An ON–OFF Orientation Selective Address Event Representation Image Transceiver Chip

Thomas Yu Wing Choi, *Student Member, IEEE*, Bertram E. Shi, *Fellow, IEEE*, and Kwabena A. Boahen

***Abstract*—This paper describes the electronic implementation of a four-layer cellular neural network architecture implementing two components of a functional model of neurons in the visual cortex: linear orientation selective filtering and half wave rectification. Separate ON and OFF layers represent the positive and negative outputs of two-phase quadrature Gabor-type filters, whose orientation and spatial-frequency tunings are electronically adjustable. To enable the construction of a multichip network to extract different orientations in parallel, the chip includes an address event representation (AER) transceiver that accepts and produces two-dimensional images that are rate encoded as spike trains. It also includes routing circuitry that facilitates point-to-point signal fan in and fan out. We present measured results from a 32×64 pixel prototype, which was fabricated in the TSMC0.25-μm process on a 3.84 by 2.54 mm die. Quiescent power dissipation is 3 mW and is determined primarily by the spike activity on the AER bus. Settling times are on the order of a few milliseconds. In comparison with a two-layer network implementing the same filters, this network results in a more symmetric circuit design with lower quiescent power dissipation, albeit at the expense of twice as many transistors.**

***Index Terms*—Address event representation (AER) , analog circuits, asynchronous logic, Gabor filter, image processing, neuromorphic engineering, nonlinear circuits, visual cortex.**

I. Introduction

MOVING from the retina to higher levels of visual processing in the cortex, neurons become progressively more selective to more complex stimuli. Cells in the retina are sensitive along stimulus dimensions of position, spatial frequency (size), temporal frequency and color. In the primary visual cortex (V1), additional selectivity along the dimensions of orientation, direction of motion, and binocular disparity emerges. Subsequent areas are selective to higher order combinations of previous dimensions, e.g., curvature and illusory contours. Concurrently, there is a progressively more invariance along stimulus dimensions established earlier. For example, neurons in V2 respond to visual stimuli over a much larger spatial area than ganglion cells in the retina.

A functional model that seems to account for the responses of a large proportion of cells in the primary visual cortex consists of a linear spatio-temporal filtering stage and three nonlinear mechanisms: half-wave rectification, expansive exponentiation, and contrast normalization [1]–[3]. Linear spatio-temporal filtering determines the neural selectivity along different stimulus dimensions. Half-wave rectification conserves metabolic energy by mapping mean levels to a low quiescent spike rate. Expansive exponentiation sharpens selectivity. Contrast normalization enables neurons to retain stimulus selectivity over a wide input contrast range.

This paper describes a VLSI chip that implements two components of this model: linear orientation selective spatial filtering followed by half-wave rectification. Orientation selectivity is a predominant characteristic of neurons in the primary visual cortex [4]. The implementation of orientation selective neurons is an appropriate starting point in building a silicon model of the selectivity of neurons in the visual cortex, since orientation selective neurons are used in neural models of selectivity along other stimulus dimensions such as direction of motion [5], [6], and binocular disparity [7].

This chip implements neurons with spatial receptive field (RF) profiles that are similar to a Gabor function. In the functional model, the RF profile is the filter's impulse response reflected around the x and y axes. Gabor functions fit the RF profiles measured from orientation selective cortical neurons well [8]–[10]. A Gabor function is a sinusoidal grating with frequency Ω and orientation θ modulated by a Gaussian envelope f

$$h(m', n') = f(m', n') \cos(\Omega n' + \phi) \tag{1}$$

where (m', n') represents the original coordinate function translated by (m_0, n_0) and rotated by θ. The parameter ϕ determines the spatial phase of the sinusoid with respect to the center of the Gaussian. The RF profiles of the neurons on this chip are Gabor-type, since their modulating function is not Gaussian.

The chip described here processes a 32 by 64 pixel input image with two orientation selective filters using a continuous time analog processing network. The filters have even and odd symmetric impulse response that are said to be in phase quadrature, since they differ in phase by $\pi/2$. Physiological measurements in cortex indicate that neighboring neurons often differ in phase by $\pi/2$ [11], with the distribution of phases clustering around even and odd symmetric RF profiles [12]. Energy models of motion and binocular disparity selectivity rely heavily on the existence of neurons with phase quadrature RF profiles [5], [7].

The differential ON–OFF channels used to represent all input, output and internal signals differentiates this chip from previous electronic implementations of orientation selective filtering

Manuscript received January 2, 2003; revised May 31, 2003. This work was supported in part by the Hong Kong Research Grants Council under Grant HKUST6218/01E, and in part by the National Science Foundation CAREER Grant ECS00-93 851. This paper was recommended by Associate Editor P. Arena.

T. Y. W. Choi and B. E. Shi are with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: eethomas@ee.ust.hk; eebert@ee.ust.hk).

K. A. Boahen is with the Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104-6392 USA (e-mail: boahen@seas.upenn.edu).

Digital Object Identifier 10.1109/TCSI.2003.822551

networks, which mostly used single-ended representations [13]–[16]. Serrano–Gotarredona *et al.* propose an architecture that uses an internal differential representation to accumulate signals, but the input and output are single ended [17]. Liu *et al.* constructed orientation selective neurons from the output of a silicon retina that included both ON and OFF channels, but only used the OFF channels [18]. Although the ON–OFF circuit architecture requires as many transistors as a functionally equivalent single-ended design, it has several compelling advantages as we describe in the latter part of the paper.

Section II describes the four-layer cellular neural network (CNN) architecture used to establish the orientation selectivity of the neurons on the chip. Section III derives the spatial transfer functions of the orientation selective filters and proves stability. Section IV describes the chip architecture, with a high level description of the address event representation (AER) communication circuits used for input and output. Section V describes in detail the pixel level processing circuits, including the analog circuits for the orientation selective filtering network, as well as the circuits converting between the digital asynchronous spike train representation used at the periphery and the continuous time current-mode representation used internally. Section VI reports experimental measurements from the chip, and compares the design here with a previous two-layer design. Section VII concludes with a summary and discussion of future directions. Preliminary reports of this work have appeared in [19]–[21].

II. Network Architecture

Biological systems use ON and OFF channels to encode signal variations around a background level efficiently. While single neurons can encode positive and negative signals as variations around a quiescent firing rate, this representation is inefficient as each spike consumes metabolic resources. With separate ON and OFF channels, background signals correspond to low quiescent spike rates on both channels. Positive signals are encoded by increases in the ON-channel spike rate, negative signals by increases in the OFF-channel spike rate. For example, ON-center and OFF-center retinal ganglion cells respond to positive and negative contrast of the center with respect to the surround. Diffuse illumination applied to both center and surround elicits little response from either cell.

