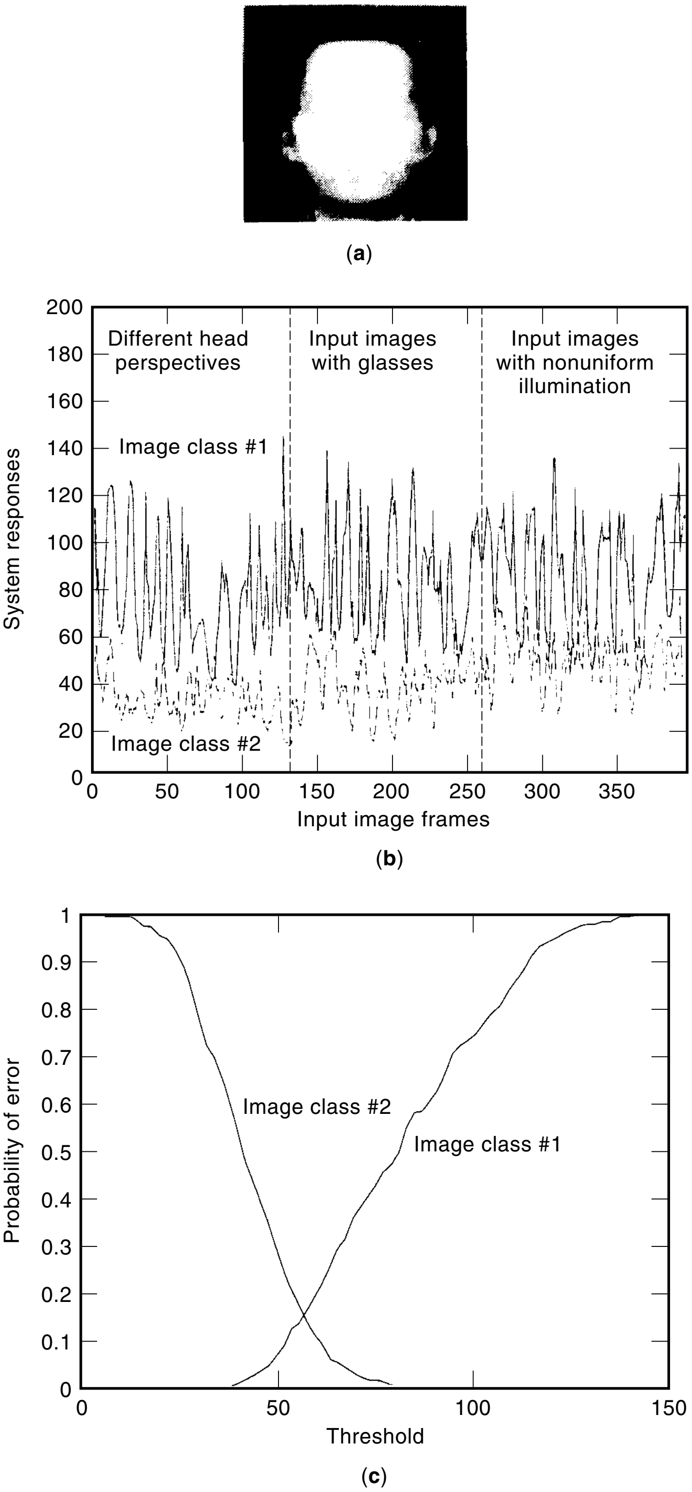


Figure 15. Performance of a pattern recognition system in which a correlator is used instead of a neural network system. Correlation tests are performed with a composite image obtained by averaging 48 input training images of class 1: (a) Composite image, (b) correlator response to facial images of class 1 and class 2 with different head perspectives and various distortions when the composite image in (a) is used, (c) plot of the correlator error probability versus the correlator output threshold level.

