Computer Generated Holograms for Optical Neural Networks

K. Kaikhah & F. Loochan
Department of Computer Science
Southwest Texas State University
kk02@swt.edu

Abstract

While numerous artificial neural network (ANN) models have been electronically implemented and simulated by conventional computers, optical technology provides a far superior mechanism for the implementation of large-scale ANNs. The properties of light make it an ideal carrier of data signals. With optics, very large and high speed neural network architectures are possible. Because light is a predictable phenomenon, it can be described mathematically and its behavior can be simulated by conventional computers. A hologram is in essence a capture of the light field at a particular moment in time and space. Later, the hologram can be used to reconstruct the three dimensional light field carrying optical data. This makes a hologram an ideal medium for capturing, storing, and transmitting data in optical computers, such as optical neural networks (ONNs). Holograms can be created using conventional methods, but they can also be computer generated. In this paper, we will present an overview of optical neural networks, with emphasis on the holographic neural networks. We will take a look at the mathematical basis of holography in terms of the Fresnel Zone Plate and how it can be utilized in making computer generated holograms (CGHs). Finally, we will present various methods of CGH implementation in a two layer holographic ONN.

1. Introduction

A multitude of neural network models have been designed for and simulated by conventional computers and special-purpose computational hardware. They have also been implemented using digital electronic technology. However, most electronic technologies fall short of providing a suitable medium for producing circuits for very high density neural networks composed of millions of neurons with parallel processing architectures. Among other things, a prominent limiting factor for an achievable minimum size for a neural network's electronic circuitry lies in the minimum spatial requirements for physically interconnecting the neurons. The number of interconnections in a fully-connected network is equal to the square of the number of neurons. A two-layer, fully-connected network of 10^4 neurons, for example, would require 10^8 interconnections, which is beyond the state-of-the-art electronic technology. If our goal is to build large-scale artificial neural networks (ANNs) capable of producing a conglomerate of complex functions, we need to find an alternative form of implementation.

1.1 Background in Optics

The solution lies in the optical implementation of ANNs. Optics is the study of the generation, propagation, and detection of electromagnetic radiation in the visible spectrum of light and its practical applications. Light is a predictable phenomenon with definite properties of reflection, refraction, and diffraction, in different mediums. The predictable properties of light can be calculated, using the laws of physics, for any given medium and its geometry. The light may be coherent, as in laser light, or incoherent, as in normal white light. Light can be easily controlled in its intensity and direction, and it can be focused onto microscopically small regions. Thus, light becomes a natural medium for data signals that can be used to build computing mechanisms, including neural networks. While optical technology has been utilized in a broad range of industrial applications, it has particularly been responsible for major advances in the computer industry. From laser CDs to holographic storage devices, from high bandwidth fiber optic lines to optical analog networking, optics has pushed the computer industry to ever greater heights, even while optical computing is still in its infancy.

1.2 Holographic ONNs and the Focus of this Paper

Much progress in the area of optical ANNs has been made in the last decade. This paper serves to provide an overview of the field of optical neural networks (ONNs), focusing on the holographic ONNs. The main component of any holographic network is the hologram. In addition to conventional methods, holograms can also be generated using a computer. The main focus of this paper is the formulation of such computer generated holograms (CGHs) and the mathematical derivation necessary in building an application that will allow generation of CGHs of a single point to as many points that may comprise the object. After exploring the physics of holograms, we will explore the programming issues concerning computer generated holograms, and describe an architecture for the implementation of plane CGH within an ONN.

2. Benefits of Holography

Many of the optical and opto-electronic devices incorporate holography, which provides further advantages and allows model designs that otherwise would be very expensive or altogether not even possible. With the holographic technology at hand, establishing and cloning high order interconnection weight matrices can be an inexpensive operation. In many of the integrated opto-electronic systems where the interconnections are confined to one dimension, interconnections can easily be programmed into the system using holograms in the plane of the waveguide [Feldman, et. al., 1988]. Holograms are also utilized in optically programmed electronic interconnection systems where the interconnections are controlled by photoconductivity. As many as 10^6 interconnection patterns can be stored on holograms within a Page Oriented Holographic Memory and called forth at will to reconfigure the interconnections [Agranat, et. al., 1988]. Also, the read/write ability of volume holograms, implemented with volumetric photorefractive crystals, make volume holograms ideally suited for real-time implementations of neural network models. Another area in which holography plays a significant role is in modular neural networks where an assembly of specialized neural network models compete for dominance in an environment of various problem domain classes [Minsky 1986]. Thus holographic memories used in ONNs architectures show a high degree of reconfigurability and parallelism unmatched by any electronic system.

3. Holography: An Overview

Holograms have been used extensively in the optical implementation of ANNs. It is vital to have a thorough understanding of the principles of holography in order to appreciate their use in various technologies including ONNs. While the mathematical derivations of holograms will be covered in detail in a later section, the objective here is to provide an overview of the general aspects of holography.

3.1 Historical Background

Holography, a technique invented by Gabor in 1948 and further developed by Leith and Upatnieks, makes it possible to record the amplitude as well as the phase of the light field on to a high resolution photographic film [Gabor 1948], [Leith and Upatnieks 1963] [Kock 1975]. A hologram is simply an ultra high resolution photographic image recorded using special techniques. Instead of using ordinary white light as in conventional photography, holography uses coherent laser light to illuminate the object during the photo recording process. The same wavelength laser is necessary to view the recorded hologram.

3.2 Holography Versus 2D Photography

When light reflects off an object it travels outward in a straight line in a non-obtrusive free space. This light carries with it the 3D information defining the object. An observer's eye intercepts this light and from it abstracts 3D information about the object. Holography, in essence, is a technique devised to freeze the travelling light in its path at the film plane. Then during the viewing of the hologram it is as if the original light is released and allowed to continue its journey just as it was prior to the capture. Thus, an observer's eyes viewing the hologram are in essence intercepting the same reflected light propagating through space as if coming directly from the original object just as it would have been at the exact moment of the holographic recording. This is the most important characteristic of holography--the ability to reconstruct the original object light wave. This is possible because in the holographic process the film records the amplitude (intensity) as well as the phase of the light field at the film plane. In particular, the phase information preserves the data about the travelling direction of light beams. This, in turn, pertains to the depth and orientation information, derived by the observer's visual system, of the various points of reflection on the surface of the object.   However, ordinary photographs are different in that they capture only the point intensities (2D information) of a 3D object. Thus, an ordinary photograph is simply a new object itself: a picture--a painted abstraction of an object on paper. This means when we look at an ordinary photograph, we are simply looking at the document describing the object because this light reflecting off the photograph carries to the eye information about the photograph itself. It does not reconstruct the original light waves that carried the object information to the photographic plane at the time of recording. The holographic film on the other hand, records the interference patterns caused by interfering wavefronts of coherent light beams at the film plane. The physics of this phenomenon of wave interference at the holographic recording plane is further illustrated in a later section on the physics of holography. The interference pattern is recorded on the holographic plate as pixel intensity variations forming an ultra high density data matrix with pixel size close to the wavelength of light used. As a result, the recorded information implicitly contains bothamplitude and phase information. Having both, the amplitude as well as the phase, allows the reconstruction of the original light field carrying the object information. When the processed holographic plate is illuminated with a laser equivalent to the reference beam of the same original wavelength, a three-dimensional image of the original object reappears. Actually, what happens is that the reference (reconstruction) beam propagates through the interference gratings in the film and by reverse convolution the original object wave is reconstructed at the holographic plane. This object wave travels on outward from the plate just as it would have at the time of the recording. An observer in the path of this wave can thus view the original 3D object from varying angles of perspective--a phenomenon known as optical parallax. Instead of a planar reference beam, the hologram could just as well be constructed with two different object waves. In that case, when one object wave is reflected onto the hologram, the other object wave can be derived. This way various images can be holographically associated. The great ability of holograms to be able to store vast amounts of information and then recall immediately on demand suggests that a hologram is an ideal medium for neural network applications.

