ARTICLE **Communicated by Tobias Delbruck**

A Multichip Neuromorphic System for Spike-Based Visual Information Processing

R. Jacob Vogelstein
jvogelst@jhu.edu
Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, U.S.A.

Udayan Mallik
udayan@gmail.com
Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, U.S.A.

Eugenio Culurciello
eugenio.culurciello@yale.edu
Department of Electrical Engineering, Yale University, New Haven, CT 06511, U.S.A.

Gert Cauwenberghs
gert@ucsd.edu
Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Ralph Etienne-Cummings
retienne@jhu.edu
Department of Electrical Engineering, Yale University, New Haven, CT 06511, U.S.A.

We present a multichip, mixed-signal VLSI system for spike-based vision processing. The system consists of an 80 $\times$ 60 pixel neuromorphic retina and a 4800 neuron silicon cortex with 4,194,304 synapses. Its functionality is illustrated with experimental data on multiple components of an attention-based hierarchical model of cortical object recognition, including feature coding, salience detection, and foveation. This model exploits arbitrary and reconfigurable connectivity between cells in the multichip architecture, achieved by asynchronously routing neural spike events within and between chips according to a memory-based look-up table. Synaptic parameters, including conductance and reversal potential, are also stored in memory and are used to dynamically configure synapse circuits within the silicon neurons.

1 Introduction

The brain must process sensory information in real time in order to analyze its surroundings and prescribe appropriate actions. In contrast, most simulations of neural functions to date have been executed in software programs that run much more slowly than real time. This places fundamental limits on the kinds of studies that can be done, because most software neural networks are unable to interact with their environment. However, software models have the advantages of being flexible, reconfigurable, and completely observable, and much has been learned about the brain through the use of software.

Neuromorphic hardware aims to emulate the functionality of the brain using silicon analog of biological neural elements (Mead, 1989). Typically, unlike most software, these hardware models can operate in real time (or even faster than their biological counterparts), providing the opportunity to create artificial nervous systems that can interact with their environment (Horiuchi & Koch, 1999; Indiveri, 1999; Simoni, Cymbalyuk, Sorensen, Calabrese, & DeWeerth, 2001; Indiveri, Murer, & Kramer, 2001; Jung, Brauer, & Abbas, 2001; Cheely & Horiuchi, 2003; Lewis, Etienne-Cummings, Hartmann, Cohen, & Xu, 2003; Zaghloul & Boahen, 2004; Reichel, Leichti, Presser, & Liu, 2005). Unfortunately, silicon designs take a few months to be fabricated, after which they are usually constrained by limited flexibility, so fixing a bug or changing the system's operation may require more time than that required for an equivalent software model (although a mature hardware design can be reused in many different systems; see, e.g., Zaghloul & Boahen, 2004). Additionally, the models are not usually as detailed as software models due to the limited computational primitives available from silicon transistors and the deliberate use of reductionist models to simplify the hardware infrastructure by reducing the dimensionality of parameter space.

Reconfigurable neuromorphic systems represent a compromise between fast, dedicated silicon hardware and slower but versatile software. They are useful for studying real-time operation of high-level (e.g., cortical), large-scale neural networks and prototyping neuromorphic systems prior to fabricating application-specific chips. Instead of hardwiring connections between neurons, most reconfigurable neuromorphic systems use the address-event representation (AER) communication protocol (Sivilotti, 1991; Lazzaro, Wawrzynek, Mahowald, Sivilotti, & Gillespie, 1993; Mahowald, 1994). In an address-event system, connections between neurons are emulated by time-multiplexing neural events (also called action potentials, or spikes) onto a fast serial bus, and AER "synapses" are implemented with encoders and decoders that monitor the bus and route incoming and outgoing spikes to their appropriate neural targets. These systems can be reconfigured by changing the routing functions, and multiple authors have demonstrated versions of AER that use memory-based

projective field mappings toward this end (Deiss, Douglas, & Whatley, 1999; Higgins & Koch, 1999; Goldberg, Cauwenberghs, & Andreou, 2001; Häfliger, 2001; Liu, Kramer, Indiveri, Delbrück, & Douglas, 2002; Indiveri, Chicca, & Douglas, 2004; Ros, Ortigosa, Agis, Carrillo, & Arnold, 2006).

We have developed a reconfigurable multichip AER-based system for emulating cortical spike processing of visual information. The system uses one AER subnet to communicate spikes between an 80×60 pixel silicon retina (Culurciello, Etienne-Cummings, & Boahen, 2003; Culurciello & Etienne-Cummings, 2004) and a 4800-neuron silicon cortex (Vogelstein, Mallik, & Cauwenberghs, 2004), and a second AER subnet to communicate spikes between cortical cells (see Figure 1). Each cell in the silicon retina converts light intensity into spike frequency (see section 2.2). Each cell in the silicon cortex implements an integrate-and-fire neuron with conductance-like synapses (see section 2.1). Neural connectivity patterns and synaptic parameters are stored in digital memory, allowing "virtual synapses" to be implemented by routing spikes to one or more locations on the silicon cortex.

A number of multichip reconfigurable neuromorphic systems have been described in the literature (Goldberg et al., 2001; T. Horiuchi & Hynna, 2001; Taba & Boahen, 2003; Arthur & Boahen, 2004; Indiveri et al., 2004; Paz et al., 2005; Riis & Häfliger, 2005; Serrano-Gotarredona et al., 2006; Zou, Bornat, Tomas, Renaud, & Destexhe, 2006), but ours differs in some important ways. First, the 4800-neuron silicon cortex is the largest general-purpose neuromorphic array presented to date. Second, unlike the other systems, our silicon cortex has no local or hardwired connectivity, and each neuron implements a synapse with programmable weight and equilibrium potential, so all 4800 neurons can be utilized for any arbitrary connection topology, limited only by the capacity of the digital memory (and bandwidth if real-time operation is desired). Third, the hardware infrastructure supports up to 1 million address events per second and allows real-time operation of large networks. Finally, the silicon cortex can function as a standalone AER transceiver, enabling the creation of even larger networks by connecting multiple cortices together.

