Reprinted from

Optics in Computing 2000

18–23 June 2000
Quebec City, Canada
Vertically integrated photonic multichip module architecture for vision applications

Armand R. Tanguay, Jr., B. Keith Jenkins, Christoph von der Malsburg, Bartlett Mel, Gary Holt, John O’Brien, Irving Biederman, Anupam Madhukar, Patrick Nasiatka, and Yunsong Huang

Departments of Electrical Engineering–Electrophysics, Electrical Engineering–Systems, Materials Science, Biomedical Engineering, Computer Science, Neurobiology, Physics, and Psychology; and Program in Neural, Informational, and Behavioral Sciences; University of Southern California, University Park, MC-0483, Los Angeles, CA 90089-0483

ABSTRACT

The development of a truly smart camera, with inherent capability for low latency semi-autonomous object recognition, tracking, and optimal image capture, has remained an elusive goal notwithstanding tremendous advances in the processing power afforded by VLSI technologies. These features are essential for a number of emerging multimedia-based applications, including enhanced augmented reality systems. Recent advances in understanding of the mechanisms of biological vision systems, together with similar advances in hybrid electronic/photonic packaging technology, offer the possibility of artificial biologically-inspired vision systems with significantly different, yet complementary, strengths and weaknesses. We describe herein several system implementation architectures based on spatial and temporal integration techniques within a multilayered structure, as well as the corresponding hardware implementation of these architectures based on the hybrid vertical integration of multiple silicon VLSI vision chips by means of dense 3-D photonic interconnections. In one such approach, vertical signal transmission is accomplished by using a planar-waveguide-based optical power bus to provide a parallel array of readout beams to an inverted cavity III-V compound semiconductor multiple quantum well modulator array that is flip-chip bonded to each silicon substrate. Local and quasi-local connectivity is accomplished between chips stacked in the vertical dimension by using diffractive optical element arrays that provide for both point-to-point interconnections and weighted fan-out within a local neighborhood. Global connectivity between multichip modules is incorporated by means of computer-generated volume holographic optical elements that can be fabricated as multilayer diffractive optical elements.

Keywords: Biologically-Inspired Vision Models, Hardware Implementations of Vision Models and Algorithms, Photonic Implementations of Neural Networks, Hybrid Electronic/Photonic Multichip Module Integration, Hybrid Analog/Digital Silicon VLSI, Inverted Cavity Fabry-Perot Multiple-Quantum-Well Modulator Arrays, Vertical Cavity Surface Emitting Laser Arrays, Flip Chip Bonding, Diffractive Optical Element Arrays, Stratified Volume Diffractive Optical Elements

1. INTRODUCTION

Over the past decade, significant advances in the psychology and physiology of vision, in computational neurobiology, in hybrid analog/digital VLSI technology and parallel processing systems, in micro-optics, in photonic technology, and in hybrid electronic/photonic packaging have made the long-sought goal of an adaptive vision sensor appear feasible$^{1-7}$. The development of both a generic theoretical understanding of, and a technology implementation platform for, adaptive vision sensors could enable the potential development of a wide range of advanced smart camera and smart display systems. In addition, the rapid emergence of multimedia applications has both created a critical need for advanced vision sensors and generated increased interest in both hybrid analog/digital and hybrid electronic/photonic systems implementations.

In parallel with these advances, adaptive vision applications that involve rapid object identification, moving object tracking, and pose estimation functions, such as are envisioned for augmented reality systems, increasingly place stringent upper bounds on processing latency and therefore computational throughput in order to accomplish the desired visual task in times significantly less than human perception and/or reaction times. In many

In Optics in Computing 2000, Roger A. Lessard, Tigran Galstian, Editors, SPIE Vol. 4089 (2000) 0277-780X/00/$15.00
such applications, several hierarchical stages of processing must be accomplished within the requisite latency, including (for example) image acquisition, image pre-processing, feature extraction, object recognition, extraction of image orientation with respect to environmental coordinates, determination of course of action, and in some cases precise registration of computer animation and/or graphics with respect to structured or natural environments.

In addition to the requirement for low latency, many emerging vision models and algorithms involve operations that are parallel in nature, nonlinear in functionality, and both local and non-local in structure. The resulting computational complexity places correspondingly complex demands on envisioned hardware implementations.

In order to satisfy these conjoint requirements, we are investigating a hybrid electronic/photonics multichip module architecture that comprises multiple layers of silicon VLSI detection and processing circuitry, coupled in the vertical dimension with dense photonics fan-out/fan-in interconnections. The interconnections are implemented by means of 2-D arrays of MQW modulators (illuminated by an integrated optical power bus) or vertical cavity surface-emitting lasers (VCSEL's) that are flip-chip-bonded on a pixel-by-pixel basis to the silicon VLSI detector/processor array, in conjunction with proximity-coupled diffractive optical element and microlens arrays that implement 2-D weighted space-invariant or space-variant fan-out patterns. As a consequence of the layer-to-layer fan-out, the effective receptive field of a given pixel increases with the number of preceding layers.

This layered, densely interconnected architecture is inspired by, but not emulative of, biological vision systems, in which multiple concatenated nonlinear operations are interspersed with the weighted interlayer fan-out and fan-in of information. For example, biological vision systems commonly exhibit very high spatial complexity but relatively high layer-to-layer latencies, whereas photonic multichip modules are expected to exhibit more modest spatial complexities in conjunction with much lower layer-to-layer latencies. In order to effectively map emerging vision models and algorithms, such as wavelet decomposition, elastic graph matching, feature conjunctions, and geon formation onto this emerging hardware platform, we have investigated several approaches for using this intrinsic space-bandwidth tradeoff to advantage.