4. ANNs and ONNs

In the recent years there has been a huge influx of interest toward artificial neural network technologies, and optical neural networks have received much attention. Much literature is available to aid in a thorough understanding ANNs [Yu 1993], [Simpson 1992], [Hopfield 1982], as well as ONNs [Yu 1993], [Farhat 1989], [Midwinter and Selviah 1989], [Hsu, et. al., 1988], [Farhat 1987], [Soffer, et. al., 1986], [Farhat, et. al., 1985]. In order to illustrate several key features of optical implementation of neural networks, it is useful to first explore the main features of a typical artificial neural network. With this background, a straightforward transition to ONNs is easily made.

4.1 The Artificial Neural Network

Neural networks are intelligent systems. They are information processing systems that accept inputs and produce outputs. Any neural network is comprised of at least two physical components, namely, the neurons or processing elements (PEs) and the interconnections. All PEs communicate with each other via weighted interconnections. The PEs are configured in layers such as input, output, and hidden layers. Figure 1 illustrates a typical multi-layer neural network.

In addition to the input and the output layers, there can be zero to many hidden layers depending on the type and model of the network. In addition to the input and output layers, hidden layers provide the necessary non-linearity to the solution space required in resolving many multidimensional non-linear problem spaces. The dimensionality (number of PEs) of any layer depends on its use within the problem-solution space of the network. It is important to note that all the PEs of a given layer operate in parallel and process their data synchronously, which in turn implies a simultaneous parallel data transmission across the weighted interconnections between any two layers. The PEs can be interconnected in any desired pattern set forth by the design architecture for a given neural network model.

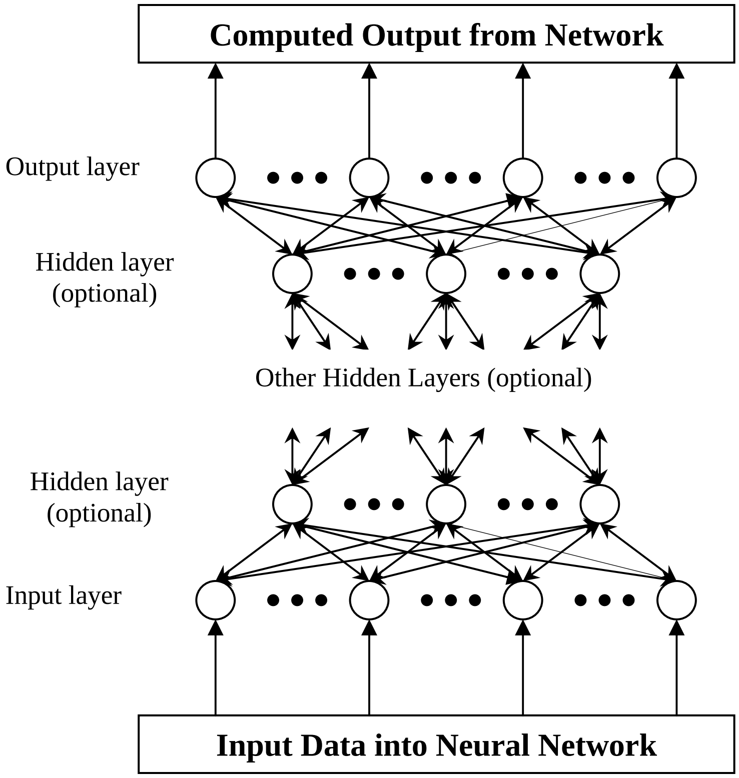
Computed Output from Network
Output layer
Hidden layer (optional)
Other Hidden Layers (optional)
Hidden layer (optional)
Input layer
Input Data into Neural Network

Figure 1. Artificial neural network architecture.

While the mathematical function performed by a PE may depend on the particular neural network model architecture, the basic operation of any PE can be generalized by the representation in Figure 2. Each PE in a given layer collects the values from all of its weighted input connections simultaneously and performs a predefined mathematical operation (typically a dot product followed by a thresholding function), and produces a single output value transmitted simultaneously to all the PEs whose inputs it is connected to.

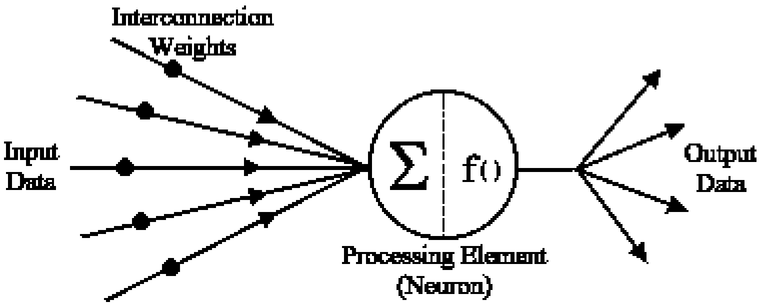

Figure 2. Operation of a neural network processing element (neuron).

In a given state, a layer of PE (e.g., an input layer) in the neural network is an abstraction of a data pattern describing the problem space. The most outstanding characteristic of a neural network is its ability to associate a given input pattern with the desired output pattern. An ANN is able to capture (learn) each distinctive associative mapping of input-output data pattern pairs and later quite efficiently reproduce (recall) the paired output pattern given its associated input. Such systems are called associative networks. One-to-one association is very useful in various problem solutions; however, many solution domains demand the extraction of the non-linear relationships hidden within the complex associations of a problem domain. Such non-linear problem-solution capabilities allow accurate predictions and emerging foresights we so relate to intelligence. ANNs are able to de-coagulate the hidden non-linear relationships within a problem domain and capture such characteristics within their weight-matrix state. This non-linear hyper-dimensional mapping capability can then be utilized by propagating an input pattern existing within the bounds of problem domain to receive its non-linearly associated output pattern.

4.2 Limitations of Electronic Implementation of ANNs

In implementing any neural network, the essential factors to address are how to establish the physical connections between PEs, storage and accessibility of weight values, process for weighting the inputs, the PE function (e.g., dot product computation of weighted inputs) and thresholding, and finally the parallel process synchronization issues. In electronic implementation of a neural network, all connections have to be hardwired, the weights have to be stored separately on electromagnetic storage devices, and all the weighted inputs computed either by another intermediate processor or integrally processed by the neuron (PE) itself. Real-time process synchronization, then, becomes a challenging issue in achieving computational efficiency. All of this, in addition to numerous other factors discussed earlier, poses size-limiting design constraints and produces high order computation latencies.

4.3 Optical and Opto-electronic Implementation of ANNs

Optics offers many advantages in implementing artificial neural networks as opto-electronic systems. The interconnections can be established with focused light beams from one element to another. Light does not suffer from cross-talk, so light beams can, in fact, intersect and superimpose each other without being affected. Thus, numerous connections can be very densely packed. Typically, laser beams are used for such connections and the beam intensity (proportional to amplitude^2) can be analogous to the data value being transmitted. At the transmitting end is a light source element and at the receiving end is a photo-detector that computes the intensity value of the intercepted light beam. This information can then be further processed into the non-linear function and threshold computations via electronic calculating devices. Thus, the most important function that light performs within an optically-implemented neural network is to establish free-space interconnections between processing elements.

Another function of light within the implementation of ONNs results from its wave behavior. Light, as a wave, has both phase and amplitude. It is through the properties of phase and amplitude that light is able to convey information about the object off of which it is reflected, and this information can be captured in a holographic medium and used in the implementation of ONNs. The information-conveying capacity of holograms can play several key roles in optical neural networks.

Light, as data transmitting medium, can be introduced into an opto-electronic neural network as a pattern of light modulated by electrical signals. Spatial Light Modulators (SLMs) are opto-electronic devices comprised of an array of modulators that can be used to insert modulated pattern of light at any layer in the network. While SLMs are ideal in serving as integrated opto-electronic PEs, very high density SLMs can be used for deriving dynamic holograms that serve as interconnection weight filters or associative memories. Most ONN implementations, however, utilize ultra-high-density films as the 2D holographic medium or photorefractive crystals for volumetric holograms [Hsu, et. al., 1988], [Soffer, et. al., 1986].