To explicate all of these features, we conducted a series of four experiments within the common framework of a hierarchical model of visual information processing (see Figure 2). Section 3.1 demonstrates the speed and scale of the hardware by operating all 4800 cortical cells in real time. Sections 3.2 and 3.4 highlight the versatility of the hardware by reconfiguring the cortex into both feedforward and feedback networks. Section 3.3 uses the neurons' dynamically programmable synapses to multiplex a wide range of synaptic connections onto individual cells. Finally, section 4 details how the complete hierarchical model could be implemented in real time by partitioning the network into functional units, each organized around one silicon cortex.

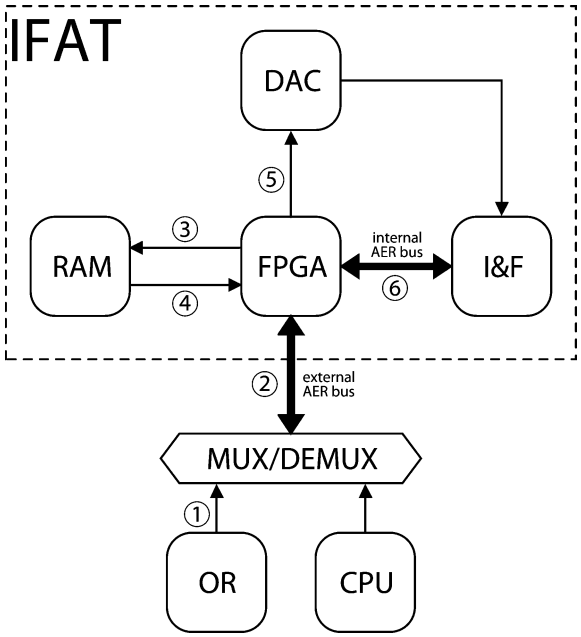
2 Hardware

Every neuron on the silicon retina and cortex is assigned a unique address at design time, which is transmitted as an address event (AE) over an AER bus when that neuron fires an action potential. All of the address-event transactions in the multichip system illustrated in Figure 1 are processed by a field programmable gate array (FPGA) located within the integrate-and-fire array transceiver (IFAT) component (Vogelstein, Mallik, & Cauwenberghs, 2004). In addition to the FPGA, the IFAT contains 128 MB of nonvolatile digital memory in a 4 MB $\times$ 32-bit array (RAM), the 4800-cell silicon cortex, and an 8-bit digital-to-analog converter (DAC) required to operate the silicon cortex (see section 2.1).

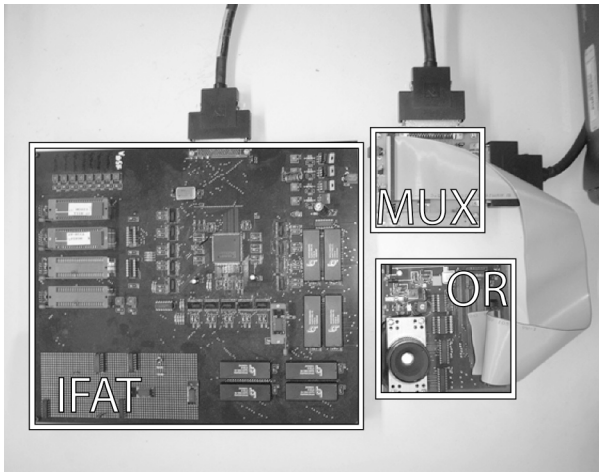
The path of an AE through the system is illustrated in Figure 1. In this example, an outgoing presynaptic address from the silicon retina is placed on the external AER bus and captured by the FPGA, which uses the neuron's address as an index into the RAM. Each of the 4,194,304 lines of RAM stores information on a single synaptic connection, including its equilibrium potential, its synaptic weight, and the destination (postsynaptic) address (see Figure 3; Deiss et al., 1999). This information is then used by the FPGA to activate a particular cell in the silicon cortex. Divergent connectivity is achieved by allowing the FPGA to access sequential lines in memory until it retrieves a stop code, as well as by implementing reserved address words that are used to activate multiple cells on one or more chips simultaneously.

Each application in section 3 requires a different number of synapses. Full-field spatial feature extraction (see section 3.1) and salience detection (see section 3.2) can be implemented with approximately 19,200 synapses each. Spatial acuity modulation (see section 3.3) with a 16 $\times$ 16 fovea surrounded by three concentric rings of geometrically decreasing resolution uses 60,736 synapses. And computing the maximum of N neurons (see section 3.4) relies $N + N^2$ synapses (903 in the example shown here, or 90,300 to compute the maximum of all local salience estimates for an 80 $\times$ 60-pixel visual field).

Figure 1: (a) Block diagram of the multichip system with silicon retina (OR) and cortex (I&F). The silicon cortex is located within the IFAT subsystem, which also contains a field programmable gate array (FPGA), digital memory (RAM) for storing the synaptic connection matrix, and a digital-to-analog converter (DAC) required to operate the I&F chips. The FPGA controls two AER buses: one internal bus for events sent to and from the silicon cortex and one external bus for events sent to and from external neuromorphic devices or a computer (CPU). Circled numbers 1–6 highlight the path of incoming events from the OR (see section 2.2). (b) Photograph of the system.