ON–OFF signal representations prevail in computational models of visual cortical neurons. Most cortical neurons exhibit low spontaneous spike rates. Hubel and Weisel proposed that the oriented excitatory and inhibitory regions of the RF profiles arise from linear summation of feedforward input from corresponding ON-center and OFF-center cells in the lateral geniculate nucleus [4]. This basic model has been preserved in most subsequent work, which has extended it to include push–pull inhibition to account for contrast invariance or cortical feedback to sharpen orientation selectivity. Ferster and Miller give a review of recent work in [22].

Our network also adopts an ON–OFF signal representation. Each pixel in the image is associated with four neurons. Two neurons carry the positive (ON) and negative (OFF) half-wave rectified outputs of the even symmetric filter. The other two carry the output of the odd symmetric filter. We refer to the neurons as EVEN ON$(\mathbf{e}+)$, EVEN OFF$(\mathbf{e}-)$, ODD ON$(\mathbf{o}+)$, and ODD OFF$(\mathbf{o}-)$.

Our network establishes orientation selectivity through local recurrent interconnections between neurons, which facilitate implementation in VLSI while enabling the resulting RF profiles to extend over many pixels. We model the network as a four-layer CNN whose layers are indexed by $k \in \{e+, e-, o+, o-\}$. Each layer consists of an M by N array of cells, each with real valued input $u_k(m,n)$, state $x_k(m,n)$, and output $y_k(m,n)$, where (m,n) indexes the array location. We drop the (m,n) when referring to an entire layer. We assume that the input to all layers is strictly positive. The outputs of the ON and OFF layers are equal to the difference between their states positive and negative half wave rectified

$$\begin{aligned} y_{e+} &= |x_{e+} - x_{e-}|^+ \quad y_{o+} = |x_{o+} - x_{o-}|^+ \\ y_{e-} &= |x_{e+} - x_{e-}|^- \quad y_{o-} = |x_{o+} - x_{o-}|^- \end{aligned} \tag{2}$$

where $|x|^+ = \max\{x, 0\}$ and $|x|^- = -\min\{x, 0\}$.

The state evolves according to the differential equation

$$\begin{aligned} \dot{x}_k = x_k \bullet \Bigg(-x_k + \sum_l A^s_{kl} \star x_l \\ + \sum_l A_{kl} \star y_l + \sum_l B_{kl} \star u_k \Bigg) \end{aligned} \tag{3}$$

where the summation is over all layers. The $\bullet$ denotes the elementwise product of two arrays. The $\star$ denotes correlation, e.g., $A^s_{kl} \star x_l = \sum_{o,p} A^s_{kl}(o,p) x_l(m+o, n+p)$. The coefficient matrices A^s_{kl}, A_{kl}, and B_{kl} are the state feedback, output feedback and feedforward cloning templates.

This network differs from the classical multilayer CNN [23] in three ways. First, it adds the elementwise product with the state. Second, it contains an additional state feedback template A^s_{kl}, through which each cell's state influences its neighboring cells' states. Third, the output of each cell is a nonlinear function of the states in cells of *two* layers, which introduces additional coupling between layers.

The nonzero cloning templates for orientation selective filtering are

$$\begin{aligned} A^s_{e+e+} &= A^s_{e-e-} = A^s_{o+o+} = A^s_{o-o-} = A_{\nabla^2} \\ A_{e+o-} &= A_{e-o+} = A_{o+e+} = A_{o-e-} = A_{-1} \\ A_{e+o+} &= A_{e-o-} = A_{o+e-} = A_{o-e+} = A_{+1} \\ B_{e+e+} &= B_{e-e-} = B_{o+o+} = B_{o-o-} = 1 \end{aligned}$$

where

$$A_{\nabla^2} = \begin{bmatrix} 0 & \alpha_{1n} & 0 \\ \alpha_{1m} & -(2\alpha_{1m} + 2\alpha_{1n}) & \alpha_{1m} \\ 0 & \alpha_{1n} & 0 \end{bmatrix}$$

$$A_{-1} = \begin{bmatrix} 0 & 0 & 0 \\ \alpha_{2m} & 0 & 0 \\ 0 & \alpha_{2n} & 0 \end{bmatrix}$$

$$A_{+1} = \begin{bmatrix} 0 & \alpha_{2n} & 0 \\ 0 & 0 & \alpha_{2m} \\ 0 & 0 & 0 \end{bmatrix}.$$

In each template matrix, the central element indicates the (0, 0) term. The α parameters are nonnegative reals and determine the

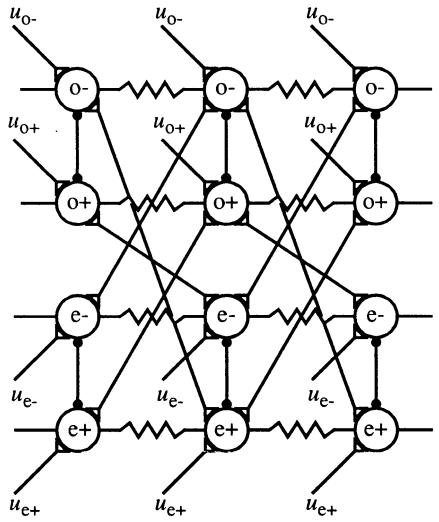


Fig. 1. Cell interconnections in a 1-D version of the four-layer network described by (2). Interconnections terminating with filled circles are inhibitory. Interconnections terminating with open triangles are excitatory. Interconnections by resistors are diffusive. The mutually inhibitory interconnections between positive and negative layers of the same type are introduced by the output nonlinearity.

strength of the interconnections between cells. Substituting the template expressions into (3)

$$\begin{aligned}
\dot{x}_{e+} &= x_{e+} \bullet (-x_{e+} + A_{\nabla^2} \star x_{e+} + A_{+1} \star y_{o+} \\
&\quad + A_{-1} \star y_{o-} + u_{e+}) \\
\dot{x}_{e-} &= x_{e-} \bullet (-x_{e-} + A_{\nabla^2} \star x_{e-} + A_{-1} \star y_{o+} \\
&\quad + A_{+1} \star y_{o-} + u_{e-}) \\
\dot{x}_{o+} &= x_{o+} \bullet (-x_{o+} + A_{\nabla^2} \star x_{o+} + A_{-1} \star y_{e+} \\
&\quad + A_{+1} \star y_{e-} + u_{o+}) \\
\dot{x}_{o-} &= x_{o-} \bullet (-x_{o-} + A_{\nabla^2} \star x_{o-} + A_{+1} \star y_{e+} \\
&\quad + A_{-1} \star y_{e-} + u_{o-}).
\end{aligned}$$

Fig. 1 shows the interconnections between cells in a one-dimensional (1-D) array. Correlation with the A_{∇^2} is a discrete approximation to a Laplacian operator. The template A_{-1} reflects connections to a cell from cells in another layer to the left and below. The template A_{+1} reflects connections from the right and above.