The use of holographic techniques in the development of optical ANNs has been well documented and numerous implementations have been discussed in the literature [Liu 1994], [Mikaelian 1994]. As discussed earlier, in neural networks communication between neurons is conducted through weighted connections. In ONNs, it is necessary to have a means to store the synaptic weight values and perform weight related calculations somewhere along the light path between the transmitting input element and the receiving output element so as to determine the test for activation potential of the output neuron. A hologram is ideally suited to provide a medium for not only the storage of weight values between each pair of input-output neurons but is inherently able to conduct the necessary calculations for weighted data throughput simply by the light transmitting through the hologram during the reconstruction phase. While the weight values are directly a product of the varying intensities (a property of light associated with the distance from the emitting source), the calculations are performed by the mere holographic reconstruction process of light simply propagating through the recorded interference pattern gratings. In a pure optical neural network, therefore, all operations can theoretically be performed at the speed of light.

Various types of 2D and 3D holographic materials and techniques are used in holographic ONNs. The nature of holographic wiring of the interconnections is only constrained by the optical properties of the material being used as the hologram storage medium. Different materials offer different index changes and mechanical stability and power scattering efficiency. For example, typical volume holographic materials offer a maximum index change of 10^{-2}, which translates into an upper limit of 100 layers (each layer storing an individual 2D hologram) within a crystal of 1 cm thickness by an allowed change in phase thickness of approximately 100 wavelength units for the material [Midwinter and Selviah 1989].

In considering the formation of modifiable optical interconnections, the holographic photorefractive crystals can meet the demand of high storage capacities and very fast response speeds while offering the much needed optical non-linearities between neural layers [Hsu, et. al., 1988], [Hopfield 1982]. While volumetric photorefractive crystals are best suited for volumetric holograms, the readily available two dimensional spatial light modulators (SLMs) are the opto-electronic devices that offer the speed and resolution, as well as the necessary non-linearities mediated by electronic effects. High resolution SLMs are also well suited in implementing holograms into the network.

5. Spatial Light Modulators in Opto-electronic ONNs

In electronics, the processing elements are electronic circuits built out of logical gates. These circuits perform the neuronal calculations of activation functions, including the weight calculation and the output values. The electronic processing elements have to be connected by a charge carrying material, e.g. a wire. In optics, a beam of light is used to connect any two elements. In a true optical ANN, light is emitted and received by elements that are in themselves the processing elements (neurons). The means by which these optical processing elements perform neuronal calculations is inherent within the intensity amplitude and phase properties, as well as the refractive, defractive, and reflective properties, of the light-transmitting and photo-detective materials from which these elements are made. In exploring the opto-electronic implementations of ANNs, it is useful to discuss the most commonly implemented component in optical ANNs, namely, the SLM.

5.1 The Function of an SLM within an ONN

A spatial light modulator (SLM) is simply a device used to modulate the light being transmitted through it. There are numerous types of SLMs designed for a variety of tasks. However, all of the SLMs play a basic role in optical computing, such that they are used to propagate a desired pattern of a laser beam either by passing that laser beam through the SLM or by reflecting the laser beam off of the SLM. Figure 3 is a cross-sectional diagram of a typical SLM utilizing a liquid crystal light modulator and a photoconductive detector sandwiched between transparent electrodes [Midwinter 1989].

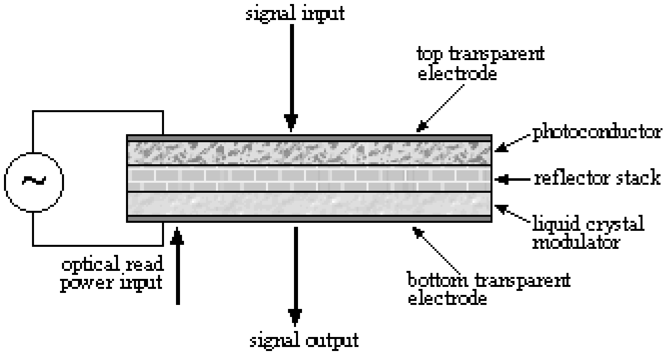

Figure 3. Cross-section of a spatial light modulator.

5.2 Traditional Classification of SLMs

A variety of SLMs are currently available and can be traditionally classified into two types: electrically addressed SLMs (EA-SLMs) and optically addressed SLMs (OA-SLMs).

5.2.1 Electronically-Addressed SLMs

EA-SLMs are used in electro-optic hybrid systems. In general, an EA-SLM receives an electronic signal to modulate the light accordingly. The EA-SLM used in a particular implementation must be chosen to at least match the speed of the independent electrical system from which it receives its signals. For example, an EA-SLM may have to work synchronously in conjuction with a video monitor (LCTV) or LEDs with lenslet array system, a CCD camera or special photo-detector systems, and a computer, each of which may require different EA-SLM response times because these information-relaying devices transmit information at different speeds.

Liquid crystal displays (LCDs) are commonly used as EA-SLMs. Recent advances in thin-film-transistor LCDs (TFT-LCDs) have resulted in high-resolution, high-contrast, full-color or grayscale displays. TFT-LCDs, however, cannot function at speeds much higher than the standard TV video rate. Alternative liquid crystal materials and implementations provide greater speeds. EA-SLMs can also be produced using materials such as an opto-electronic thin film of PLZT or a magneto-optical material like iron garnet. Deformable mirror devices are another type of EA-SLM which are especially promising in achieving high-processing speeds [Ichioka 1996].

5.2.2 Optically-Addressed SLMs

OA-SLMs receive optical signals to modulate the transmitted light. The two most common materials used in OA-SLMs are liquid crystal and photorefractive crystal. Because of their speed, resolution, and contrast, ferroelectric crystal devices have become one of the most important developments in OA-SLMs. OA-SLMs developed for optical parallel processing have been made using photorefractive crystals [Ichioka 1996]. Innovations in OA-SLMs have been the largest advancement in achieving realtime modifiable volumetric holograms used as the main component in very large holographic associative memory neural networks.

5.3 Functional Classification of SLMs

Ichioka, et. al. suggest that as the technology for SLMs develops further, we may want to use a more functional type of classification--input SLMs, output SLMs, and processor SLMs. They also propose a model for an adaptive SLM that can modulate both digital and analog signals. Also an elaborate summary chart comparing many of the state-of-the-art SLMs is given in [Ichioka 1996].

6. A Holographic ONN Architecture

A neural network model may be comprised of several interconnected layers. The principle behind any two layers is the same and is similarly applied to all subsequent layers implemented in a specific ONN architecture. Therefore, a single layer ONN will be the architecture under consideration and will serve as the master base model for any additional ONN layers comprising the system.