IFAT
DAC
⑤
③
RAM
FPGA
internal
AER bus
I&F
④
⑥
②
external
AER bus
MUX/DEMUX
①
OR
CPU

(a)


MUX
OR
IFAT

(b)

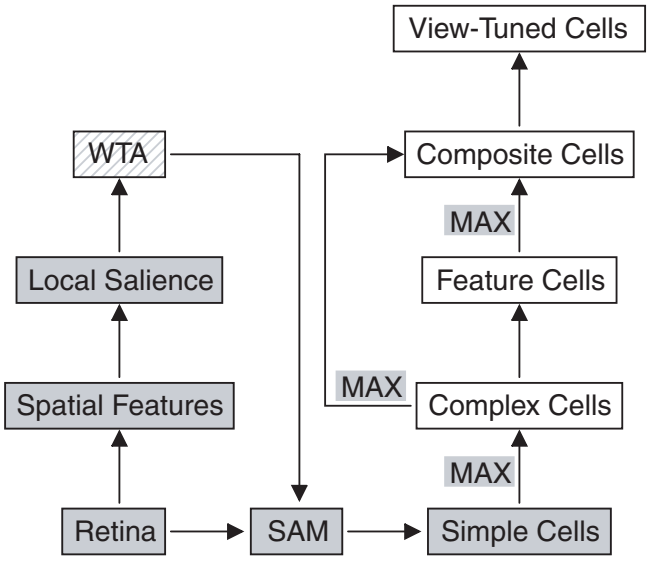


Figure 2: Hierarchical model of visual information processing based on work by Riesenhuber and Poggio (1999). Spatial features are extracted from retinal images using a small set of oriented spatial filters, whose outputs are combined to form estimates of local salience. The region with maximum salience is selected by a winner-take-all network (WTA) and used to foveate the image by spatial acuity modulation (SAM). A large set of simple cells with many different preferred orientations is then used to process this bandwidth-limited signal. The simple cells' outputs are combined with a MAX function to form spatially invariant complex cells, and the resulting data are combined in various ways to form feature cells, composite cells, and, finally, "view-tuned cells" that selectively respond to a particular view of an object. Shaded boxes indicate functions computed by the IFAT. The WTA function is not explicitly computed but is generated as an intermediate result of the MAX operation.

2.1 Silicon Cortex. The silicon cortex used in this system is composed of 4800 random-access integrate-and-fire (I&F) neurons implemented on two custom aVLSI chips, each of which contains 2400 cells (Vogelstein, Mallik, & Cauwenberghs, 2004). All 4800 neurons are identical; every one implements a conductance-like model of a general-purpose synapse using a switched-capacitor architecture. The synapses have two internal parameters—the synaptic equilibrium potential and the synaptic weight—that can be set to different values for each incoming event. Additionally, the range of synaptic weights can be extended by two dynamically controlled external parameters: the probability of sending an event and the number of postsynaptic events sent for every presynaptic event (Koch, 1999; Vogelstein, Mallik,

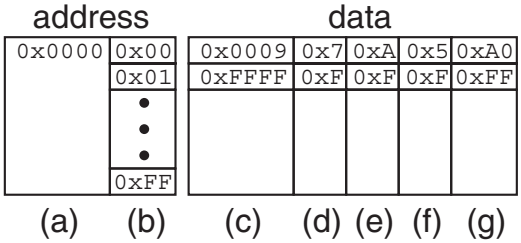


Figure 3: Example of IFAT RAM contents. Each line stores parameters for one synaptic connection. The presynaptic neuron's address is used as a base index (a) into the lookup table, while the FPGA increments an offset counter (b) as it iterates through the list of postsynaptic targets (c). Synaptic weight is represented as a product of the three values stored in columns d–f, which represent the size of the postsynaptic response to an event, the number of postsynaptic events to generate for each presynaptic event, and the probability of generating an event, respectively (Koch, 1999; Vogelstein et al., 2005). The synaptic equilibrium potential is stored in column g and is used to control the DAC (see Figure 1). The reserved word shown at offset 0×01 is used to indicate the end of the synapse list for presynaptic neuron 0×0000, so the data at offsets 0×02–$0 \times$ FF is undefined.

Cauwenberghs, Culurciello, & Etienne-Cummings, 2005). By storing values for these parameters along with the pre- and postsynaptic addresses in RAM (see Figure 3), the FPGA on the IFAT can implement a different type of synapse for every virtual connection between neurons. The maximum rate of event transmission from the silicon cortex and its associated IFAT components is approximately 1,000,000 AE per second and is primarily limited by the speed of the internal arbitration circuits.

2.2 Silicon Retina. The silicon retina used in this system is called the octopus retina (OR) because its design is based on the phototransduction mechanism found in the retinae of octopi (Culurciello, et al., 2003; Culurciello & Etienne-Cummings, 2004). Functionally, the OR is an asynchronous imager that translates light intensity levels into interspike interval times at each pixel. However, unlike a biological octopus retina, in which each photosensor's output travels along a dedicated axon to its target(s), all of the OR's outputs are collected on its AER bus and transmitted serially off-chip to the IFAT. Under uniform indoor lighting (0.1 mW/cm^2), the OR produces an average of 200,000 address events per second (41.7 effective fps) while consuming 3.4 mW. However, most visual scenes do not have uniform lighting, so the typical range of event rates for this application is approximately 5,000 to 50,000 address events per second.

3 Results: Spike Domain Image Processing

As described in section 1, we chose to exploit different aspects of the reconfigurable multichip system in a series of experiments organized around the common framework of a hierarchical model of visual information processing (see Figure 2). This model was selected to showcase the system's versatility because each processing stage places different requirements on the fundamentally similar neurons within the silicon cortex, just as sensory processing in the human cortex requires fundamentally similar pyramidal cells in different locations to execute different functions (Kandel, Schwartz, & Jessell, 2000).

In the model (see Figure 2), retinal outputs are first processed through oriented spatial filters that highlight regions of high contrast (Mallik, Vogelstein, Culurciello, Etienne-Cummings, & Cauwenberghs, 2005). This information is then used by a salience detector network that focuses attention on a region of interest and decreases the resolution in surrounding areas to reduce the number of data being used for computations and transmission (Vogelstein et al., 2005). Within the foveated center, data from the local spatial filters are combined with a nonlinear pooling function to form global spatial filters, which are subsequently combined to create feature cells, composite cells, and view-tuned cells (Riesenhuber & Poggio, 1999).

Results from implementations of the first few stages of this network computed entirely in the spike domain on our multichip system are described below. Because this reconfigurable system is optimized not for any particular application but for flexibility, these data are primarily intended to illustrate the breadth of computations that can be performed and confirm the general functionality of the proposed network architecture.