III. Network Analysis

This section derives the spatial transfer function of the network that relates a constant input image with the steady state output and examines the stability of the network. The first part summarizes the main results. The subsections contain the detailed proofs, which can be skipped without loss of continuity.

Because the analysis of this network is complicated by the element-wise product, we consider a **simplified network** without the product

$$\dot{x}_k = -x_k + \sum_l A^s_{kl} \star x_l + \sum_l A_{kl} \star y_l + \sum_l B_{kl} \star u_k \quad (4)$$

Fig. 2(a) shows a block diagram of the interactions between the layers. This network is easier to analyze, but has much in common with the **original network** in (3). First, the two networks share a unique common equilibrium point where the state of all cells is positive. Second, any additional equilibrium point in the original network is unstable. Finally, we conjecture that stability of the common equilibrium point in the simplified network implies its stability in the original network.

We derive the transfer function from a constant input image to the common equilibrium by expressing (4) in terms of the sum and difference of the ON and OFF input and state variables, e.g., $x_{ed} = x_{e+} - x_{e-}$ and $x_{es} = x_{e+} + x_{e-}$. We find that both the sum and difference components evolve according to linear differential equations, but that the sum components are driven by a nonlinear function of the difference components. The difference components evolve independently, and are related to the input difference components by the transfer function

$$\begin{aligned}
H(\omega_m, \omega_n) &= \frac{X_d(\omega_m, \omega_n)}{U_d(\omega_m, \omega_n)} \\
&= \frac{H_\Omega}{1 + \frac{2-2\cos(\omega_m - \Omega_m)}{(\Delta\Omega_m)^2} + \frac{2-2\cos(\omega_n - \Omega_n)}{(\Delta\Omega_n)^2}} \quad (5)
\end{aligned}$$

where X_d and U_d are the discrete Fourier transforms of $x_d = x_{ed} + jx_{od}$ and $u_d = u_{ed} + ju_{od}, \omega_m$ and ω_n are spatial-frequency variables, and

$$\begin{aligned}
H_\Omega &= \left(1 + 2\alpha_{1m} - 2\sqrt{\alpha_{1m}^2 + \alpha_{2m}^2} + 2\alpha_{1n} \right. \\
&\quad \left. -2\sqrt{\alpha_{1n}^2 + \alpha_{2n}^2}\right)^{-1} \\
\Omega_m &= \operatorname{atan}(\alpha_{2m}/\alpha_{1m}) \\
(\Delta\Omega_m)^2 &= \left(H_\Omega \sqrt{\alpha_{1m}^2 + \alpha_{2m}^2}\right)^{-1} \\
\Omega_n &= \operatorname{atan}(\alpha_{2n}/\alpha_{1n}) \\
(\Delta\Omega_n)^2 &= \left(H_\Omega \sqrt{\alpha_{1n}^2 + \alpha_{2n}^2}\right)^{-1}.
\end{aligned}$$

This transfer function reaches its maximum value of H_Ω at $(\omega_m, \omega_n) = (\Omega_m, \Omega_n)$, corresponding to orientation $\theta = \operatorname{atan}(\Omega_m/\Omega_n)$ and spatial frequency $\Omega = \sqrt{\Omega_m^2 + \Omega_n^2}$. Since the transfer function drops by approximately one half at $(\omega_m, \omega_n) = (\Omega_m \pm \Delta\Omega_m, \Omega_n)$ and $(\Omega_m, \Omega_n \pm \Delta\Omega_n)$, we refer to $\Delta\Omega_m$ and $\Delta\Omega_n$ as the 6-dB half bandwidth in the m and n directions.

Although there are five parameters that determine the filter shape, only four can be specified independently by the α parameters. The filter gain H_Ω is fixed by the choice of $\Omega_x, \Omega_y, \Delta\Omega_x$, and $\Delta\Omega_y$ according to

$$H_\Omega = 1 + \frac{2 - 2\cos\Omega_m}{(\Delta\Omega_m)^2} + \frac{2 - 2\cos\Omega_n}{(\Delta\Omega_n)^2}. \quad (6)$$

Because the α parameters are positive, both Ω_m and Ω_n are nonnegative, corresponding to orientations between 0 and $\pi/2$. Orientations outside this range can be obtained by reflecting the templates around the horizontal and/or vertical axes. Alternatively, we can flip the image from left to right before input to the network.