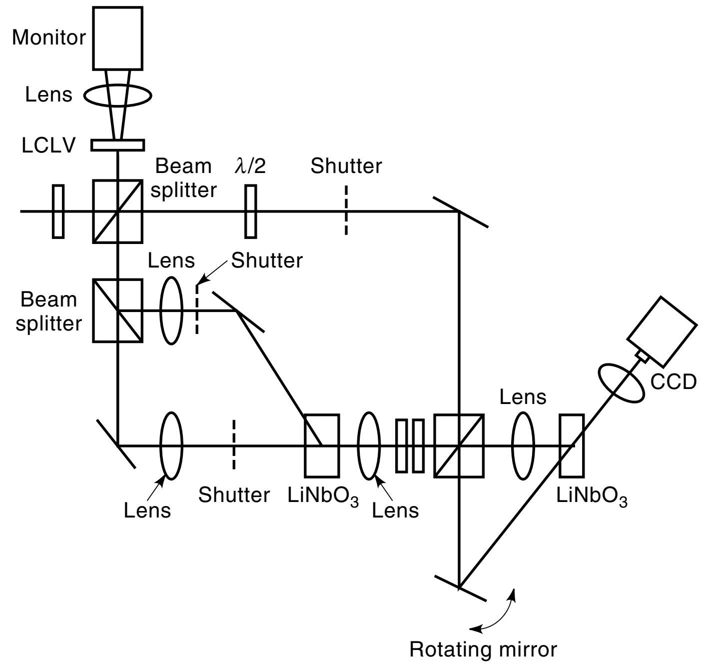


Figure 16. A two-layer optical network for character recognition.

TV is illuminated with collimated light. Lens L_1 produces the Fourier transform of the input image at plane P_2. Not shown in the figure is a filter which blocks the low frequency components of the input image that enhances the edge of the input image and improves the ability of the system to discriminate between inputs from different classes.

A single hologram is recorded in the crystal at a particular angle of the reference beam. Lens L_3 is a Fourier transform lens that produces an image of the edge enhanced input image on CCD for visual assessment. Lens L_2 is also a Fourier transform lens that produces at the output plane P_4 the response of the first layer where it is sensed by a linear detector

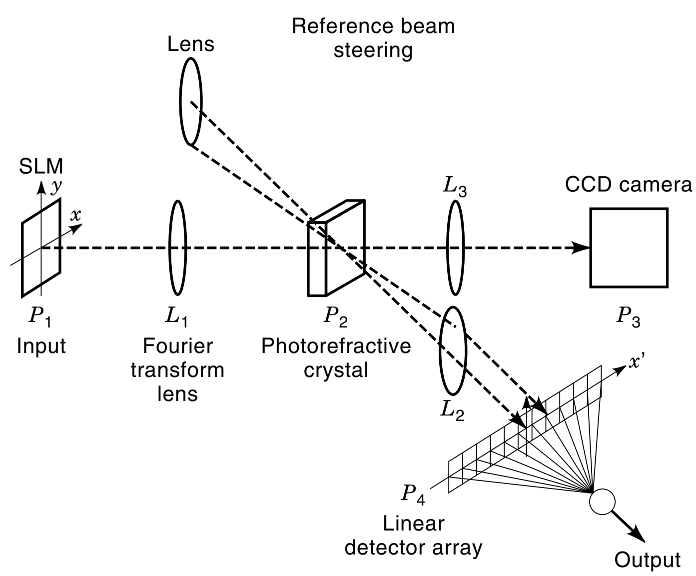


Figure 17. Optical setup of a two-layer photorefractive pattern recognition neural system.

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. The system from the input plane P_1 to P_4 is an array of image correlators. For one filter, the system is a correlator. If we change the angle of the reference beam and record a different hologram at each angle, then the one-dimensional strip of the two-dimensional correlation function will be produced at a different horizontal location. The role of the second layer is to nonlinearly combine the outputs of the correlators and make the final classification.