3.1 Spatial Feature Extraction. In the human visual cortex, the first stage of processing is spatial feature extraction, performed by simple cells (Kandel e al., 2000). Simple cells act as oriented spatial filters that detect local changes in contrast, and their receptive fields and preferred orientations are both functions of the input they receive from the retina. Spatial feature extraction is used twice in the hierarchical model of visual information processing in Figure 2—first coarsely over the entire visual field to estimate salience and then more finely within a small region of interest.

Figure 4 illustrates how the silicon cortex can be used to perform spatial feature extraction by emulating eight different simple cell types (see Figure 4B1–I1) with overlapping receptive fields (Mallik et al., 2005). Note that because the OR output is proportional to light intensity, these simple cells respond to intensity gradients, not contrast gradients. In this example, each cortical cell integrates inputs from four pixels in the OR, two of which make excitatory synapses and two of which make inhibitory synapses. The excitatory and inhibitory synaptic weights are balanced so that there is no net response to uniform light.

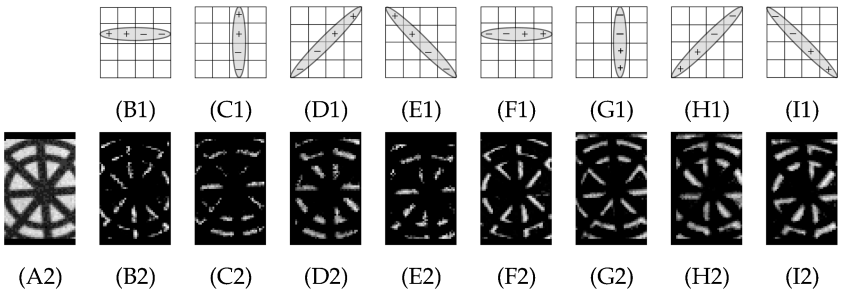


Figure 4: (B1–I1) Orientation-selective kernel compositions in the simple cell network. Each simple cell has a 4×1 receptive field and receives two excitatory (+) and two inhibitory (−) inputs from the silicon retina. (A2) Original image captured by silicon retina. (B2–I2) Frames captured from real-time video sequences of retinal images processed by simple cell networks implemented on the silicon cortex. Each frame is composed from the output of 4800 simple cells that were all configured in the orientation shown above (e.g., B2 shows output from cells with receptive field drawn in B1) (Mallik et al., 2005).

Figures 4B2 to I2 show a few sample frames from real-time video images generated by a simple cell network implemented on the silicon cortex (Mallik et al., 2005). Because both the silicon cortex and the silicon retina contain 4800 neurons, there is necessarily a trade-off between the spacing of similarly oriented simple cells throughout the visual field and the number of differently oriented simple cells with overlapping receptive fields. For the images in Figure 4, this trade-off was resolved in favor of increased resolution: each frame was captured from a different configuration of the system wherein all 4800 simple cells had identical preferred orientations. However, we have also generated similar results with lower resolution when the cortex is configured to simultaneously process two or four different orientations (data not shown). In addition to illustrating the principle of spatial feature extraction, these data demonstrate that the multichip system is capable of executing large-scale networks in real time.

3.2 Salience Detection. Salient regions of an image are areas of high information content. In the hierarchical model of visual information processing, estimates of salience are used to select a region of interest that will undergo further processing. There are many ways to compute salience; one simple technique uses the magnitude of spatial derivatives of light intensity within a given region as an approximate measure. A neural network architecture for computing this metric is illustrated in Figure 5. In this scheme, outputs from simple cells with overlapping receptive fields and different preferred orientations are linearly pooled by second-level cells to form estimates of local salience. A winner-take-all (WTA) circuit with one

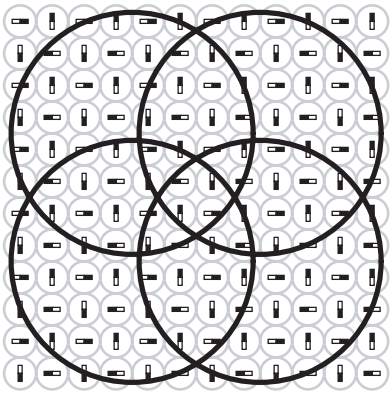
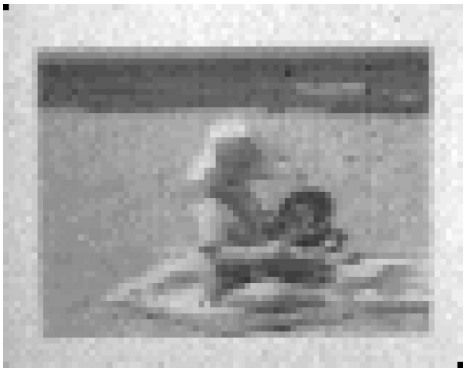


Figure 5: Pictorial representation of network for computing salience. Each local salience detector cell (large black circle) integrates inputs from a neighborhood of simple cells (small gray circles) with multiple different preferred orientations. A large output from a salience detector cell indicates a strong change in spatial image intensity, which frequently coincides with high information content.

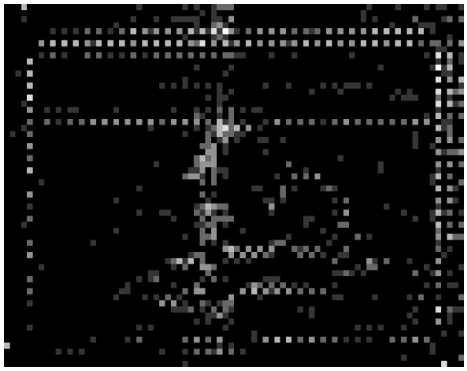
input from each second-level cell could then be used to detect the region of overall greatest salience (see section 3.4).