Initial results of a multidisciplinary research effort are described herein on an analysis of the biological imperative for layering throughout the mammalian visual system; on the mapping of biologically-inspired vision models and algorithms onto the emerging hybrid electronic/photonics multichip module platform; on the incorporation of spatial and temporal multiplexing approaches to implement computationally complex operations; and on the current status of the corresponding hardware component design, fabrication, testing, and integration.

2. FUNCTIONALITY REQUIREMENTS FOR ADAPTIVE VISION SENSORS

The set of functionality requirements that describe the envisioned adaptive vision sensors is, of course, application-dependent. In this section, we examine the features of one particular application, that of enhanced augmented reality systems, in order to illustrate the specific functionality requirements that define our technological approach.

In virtual reality applications, computer animation and graphics are employed to create an artificial world, and in some cases to simulate reality as closely as possible. In augmented reality applications, computer animation and graphics are employed to supplement and annotate real world images in real time. For example, a head-mounted display for a pilot in the commercial cockpit environment could potentially provide navigational and threat-warning (e.g., collision-avoidance) information that is precisely registered with environmental features, independent of pose (head orientation) or vehicle attitude.

A key issue in the implementation and deployment of augmented reality systems is the requirement for a miniature (both compact and light weight) smart camera or vision sensor that is capable of providing pose estimation (orientation of the user's head with respect to naturally-occurring or artificially-dispersed (fiducial) environmental features) as well as both stationary and moving object acquisition and identification at very high frame rates with low processing latency. The requirement for low processing latency allows extracted features within the field-of-view to be identified, labeled, and graphically annotated in such a manner that they appear properly oriented and stable at all times as seen by the user, irrespective of relative head orientation or motion. The function of low latency video-based tracking alone is already difficult to implement robustly even in highly structured environments. The additional functions of object identification and tracking in natural environments are considerably more complex. In current augmented reality applications, therefore, reliance is necessarily placed to a
much larger degree than desired on the acuity and dynamic range of the human visual system for provision and interpretation of a concurrent view of the natural environment, as the human visual system is as yet unsurpassed in critical dynamic image acquisition and interpretation functions. As such, although currently envisioned augmented reality systems use the human visual system to advantage where it has superior capabilities and adaptivity, thus avoiding the shortcomings inherent in current sensor-based image acquisition and display (e.g., limited dynamic range, limited contrast ratio, limited field-of-view, lack of gaze-dependent autofocus, and lack of adequate stereoscopic cues), the degree to which an augmented reality system can minimize distractions and maximize user concentration on critical tasks is strongly limited.

The emerging technology area of enhanced augmented reality represents a significant extension of the basic capabilities of augmented reality (such as head tracking and feature labeling, as described above). Enhanced augmented reality systems may include, for example, the additional functions of contrast enhancement, dynamic range compression, object recognition, object tracking, camouflage differentiation, and sensor fusion to overcome one or more of the limitations in augmented reality systems described above. In addition, supplemental video feeds or annotations from remote sensors can be superimposed within the display in registry with the features of the natural environment to create orientational and positional context. Incorporation of this latter feature yields a form of immersive augmented telepresence, in which both autonomous and user-directed remote sensor configurations are of interest.

Figure 1 illustrates a sample mock-up of what an enhanced augmented reality system user might see looking through a head-mounted display in an outdoor setting, and includes the identification and labeling of objects in the field of view, the display of remote sensor data that is visually obscured from the user's point of view, and the labeling of possible chemical threats. Known important geographical features, buildings, roads, and objects are labeled using features extracted from the scene and other data such as GPS coordinates and a stored map. Remote sensor data (the IR insert) provides information otherwise obscured to the user (by mountains in this case) and is fused into the scene in the correct direction, and labeled with informative data. Possible textural and/or metameric camouflage is also detected and labeled.

Several fundamental scientific and technological hurdles must be overcome in order to provide enhanced augmented reality technology in the field. First and foremost, efficient techniques for salient feature identification and tracking must be developed, along with a hardware platform capable of low-power, compact implementation. In order for these techniques to be robust enough to operate in natural environments, they must include at the very least provision for dynamic range compression, contrast enhancement, and color constancy over multiple image scales. From this perspective, these latter three features are all nonlinear, non-local, and adaptive functions that require complex, hierarchical operations. Second, the latency associated with the image acquisition and processing functions required to determine pose from identified salient environmental features must be reduced to much less than a human response time (of order 10 to 100 msec). In addition, the concomitant latencies of computation, registration, and display of graphical annotations must also be of this order. Even more stringent constraints may apply for fast-moving targets and threats, as evident for example in commercial aviation and semi-autonomous highway vehicular navigation. Third, highly parallel processing techniques for specific object or feature recognition and annotation must be developed that can be mapped onto a compatible hardware platform. Finally, provision must be made in the combined software/hardware suite to naturally support the fusion and interpretation of inputs from multiple sensors, all in very much less than real time. The overall latency constraint is further accentuated by the desire to make several, potentially contingent, queries of a given incoming image stream, followed by higher level processing before a given annotation is committed or warning issued.