OPTICAL ASSOCIATIVE MEMORY

An associative memory processor stores signals or patterns in the memory (17,20,24,25). It is capable of producing an output that is a reproduction of the stored input pattern in response to an input that is a partially obscured or distorted version of one of the stored patterns. For one-dimensional signals, the patterns are stored in a matrix W. The output of the associative memory is the input vector multiplied by the matrix W followed by nonlinear transformations. The associative processors include autoassociative memory and heteroassociative memory. In autoassociative memory, the output recollection is the same as the stored pattern associated with the input. If x is one of the stored patterns, and x' is a distorted version of x, then the output is x. An important property of associative memory is fault tolerance, which means being able to produce the original pattern in response to an input that is a noisy or partially obscured version of a stored pattern. There are also applications to content-addressable memory. When the associative memory is used for pattern recognition, the output is compared to the stored signals to determine if there is a match. The heteroassociative memory produces outputs that are arbitrarily associated to a given input x.

The Hopfield associative memory used with binary inputs such as black and white images or binary data does not perform as well when continuous grayscale input values are used. The network contains N nodes with hard limiting nonlinearities $f(\cdot)$ and binary outputs. The output is fed back to the input. The weights t_{ij} are fixed using the M associative signals x^k:

$$t_{ij} = \begin{cases} \sum_{k=0}^{M-1} x_i^k x_j^k \\ 0, \quad \text{for } i = j, j \le M - 1 \end{cases}$$

The output at the time t is μ_i and at time $t + 1$ is

$$\mu_i(t+1) = f\left(\sum_{i=0}^{n-1} t_{ij}\mu_i\right)$$

The input is an unknown pattern x' and $\mu_i(0) = x'$. The process is repeated until the output remains unchanged. The output is forced to match one of the patterns x^k stored in the weight. The network converges to a correct solution if the output is the correct version of the distorted input. Graded nonlinearities improve the performance of the Hopfield network.

The Hopfield network is a single-layer network. It can be used as an autoassociative memory. In this network, neurons accept the input and present the output, and each neuron is connected to all other neurons via the interconnection weights. The weights form a matrix that is called the memory matrix. The matrix is generated by summing all the outer-products of the exemplars, which are used to represent the pattern to be retrieved or recognized, in the learning process (an outer product of two vectors is calculated by the matrix multiplication of one column vector with transpose of the second column vector). The diagonal elements of the matrix are set to zero to prevent the connection of each node to itself. In the recall process, when an unknown input vector is presented to the network, the output is obtained by performing the matrix multiplication of the input vector and the memory matrix, which is a summation of the stored exemplars weighted by the inner-product between the input and the corresponding stored exemplars. The iterations will be repeated until the output converges to a stable state, which is one of the stored exemplars that has the least different bits from the input.

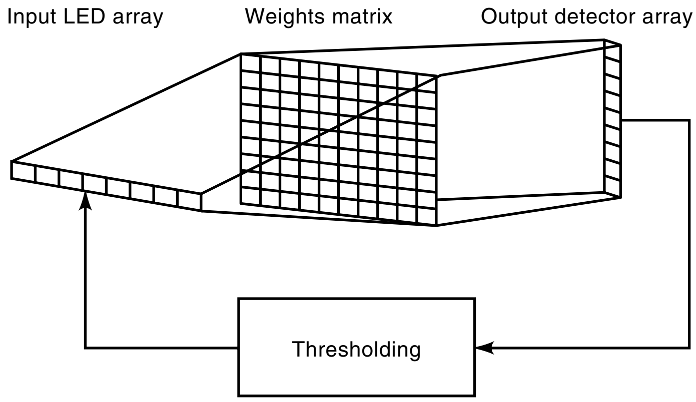


Figure 18. A scheme of an optical vector-matrix multiplier with nonlinear feedback.

The Hopfield network has many major limitations. First, the storage capacity is limited. If many patterns are stored, the network may converge to a false memory pattern different from all stored patterns, which will produce a no-match output when used as a classifier. This problem can be remedied if the patterns are generated randomly and the number of classes M is less than 0.15 times the number of the input nodes N. The second limitation is that the network may not converge to a correct solution if the stored patterns are too similar to one another. In this case, the stored patterns are considered to be unstable. This problem can be remedied by orthogonalizing the patterns before storing them in the network (20).