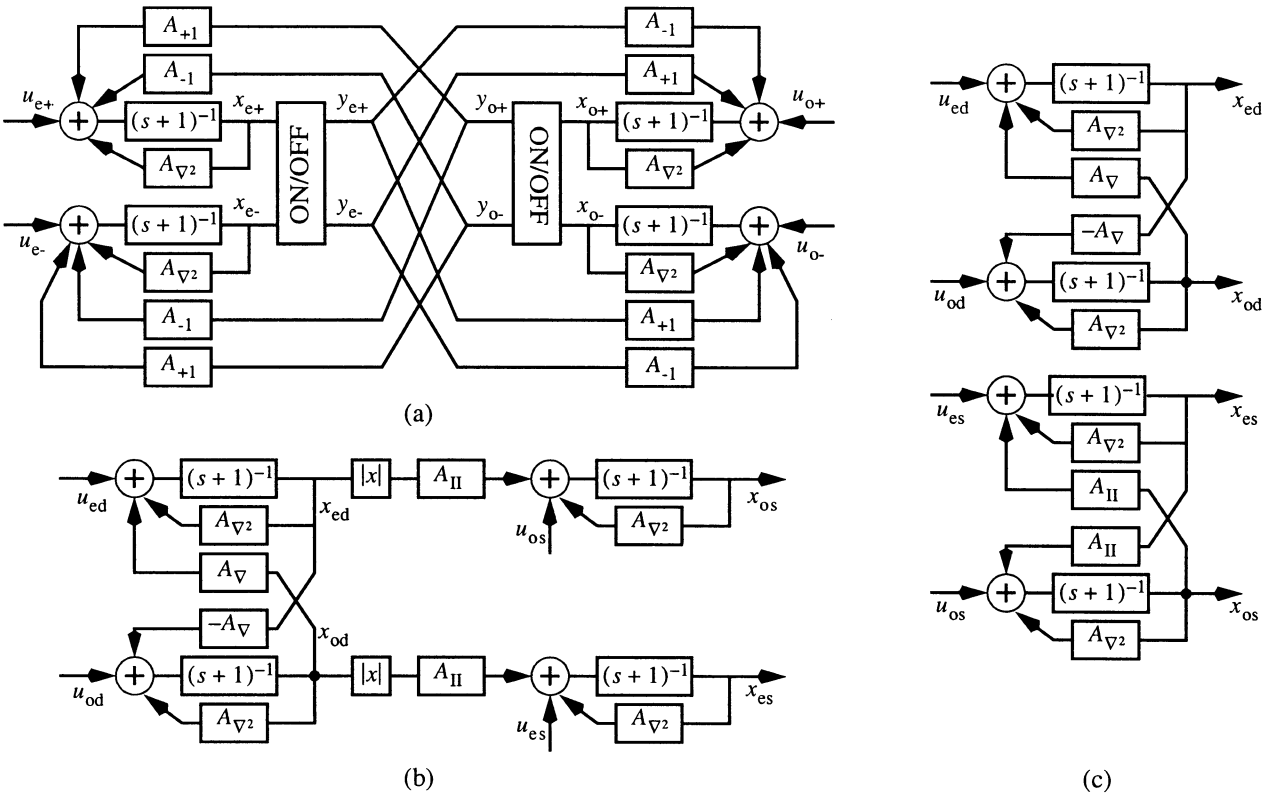


Fig. 2. Block diagram (a) showing the interconnections between and among layers due to the templates and the output nonlinearity, (b) of the interaction between the difference and sum components, and (c) of the difference and sum components when the output nonlinearity is removed.

To relate this complex valued filter to real valued Gabor filter described previously, observe that

$$\begin{bmatrix} X_{ed} \\ X_{od} \end{bmatrix} = \begin{bmatrix} H_e & -H_o \\ H_o & H_e \end{bmatrix} \begin{bmatrix} U_{ed} \\ U_{od} \end{bmatrix}$$

where $H_e(\omega_m, \omega_n) = (H(\omega_m, \omega_n) + H(-\omega_m, -\omega_n))/2$ and $H_o(\omega_m, \omega_n) = (H(\omega_m, \omega_n) - H(-\omega_m, -\omega_n))/(2j)$ are the transforms of (1) with $\phi = 0$ and $\phi = \pi/2$. The modulating function $f(m, n)$ can be approximated by a Laplacian function in 1-D and by a Bessel function in two-dimensional (2-D) [24].

Because the network connections are spatially invariant, the Fourier modes evolve independently [25]. We establish the stability of the sum and difference components by showing that the eigenvalues of the feedback matrices lie in the left half plane for all α parameters corresponding to valid filter parameters. For unstable α parameters, the network exhibits spatially oriented Turing patterns [26]. Our implementation uses a transistor analog of a network of conductances, which Poggio and Koch suggested for solving problems in computational vision [27], [28]. The stability of conductance networks can be established by viewing the dynamics as gradient descent on a suitably defined cost function [29]–[33], and a similar approach can be taken for this network [24]. Incorporating the half wave rectifying nonlinearity into the feedback is critical in ensuring network stability.

A. Original Versus Simplified Network

Clearly, any equilibrium point of (4) is also an equilibrium point of (3). The existence of the spatial transfer function indicates that the equilibrium point of (4) is unique. The state of all cells is positive at this equilibrium point because the feedback loops containing the blocks $(s+1)^{-1}$ and A_{∇^2} correspond to a lossy diffusion process driven by a strictly positive input. Formally, assume that x_{e+} assumes its minimum at pixel (m, n). The first equation in (4) evaluated at equilibrium gives $x_{e+}(m, n) = K_{e+}(m, n)$ where

$$\begin{aligned} K_{e+}(m,n) =& (A_{\nabla^2} \star x_{e+})(m,n) + (A_{+1} \star y_{o+})(m,n) \\ &+ (A_{-1} \star y_{o-})(m,n) + u_{e+}(m,n). \end{aligned} \tag{7}$$

The term $(A_{\nabla^2} \star x_{e+})(m, n) \geq 0$ since $x_{e+}(m, n)$ is a minimum. The next two terms $(A_{+1} \star y_{o+})(m, n) + (A_{-1} \star y_{o-})(m, n) \geq 0$ since the outputs and the α parameters are nonnegative. The last term $u_{e+}(m, n) > 0$ by assumption. In the electronic implementation, this input is represented by the current through a transistor in saturation, which will always be positive due to leakage. Thus, $\min\{x_{e+}\} > 0$. Similar arguments hold for the other layers.

Any additional equilibrium point in the original network is unstable. First note that the state of any neuron at equilibrium must be nonnegative. If the minimum state is nonzero, it must satisfy (7), which implies that it must be positive. Any additional equilibrium point must have $x_k(m, n) = 0$ for some k, m, n. Linearizing around this equilibrium and letting $\Delta x_k(m, n)$ denote a small perturbation around it, we have that $\Delta \dot{x}_k(m, n) = K_{e+}(m, n) \cdot \Delta x_k(m, n)$ where $K_{e+}(m, n) > 0$ by the arguments above.

It seems reasonable that stability of the common equilibrium point in the simplified network should imply stability for the

original network, since the element-wise product operation does not change the slope of the derivative at the equilibrium point, as the state is strictly positive at equilibrium. In addition, our numerical simulations and experimental measurements from the chip have not revealed any unexpected instability.

B. Spatial Transfer Function

Expressing (4) in terms of the sums and differences of the ON and OFF variables

$$\begin{aligned}\dot{x}_{ed} &= -x_{ed} + (A_{\nabla^2} \star x_{ed}) + (A_{\nabla} \star x_{od}) + u_{ed}\\ \dot{x}_{od} &= -x_{od} + (A_{\nabla^2} \star x_{od}) - (A_{\nabla} \star x_{ed}) + u_{od}\\ \dot{x}_{es} &= -x_{es} + (A_{\nabla^2} \star x_{es}) + (A_{\mathrm{II}} \star |x_{od}|) + u_{es}\\ \dot{x}_{os} &= -x_{os} + (A_{\nabla^2} \star x_{os}) + (A_{\mathrm{II}} \star |x_{ed}|) + u_{os} \qquad (8)\end{aligned}$$

where $A_{\nabla} = A_{+1} - A_{-1}, A_{\mathrm{II}} = A_{+1} + A_{-1}$ and $|x|$ denotes the element-wise absolute value of x. Correlation by A_{∇} is a discrete approximation to a directional derivative. The A_{II} template is a combination of even impulse pairs in the horizontal and vertical directions. Fig. 2(b) illustrates that the both the sum and difference components evolve linearly. The difference components are unaffected by the sum components, but the sum components are driven by the absolute value of the difference components.