A key enabling component in any envisioned enhanced augmented reality system is, therefore, a feature extraction and object identification module that provides extremely high computational throughput and low latency in a small, lightweight package, and is proximity coupled to one or more detector arrays. The computational requirements for augmented reality and enhanced augmented reality modules are severe, and are shown schematically relative to other applications in Fig. 2. These concomitant requirements of low latency and high computational complexity, combined with the environmental restrictions on size, power, and weight, are unlikely to be satisfied by conventional systems approaches based on microprocessor and DSP chips. In order to satisfy these requirements, we are investigating an implementation architecture and associated technology based on a hybrid electronic/photonic multichip module approach, as described in the next section.
Fig. 1. Hypothetical annotated view through a head-mounted display demonstrating enhanced augmented reality.

Fig. 2. Application domain in terms of computational complexity vs. latency. The domains of augmented reality and particularly enhanced augmented reality are the most difficult to realize.
Fig. 3. Conceptual diagram of 3-D optoelectronic structure, showing silicon analog/digital VLSI chip and optical fan-out/fan-in interconnections.

Fig. 4. Schematic diagram of multilayer hybrid electronic/photonics computation/interconnection element, showing novel optical power bus and diffractive optical element.

Fig. 5. Schematic diagram of multilayer hybrid electronic/photonics computation/interconnection element, showing vertical cavity surface emitting laser array and diffractive optical element.

Fig. 6. Densely interconnected 3-D electronic/photonics computational module, showing both local (DOE) and global (VHOE) optical interconnections.
3. ENVISIONED ARCHITECTURE AND IMPLEMENTATION STRATEGY

A conceptual diagram of the 3-D integrated electronic/photonic multichip module (PMCM) structure is shown in Fig. 3. Multiple layers of pixelated silicon VLSI chips are densely interconnected by a combination of electronic, optical, and photonic devices to produce either a space-invariant or space-variant degree of fan-out and fan-in to each individual pixel (processing node).

In one such implementation\(^{2-4}\), shown schematically in Fig. 4, 2-D arrays of inverted cavity InGaAs/AlGaAs multiple quantum well (MQW) modulators are flip-chip-bonded on a pixel-by-pixel basis to the silicon VLSI chips, which act as local detectors (for inputs from the previous layer), processors (either acting alone or in concert with electrical inputs from nearest and next-nearest neighbors within the plane), memory elements (in the analog or digital domain) and modulator drivers. Alternatively, the detectors can be co-integrated with the modulator elements, but at the cost of twice as many bump bonds per pixel. In this implementation, an optical power bus, which principally comprises an array of 1-D rib waveguides with a 2-D array of outcoupling gratings spaced to match the pitch of the modulator array, is powered by a semiconductor laser source or source array to provide a 2-D array of uniform intensity modulator readout beams. Proximity-coupled diffractive optical element (DOE) arrays, designed to incorporate both focal power and weighted fan-out functions, are used to establish interconnections that are modulated in intensity by each individual modulator element and its associated Si driver circuit. Alternatively, either combined refractive/diffractive elements or separate concatenated DOE and microlens arrays can be used for this interconnection function.

In another such implementation\(^{5-7}\), shown schematically in Fig. 5, 2-D arrays of vertical-cavity surface emitting lasers (VCSEL's) are flip-chip-bonded on a pixel-by-pixel basis to the silicon VLSI chips, which act in this case as VCSEL drivers in addition to the functions described in the modulator case. This replacement of the modulator array with a VCSEL array has the simplification of eliminating the requirement for the optical power bus, provided the power dissipation of the VCSEL array can be made low enough to allow for the requisite (aggregate) operational bandwidth.

The design of the individual Si VLSI chips, and in particular the use of spatio-temporal multiplexing techniques for network implementation and signal processing functions, is motivated by the recent development of several promising biologically-inspired vision algorithms that can potentially be mapped into the emerging 3-D PMCM platform. Software-based implementations of these vision algorithms, which collectively include low-level, mid-level, and high-level visual processing functions, have in a number of cases been directly tested against human (and in some cases trained human) observers, with the result that the recognition rates and even confusion matrices are surprisingly well correlated\(^{8}\).

As the level of representation extends from low-level-vision through mid-level-vision to high-level-vision operations, interconnections tend to become both more sparse and more global; in some cases, particularly between functionally-partitioned vision processing modules, dense global interconnections may be required. Both local and global interconnection cases can be accommodated within the PMCM architecture by incorporating novel stratified volume diffractive optical elements (SVDOE's)\(^{9-11}\) as shown schematically in Fig. 6, which consist of multiple layers of proximity-coupled and aligned DOE's that implement either space-variant or space-invariant interconnection patterns with properties characteristic of volume holograms. These new devices offer the advantages of planar fabrication methods compatible with VLSI design rules, and thus can potentially circumvent the difficulties inherent in optical recording of volume holographic optical elements.

4. GENERAL PRINCIPLES EXTRACTED FROM BIOLOGICAL VISION SYSTEMS

Since the base technology substrates exhibit wide-ranging differences\(^{12}\), the implementation of advanced sensors in hybrid electronic/photonic multichip modules for vision applications must be based, if at all, on mammalian vision systems primarily for functional guidance, architectural novelty, and as a performance metric, but without resort to straight biological emulation. As a matter of fact, useful guidance can also be obtained from the visual systems of non-mammalian organisms, as was pointed out eloquently by Veldkamp\(^\dagger\). In this section, we describe several general principles that can be usefuly extracted from biological vision systems and employed to advantage.