An optical implementation of the Hopfield network (24) is shown in Fig. 18 (17,24). In the architecture shown in Fig. 18, a light-emitting diode array is used to represent the input array, a photo-diode array is used to represent the output array, and a programmable spatial light modulator is used to store the weights. A lens is used between the input and the weight mask to perform the multiplication of the input and the memory matrix, and another lens is used between the weight mask and the output array to carry out the summation of the multiplication results in the row direction and to generate the output. When an unknown input is imposed on the network, the product of the input and the memory matrix is obtained at the output array, and the output is fed back optically to the input through thresholding and gain.

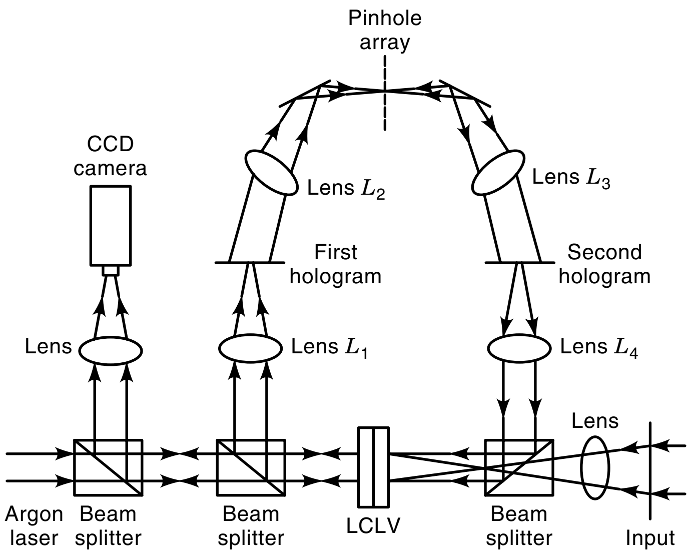


Figure 19. Schematic diagram of a double-correlator neural network with optical loop.

OPTICAL CORRELATOR ASSOCIATIVE MEMORY

Optical correlators are inherently two-dimensional systems. They can perform some operations of two-dimensional images. Psaltis and Farhat proposed using optical correlator to implement the Hopfield model for two-dimensional images (26). An autoassociative content addressable memory (27) using optical correlator is shown in Fig. 19. Originally, the Hopfield network is based on one-dimensional vectors and outer-products of the vectors were used. To implement the outer-product between two-dimensional images, spatial-frequency multiplexing is introduced by using two optical correlators. The first correlator is used to obtain the cross-correlations between the input image and the stored images. The two-dimensional images are stored holographically by means of the Fourier transform hologram. The hologram is recorded on a thermoplastic plate. Each image is recorded with a different reference beam angle, which is called angular multiplexing. In the retrieval process, the input is imaged onto a liquid crystal light valve (LCLV). An argon laser is used to read out the input image from the LCLV. The lens L_1 is used to perform the Fourier transform of the input image. On the back focal plane of lens L_1, the Fourier transform of the input image is multiplied by the transmittance of the hologram that records the complex conjugate of the Fourier transforms of the stored images. The product is then passed through another Fourier transform lens L_2. On the focal plane of lens L_2, a pinhole array is placed to sample the correlation signals between the input image and the stored images recorded on the hologram. The separation between the pinholes is determined by the focal length of L_2 and the reference beam angles used for recording the images on the hologram. In the second correlator architecture, the output light of the pinhole array illuminates another hologram that is similar to the first one. The output at the back focal plane of lens L_4 is a weighted superposition of the images stored in the second hologram, and the weights are proportional to the correlations of the input and the images stored in the first hologram. By doing this, the outer-product model with associative property is realized for two-dimensional images. The same LCLV is used to obtain the output and feed it back into the first correlator to form the iteration loop. The iterations are repeated until the output is stable. A facial image retrieval test with four people images stored on the hologram and with partial and distorted images as inputs was shown in Hsu et al. (27).