6.1 Single Layer Holographic ONN Architecture

Figure 4 shows an ONN architecture with an input layer connected to the output layer through a holographic interconnection scheme. The optical interconnections are achieved with a hologram of spatially variant light objects at varying depths from the output layer abstracted as weighted interconnections between the input and output neurons. Each input neuron is a simple contraption that either blocks (neuron off) the input light beam (same as the coherent collimated reference light beam) or allows it to pass through (neuron on) and subsequently illuminates its associated hologram. The input neuron can therefore be a simple optical shutter or an SLM driven by an externally controlled input mechanism such as a computer. Each output neuron is a fan-in contraption that registers the collective intensities of the various point objects in its projected field of view. A cylindrical lens can be used to collect the light rays and project them collectively onto the output neuron. The output neuron can therefore be a simple photo-intensity sensing device such as a commercial photo-detector specially designed for ONNs or a charged coupled devise (CCD) camera. Finally, the interconnection hologram plane consists of a hologram of points matrix comprising of points at varying depths pertaining to their relative weighted interconnections. However, using a single hologram for the entire set of interconnections may result in a reduced resolution through reverse convolution in the

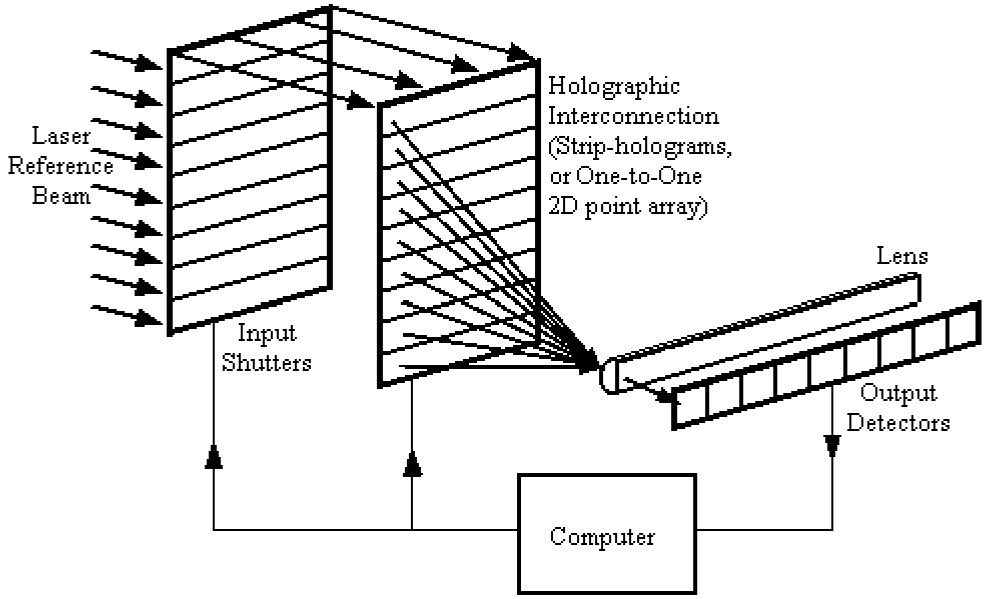

Figure 4. A Holographic ONN architecture where the input layer is connected to the output layer through a holographic array. Each input element is a shutter mechanism (controllable transparency) that may or may not let the reference beam through. The holographic interconnection is either a 1D array of strip holograms or 2D array of point holograms which directs the light onto the output layer detector through a cylindrical lens, thus establishing weighted connections between every input and output elements.

reconstruction phase. As each input neuron is independent of the other neurons in the input layer and in its interconnections to the neurons of the output layer, each input neuron is connected to the output layer via its own hologram hereby called the strip hologram. While the interconnection scheme may be established through a 1D array of one-input-to-many-outputs connection strip holograms (shown in Figure 5a), it is also suitable to implement a 2D array of one-input-to-one-output interconnection single-point holograms for each input-output interconnection (shown in Figure 5b).

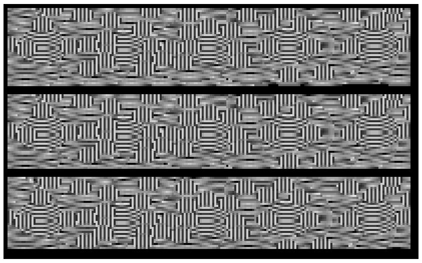

Figure 5a. 3 strip holograms in a 1 x 3 array.

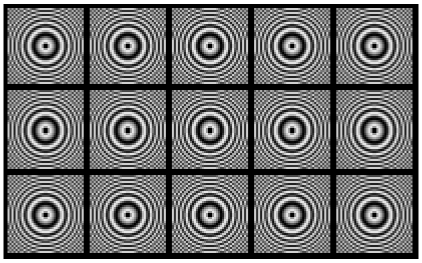

Figure 5b. 15 point holograms in a 5 x 3 array.

6.2 Strip Holograms

A strip hologram interconnects a single input neuron to all the output neurons. Thus a single strip hologram contains object points of varying depths, each of which pertains to the weighted interconnection between the input and the particular output it is connected to. A strip hologram, therefore, will be a hologram of as many points as there are output neurons if the input is fully connected to the output layer. If, however, the synaptic weights of all such connections are zero, then the strip hologram would be a blank hologram (i.e. solid black). For a scenario in which all the connections between an input and all the outputs are weighted equally, the strip hologram would contain object points at equal distances from their respective output neurons.

To create a strip hologram, it is first necessary to evaluate the minimum separation between objects of different input neurons' perspectives based on holographic plate size, diffraction noise created by the SLM (or emulsion), and other optical aberrations caused by the medium being used. These could only be studied for specific physical devices used during the actual experiment. Nonetheless, they are extremely important because they can be the limiting factors for the number of input-output neurons that can be used within the system without superceding an acceptable noise level that does not affect the neural network operations. One way to overcome these limitations is to use a 2D matrix of single point object holograms, one hologram for each one-to-one interconnection. Individual blocks of SLMs would then display each point hologram in a real-time ONN implementation. This method will also help to control the noise that would otherwise be created by light diffracting into adjacent output neurons. The minimum size of each hologram is limited only by the desired degree of optical resolution vs. aberration. However, keeping the hologram size to the minimum, and spacing the output neurons (photo-detectors) so that they are seated exactly opposite their relative interconnection hologram axis, as well as distancing the object point behind the hologram plane at least some distance away (pertaining to the highest weight value) will further help in the noise reduction. The reason noise is reduced is because such a method would be analogous to looking through a tube at the point situated inside the tube at the other end; thus its light rays are prevented from reaching the neighboring output photo-detectors.

6.3 Mutiple Layer Holographic ONN Architecture

To implement a holographic optical neural network of additional consecutively connected layers, the inner layers would be implemented with SLMs, rather than photo-detectors, so that the light received by the output SLM can be modulated to become the input for the next layer. Each layer would have its own holographic interconnection weight matrix also implemented with SLMs. Such an architecture can accommodate any of the ANN algorithms such as those for Backpropagation networks, Hamming nets, Bidirectional associative memories, etc. Special measures would have to be taken, however, to design some of these networks that learn in a feedback process. The feedback could either be handled electronically by computer control or by establishing additional loop backed layers to serve as the feedback network. The important ingredient in any holographic optical network, therefore, is obviously the CGH generation and implementation process, which is described in the next section.

7. Computer Generated Holograms

Numerous methods have been used in synthesizing holograms [Siemens-Wapniarski and Givens 1968], [Waters 1966]. Synthetic binary holograms of nonexistent 3D objects was successfully produced as early as 1966 by James P. Waters by plotting black and white dots on a plotter and using photo-reduction to create the hologram [Waters 1966]. The biggest concern for Waters was to introduce a method that would reduce the computational time in generating the CGH as well as in compensating for the negative to positive transfer process and thus he generated a sampling of only purely opaque or transparent portions comprising his synthetic holograms. However, today's technologies offer much faster computation speeds allowing inexpensive grayscale CGH generations which is the technique implemented in the HoloGen application presented in the next chapter.

7.1 Fresnel Zone Plate Holograms

In 1950, G. L. Rogers pointed out the similarity between a Gabor hologram and a Fresnel Zone Plate (FZP) [Rogers 1950]. Since then computer generated holograms have relied heavily upon the use of FZP holograms of single point objects at various depths to compute a superimposed multi-point CGH of a finite 3D object defined by the multiple points thereof. A digital computer is used to generate the necessary intensity values for each pixel or emulsion grain. These intensity points are then plotted in grayscale on a high definition printer to create the holographic negative of the interference pattern. Then an ultra high density photo-reduction process is employed to create the final positive film of the hologram. Similarly, in real-time optical implementations, the intensity values are used as inputs into the SLM array comprising the hologram [Poon 1993], [Hasmoto 1991]. Recently, similar attempts have been made in successfully demonstrating holographic TV [Macovski 1971], holographic video [Onural, et. al., 1994], and holographic scanner technologies [Poon, et. al., 1996].

It is therefore important to understand the principles of physics that describe holographic aspects employed in generating typical FZP holograms. The following section describes the mathematical derivations necessary in formulating the computer program of a CGH application.