Biological vision systems exhibit a number of common themes, including (1) a propensity for layering of the processing architecture\(^{13-17}\), (2) the employment of massive parallelism with simple local processing units and
minimal (if any) local storage within each processing unit, (3) the use of a multiplicity of neuron unit types and associated fan-out and fan-in patterns (which gives rise to a set of interpenetrating neural network topologies), (4) the incorporation of dense interconnections at all scales (from local to global, among multiple modules) with a high degree of fan-out and fan-in at each processing node within the visual cortex, (5) adaptivity on multiple time scales as exhibited by both short term and long term plasticity, (6) distributed storage of information as exemplified by the plasticity of neuronal interconnection weights at individual synapses, (7) an associative memory organizational construct; (8) the importance of temporal correlations for both synaptic plasticity and neuron activation, and (9) the incorporation of adaptive temporal dynamics within each synapse (giving rise to nonlinear dynamical system properties throughout). This latter feature may prove key to the understanding of vision system responsivity to continuous rather than framed motion, as well as speech and sound recognition in the auditory system.

Primate visual systems, for example, use dense layers of photoreceptors and neurons at the lowest levels of vision processing, with primarily local, fixed, weighted interconnections among multiple pre-processing layers within the retina. The density of photoreceptors can be extremely high, ranging from about $1 \times 10^7$ cm$^{-2}$ in the fovea (corresponding to a cone diameter of order 1 µm) to $4 \times 10^8$ cm$^{-2}$ in the periphery (with a mixture of 4 to 10 µm diameter cones separated by a much higher density of 1 µm diameter rods). This density can be instructively compared with the current pixel densities of solid state imaging sensor arrays (including local plane arrays in the visible, IR, and UV; CCD arrays; and active pixel sensor (APS) arrays), which range from about $1 \times 10^3$ cm$^{-2}$ to $4 \times 10^4$ cm$^{-2}$. Current smart pixel arrays do not come close to achieving even these densities as a result of the incorporation of local processing circuitry within each pixel of the array.

For the stages of early vision implemented within the retina and the lateral geniculate nucleus, and extending into the lowest level of the visual cortex (region V1), the interconnection mappings tend to be local (restricted neighborhoods), highly regular (retinotopic), and only partially adaptive. Higher up the biological processing stream (within the primate visual cortex), interconnections tend to become gradually less local, less regular, and more adaptive, with a degree of interconnectivity (fan-out from and fan-in to a given neuron) that is typically $10^3$ to $10^4$. Throughout the biological vision system, color information is merged and re-merged with spatial information, as well as with inputs from other sensory modalities (beyond a certain stage of the visual system). This variety and degree of interconnectivity is difficult if not impossible to achieve in a current VLSI implementation within a given plane, as a result of limitations both in the number of metallization layers allowed in a given process (5 to 7) and in the number of following devices that can be driven from a given device without intervening buffers and signal amplifiers.

As the extension of VLSI chip implementations of smart cameras and biologically-inspired vision systems from single-chip 2-D arrays to 3-D hybrid electronic/photonics multichip module implementations is a principal focus of our technical approach, it is of considerable interest to examine the biological imperative for layering. In primate visual systems, layering accomplishes a number of important functions. (1) Layering provides a convenient mechanism for the implementation of multiple concatenated operations comprising nonlinearities and weighted fan-out/fan-in functions. The latter operation can be viewed as the convolution of a 2-D input function with a set of 2-D kernels (weighting functions), and provides the basis for implementing both space-invariant and space-variant non-local operations across multiple spatial scales. The separation of a given complex operation into several sequential steps of nonlinearity/convolution also allows access to intermediate scale results for both feedforward and feedback connections that project beyond intervening layers, as observed throughout biological vision systems. (2) Layering also provides for the implementation of higher-order complexity (hierarchical) operations that can be derived from simple primitives implemented over multiple spatial scales. (3) Layering naturally provides for the hierarchical buildup of the size of the receptive field, so that non-local operations such as contrast enhancement and color constancy can be implemented with invariance to object size. (4) Finally, layering carries with it the potential for increased algorithmic efficiency, in that certain operations (e.g., even certain linearly decomposable convolutions at a given kernel size) can be performed in multiple layers with less cost in computational resources (e.g., fan-out from neurons via axons, synapses, and dendrites; total number of equivalent primitive operations; computational energy).

From a systems perspective, an unresolved question of considerable interest is the overall efficiency of representation in biological vision systems, as defined by the efficiency with which higher level representations are generated from the input visual field through the use of lower level primitives. Key to the understanding of this question is the related efficiency of representation from the perspective of memory organization (storage and recall). These coupled questions are crucial to the effective design of a biologically-inspired vision system, particularly one
implemented in a technology base with vastly different capabilities and characteristics from that provided by biological wetware.

5. THE HIERARCHICAL ORGANIZATION AND MAPPING OF COMPLEMENTARY VISION MODELS

Computational vision algorithms that work well in the real world are (usually severely) constrained by the computational power available. As computational resources improve in the future, both problem domain size and algorithm generality will increase. One approach that may enable such continued progress is threefold: an emphasis on biologically inspired algorithms; the incorporation of architectures that are at a degree biologically derived but at the same time respect the constraints imposed by the hardware; and a reductionist (and therefore generalizable) approach to mapping algorithms onto the hardware. Such a reductionist approach involves the use of tools for implementing basic operations that are common to vision models and algorithms. The differences among vision models and algorithms tend to reduce or even dissolve when viewed from the standpoint of the basic operations needed to implement them. When developed from a biologically inspired viewpoint, these basic operations are more likely to be efficiently implementable on such a parallel hardware structure.