Paek et al. proposed a holographic memory (28) shown in Fig. 20. A photorefractive crystal lithium niobate ($LiNbO_3$) is used to store multiple holograms. An argon laser with λ = 514.5 nm is used to record the holograms. A low-threshold, electronically pumped, vertical-cavity, surface-emitting microlaser diode arrays (VCSELs) is used to retrieve the images stored in the holograms. The system has a 10^4 bits/s access speed.

HOLOGRAM-BASED ASSOCIATIVE MEMORY

Owechko et al. proposed an autoassociative memory (29) shown in Fig. 21. This network uses the outer-product neural network model. The desired images are stored in Fourier transform holograms with different reference beams in different angles. A thermoplastic film is used to record the hologram. An input is Fourier transformed by the first lens, and the Fourier transform of the input is multiplied by the hologram that records the conjugate of the Fourier transforms of the stored images. The correlations between the input and the stored images are obtained on the focal back plane of the second lens. A phase conjugate mirror is used to reflect the correlation signal with a conjugated phase back to the hologram. In this way, the hologram acts as a memory that, with the illumination of the correlation signals, generates a weighted superposition of the stored images. This newly formed image is then reflected back by another phase-conjugate mirror as a new input. Thus, the two phase-conjugate mirrors provide a resonator cavity for feedback, and all the stored images oscillate inside the cavity. With the threshold and the nonlinear reflectivity of the phase-conjugate mirror,

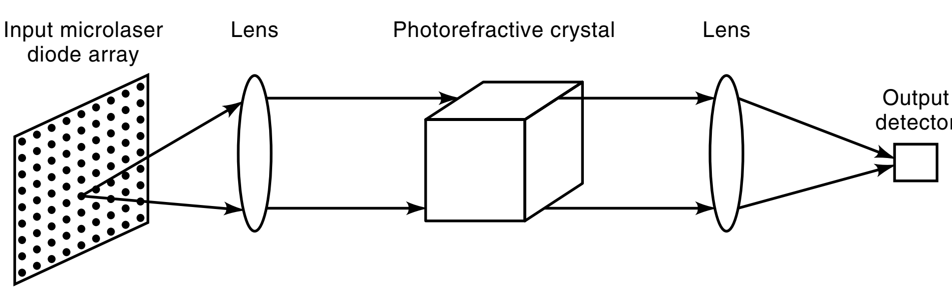


Figure 20. A holographic memory system.

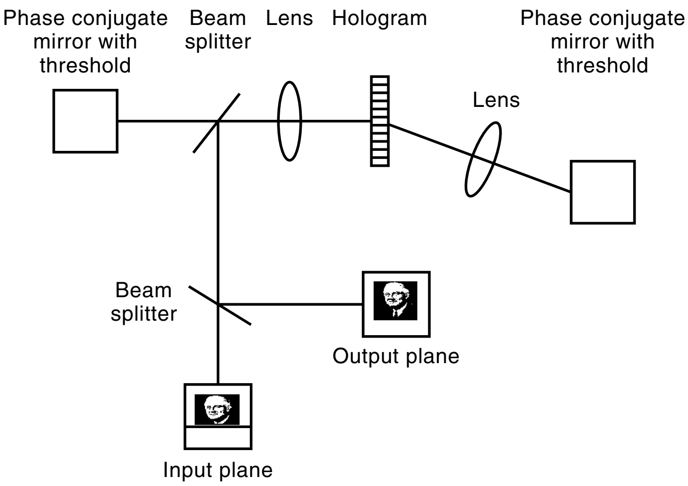


Figure 21. Schematic diagram of a nonlinear holographic associative memory.

the system will converge to the strongest correlated stored image while the other images will vanish.