7.2 The Physics of Holography

To be able to construct a CGH, it is necessary to understand the hologram recording process mathematically [Poon 1996], [Banerjee and Poon 1991], [Feitelson 1988], [Lee 1978], [Smith 1969], [Siemens-Wapniarski 1968], [Goodman 1968], [Waters 1966]. Let the holographic recording medium be a two-dimensional plane (x, y), with its center as the origin of a Cartesian (x, y, z) coordinate system. Since any 3D object can be described as a collection of points in 3D space, only a single point object will suffice in understanding the mathematical aspects of the holographic process. The point object, denoted by (x_o, y_o, z_o), can be abstracted as a plane light beam passing through a pinhole aperture. Figure 6 shows a general setup for an on-axis recording of a point object hologram. The setup consists of a collimated laser beam split by a beam splitter into two beams, one of which is deflected onto the pinhole aperture while the other is guided directly onto the hologram plane. The light passing through the aperture forms a diverging spherical wave known as the object wave. The wave projected directly onto the hologram plane is know as the reference (or reconstruction) wave and is denoted by (x_r, y_r, z_r). For an on-axis plane, reference wave z_r is infinity.

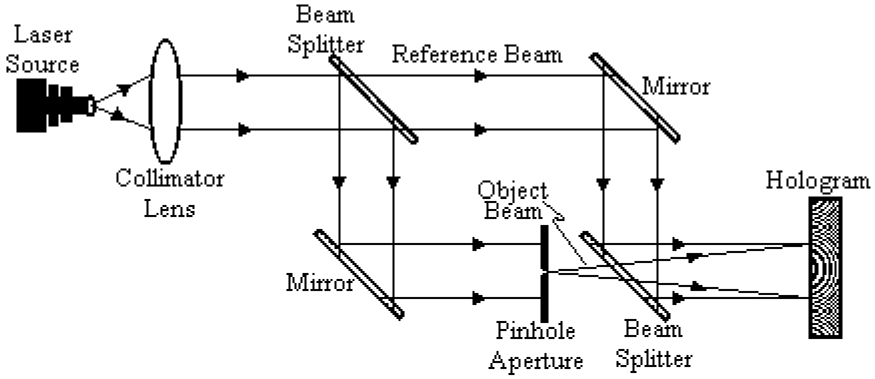

Figure 6. General setup for an on-axis point object hologram production.

The two waves being mutually coherent interfere and form an interference pattern on the hologram plane. The holographic recording of an interference pattern between the mutually coherent spherical object wave and the plane reference wave is illustrated in Figure 7. On the holographic plate, where the two wave fields are in phase, they reinforce, giving rise to constructive interference, and where the two wave fields are out of phase, they cancel, giving rise to destructive interference. At each point (emulsion grain or pixel) on the holographic plate, the wave fields are either added or subtracted algebraically, for an occurring constructive interference or a destructive interference, respectively. These constructive and destructive interferences at the hologram plane form ripple like patterns and are recorded as light brightness or intensity variations on the holographic film emulsion plate. Brightness or intensity of light is directly proportional to the square of the light's amplitude [Kock 1975].

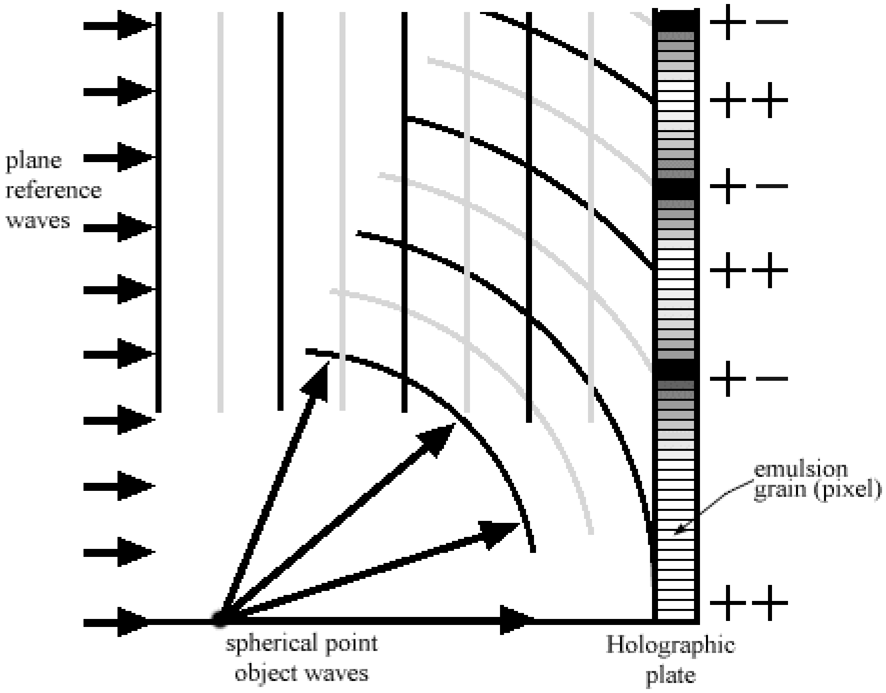

Figure 7. Interference pattern created on the holographic plate by the plane reference wave and the spherical point object wave.

7.3 The Mathematical Basis of CGHs

If O(x, y; z) is the field distribution of the spherical object wave on the hologram plane and R(x, y; z) is the field distribution of the reference plane wave on the hologram plane, then the field intensity distribution being recorded on the hologram is the square modulus of the interference given by:

$$I(x, y) = |O(x, y; z) + R(x, y; z)|^2. \tag{1}$$

The light field in terms of its amplitude and phase at a given plane (x, y; z=0) by a complex function F(x, y), the light field F(x, y; z) at any other plane z distance away can be described as [Goodman 1968], [Yu 1983], [Banerjee and Poon 1991]:

$$F(x, y; z) = F(x, y) * h(x, y; z) \qquad (2)$$

The symbol * in (2) denotes convolution and h(x, y; z) is the free space impulse response given by:

$$h(x, y; z) = \exp [i (2\pi/\lambda) r] \qquad (3)$$

where $r = \mathrm{sqrt}(x^2 + y^2 + z^2)$. Substituting r into (3), we get:

$$h(x, y; z) = \exp [i (2\pi/\lambda) \mathrm{sqrt}(x^2 + y^2 + z^2)] \qquad (4a)$$
$$h(x, y; z) \sim \exp [i (2\pi/\lambda) (x^2 + y^2)/2z] \qquad (4b)$$

The complex function F(x, y), for the point-object-offset at (x_o, y_o) from the origin and at a distance z from the hologram plane, can be described by an offset delta function:

$$F(x, y) = \delta(x - x_o, y - y_o). \qquad (5)$$

Using (4b) and (5), we can rewrite (2) as:

$$F(x, y; z) = \exp [i (2\pi/\lambda) ((x - x_o)^2 + (y - y_o)^2)/2z]. \qquad (6)$$

The point object field O(x, y; z) will be:

$$O(x, y; z) = F(x, y; z) = \exp [i (2\pi/\lambda) ((x - x_o)^2 + (y - y_o)^2)/2z] \qquad (7)$$

while the reference field distribution, in case of the on-axis recording, is uniform over the entire hologram plane and is given as :

$$R(x, y; z) = a \qquad (8)$$

where '*a*' is a constant. For simplicity, *a* is set to 1. Finally, substituting (7) and (8) into(1), we get:

$$\begin{aligned} I(x, y) &= |O(x, y; z) + R(x, y; z)|^2 \\ &= |\exp [i (2\pi/\lambda) ((x - x_o)^2 + (y - y_o)^2)/2z] + a|^2 \\ &= 2 + 2\cos[(2\pi/\lambda) ((x - x_o)^2 + (y - y_o)^2)/2z] \\ &= 2 + 2\cos[(\pi/z\lambda) ((x - x_o)^2 + (y - y_o)^2)] \qquad (9) \end{aligned}$$

The holographic field distribution defined by (9) is called the Fresnel Zone Plate (FZP), with its center displaced at (x_o, y_o) from the center (0, 0) of the holographic plate (x, y). If such a Fresnel hologram is illuminated by an on-axis plane wave (λ), the point image will be created. Figures 8a, 8b, and 8c show examples of on-axis FZP holograms for points (x_o = 0 mm, y_o = 0 mm, z = 5 mm), (x_o = 0.1 mm, y_o = 0.1 mm, z = 5 mm), and (x_o = -0.1 mm, y_o = -0.1, z = 10 mm mm), respectively.