Data from the silicon cortex configured to compute salience are illustrated in Figure 6 (Vogelstein et al., 2005). Figure 6a shows a raw image generated by the silicon retina (focused on a photograph) under normal indoor lighting. This image is then processed by the coarse-oriented spatial filtering network described in section 3.1, with four sets of 1200 simple cells simultaneously processing horizontal and vertical intensity changes (cell types RF5–RF8 as designated by Figure 4; see Figure 6b for simple cell output). To compute the local salience estimates (see Figure 6c), outputs from 64 simple cells of various orientations spanning an 8×8-pixel visual space are pooled by a single second-level cell. Smooth transitions between adjacent estimates are ensured by shifting each second-level cell's receptive field by four pixels in either the horizontal or vertical direction (see Figure 5).

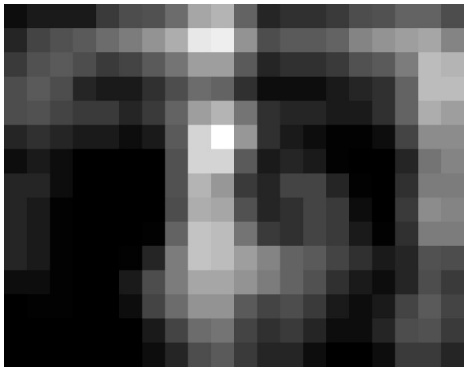
Because the silicon cortex contains only 4800 neurons, the spatial filtering and salience detection cannot both be implemented for the entire visual field simultaneously. Therefore, to generate the images in Figure 6, each stage of network processing was executed serially. This was achieved by using a computer to log the output of the silicon cortex configured as spatial filters, changing the cortical network to pool simple cell outputs, and then playing back the sequence of events to the silicon cortex to compute the local salience estimates. This strategy highlights the versatility of the hardware; the same approach could also be used to perform a WTA or MAX operation on the local salience estimates (see section 3.4). Moreover, this



(a)



(b)



(c)

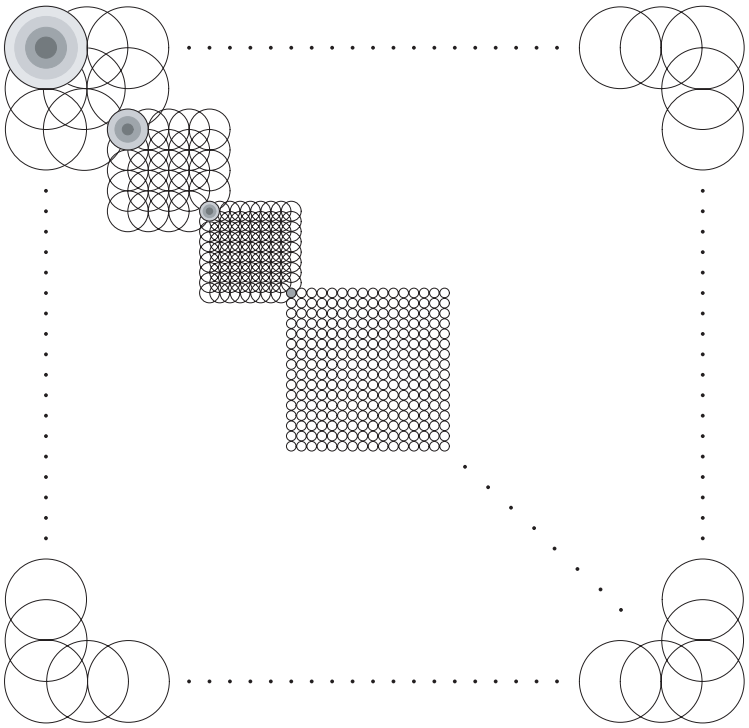
Figure 6: (a) Frame capture of image generated by silicon retina. (b) Output of feature detectors (simple cells) using silicon retina data as input. (c) Output of salience detectors using simple cell data as input (Vogelstein et al., 2005).

technique faithfully simulates the operation of any hierarchical feedforward network (feedback can be implemented within a given processing stage), while allowing analysis of each stage's output independently.

3.3 Spatial Acuity Modulation. In a human retina, there is a natural distribution of photoreceptors throughout the visual field, with the highest density of light-sensitive elements in the center of vision and lower numbers of photoreceptors in the periphery (Kandel et al., 2000). In combination with reflex circuits that guide the center of the eye to salient regions in space, this configuration conserves computational resources and transmission bandwidth between levels of the network (see Figure 2). The same principles of conservation are important in our multichip hardware system. However, because the silicon retina used as the frontend to our visual system has a fixed position and uniform resolution throughout its field of view, we modulate the spatial acuity of the image in the address domain.

Spatial acuity modulation is performed by pooling the outputs from neighboring pixels in the retina onto single cells in the silicon cortex, using overlapping gaussian kernels with broad spatial bandwidths in the periphery and narrow bandwidths in the center of the image (see Figure 7a). Because synaptic weights in the IFAT can be dynamically configured using multiple degrees of freedom, these kernel functions can be reasonably approximated using discrete changes to the internal weight variable, the number of output events sent per input event, and the synaptic equilibrium potential. To relocate the center of vision (called the *fovea*) to an area of interest in the visual field, the system could reprogram the RAM with a different connectivity pattern, but instead, the FPGA performs simple arithmetic manipulations to incoming address events, adding or subtracting a fixed value from the row and column addresses to offset their position (Vogelstein, Mallik, Culurciello, Etienne-Cummings, & Cauwenberghs, 2004).

Figure 7: (a) Pictorial representation of the spatial acuity modulation network with fovea positioned over the center of the image. Circles represent cortical cells, with the diameter of the circle proportional to the size of the spatial receptive field. The outermost cortical cells integrate inputs from 64 retinal cells, while the innermost cortical cells receive inputs from a single retinal cell. One cell within each group is shaded to illustrate the pattern of synaptic weights onto the cells in that group. Light shading represents a small synaptic weight, and dark shading represents a large synaptic weight. (b) Example image output from silicon retina with spatial acuity modulation performed by silicon retina. The nine subfigures show how the output varies as the center of vision (fovea) moves from the top left corner of the image to the bottom right corner (Vogelstein, Mallik, Culurciello, et al., 2004).