A single-image experiment is performed by storing a portrait in the hologram and retrieving it from a partial version of the original image.

Paek and Lehmen realized a holographic associative memory capable of identifying individual words and inserting word breaks into a concatenated word string (30). The architecture is quite similar to the two-correlator architecture described earlier, except that electronics was used at the correlation plane to find the correlation peaks and stretch them in the horizontal direction. The stretched correlation peaks were forwarded to the second correlator to restore the desired words with proper space between them.

In a hybrid optoelectronic system (31) shown in Fig. 22, two liquid crystal light valves are used to form the oscillation cavity, and a photorefractive crystal is used to store the volume phase holograms. It is a modification of the system described in the previous section (29). Computers are used to realize the feedback loops. It is claimed that 7×10^5 neurons and 7×10^7 interconnections can be achieved with this system. The neuron update rate can be 10^7 neurons per second, and the data rate can be 2×10^9 interconnects per second.

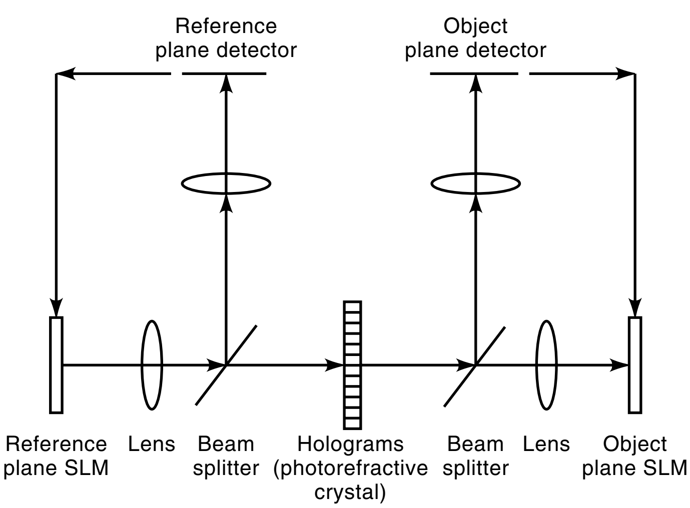


Figure 22. Schematic diagram of a programmable optical neural network.

SUMMARY

This article presents a brief overview of optical information processing systems and devices for developments of neural networks. The field has made significant advances over the last 20 years with the availability of optical input–output devices or spatial light modulators for high-speed information processing, such as commercially available compact liquid crystal display panels with one million pixels. Further improvements are occurring rapidly in the spatial light modulators technology and in new areas of applications, such as the use of optical systems for law enforcement, security, and anticounterfeiting (32). However, progress is needed in developing reliable active optical devices and materials to realize low-cost optical systems.

For more information on the topics discussed in this paper, we refer the reader to the publications of the IEEE Lasers and Electro-Optics Society, the Optical Society of America, and the International Society for Optical Engineering. Each society publishes monthly journals on these topics. For example, the Optical Society of America publishes a separate monthly research journal on information processing alone. We also refer the reader to the proceedings of conferences of these societies on more specialized areas of optical information processing devices and system.