In [24], we studied the dynamics of the differential components. For completeness, we recapitulate that analysis here. The steady state response and the stability can be analyzed easily in the spatial-frequency domain. Assume a doubly infinite array and define $U(\omega_m, \omega_n), X(\omega_m, \omega_n)$, and $Y(\omega_m, \omega_n)$ to be the 2-D discrete Fourier transforms of the input, state, and output, e.g., $Y(\omega_m, \omega_n) = \sum_{m,n} y(m,n) e^{-j\omega_m m} e^{-j\omega_n n}$. Taking the discrete Fourier transform of the first two equations in (8), correlation by A_{∇^2} and A_{∇} correspond to multiplication by

$$\begin{aligned}\tilde{A}_{\nabla^2}(\omega_m, \omega_n) &= -2\alpha_{1m} + 2\alpha_{1m}\cos\omega_m\\ &\quad - 2\alpha_{1n} + 2\alpha_{1n}\cos\omega_n\\ \tilde{A}_{\nabla}(\omega_m, \omega_n) &= j2(\alpha_{2m}\sin\omega_m + \alpha_{2n}\sin\omega_n)\end{aligned}$$

where $j = \sqrt{-1}$. Thus

$$\begin{bmatrix}\dot{X}_{ed}\\ \dot{X}_{od}\end{bmatrix} = \begin{bmatrix}-1+\tilde{A}_{\nabla^2} & \tilde{A}_{\nabla}\\ -\tilde{A}_{\nabla} & -1+\tilde{A}_{\nabla^2}\end{bmatrix}\begin{bmatrix}X_{ed}\\ X_{od}\end{bmatrix} + \begin{bmatrix}U_{ed}\\ U_{od}\end{bmatrix}. \quad (9)$$

We suppress the dependence on ω_m and ω_n to avoid clutter. If we define $x_d = x_{ed} + jx_{od}$ and $u_d = u_{ed} + ju_{od}$, then $\dot{X}_d = (-1+\tilde{A}_{\nabla^2} - j\tilde{A}_{\nabla})X_d + U_d$. Letting $\dot{X}_d = 0$ and defining $H = X_d/U_d$ to be the spatial transfer function at temporal steady state, we obtain (5).

C. Stability

Stability of the difference components is guaranteed for any set of α parameters corresponding to valid filter parameters. Equation (6) implies $H_{\Omega} \geq 1$, which implies $-1 + \tilde{A}_{\nabla^2} - j\tilde{A}_{\nabla} < 0$ for all (ω_m, ω_n). However, the network is unstable for combinations of α parameters that imply $H_{\Omega} < 0$. It is impossible for $0 \leq H_{\Omega} < 1$, because the α parameters are nonnegative. Fig. 3 shows that the network is unstable if the cross coupling between the even and odd layers, which is determined by the α_2 parameters, is large enough.

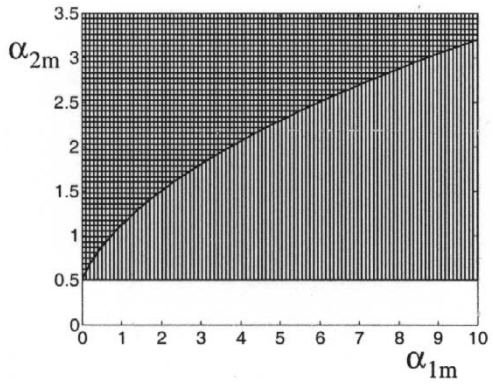


Fig. 3. Unstable regions in the network parameter space. Horizontal hatching indicates the region in the $\alpha_{1m} - \alpha_{2m}$ parameter space where the where the difference components are unstable. Vertical hatching indicates the region where the sum components of the network that does not include the ON–OFF nonlinearity are unstable. Cross hatching indicates regions where both components are unstable.

The sum components evolve according to a discrete approximation to a lossy diffusion equation driven by the sum of the full-wave rectified difference component and the input sum component. In the spatial-frequency domain

$$\begin{bmatrix}\dot{X}_{es}\\ \dot{X}_{os}\end{bmatrix} = (-1+\tilde{A}_{\nabla^2})\begin{bmatrix}X_{es}\\ X_{os}\end{bmatrix} + \begin{bmatrix}0 & \tilde{A}_{\mathrm{II}}\\ \tilde{A}_{\mathrm{II}} & 0\end{bmatrix}\begin{bmatrix}X_{|ed|}\\ X_{|od|}\end{bmatrix} + \begin{bmatrix}U_{es}\\ U_{os}\end{bmatrix}$$

where $\tilde{A}_{\mathrm{II}} = 2(\alpha_{2m}\cos\omega_m + \alpha_{2n}\cos\omega_n)$. $X_{|ed|}$ and $X_{|od|}$ denote the discrete Fourier transforms of the absolute values of the even and odd difference components. Stability is ensured since $-1 + \tilde{A}_{\nabla^2} < -1$ for all (ω_m, ω_n).

Incorporating the half wave rectifying nonlinearity into the dynamics is essential to ensure network stability. To see this, suppose that we remove the nonlinearity and instead let $y_{e+} = x_{e+}$ in (4) and make similar substitutions for y_{e-}, y_{o+}, and y_{o-}. We find that

$$\begin{aligned}\dot{x}_{ed} &= -x_{ed} + (A_{\nabla^2} \star x_{ed}) + (A_{\nabla} \star x_{od}) + u_{ed}\\ \dot{x}_{od} &= -x_{od} + (A_{\nabla^2} \star x_{od}) - (A_{\nabla} \star x_{ed}) + u_{od}\\ \dot{x}_{es} &= -x_{es} + (A_{\nabla^2} \star x_{es}) + (A_{\mathrm{II}} \star x_{os}) + u_{es}\\ \dot{x}_{os} &= -x_{os} + (A_{\nabla^2} \star x_{os}) + (A_{\mathrm{II}} \star x_{es}) + u_{os}.\end{aligned}$$

Fig. 2(c) shows that the evolution of the difference components is identical to that in (8), but the sum component now evolves independently of the difference component.

The sum components are unstable for some α parameters that correspond to valid filter parameters. In the spatial-frequency domain

$$\begin{bmatrix}\dot{X}_{es}\\ \dot{X}_{os}\end{bmatrix} = \begin{bmatrix}-1+\tilde{A}_{\nabla^2} & \tilde{A}_{\mathrm{II}}\\ \tilde{A}_{\mathrm{II}} & -1+\tilde{A}_{\nabla^2}\end{bmatrix}\begin{bmatrix}X_{es}\\ X_{os}\end{bmatrix} + \begin{bmatrix}U_{es}\\ U_{os}\end{bmatrix}.$$

For stability of the eigenvalues of the feedback matrix, $s = -1 + \tilde{A}_{\nabla^2} \pm \tilde{A}_{\mathrm{II}}$, must be negative for all (ω_m, ω_n). Since the α parameters are nonnegative, the largest eigenvalue occurs for $(\omega_m, \omega_n) = (0,0)$. This implies that for stability, we must have $\alpha_{2m} + \alpha_{2n} < 1/2$. Fig. 3 shows that the stable parameter region is only a subset of the stable parameter region for the difference components.