(a)

(b)

An example image with nine different foveations is shown in Figure 7b. With a 16 × 16-pixel fovea surrounded by k concentric rings of geometrically decreasing resolution, the number of cortical neurons (M) required to represent the foveated image from a $N \times N$-pixel retina is given by $M = 16^2 + 2^{(k+2)} - 4 \ll N^2$. This allows for a significant reduction (75% for the example shown here) in the number of address events processed by the hardware as well as a reduced communication cost of transmitting an image "frame" (Vogelstein, Mallik, Culurciello, et al., 2004).

3.4 MAX Computation. The cortical process of object recognition is modeled in Figure 2 with a series of linear and nonlinear poolings of simple cell outputs (Riesenhuber & Poggio, 1999). In the first set of computations, outputs from simple cells with similar preferred orientations and different receptive fields are combined with a maximum operation to form complex cells, which essentially act as position-invariant-oriented spatial filters. Because of the large bandwidth required, the maximum is taken over only a subset of the image with high salience. This is similar to the attention spotlight model of human perception (Posner, Snyder, & Davidson, 1980; Eriksen & St. James, 1986).

The maximum operation (MAX) is defined here as a nonlinear saturating pooling function on a set of inputs whose output codes the magnitude of the largest input, regardless of the number and levels of lesser inputs. A neural implementation of the MAX is illustrated in Figure 8a, where a set of input neurons $\{x\}$ causes the output neuron z to generate spikes at a rate proportional to the input with the fastest firing rate. The MAX operation is closely related to the WTA function, except that a standard WTA network allows one of many potential output neurons to be active, and that neuron's activity level is dependent on only the relative magnitude of the inputs, not their absolute value. (For distractor input spike frequencies up to about 80% of the maximum, the y neurons in the MAX network compute a WTA as an intermediate step toward computing the maximum. Higher distractor input spike frequencies can be accommodated by increasing the reciprocal inhibitory feedback between y neurons at the expense of the accuracy of the z neuron.) When used in a neural network to pool responses from different feature detectors, such as simple cells, a MAX neuron can simultaneously achieve high feature specificity and invariance (Riesenhuber & Poggio, 1999).

We implemented a MAX network model originally proposed by Yu, Giese, and Poggio (2002), shown in Figure 8a. The network is highly interconnected with all-to-all reciprocally inhibitory feedback connections between y neurons, confirming the ability of the silicon cortex to implement recurrent networks. The invariance of the network to the number of inputs n is illustrated in Figure 8b. Thirty configurations, with $n \in [1, 30]$, were tested on the silicon cortex. In each configuration, the networks were allowed to run for 60 seconds, with the x cells' inputs generated by

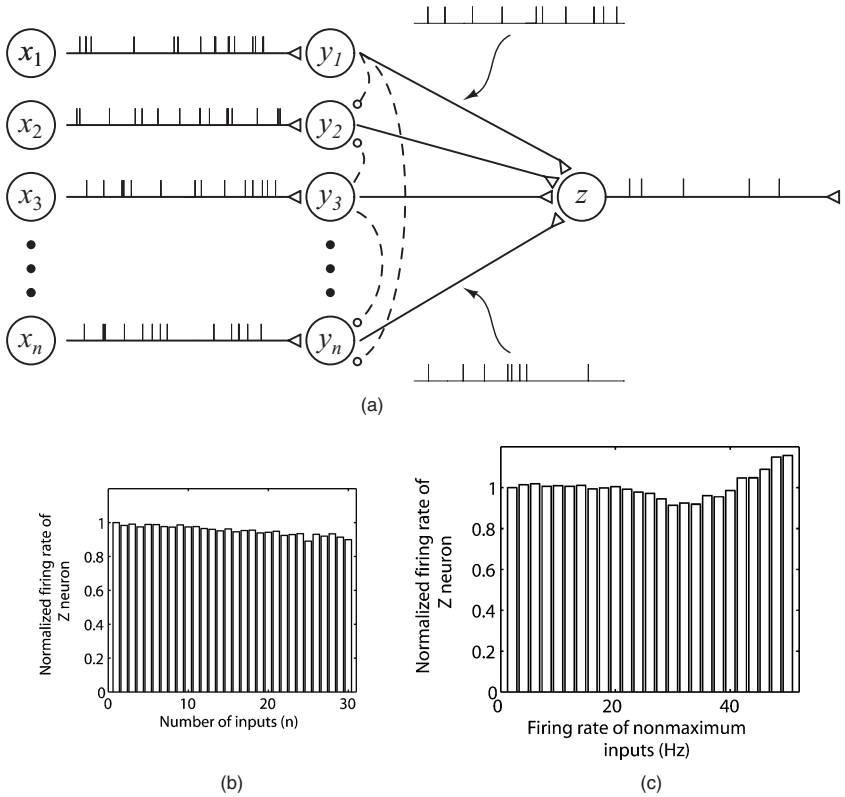


Figure 8: (a) Pictorial representation of MAX network. Excitatory connections are shown by solid lines and triangular synapses. Inhibitory connections are shown by dashed lines and circular synapses. Input to the x neurons is provided by a computer that generates independent homogeneous Poisson processes. Each y neuron makes inhibitory synapses with all other y neurons, but only some connections are shown for clarity. The output of the z neuron is monitored by a computer. (b) Invariance of the silicon cortex-based MAX network to the number of inputs, n. (c) Invariance of the silicon cortex-based MAX network to the firing rate of nonmaximum inputs.

independent homogeneous Poisson processes with parameter $\lambda = 30$ Hz for nonmaximum inputs and $\lambda = 50$ Hz for $x_{\max}$. As shown in Figure 8b, the firing rate of the output neuron z is approximately the same for any value of n. In addition to invariance toward the number of inputs n, the MAX network is invariant to the firing rate of nonmaximum inputs. This was tested by fixing $n = 25$ and allowing λ to vary for nonmaximum inputs from 2 Hz to 50 Hz. The results are shown in Figure 8c, where the firing

rate of the output neuron z is seen to be roughly constant for rates up to 40 Hz. Because the inputs to the network are stochastic, its performance is weakened by very high firing rates of nonmaximal inputs or very large numbers of nonmaximal inputs.