BIBLIOGRAPHY

1. J. W. Goodman, *Introduction to Fourier Optics,* New York: McGraw-Hill, 1968.
2. B. E. A. Saleh, *Fundamentals of Photonics,* New York: Wiley, 1991.
3. A. D. McAulay, *Optical Computer Architecture,* New York: Wiley, 1991.
4. D. Casasent, *Optical Data Processing: Applications,* Berlin: Springer Verlag, 1981.
5. A. VanderLugt, *Optical Signal Processing,* New York: Wiley, 1992.
6. B. Javidi and J. L. Horner, *Real Time Optical Information Processing,* Boston: Academic Press, 1994.
7. J. L. Horner, *Optical Signal Processing,* Boston: Academic Press, 1987.
8. H. J. Caulfield, *Handbook of Optical Holography,* Boston: Academic Press, 1979.
9. B. Javidi and J. L. Horner, Signal processing, optical, *Encyclopedia of Applied Physics,* vol. 18, pp. 71–100, Amer. Inst. Phys., VCH, 1997.
10. A. Yariv and P. Yeh, *Optical Waves in Crystals,* New York: Wiley, 1984.
11. N. Berg and J. Lee, *Acousto-Optic Signal Processing,* New York: Dekker, 1983.
12. D. L. Flannery and J. L. Horner, Fourier optical signal processors, *Proc. IEEE,* **77**: 1511, 1989.
13. D. Psaltis, Two dimensional optical processing using one dimensional input devices, *Proc. IEEE,* **72**: 962, 1984.
14. D. Casasent, Unified synthetic function computation formulation, *Appl. Opt.,* **23**: 1620–1627, 1984. Also see D. Casasent and W. Chang, Correlation sdfs, *Appl. Opt.,* **25**: 1032–1033, 1986.

15. J. L. Turin, An introduction to matched filters, *IEEE Trans. Inf. Theory,* **IT-6**: 311–329, 1960.
16. B. Javidi and J. Wang, Design of filters to detect a noisy target in non-overlapping background noise, *J. Opt. Soc. Amer. A,* **11**: 2604, 1994.
17. Y. S. Abu-mostafa and D. Psaltis, Optical neural computers, *Sci. Amer.,* **256** (3): 66–73, 1987.
18. D. Psaltis and F. Mok, Holographic memories, *Sci. Amer.,* **273** (5): 52–58, 1995.
19. J. F. Heanue, M. C. Bashaw, and L. Hesselink, Volume holographic storage and retrieval of digital data, *Science,* **265**: 749–752, 1994.
20. R. P. Lippmann, An introduction to computing with neural nets, *IEEE ASSP Magazine,* **4** (2): 4–22, 1987.
21. B. Javidi, J. Li, and Q. Tang, Optical implementation of neural networks for face recognition using a nonlinear joint transform correlator, *Appl. Opt.,* **34**: 3950–3962, 1995.

21a. D. Psaltis and Y. Quio, Optical neural networks, *Optics Photonics News,* **1**: 17, 1990.

22. P. Kanerva, in J. S. Denker (ed.), *Neural Networks for Computing,* New York: American Inst. Physics, 1986.
23. H. Y. Li, Y. Qiao, and D. Psaltis, Optical networks for real time face recognition, *Appl. Opt.,* **32**: 5026–5035, 1995.
24. N. H. Farhat et al., Optical implementation of the Hopfield model, *Appl. Opt.,* **24**: 1469, 1985.
25. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA,* **79**: 2554, 1982.
26. D. Psaltis and N. H. Farhat, Optical information processing based on an associative-memory model of neural nets with thresholding and feedback, *Opt. Lett.,* **10**: 98, 1985.
27. K-Y. Hsu, H-Y. Li, and D. Psaltis, Holographic implementation of a fully connected neural network, *Proc. IEEE,* **78**: 1637, 1990.
28. E. G. Paek et al., Compact and ultrafast holographic memory using a surface-emitting microlaser diode array, *Opt. Lett.,* **15**: 341, 1990.
29. Y. Owechko et al., Holographic associative memory with nonlinearities in the correlation domain, *Appl. Opt.,* **26**: 1900, 1987.
30. E. Paek and A. Lehmen, Real-time holographic associative memory for identifying words in a continuous letter string, *Opt. Eng.,* **28**: 519, 1989.
31. Y. Owechko and B. H. Soffer, Programmable multi-layer optical neural networks with asymmetrical interconnection weights, *Proc. IEEE Int. Conf. Neural Netw.,* 1988, p. 385.
32. B. Javidi, Encrypting information with optical technologies, *Phys. Today,* **50** (3): 27–32, 1997.

BAHRAM JAVIDI
University of Connecticut

OPTICAL PARAMETRIC OSCILLATORS. See OPTICAL HARMONIC GENERATION PARAMETRIC DEVICES.