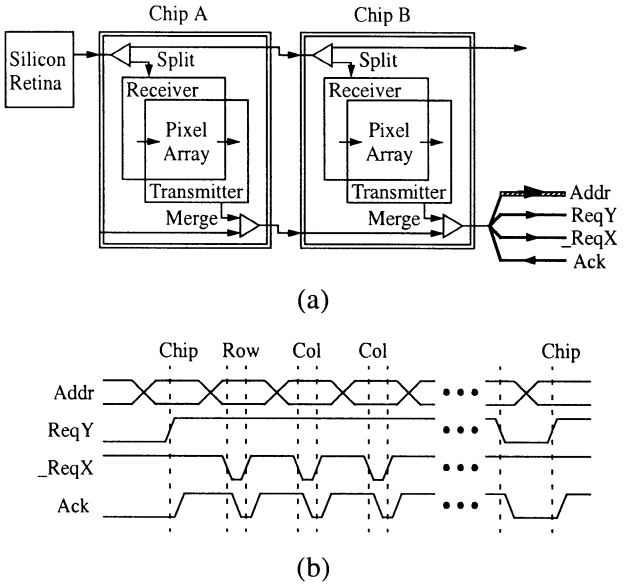


Fig. 4. (a) Three-chip system where the output of a silicon retina is fanned out to two orientation selective chips (Chip A and Chip B) tuned to different orientations. (b) Addressing scheme at the merge output of chip B.

IV. Chip Architecture

The visual cortex processes each region of the visual field with neurons selective to many different orientations, which are grouped into a hypercolumn. To replicate this organization, we require a set of chips, each processing the same image but tuned to different orientations. Fig. 4(a) depicts a three-chip network.

To enable the construction of this multichip system, each chip is a transceiver, containing both a receiver to receive input images and transmitter to transmit output images. Each chip also includes asynchronous routing circuits to facilitate signal fan out and fan in, which will be described in detail in a forthcoming publication. Briefly, the **split** replicates its input, sending one copy into the processing array via a receiver and sending the other off chip. Fanning the output of a silicon retina (e.g., [34]) out to a set of chips by cascading the split output of one with the split input of the next, we can build an array of orientation selective hypercolumns. The **merge** circuit combines its input with the array output from the transmitter and sends the combined stream off chip. The encoding we use enables us to distinguish the different images at the merge output, but we can also use the merge to combine images additively, implementing signal fan in. Combining input signals from a retina with output signals from other chips, we can implement intracortical interconnections, as well as feedback interconnections from later processing stages. The vast majority of inputs to cortical neurons come from other nearby cortical neurons, i.e., neurons tuned to similar orientations [35], [36]. Feedback from extrastriate areas appears to modulate the responses of neurons in V1 [37].

Input and output images are rate encoded as arrays of spike trains, which are communicated using the AER protocol [38]. The AER protocol communicates continuous time spike activity from an array of silicon neurons in one chip to another chip over an asynchronous digital bus. It is more efficient than scanning when the spike activity within the array is sparse, as we expect here since only a few image locations will contain edges near the orientation selected by each chip.

The transmitter signals a spike occurrence by placing the location (address) of the spiking neuron onto the bus. The receiver takes the address that appears on the bus and feeds a spike to the corresponding neuron in its array. The protocol is asynchronous, with the time that the address appears on the bus encoding the spike time directly. Collisions between simultaneous spikes from two neurons are handled by arbitration.

Addresses are placed onto the bus in "bursts," where each burst encodes all of the simultaneous spikes from neurons within a given row and a given chip. We use a word serial format, where each burst is a sequence of addresses. As shown in Fig. 4(b), the transmitter signals the start of a burst by placing an address identifying the source chip onto the address lines (Addr) and taking the request signal ReqY high. Subsequent addresses are signalled by taking _ReqX low. The second address identifies the row. Each of the remaining addresses identifies one of the columns containing a neuron that spiked. The transmitter signals the end of the burst by taking ReqY low. The receiver acknowledges receipt of each address by a transition on the Ack line.

We use absolute addressing to identify rows and columns within a chip, but relative addressing to identify each chip. Each chip signals its own activity with bursts whose chip addresses are set to zero. Every time a chip relays a burst from its split or a merge input, it increments the chip address. For example, a chip address of 1 at the merge output of Chip B in Fig. 4(a) indicates the spikes in the burst come from Chip A.

For each pixel, the four neurons are addressed using the least significant bit (LSB) of the row and column addresses. EVEN and ODD neurons are indexed by row addresses with the least significant bit (LSB) at 0 and 1. ON and OFF neurons are indexed by column addresses with LSB 0 and 1. Thus, the network for processing an M by N pixel image actually contains a $2M$ by $2N$ array of neurons arranged into 2×2 blocks.

V. Pixel Processing Circuits

Each pixel in the array contains the circuits necessary for processing four neurons. This includes four leaky integrators that convert input spike trains to continuous currents, current-mode analog processing circuits that implement the filtering/rectification network and four spiking neuron circuits that convert the current outputs of the network to spike trains.

We represent each state variable array as the drain currents in an array of nMOS transistors with fixed gate voltage V_0, as shown in Fig. 5(a). The sources are connected through capacitors to the ground. We assume all transistors operate in weak inversion and are saturated, so the drain currents representing the $e+$ layer are given by

$$x_{e+}(m,n) = I_0 \exp(\kappa V_0/U_T) \exp(-\nu_{e+}(m,n)/U_T) \quad (10)$$

where V_0 and $\nu_{e+}(m,n)$ are the gate and source voltages referenced to the bulk node, U_T is the thermal voltage, I_0 is a process and geometry dependent current and $0 < \kappa \leq 1$ is a process dependent constant. Representative parameters for the TSMC0.25 um process are $I_0 = 3.1$ pA and $\kappa = 0.68$. Differentiating with respect to time, $CU_T \dot{x}_{e+}(m,n) = x_{e+}(m,n)(-i_C(m,n))$, where $i_C(m,n)$ is the current entering the capacitor C.