Consider, for example, three key yet complementary biologically inspired vision models: recognition of highly disparate objects based on their constituent primitive features (Mel's SEEMORE object/scene recognition system and its variants, based on taking selected combinations (conjunctions and disjunctions) of extracted primitive features); robust recognition and evaluation of similarity and differences of objects (e.g., faces) based on elastic graph matching (von der Malsburg's Dynamic Link Architecture, based on wavelet decomposition); and recognition based on features that are invariant over 3-D transformations (Biederman's Recognition by Components, based on the extraction of geometrically simple 3-D components, or geons). Together, these three models exhibit the potential for invariance not only to scale and orientation but also to object deformations and 3-D viewpoints. A schematic depiction of the key features of these three complementary models as they might be combined in a hierarchical vision system is shown in Figs. 7 to 9, spanning representational levels from the input scene(s) to a higher level, invariant representation that can be used by a subsequent processor for decision-making, initiation of actions, storage in an associative memory, or passed on to processing modules that perform specialized functions.

From this viewpoint, the mapping problem can be performed as a sequence of two steps. The first step, algorithm modeling, represents the original algorithm (such as SEEMORE object recognition, wavelet transformation, or elastic graph matching) as a set of lower level operations (such as repeated comparisons, inner products, and point nonlinearities). The second step, which uses implementation tools, converts these to a physical representation (such as forms of optical interconnection patterns, interconnection weights, and electronic shift registers). Because the implementation-tools step begins with low level operations, in a sense it is blind to the vision algorithm being implemented and therefore is generalizable across vision algorithms that can be decomposed into appropriate lower level operations. Its function is to map these low level operations onto the hardware in a way that is parallel, provides sufficiently low latency, and is reasonably hardware efficient. As an illustration of the issues that arise in the mapping of vision models onto the photonic multichip module hardware, the possible synergistic combination of biologically-inspired spatial multiplexing with electronic- and photonics-technology-inspired temporal multiplexing is described in the next section.

6. THE INCORPORATION OF TEMPORAL MULTIPLEXING APPROACHES

The envisioned hybrid electronic/photonic hardware platform is in some ways biologically inspired (with its capability for weighted fan-out/fan-in interconnections, layered structure, and parallelism), and in some ways markedly different (with its much higher anticipated temporal bandwidth and much lower anticipated spatial parallelism, compared with biological systems). These characteristics beg for some form of temporal multiplexing that can make use of both the parallel nature of the algorithms and the spatio-temporal nature of the hardware. In addition, when applied to vision problems in which the first hardware layer receives an input image from the physical world, as shown in Fig. 10 at the top of the figure, a large mismatch in bandwidth exists between the input image stream (~100 Hz frame rate) and the available processing rate that can be achieved in subsequent layers (~50 to 100 MHz analog bandwidth; ~200 to 500 MHz digital synchronous clock rate or asynchronous operation rate; ~1 to 100 MHz layer-to-layer interconnection bandwidth per pixel). Again, some form of multiplexing is desirable in subsequent layers to make use of the available processing power.
Fig. 7. Overview of a vision system employing aspects of three complementary biologically inspired vision models (all levels).

Fig. 8. Overview of a vision system employing aspects of three complementary biologically inspired vision models (lower levels).

Fig. 9. Overview of a vision system employing aspects of three complementary biologically inspired vision models (higher levels).

Fig. 10. Example optoelectronic eye-and-vision processor layout, shown in cross section.
Because of their biological inspiration, these vision algorithms, when implemented using parallel hardware, are likely to be most efficiently implemented using some weighted interconnections that are fixed, and some that are adaptive. The hardware described herein currently employs optical (physical) interconnection weights that are designed a priori and fixed once fabricated, while the in-plane electronic interconnections (nearest neighbor or next-nearest-neighbor) can be either fixed or adaptive. As such, the multilayered network depicted schematically in Fig. 10 can be represented as a set of interpenetrating network topologies, with both fixed and adaptive weights. At this point, it is an open question as to whether or not this combined topology contains sufficient adaptive capability for the full range of envisioned applications. Consequently, a second set of useful temporal multiplexing techniques could provide means for effectively achieving adaptive or programmable interconnection weights at the functional level, even for the layer-to-layer interconnections.

As an example, consider the Gabor wavelet transform\textsuperscript{20,21}. This transform can be modeled as a set of convolutions, each with a kernel of the type shown in Fig. 11, but at different scale sizes and orientations. (For example, the Dynamic Link Architecture\textsuperscript{2} typically uses eight different orientations and five different scale sizes, yielding 40 different kernels.) Once the transform is so modeled, a variety of physical implementations are possible. One such implementation uses a “direct mapping” approach that involves representing the input image data in analog form, laid out spatially (topographically, or retinotopically) in a first hardware plane. Each convolution kernel is then laid out as an optical weighted interconnection, i.e., as a fan-in pattern to a processing element (pixel) in a second, or receiving, hardware plane. (The use of bipolar or complex numbers and arithmetic, if desired, can be handled using any of a variety of known representations.) Neighboring processing elements (pixels) in the receiving plane correspond to different kernels (three kernels are depicted in Fig. 12), so that an \( n \times m \) region is used to implement all \( nm \) kernels. The \( n \times m \) array of fan-in kernels is repeated across the array. Thus, in one time step the transform is performed as the data moves from one plane through the interconnections (kernels) and is summed optically at the detectors (for example) as it is received in the subsequent plane. Although one processing element in the receiving plane corresponds to just one kernel, mappings of this type can be sampled sufficiently to avoid any loss of meaningful information. The transform is performed in essentially one time step.