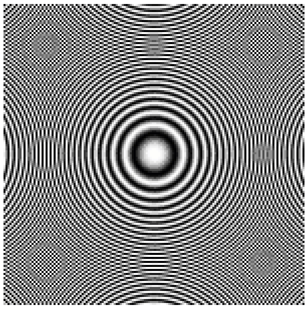

Figure 8a. Hologram of a point at center.

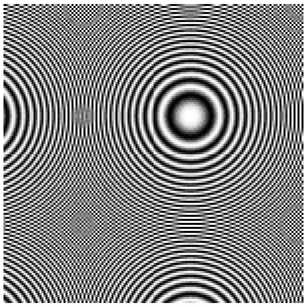

Figure 8b. Hologram of an off-set point.

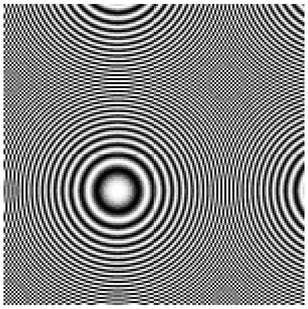

Figure 8c. Hologram of an off-set point.

The FZP of (9) is an on-axis hologram. To derive I(x, y) for an off-axis hologram, where the reference beam is at an angle θ to the hologram plane, the R(x, y; z) is given by:

$$R(x, y; z) = \exp [i (2\pi/\lambda) \text{Sin } \theta x]. \qquad (10)$$

In this case, the hologram I(x, y) is described as:

$$\begin{aligned} I(x, y) &= |O(x, y; z) + R(x, y; z)|^2 \\ &= |\exp [i (2\pi/\lambda) (x^2 + y^2)/2z] + \exp [i (2\pi/\lambda) \text{Sin } \theta x]|^2 \\ &= 2 + 2\cos[(2\pi/\lambda) ((x^2 + y^2)/2z - x \text{ Sin } \theta)] \qquad (11) \end{aligned}$$

Given the parameters in (11), each point (x, y) on the computer generated hologram image can be plotted. I(x, y) would be directly correlated as the white intensity or transparency of the pixel in the image.

For a multi-point object hologram, a FZP hologram for each individual coordinate object point is created separately and the (x, y) pixels of each image are superimposed in constructing the final image. The I(x, y) of the final image is therefore an accumulative average $I(x, y) = \{\text{Sum}[I(x, y)_1 + I(x, y)_2 + \ldots I(x, y)_n]/n\}$ of the *n* points comprising the object. A hologram of any non-existing object can be constructed using this method.

7.4 Advantages and Limitations of CGHs

Computer-generated holography has both advantages and disadvantages over conventional photographic holography in the construction of holograms suitable for ONN architecture. One advantage offered by CGH is the ability to construct non-real objects such as a single point. In implementing ONNs, the lower-bound of the minimum size for an optical processing element is then only confined by the wavelength of the light used, where a single point would be the abstraction of a single particle-wave beam of light.

A second advantage innate to only a CGH process is the ability to defy the physical limitations of conventional holography such as light path obstructions. For example, in conventional photographic on-axis holography, the reference beam and the object beam cannot be incident along the same path normal to the holographic plane without the use of a beam splitter. The use of a beam splitter, however, will often produce unnecessary noise that is absent in the mathematically derived CGH.

The most important disadvantage of a CGH is the immense demand it can make on both memory and computational power. In conventional photographic holography, the film resolution is in the order of about 1500 lines per millimeter. To produce a 10 cm square CGH of this same resolution would require 2.25×10^{10} pixel calculations. For a 256 grayscale composition, this would necessitate 1.8×10^{11} bits, or 22.5 Gigabytes, of memory. This translates further into several folds more of computing cycles that would in most cases take many days of calculations.

7.5 The Implementation of CGHs within ONN architectures

A CGH is in reality nothing more than data. Thus, a CGH, unlike a photographic hologram, can be transmitted through a network, even the Internet. In a world which is, each day, becoming increasingly connected, the ability to transmit data at a very high speed over networks can be a real advantage, especially in the near-future modular network-connected ONN technologies. However, a CGH in itself cannot be implemented into an ONN architecture; the CGH data must be transferred to a photographic hologram or an SLM array. The process of transferring a CGH to film involves creating a printout of the data, from which a photograph, reduced in size, can be made. The resolution of print-based materials is quite low, so the printout that is made must be quite large, even for a hologram of less than a centimeter. The plane of the hologram must be kept flat, so the lens of the camera used to take the photograph must be able to capture the image with no curvature or distortion of the image. Either curvature or distortion introduced into the final holographic image would then require a processing method to readjust for the discrepancy, but such readjustment would be at the expense of compromised data. Another constraint is that most printers are not capable of exact and consistent grayscaling for every pixel, which also results in further noise. A similar noise is again introduced during the photo-reduction process. Beyond the above constraints, to generate a film-based hologram from a CGH is an involved process; it cannot be accomplished in real-time.

The second means by which a CGH can be implemented into an ONN is through the use of SLMs to generate a hologram. The main constraints in using SLMs is that each pixel of the hologram is mapped onto an SLM--an opto-electronic light-modulating device--which is bounded in its minimum size by the available SLM technology [Hasmoto, et. al., 1991], [Poon, et. al., 1993]. The current technology of SLMs is making great strides at obtaining highly compacted resolutions in the order of 300 lp/mm. An up-to-date chart of the current state of the art SLMs is given in [Ichioka 1996]. With SLM technology, CGH generated holographic weight matrices can be implemented in real-time, which is its greatest attribute to the ONN technology.

7.6 Single Point Versus Multiple Point Holograms

The theoretical basis that is used to synthesize a CGH is the ability to mathematically derive the formulae to construct a single point object. In practical holography, however, a hologram of a true point object cannot be physically constructed since a point is dimensionless. Only hologram of a disk with finite diameter can be physically constructed, and in the limit, as the diameter of the disk approaches zero, the intensity distribution is found to be the same as calculated from the relationship in (11) [Waters 1966]. On the other hand, in using the points method for CGH generation, we must depend upon the optical aberrations of the CGH in the reconstruction phase to give points physical dimension. Although this should not pose any problem when considering its implementation within an ONN, it is important however, to have a consistent degree of light sensitivity observable across all photo-detectors used within the architecture. Therefore, it is necessary to establish precise correlation between the range of detectable photo-intensities and the sets of point object holograms to be used as the synaptic weight values for the interconnections in the ONN. The weight values would be proportional to the depth intensity of detected light. The design must entail compatibility among the SLM setups in conjunction with various photo-detectors available on the market for ONN implementations. The following one-point (Figure 9a), four-point (Figure 9b), five-point (Figures 9c-d) and nine-point (Figure 9e) patterns for the CGH base weight models may be used to accommodate the needs of a given design.