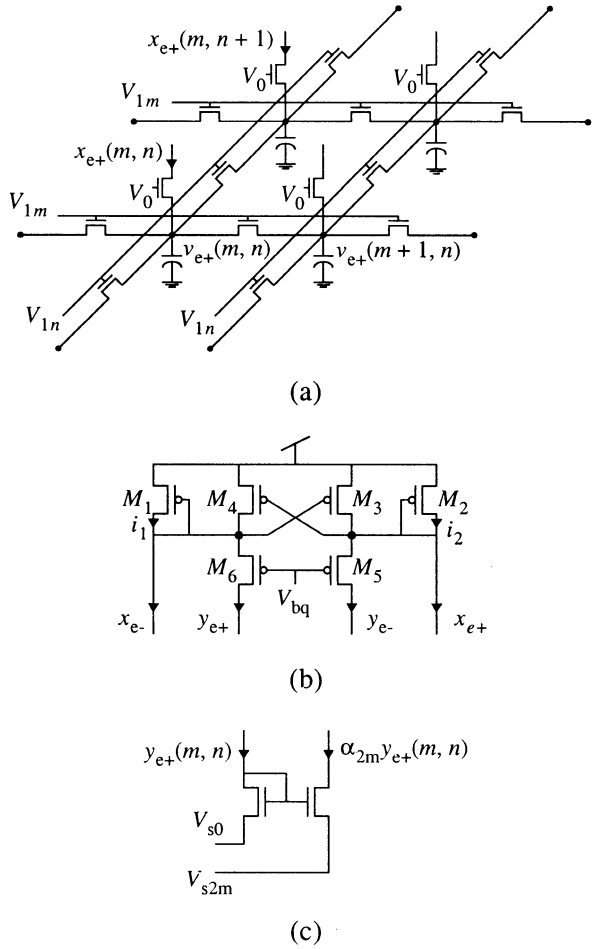


Fig. 5. (a) 2-D diffuser network. (b) Circuit mapping the state variables, represented by currents, to outputs, represented by currents. All transistors operate in saturation. (c) Circuit implementing the spatial shift operator with gain.

To implement the network, we equate the current flowing *out* of the capacitor with the sum on the right hand side of the elementwise product operator in (3). Each sum can be grouped as three currents i_{fb}, i_{xc}, and i_u. For the first equation, $i_{fb} = -x_{e+} + A_{\nabla^2} \star x_{e+}$ describing the intralayer state feedback, $i_{xc} = A_{+1} \star y_{o+} + A_{-1} \star y_{o-}$ describing the cross coupling between layers and $i_u = u_{e+}$ describing the input. In the fabricated design, we do not implement the capacitor explicitly. However, the source nodes will invariably have some parasitic capacitances associated with them. The remainder of this section describes the circuits generating these three currents, as well as the spiking neuron circuit that converts each output current into a spike train.

A. Layer Self-Feedback

We use the diffuser/pseudoresistor network [39], [40] in Fig. 5(a) to implement i_{fb}. In total, the circuit contains four diffuser networks, one for each layer.

In weak inversion, the drain current flowing through the horizontal nMOS transistor from node $(m+1, n)$ into node (m, n) is

$$i_d = I_0 \exp((\kappa V_{1m})/U_T)(\exp(-\nu_{e+}(m, n)/U_T) - \exp(-\nu_{e+}(m+1, n)/U_T)).$$

Substituting (10), we get $i_d = \alpha_{1m}(x_{e+}(m, n) - x_{e+}(m+1, n))$ where $\alpha_{1m} = \exp(\kappa(V_{1m} - V_0)/U_T)$. The total current flowing out of the capacitor at each node due to the five transistors connected in the diffuser network implements i_{fb} where $\alpha_{1n} = \exp(\kappa(V_{1n} - V_0)/U_T)$.

B. Cross-Coupling Circuits

Layers are coupled through the cell outputs. The mapping from state to output in (2) can be specified by the implicit equations

$$y_{e+} - y_{e-} = (x_{e+} - x_{e-})$$
$$\min\{y_{e+}, y_{e-}\} = 0.$$

The ON–OFF circuit [34] in Fig. 5(b) implements a similar mapping

$$y_{e+} - y_{e-} = (x_{e+} - x_{e-}) \quad (11)$$
$$\min\{y_{e+}, y_{e-}\} \leq 2^{\frac{1}{1+\kappa}} I_{bq} \quad (12)$$

where I_{bq} is a small current set by V_{bq}.

Kirchoff's current law applied at the sources of transistors M_5 and M_6 gives $y_{e+} + x_{e-} = i_1 + i_2 = y_{e-} + x_{e+}$. Rearranging the left and right sides gives (11). The translinear principle applied to the loop $M_1 \rightarrow M_6 \rightarrow M_5 \rightarrow M_2$ gives $i_1 y_{e+}^{\kappa} = i_2 y_{e-}^{\kappa} = I_{bq}^{1+\kappa}$ where

$$I_{bq} = I_0 e^{\frac{\kappa^2}{\kappa+1}(\frac{V_{\mathrm{DD}} - V_{bq}}{U_T})}. \quad (13)$$

Combining these equations with $i_7, i_8 \geq 0$

$$\frac{2 I_{bq}^{1+\kappa}}{\min\{y_{e+}, y_{e-}\}^{\kappa}} \geq \frac{I_{bq}^{1+\kappa}}{y_{e+}^{\kappa}} + \frac{I_{bq}^{1+\kappa}}{y_{e-}^{\kappa}} \geq \min\{y_{e+}, y_{e-}\}$$

which yields (12). The upper bound in (12) is equal to the zero input quiescent output current in both y_{e+} and y_{e-}.

Each spatial shift operator is implemented by a tilted current mirror shown in Fig. 5(c). The difference in the source voltage controls the gain: $\alpha_{2m} = \exp((V_{s0} - V_{s2m})/U_T)$. The shift is implemented by connecting the drain voltage of the output transistor to the appropriate node of the diffuser network. The entire cross coupling between the even and odd arrays, i_{xc} requires one diode connected transistor with source voltage V_{s0} and four mirror transistors, two with source voltage V_{s2m} and two with V_{s2n}.

C. Current-Mode Integrator

Four current-mode integrators at each pixel convert the incoming spike trains to input currents i_u. Fig. 6(a) shows the schematic of one integrator [41]. The inputs _RSelX and _RSelY are shared by one row or column of cells. The receiver takes both inputs low when an address event with the corresponding row and column address is received. This injects a charge packet into the diode-capacitor integrator formed by M_1 and M_C, and pulls the acknowledge signal _Ack low, signalling the receiver that the spike has been delivered. The bias voltage V_{int} controls the magnitude of the current pulse and the communication cycle-time determines its duration. The difference between the source voltages of the current mirror, $V_{\mathrm{is1}} - V_{\mathrm{is0}}$, controls the gain and the time constant of the integrator.