While this direct mapping approach uses space efficiently, it may result in inefficient use of the temporal domain (because one useful time step may be followed by a series of idle ones). A second technique can be implemented with a similar set of \( n \times m \) fan-in patterns, but repeated more sparsely across the array. In the interleaving regions, interconnections for other operations are implemented. Upon operation, the input image data is shifted across the first plane in time\textsuperscript{31,32}, viewed from a given interconnection kernel (and therefore from a given processing element in the second plane), different portions of the input image data are input to this interconnection kernel at different instances in time. After a sufficient number of time shifts, each (and every) kernel has operated on the entire input image. This technique allows multiple operations to be performed (e.g., Gabor wavelet decomposition, and matching to a variety of small templates for primitive feature extraction) in the single layer. These operations are multiplexed partially in time and partially in space, resulting in a more efficient utilization of the processing power available.

Temporal multiplexing techniques can also potentially be used to significant advantage in implementing adaptive functionality. For example, the incorporation of a cellular neural network (CNN) layer (depicted schematically as layer C6 in Fig. 10) naturally allows for the implementation of programmable flexibility through the array of distributed mixed analog/digital processing units that comprise the CNN. Furthermore, to the extent that the image shifting or scrolling functions are programmable, the operations performed can be sequenced or altered in time. Because the resulting data is spatio-temporally multiplexed, appropriate design techniques must be used to ensure layer-to-layer data format compatibility.

7. PHOTONIC MULTICHIP MODULE INTEGRATION

Considerable numbers of research efforts have focused previously on the development of both biologically-inspired and biologically-emulative vision chips that implement one or more functions characteristic of biological vision systems, such as dynamic range compression, edge enhancement, and motion sensing (see, for example, Ref. 10). In each case, considerable functionality has been achieved within the constraints imposed by limited lateral connectivity and the inherent 2-D nature of the single chip substrate. In the current research effort described herein, we have focused on the problem of densely interconnecting multiple such chips in the third (out-of-plane) dimension, in order to provide additional flexibility and functionality.
Fig. 11. Example Gabor wavelet kernel plotted vs. spatial coordinates x and y.

Fig. 12. Direct parallel mapping of Gabor wavelet decomposition, showing fan-in patterns for 3 kernels.

Fig. 13. Schematic diagram of multilayer hybrid electronic/photonic computation/interconnection element, depicting flip-chip-bonding of silicon photodetector/driver chip and GaAs MQW modulator array.

Fig. 14. Photograph of a portion of the 16 x 16 array of neuron units. Shown is a single sigmoidal neuron unit.

Fig. 15. Optical power bus. Photomicrograph of 8 μm x 1 cm 660 element rib waveguide array with outcoupling gratings fabricated on an AlGaAs-GaAs waveguide.
As in the algorithm and architecture development effort outlined above, the principal goal of the hardware integration effort is to develop a flexible toolbox of individual component technologies that can be used to implement a wide variety of smart cameras and artificial vision systems. In the individual subsections below, we describe the basic required functionality and results achieved to date in each of the component technologies that comprise the photonic multichip module (PMCM) hardware integration platform, as well as efforts to achieve PMCM integration. Fabrication and performance details of several of the individual components have been in part reported elsewhere, as noted in the references where appropriate.

7.1 Silicon VLSI Detector/Processor Arrays

As shown schematically in Figs. 3–6, silicon VLSI chips are incorporated in each submodule (layer) primarily to implement as-designed processing functions (such as logarithmic transformations, sigmoidal (thresholding) transformations, differentiating operations, sample-and-hold, temporal integration, and temporal differentiation). In some implementations, the Si chips also carry photodetection capability (for example, in the input image plane) as well as device drivers for III-V compound semiconductor modulators or VCSEL’s, as shown in Figs. 4, 5, and 13. In other implementations, the photodetection functions and/or III-V device drivers can be co-integrated with the modulators or VCSEL’s.

We have focused to date on the implementation of a test chip that implements a nonlinear sigmoidal input/output transformation. The current Si chip design consists of a 16 × 16 array of neuron units placed on a 100 μm pitch; a photomicrograph of a single neuron unit is shown in Fig. 14. Each neuron unit contains Si CMOS control electronics, two Si vertical photodiode detectors, and three bonding pads to allow vertical interconnection to two MQW modulators or VCSEL’s. The control circuitry provides a −2 to −9 V nonlinear sigmoidal response characteristic to the difference in light intensity incident on the two photodetectors. A separate sigmoidal response characteristic is provided for each signed difference, thereby allowing for the representation of both excitatory and inhibitory inputs as intensity values.

The analog 16 × 16 array test chip was fabricated through the MOSIS foundry service using the 1.2 μm HP scalable CMOS n-well process. The large signal sigmoidal response characteristic has been demonstrated at frequencies up to 100 kHz, with a small signal response in excess of 1 MHz (test equipment limited); SPICE simulations indicate a small signal response in excess of 4 MHz. The estimated chip power dissipation is 2 mW per pixel or about 0.5 W per chip.