4 Discussion

The above experiments have demonstrated the operation of the primary components of the hierarchical visual processing model (see Figure 2). A full implementation would require larger numbers of neurons than can be simultaneously accommodated in the present multichip system. However, to construct the complete network, multiple IFATs could be connected together, one for each stage of processing, to form a very large silicon cortex. In addition to providing more neurons, this arrangement would reduce the constraints on bandwidth. For example, the spatial feature extraction architecture described requires each retinal cell to project to 64 simple cells at full resolution. However, if the silicon retina produces 50,000 AE per second and each IFAT is limited to processing 1,000,000 AE per second, the maximum fan-out from each retinal cell to any individual IFAT is only 20. By dividing the orientations among multiple IFATs (see Choi, Merolla, Arthur, Boahen, & Shi, 2005), a fan-out of 64 could easily be sustained without overtaxing the system. Furthermore, because the number of connections between neurons within a given level is larger than the number of connections between neurons in different levels (especially in recurrent networks like the MAX network), giving each processing stage its own IFAT will conserve energy by reducing the number of events transmitted across the external AER bus.

Connecting multiple IFATs together in a feedforward structure requires little hardware beyond that shown in Figure 1. Because each IFAT functions as an address event transceiver, sending and receiving events according to a lookup table in RAM, it needs to know only the addresses of neurons in the subsequent processing stage to communicate with them directly over its AER output bus. For recurrent connections between IFATs, a central arbiter would be required to merge incoming events from multiple AER buses and route them to their appropriate targets. This can be achieved with simple logic circuits implemented in a fast complex programmable logic device (CPLD) or FPGA.

The same hardware that supports multiple IFATs could also support multiple neuromorphic sensors. Because the silicon cortex implements a general-purpose neural model, it is well suited for multimodal computations. Even without additional hardware, the system in Figure 1 can be adapted for sensory modalities other than vision; any neuromorphic sensor using AER can be attached to the port currently occupied by the OR.

Under normal operating conditions, such as when implementing the networks described in section 3, each IFAT neuron executes approximately 1 million to 10 million operations per second (MOPS; addition, subtraction,

multiplication, and comparison are all considered single operations). The exact number of MOPS can be computed if the number of input and output spikes are known, because each input event requires approximately six basic operations per neuron, and every output event requires two or three operations (Vogelstein, Mallik, & Cauwenberghs, 2004). However, if the network architecture is optimized to take advantage of parallel activation of multiple cells (see section 2), the number of OPS increases significantly. For example, if every incoming spike is routed to an entire row of neurons simultaneously, the IFAT would perform more than 360 operations per spike, or at least 360 MOPS for 1,000,000 input spikes per second. In the current hardware, the upper bound on operations per second is 19,200 MOPS if all 2400 neurons on one chip are activated simultaneously, or 38,400 MOPS if all 4800 neurons in the silicon cortex are used in parallel (these figures will improve with technology and are not fundamental limits of our approach). To date, we have utilized only the parallel activation functions of the IFAT to implement global "leakage" events, but one can easily imagine future applications that take advantage of this feature, such as a fully connected winner-take-all network (Abrahamsen, Häfliger, & Lande, 2004; Oster & Liu, 2004).

5 Conclusion

We have described a novel multichip neuromorphic system capable of processing visual images in real time. The system contains a silicon cortex with 4800 neurons that can be (re)configured into arbitrary network topologies for processing any spike-based input. Results from the first few stages of a hierarchical model for salience detection and object recognition confirm the utility of the system for prototyping large-scale sensory information processing networks. Future work will focus on increasing the number of neurons in the silicon cortex, so that the entire hierarchical visual processing model can be tested while it interacts with the environment.

Acknowledgments

This work was partially funded by the National Science Foundation, the National Institute on Aging, the Defense Advanced Research Projects Agency, and the Institute for Neuromorphic Engineering. Additionally, R.J.V. is supported by an NSF Graduate Research Fellowship.

References

Abrahamsen, J., Häfliger, P., & Lande, T. S. (2004). A time domain winner-take-all network of integrate-and-fire neurons. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 5, pp. 361–364). Piscataway, NJ: IEEE.

Arthur, J. V., & Boahen, K. A. (2004). Recurrently connected silicon neurons with active dendrites for one-shot learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (Vol. 3, pp. 1699–1704). Piscataway, NJ: IEEE.

Cheely, M., & Horiuchi, T. (2003). A VLSI model of range-tuned neurons in the bat echolocation system. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 4, pp. 872–875). Piscataway, NJ: IEEE.

Choi, T. Y. W., Merolla, P. A., Arthur, J. V., Boahen, K. A., & Shi, B. E. (2005). Neuromorphic implementation of orientation hypercolumns. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, *52*(6), 1049–1060.

Culurciello, E., & Etienne-Cummings, R. (2004). Second generation of high dynamic range, arbitrated digital imager. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 4, pp. 828–831). Piscataway, NJ: IEEE.

Culurciello, E., Etienne-Cummings, R., & Boahen, K. A. (2003). A biomorphic digital image sensor. *IEEE Journal of Solid-State Circuits*, *38*(2), 281–294.

Deiss, S. R., Douglas, R. J., & Whatley, A. M. (1999). A pulse-coded communications infrastructure for neuromorphic systems. In W. Maass & C. M. Bishop (Eds.), *Pulsed neural networks* (pp. 157–178). Cambridge, MA: MIT Press.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, *40*, 225–240.

Goldberg, D. H., Cauwenberghs, G., & Andreou, A. G. (2001). Probabilistic synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons. *Neural Networks*, *14*(6–7), 781–793.

Häfliger, P. (2001). Asynchronous event redirecting in bio-inspired communication. In *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems* (Vol. 1, pp. 87–90). Piscataway, NJ: IEEE.