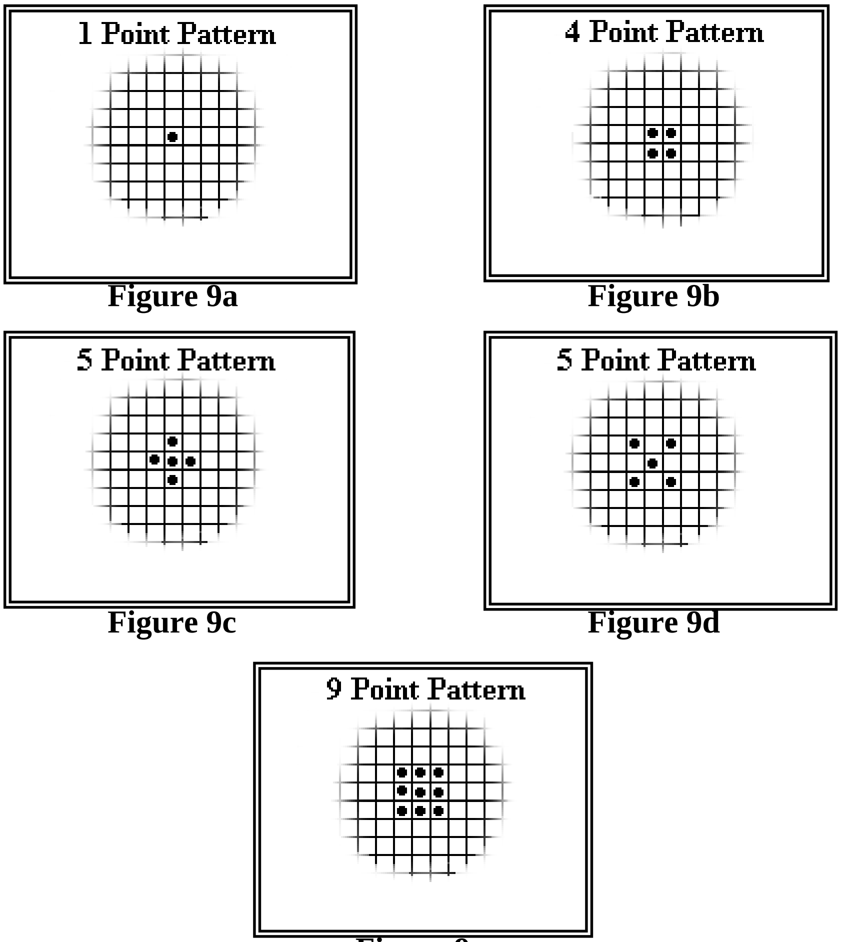

Figure 9a

Figure 9b

Figure 9c

Figure 9d

Figure 9e

Figures 9a-9e show schematic arrangements of various number of points representing a single point object.

Figures 10a - 10e are the holograms generated with the HoloGen application for the patterns of Figures 9a through 9e, respectively, with the samples generated at point object depths of 0.5 mm with an on-axis reference beam angle of zero radians and emulsion resolution of 0.003 mm/pixel.

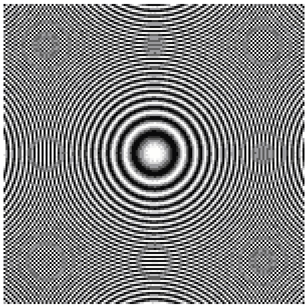

Figure 10a.

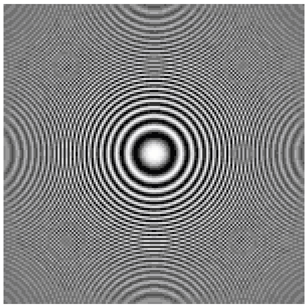

Figure 10b.

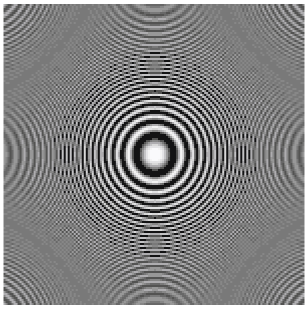

Figure 10c.

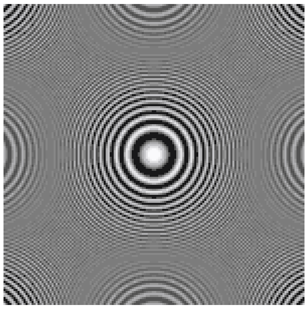

Figure 10d.

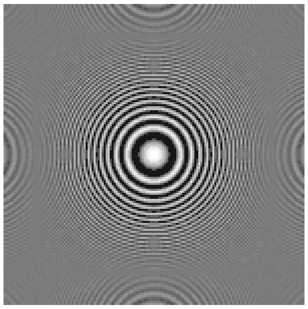

Figure 10e.

Figures 10a-10e show corresponding holograms generated by the HoloGen application for point patterns shown in Figures 9a-9e.

In most ONN implementations, the 9-point pattern of Figure 9e would be the best option for the object type in terms of achieving more stable and accurate intensity readings by the detectors under most given environments, thus allowing a better mapping from intensity to synaptic weights for the interconnections. The 9-point pattern does not impose much higher limitation than the 1-point pattern on a system's minimum size parameters. In all cases of ONN implementation, the distance between the photo-detectors and the interconnection hologram plane (along with any intermediary lens system used to focus the reconstructed object beam onto the photo-detectors) will determine the optimal size limitation for the detector plane. In any case, the observed object dimensions at the detector plane will be much smaller than the size of the object itself due to the depth of field between the object and the detector. Also, as can be seen from the comparison of holograms of Figures 10a to 10e, the minimum size requirement for a single-point-hologram will not be any greater for a comparable resolution and object definition. The optimum plate size of the hologram used in an ONN implementation is ultimately governed by the design requirements necessary in achieving a specified system sensitivity criteria that suffices the model architecture. This is so because, to obtain a holographic image, a hologram plate can be as small as only a few emulsion grains to as big a plate as one desires. Also, a holographic image can be reconstructed from any part of the hologram with a slight compromise in resolution and perspective.

7.7 Strip Holograms in ONNs

The innovatory concept of strip holograms described in this research is certainly the most significant aspect in achieveing higher resolution point intensity detections through the CGHs, which translates into more precise weight definitions within an ONN architecture. The conventional techniques of using a whole hologram for all the interconnections in the ONN architecture introduces unnecessary noise through additional wave interferences causing more diffusion and a lower resolution intensity detections. An additional advantage of using strip hologram interconnection scheme is that during the learning stage only the strip hologram, for which the interconnection weights must be modified, would have to be regenerated. This would reduce several computational latencies in comparison to the whole hologram technique and would result in a faster learning process in ONNs.

When using strip holograms in ONNs, it is not necessary that the strip used within the array be the entire hologram square plate originally created with the HoloGen application. Holograms retain complete information about the object in every part of the holographic emulsion. So a strip hologram, used within the strip hologram array, can be a rectangular cutout of the larger original square hologram plate and it will still create the entire object image with only a slight compromise in overall resolution. It is obviously best to use the central portion of the hologram to be able to better position the object along the horizontal plane, as well as to get the best resolution. The outside fringes are highly condensed and thus contribute mainly to obtain a sharper focus of the image during the reconstruction phase. However, if the outside portion is used as the master hologram, it may be at the expense of resolution.

7.8 Point Holograms in ONNs

In a one-to-one interconnection scheme, the input neurons may be interconnected with the output neurons with single point holograms. This method will provide very high resolution intensity detection across the output detectors, since each hologram behaves as a Fresnel Lens providing a very high definition focus for a single point. Another advantage of using single point hologram interconnection scheme is that during the learning stage only the few holograms, for which the interconnection weights must be modified, would have to be regenerated. This would reduce several computational latencies and would result in a faster learning process in ONNs.

The disadvantage of single-point holograms is that it imposes a size constraint for the holographic interconnection latice, since it would require more SLMs than the strip-hologram scheme for the same number of interconnections.

8. Conclusion

Neural network architectures will certainly provide some of the most exciting advances in the computer industry of the next decade. If, however, our goal is to build very large, high speed neural networks, then we must move from electronic to optical and opto-electronic implementations of neural network architectures. Optics offers numerous advantages in the implementation of ANNs: freespace, high density interconnections; absence of cross-talk; a naturally parallel and analog nature; and a very low power consumption. Holographic devices, which have already made enormous contributions to the computer industry, are particularly well-suited for various incorporations within ONN architectures, notably SLM-based architectures.