7.2 Inverted Cavity Multiple-Quantum-Well (MQW) Modulator Arrays

In the PMCM configuration shown in Figs. 4 and 13, 2-D arrays of inverted cavity multiple-quantum-well (MQW) modulators are incorporated to provide an electrical-to-optical conversion function, thereby producing a signal-encoded array of output beams that can be individually fanned out with weights by means of proximity-coupled diffractive optical elements, as described further below.

The inverted asymmetric Fabry-Perot cavity modulator arrays were fabricated by molecular beam epitaxy (MBE) on III-V GaAs substrates, and consist of a p-i-n InGaAs/AlGaAs MQW region (33 quantum wells) sandwiched between an MBE-grown low-reflectance AlAs/GaAs distributed-Bragg-reflector front mirror and an ex-situ-deposited high reflectivity multilayer dielectric back mirror. The 2-D array of InGaAs/AlGaAs multiple quantum well modulators operates at 980 ± 1 nm (with a 2-D uniformity of ± 0.3 nm), and exhibits an average contrast ratio of 13:1 at −9 V applied bias (the saturation value of the Si chip sigmoidal output channels). This operational wavelength has been chosen to allow for relatively low-loss transmission through the following (thinned) Si substrate and for acceptable Si photodetector quantum efficiency, while still retaining the high uniformity and state of technological advancement characteristic of the ternary InGaAs/AlGaAs modulator system.

7.3 Flip Chip Bonding of Si and GaAs Chips

A cold weld indium bump flip-chip bonding process has been developed and used for the hybrid integration of the Si photodetection/control/driver chips and the GaAs MQW modulator arrays, as shown schematically in Fig. 13. In this process, approximately 8 μm high indium bumps are deposited on the mating electrode pads of both chips and patterned by a lift-off photolithographic process. The thermal or electron-beam deposition parameters are set to achieve a roughened ("velcro") surface on both indium bumps, which significantly improves both adhesion and contact resistance by providing for enhanced surface penetration when the two bumps are brought into contact.
using a visible flip-chip aligner-bonder. An SEM photomicrograph of the resulting bump surfaces and bump uniformity is provided in Fig. 16.

The flip-chip bonding process is performed using a near-room-temperature cold weld bonding procedure that avoids significant heating of the modulator and Si substrates. Measurements of MQW modulator reflectivity as a function of wavelength both before and after flip-chip bonding show no signs of performance degradation, as shown in Fig. 17.

7.4 Optical Power Bus

For the modulator implementations of the PMCM described above and in Figs. 4 and 13, reflective readout of the 2-D MQW modulator array must be provided in a highly compact manner. This function is performed by an optical power bus, which comprises a 2-D array of 1-D rib waveguides with superimposed outcoupling gratings. Optical intensity provided to the set of rib waveguides from an edge-mounted semiconductor laser diode or diode array propagates along each rib waveguide, confined in the lateral and vertical dimensions. Outcoupling gratings are provided at each location that corresponds to an individual modulator element, so that light is coupled out of the plane of the optical power bus, impinges on a given modulator element within the array, and is thereby both modulated in intensity and reflected back through the optical power bus toward the proximity-coupled diffractive optical element (see Fig. 4). The outcoupling gratings are designed to operate in the very low diffraction efficiency limit (approximately $10^{-4}$), so that only a very small fraction of the light propagating in the waveguide is outcoupled below each modulator element, thereby providing for uniform illumination along the length of each rib waveguide, and minimizing any re-coupling back into the waveguide.

Optical power buses have been fabricated to date in both Ti-indiffused LiNbO$_3$ waveguides, as well as in AlGaAs/GaAs waveguides, as shown in Fig. 15. Similar outcoupling grating arrays have also been fabricated in polymer waveguides. In the case of AlGaAs/GaAs, we have fabricated arrays of up to 660 individual 1 cm long, 8 μm wide ribs with 2 μm gaps. Outcoupling gratings of both 2 μm (shown in the SEM photomicrograph in Fig. 15) and 4 μm pitch have been successfully fabricated by a double photolithographic and ion-beam-milling process, with good output uniformity.

7.5 Vertical Cavity Surface Emitting Laser (VCSEL) Arrays

The use of vertical cavity surface emitting laser arrays for implementation of the electrical-to-optical conversion function described above (as shown schematically in Fig. 5) promises increased simplicity in photonic multichip module implementation, as both the optical modulator array and optical power buses would be replaced with a single component, and at least one additional critical alignment step would be eliminated. At the relatively low operational bandwidths envisioned for these hybrid analog/digital PMCM modules, however, power dissipation considerations currently favor the optical modulator array implementation. With continued progress in ultra-low threshold VCSEL arrays, such VCSEL arrays can be employed to advantage.

In both the optical modulator case and the VCSEL case, co-integration of III-V compound semiconductor photodetectors can provide natural wavelength compatibility between the emitters or modulators, on the one hand, and the detectors on the other. This choice allows more flexibility in choice of operational wavelength, provided sufficient space can be allocated for the additional bump bonds required within each pixel to connect the photodetector outputs back to the Si control circuitry.