Higgins, C. M., & Koch, C. (1999). Multi-chip neuromorphic motion processing. In D. S. Wills & S. P. DeWeerth (Eds.), *Proceedings of the 20th Anniversary Conference on Advanced Research in VLSI* (pp. 309–323). Los Alamitos, CA: IEEE Computer Society.

Horiuchi, T., & Hynna, K. (2001). Spike-based VLSI modeling of the ILD system in the echolocating bat. *Neural Networks*, *14*, 755–762.

Horiuchi, T. K., & Koch, C. (1999). Analog VLSI-based modeling of the primate oculomotor system. *Neural Computation*, *11*, 243–265.

Indiveri, G. (1999). Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *46*(11), 1337–1347.

Indiveri, G., Chicca, E., & Douglas, R. J. (2004). A VLSI reconfigurable network of integrate-and-fire neurons with spike-based learning synapses. In *Proceedings of the European Symposium on Artificial Neural Networks* (pp. 405–410). Bruges, Belgium: D-Facto.

Indiveri, G., Murer, R., & Kramer, J. (2001). Active vision using an analog VLSI model of selective attention. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *48*(5), 492–500.

Jung, R., Brauer, E. J., & Abbas, J. J. (2001). Real-time interaction between a neuromorphic electronic circuit and the spinal cord. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *9*(3), 319–326.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of neural science* (4th ed.). New York: McGraw-Hill.

Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.

Lazzaro, J., Wawrzynek, J., Mahowald, M., Sivilotti, M., & Gillespie, D. (1993). Silicon auditory processors as computer peripherals. *IEEE Transactions on Neural Networks*, *4*(3), 523–528.

Lewis, M. A., Etienne-Cummings, R., Hartmann, M. H., Cohen, A. H., & Xu, Z. R. (2003). An in silico central pattern generator: Silicon oscillator, coupling, entrainment, physical computation and biped mechanism control. *Biological Cybernetics*, *88*(2), 137–151.

Liu, S.-C., Kramer, J., Indiveri, G., Delbrück, T., & Douglas, R. (2002). Orientation-selective aVLSI spiking neurons. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Mahowald, M. (1994). *An analog VLSI system for stereoscopic vision*. Boston: Kluwer.

Mallik, U., Vogelstein, R. J., Culurciello, E., Etienne-Cummings, R., & Cauwenberghs, G. (2005). A real-time spike-domain sensory information processing system. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 3, pp. 1919–1922). Piscataway, NJ: IEEE.

Mead, C. (1989). *Analog VLSI and neural systems*. Reading, MA: Addison-Wesley.

Oster, M., & Liu, S.-C. (2004). A winner-take-all spiking network with spiking inputs. In *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems* (pp. 203–206). Piscataway, NJ: IEEE.

Paz, R., Gomez-Rodriguez, F., Rodriguez, M. A., Linares-Barranco, A., Jimenez, G., & Civit, A. (2005). Test infrastructure for address-event-representation communications. *Lecture Notes in Computer Science*, *3512*, 518–526.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

Reichel, L., Leichti, D., Presser, K., & Liu, S.-C. (2005). Robot guidance with neuromorphic motion sensors. In *Proceedings of the IEEE International Conference on Robotics and Automation* (pp. 3540–3544). Piscataway, NJ: IEEE.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riis, H. K., & Häfliger, P. (2005). An asynchronous 4-to-4 AER mapper. *Lecture Notes in Computer Science*, *3512*, 494–501.

Ros, E., Ortigosa, E. M., Agis, R., Carrillo, R., & Arnold, M. (2006). Real-time computing platform for spiking neurons (RT-spike). *IEEE Transactions on Neural Networks*, *17*(4), 1050–1063.

Serrano-Gotarredona, R., Oster, M., Lichtsteiner, P., Linares-Barranco, A., Paz-Vicente, R., Gómez-Rodríguez, F., et al. (2006). AER building blocks for multi-layer multi-chip neuromorphic vision systems. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems, 18* (pp. 1217–1224). Cambridge, MA: MIT Press.

Simoni, M. F., Cymbalyuk, G. S., Sorensen, M. Q., Calabrese, R. L., & DeWeerth, S. P. (2001). Development of hybrid systems: Interfacing a silicon neuron to a leech heart interneuron. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances*

in neural information processing systems, 13 (pp. 173–179). Cambridge, MA: MIT Press.

Sivilotti, M. (1991). *Wiring considerations in analog VLSI systems, with application to field-programmable networks*. Unpublished doctoral dissertation, California Institute of Technology.

Taba, B., & Boahen, K. A. (2003). Topographic map formation by silicon growth cones. In S. T. S. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 1139–1146). Cambridge, MA: MIT Press.

Vogelstein, R. J., Mallik, U., & Cauwenberghs, G. (2004). Silicon spike-based synaptic array and address-event transceiver. In *Proceedings of the IEEE International Symposium on Circuits and Systems* (Vol. 5, pp. 385–388). Piscataway, NJ: IEEE.

Vogelstein, R. J., Mallik, U., Cauwenberghs, G., Culurciello, E., & Etienne-Cummings, R. (2005). Saliency-driven image acuity modulation on a reconfigurable silicon array of spiking neurons. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 1457–1464). Cambridge, MA: MIT Press.

Vogelstein, R. J., Mallik, U., Culurciello, E., Etienne-Cummings, R., & Cauwenberghs, G. (2004). Spatial acuity modulation of an address-event imager. In *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems* (pp. 207–210). Piscataway, NJ: IEEE.

Yu, A. J., Giese, M. A., & Poggio, T. A. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Computation*, *14*(12), 2857–2881.

Zaghloul, K. A., & Boahen, K. (2004). Optic nerve signals in a neuromorphic chip I: Outer and inner retina models. *IEEE Transactions on Biomedical Engineering*, *51*(4), 657–666.

Zou, Q., Bornat, Y., Tomas, J., Renaud, S., & Destexhe, A. (2006). Real-time simulations of networks of Hodgkin-Huxley neurons using analog circuits. *Neurocomputing*, *69*, 1137–1140.

Received March 29, 2006; accepted October 7, 2006.