The two most important components in neural networks are the processing elements (neurons) and the interconnection weight matrix. The output of each processing element in an ANN is computed via the sum of several products; hence, the primary ANN operation is the multiply-accumulate, which is required for every connection in the network. ANNs typically have many connections, each of which has an associative weight that must be physically represented within the hardware architecture. This presents a serious design problem as the size of the network scales up [Kaikhah 1995]. Holographic optical neural networks provide elegant solutions. While optics provides a medium for densely integrated large number of freespace interconnections and data communication at the speed of light, the hologram provides a solution for storing as well as actually implementing the weight matrix data in an efficient and highly condensed format between fully-connected neural layers.

Holograms have been conventionally produced with lasers and film. Holograms can also be generated with computers. Since light and the holographic process can be described mathematically, it is possible to derive holograms using mathematical calculation and computer image rendering. These computer generated holograms can then be incorporated within ONN architectures to establish densely packed weighted interconnections.

Optics has come to play an important role in numerous technologies of the twentieth century, including that of artificial neural networks. Already, many advances have been made in the optical and opto-electronic implementation of various ANN algorithms. Most ANN algorithms, however, have been developed with the idea that they would eventually be implemented with conventional electronic computers. However, as we come to realize the vital role that ONNs will play in neural network computing and develop a much stronger understanding of the nature and scope of ONN architectures, we can begin to develop neural network algorithms that are truly able to utilize the advantages that optics offers. We are beginning to see a trend toward the use of neural networks throughout the computer industry. It is with ANN architectures that computers will be able to move into problem domains that have hitherto only been the realm of intelligent biological systems. With the advances in optical computing that are now underway and the breakthroughs that are certain to come in the near future, we can expect to see ONNs leading the way into the twenty-first century.

9. Bibliography

A. Agranat, C. F. Neugebauer, and A. Yariv, "Parallel Opto-electronics Realization of Neural Network Models Using CID Technology," Applied Optics 27 (1988), pp. 4354-4355.

P. P. Banerjee and T. Poon, Principles of Applied Optics, Boston: Irwin, 1991.

H. J. Caulfield, J. Kinser, and S. K. Rogers, "Optical Neural Networks," Proc. IEEE 77 (October 1989), pp. 1573-83.

H. J. Caulfield and J. Shamir, "Wave Particle Duality Considerations in Optical Computing," Applied Optics 28 (12) (15 June 1985), pp. 2184-6.

K. Doh, "Twin-image elimination in optical scanning holography - special issue on hybrid optical image processing," Optics and Laser Technology, 1996.

N. H. Farhat, "Optoelectronic Analogs of Self-Programming Neural Nets: Architecture and Methodologies for Implementing Fast Stochastic Learning by Simulated Annealing," Applied Optics, 26 (23) (Dec. 1987), pp. 5093-5103.

N. H. Farhat, "Optoelectronic Neural Networks and Learning Machines," IEEE Circuits and Devices Mag., Sept. 1989, pp. 32-41.

N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical Implementation of the Hopfield Model," Applied Optics 24(10) (15 May 1985) pp. 1469-75.

D. G. Feitelson, Optical Computing: A Survey for Computer Scientists, Cambridge, Massachusetts: The MIT Press, 1988, pp. 43-315.

M. R. Feldman, S. C. Esener, C. C. Guest, and S. H. Lee, "Comparison between Optical and Electrical Interconnects Based on Power and Speed Consideratons," Applied Optics 27 (1988), pp. 1742-1751.

D. Gabor, "A New Microscopic Principle," Nature, 166 (1948), pp. 777-778.

J. W. Goodman, Introduction to Fourier Optics, F. E. Terman, Editor, McGraw-Hill Physical and Quantum Electronics Series, 1968.

N. Hasmoto, S. Morokawa, and K. Kitamura, "Real-time Holography Using the High-resolution LCTV-SLM." Proc. SPIE 1461 (1991), pp. 291-302.

J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Ablilities," Proc. Nat. Acad. Sci. 79 (April 1982), pp. 2554-2558.

K. Hsu, D. Brady, and D. Psaltis, "Experimental Demonstrations of Optical Neural Computers", Neural Information Processing Systems, New York: American Institute of Physics, 1988, pp. 377-386.

Y. Ichioka, T. Iwaki, and K. Matsuoka, "Optical Information Processing and Beyond," Proc. IEEE 84(5) (May 1996), pp. 694-719.

M. Ismail, "Analog VLSI Neural Systems: Trends and Challenges, Applications and Science of Artificial Neural Networks, Steven K. Rogers and Dennis W. Ruck, Editors, Proc. SPIE 2492(1) (1995), pp. 634-40.

K. Kaikhah, Neural Networks Class Notes, Southwest Texas State University, Fall 1995.

W. E. Kock, Engineering Applications of Lasers and Holography, New York: Plenum Press, 1975.

W. Lee, "Computer Generated Holograms: Techniques and Applications," Progress in Optics XVI (1978), pp. 119-232.

E. N. Leith and J. Upatnieks, "Reconstructed Wavefronts and Communication Theory," Journal of the Optical Society of America, 52 (1962), p. 1123-1130.

Y. Liu, "Extensions of Fractal Theory," Science of Artificial Neural Networks II, Dennis W. Ruck, Editor, Proc. SPIE 1966 (1993), pp. 255-68.

A. Macovski, "Considerations of Television Holography," Optica Acta, 18 (1971), pp. 31-39.

J. E. Midwinter and D. R. Selviah, "Digital Neural Networks, Matched Filters and Optical Implementations," Neural Computing Architectures: The Design of Brain-Like Machines, I. Aleksander, Editor, Great Britain: North Oxford Academic Publishers Ltd., 1989, pp. 258-78.

A. Mikaelian, "1-D Holographic Memory for Information Processing," Photonics for Processor, Neural Networks, and Memories II, J. L. Horner, B. Javidi, and S. T. Kowel, Editors, Proc. SPIE 2297 (1994), pp. 155-63.

M. Minsky, The Society of Mind, New York, NY: Simon and Schuster, 1986.

L. Onural, G. Bozdagi, and A. Atalar, "New High-resolution Display Device for Holographic Three-dimensional Video: Principles and Simulations," Opt. Eng. 33 (1994), pp. 835-844.

T. Poon et al., "Real-time Two-dimensional Holographic Imaging Using an Electron-beam-addressed Spatial Light Modulator," Opt. Lett., 18 (1993), pp 63-65.

T. Poon, M. H. Wu, K. Shinoda, and Y. Suzuki, "Optical Scanning Holography," Proc. IEEE 84(5) (May 1996), pp. 753-64.

G. L. Rogers, "Gabor Diffraction Microscopy: The hologram as a Generalized Zone Plate," Nature 166 (1950), p. 237.

W. L. Siemens-Wapniarski and M. Parker Givens, "The Experimental Production of Synthetic Holograms," Applied Optics 7 (1968), pp. 535-538.

P. K. Simpson, “Foundations of Neural Networks,” Artificial Neural Networks: Paradigms, Applications, and Hardware Implementations, E. S’anchez-Simencio and C. Lau, Editor, IEEE Press, 1992, pp. 3-24.

H. M. Smith, Principles of Holography, New York: Wiley-Interscience, John Wiley and Sons, Inc., 1969, p. 32-42,

B. H. Soffer, G. J. Dunning, Y. Owechko, and E. Marom, "Associative Holographic Memory with Feedback Using Phase-Conjugate Mirrors," Optics Lett. 11 (Feb. 1986), pp. 118-120.

J. P. Waters, "Holographic Image Synthesis Utilizing Theoretical Methods," Applied Physics Letters 9(11) (1 Dec. 1966), p. 405-407.

T. Yatagai, S. Kawai, and H. Huang, "Optical Computing and Interconnects," Proc. IEEE 84(6) (June 1996), pp. 828-52.

F. T. S. Yu, "Optical Neural Networks: Architecture, Design and Models," Progress in Optics XXXII, E. Wolf, Editor, New York: North-Holland 1993, pp. 61-144.