7.6 Diffractive Optical Element (DOE) Arrays

Diffractive optical element (DOE) arrays are incorporated in the PMCM architectures shown schematically in Figs. 4 and 5 in order to implement dense 3-D fan-out/fan-in interconnections with fixed (non-adaptive) interconnection weights. The DOE arrays can be designed to provide either space-invariant or space-variant convolution kernels, depending on the layer-to-layer functionality required. For this application, the DOE designs minimize the distribution of undesired diffracted orders, not only to minimize the semiconductor laser diode power requirements and overall PMCM power dissipation, but also in view of the fact that light-sensitive components are distributed throughout the PMCM stack.

DOE arrays that implement a number of different fan-out and fan-in patterns have been designed and fabricated. For example, a computer-calculated reconstruction pattern of a DOE that performs a 4:2:1 fan-out function
Fig. 16. Large scale SEM photomicrograph illustrating the thermally evaporated indium bump uniformity.

Fig. 17. Measurement of post-flip-chip bonded modulator reflectivity.

Fig. 18. Computer calculated reconstruction, and experimental reconstruction, of a DOE that implements a weighted 3 x 3 fan-out, with relative intensities of 4:2:1.

Fig. 19. Reconstruction of diffractive optical element by two VCSEL's, showing fan-out and fan-in.

Fig. 20. GaAs antireflection-coated DOE that provides a 3 x 3 fan-out.
(gaussian-like, with a 4:2:1 ratio of intensities diffracted to the vertically-displaced center pixel, four nearest neighbor pixels, and four next-nearest-neighbor pixels, respectively) is shown in Fig. 18, alongside the experimental reconstruction pattern. This commercially-fabricated quartz-substrate DOE (QPS, Inc.) had a design etch depth of 1,920 Å and a measured etch depth of 1,978 Å, and exhibited an RMS diffracted-spot-intensity error of approximately 5%. Extensive analysis of DOE fabrication tolerances and errors across multiple fabrication runs of identical DOE arrays has shown that such DOE arrays can be produced with minimal run-to-run variances in performance.

The 4:2:1 DOE has been tested for fan-in accuracy as well, by employing two vertical cavity lasers within a VCSEL array for illumination of the DOE (by means of a pair of relay lenses), with results as shown in Fig. 19. These two resulting DOE reconstruction patterns were displaced diagonally along the next-nearest-neighbor direction, producing overlapping diffracted spot intensities in the central four spots, with an RMS error of approximately 9%. This result demonstrates both the fan-out and fan-in capability of the interconnection system, and will be tested next in a proximity-coupled geometry with multiple elements within a DOE array.

In a multichip module such as that described herein, the number of optical surfaces is considerable, and antireflection coating techniques must be carefully employed. We have previously developed low-reflectivity indium-tin-oxide (ITO) antireflection coatings for high-index compound semiconductor substrates, and have recently applied them to the AR-coating of DOE's fabricated in GaAs. An example of an AR-coated GaAs DOE that provides a 3 x 3 weighted fan-out is shown in Fig. 20. Experimental measurements of the diffracted order intensities both before and after AR coating demonstrated not only the expected large improvement in optical throughput efficiency, but also a reduction in RMS diffracted-spot-error from 15% to 6%. The use of stratified volume diffusive optical elements (SVOE) for highly non-local to global interconnection functions, as described in Sect. 3, carries with it a similar AR-coating issue, particularly in view of the large number of additional optical surfaces involved in this case. Minimization of fabrication errors is important for successful implementation of the overall SVOE interconnection function.

7.7 Photonic Multichip Module Integration Issues

Given the emergence of functional components as described above, a wide range of issues pertain to the successful integration of photonic multichip modules that comprise both hybrid substrates and mixed analog/digital representations. These issues include the computational complexity implemented per unit power dissipation, the analog accuracy achievable within each processing stage as well as overall after multiple stages, the total power consumption with resultant thermal dissipation and uniformity, the manufacturability and fabrication tolerances of hybrid-integration alignment techniques, the development of integrated computer-aided design (CAD) techniques for hybrid electronic/photic systems, and the assessment of inherent design tradeoffs.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge their many collaborators in this research effort, including Prof. P. Daniel Dapkus (USC), Prof. Nicholas George (University of Rochester, Institute of Optics), Prof. Gregory P. Nordin (University of Alabama, Huntsville), Dr. Kartik Anantharayanan (USC/Motorola), Dr. Ladan Shams (USC/Caltech), Dr. Suresh Subramaniam (USC), Dr. Moshe Bar (USC), Jaeyoun Cho, Hsing-Hua Fan, Hung-Min Jen, Po-Tsung Lee, Jaw-Chyng (Lorman) Lue, Nankyoung Suh, and Joshua Wyner. Research on adaptive vision sensors at the University of Southern California is supported in part by the FY 98 Multidisciplinary University Research Initiative (MURI): "Adaptive Optoelectronic Eyes: Hybrid Sensor Processor Architectures", Office of the Undersecretary of Defense (DDR&E), administered through the Army Research Office; in part by an FY 99 Defense University Research Instrumentation Program (DURIP) grant, also administered through the Army Research Office; and in part by Teledyne Electronic Technologies, the Eastman Kodak Company, Xerox Corporation, and the TRW Foundation. One of us (ART) gratefully acknowledges the significant hospitality, collegiality, and resources afforded to him during an ongoing sabbatical leave at the California Institute of Technology and the University of Rochester, Institute of Optics.

REFERENCES