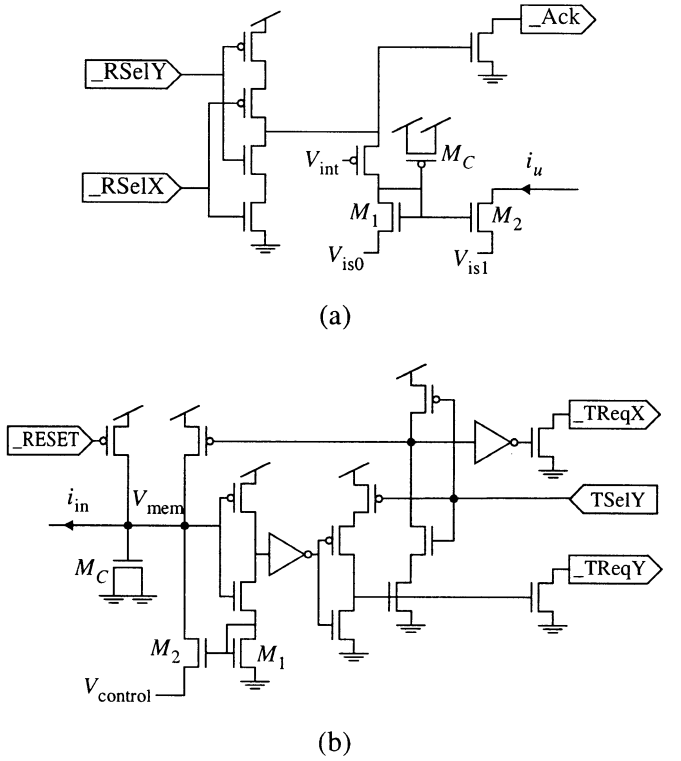


Fig. 6. Schematic of (a) the leaky integrator and (b) the firing neuron.

D. Spiking Neuron

Four spiking neuron circuits at each pixel convert the four output currents, which are obtained by mirroring the diode connected transistor of the spatial shift circuit, into spike trains. We use the design shown in Fig. 6(b), which is similar to that in [42]. The voltage $V_{\rm mem}$ is initially high, and decreases as $I_{\rm in}$ discharges M_C. Once $V_{\rm mem}$ reaches a threshold value, the inverter switches and the neuron fires, bringing the row request line, _TReqY, low to signal a spike. Once a row has been selected, the AER transmitter takes TSelX high. All of the neurons in the selected row that have generated a spike then reset and pull the column request lines, _TReqX low. Once a row has been selected, no new neurons in that row can spike.

Positive feedback through the current mirror $M_1 - M_2$ minimizes the inverter switching time, saving power. The bias voltage $V_{\rm control}$ controls the amount of feedback. If it is too high, current feedback is small and power consumption increases. If it is too low, the background firing rate is high and obscures the signal. The _RESET signal is a global signal which resets all neurons.

VI. Experimental Results

We designed and fabricated an array of 32×64 pixels in the TSMC0.25 um mixed signal/RF process available through MOSIS. This process contains five metal layers and one poly layer, uses nonepitaxial wafers, and is intended for 2.5-V applications. Chip characteristics are summarized in Table I.

We generated the array layout by tiling metapixels, which contain the circuits required for two pixels stacked vertically. Fig. 7 shows the layout of the top half of the metapixel. The bottom half is mirrored vertically so that the analog filter circuits are adjacent.

TABLE I
Chip Characteristics

Technology	TSMC 0.25µm
Chip Area	3.84mm x 2.54mm
Pixel Size	52µm x 49µm
Pixel Transistor Count	172 (48 analog, 124 digital)
Array Dimension	32 x 64 pixels

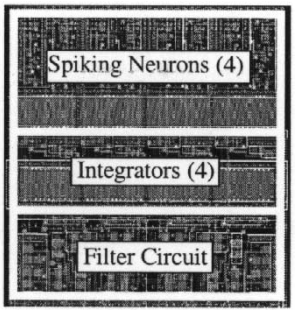


Fig. 7. Layout of the top half of one metapixel.

We laid out the metapixels to minimize cross talk from the digital communication circuits (the spiking neurons and the current-mode integrators) to the analog spatial filtering circuits. The analog filtering circuits lie in the middle of the metapixel, with the digital circuits on the top and bottom. Within the digital parts, the integrators lie next to the analog circuits. The spiking neurons, which contain the most switching transistors, lie at the top and bottom, farthest from the analog processing circuits. Guard rings, which are inserted between the Gabor cells, the integrators and the spiking neurons, provide low impedance paths to collect the minority carriers injected by digital transistors, which would otherwise lead to variations in the bulk voltage when they reach a well or substrate. The digital and analog circuits use separate power and ground lines. Bias lines connected to source voltages controlling current mirror gains run wide on the top metal layer to reduce impedance.

A. Steady-State Response

With the Gabor-type filtering circuits turned off by setting $V_{bq} = V_{\rm DD}$, the spiking neuron circuit maintains a background spike rate because the gate node of transistors M_1 and M_2 in Fig. 6(b) is not fully discharged to ground during the reset of $V_{\rm mem}$ and the residual current through M_2 discharges the gate of M_C. Increasing $V_{\rm control}$ decreases the quiescent spike rate by reducing this residual current. However, it also increases power consumption per spike by decreasing the gain of the current feedback. In a tradeoff between these two effects, we set $V_{\rm control} = 280$ mV to minimize total power consumption. At this point, the average spike rate per neuron is 5.8 Hz with a standard deviation of 6.6 Hz. We computed these statistics using a total of 392 256 spikes collected from the merge output during an 8.2-s time window.

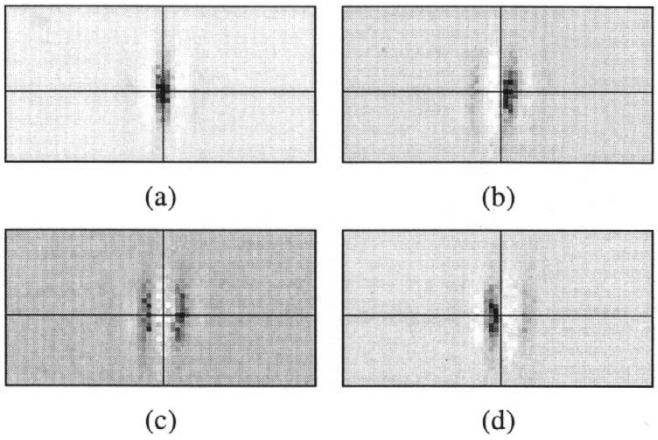


Fig. 8. Measured average spike rate (relative to the baseline firing rate) for the array tuned to vertical orientations with a spike train applied at pixel (17, 32). Crosshairs indicate the pixel location of the impulse. (a) The $e+$ output. White and black correspond to relative spike rates of -15 and 188 Hz. Spike rates outside this range are clipped. (b) The $o+$ output. White/black $= -15/117$ Hz. (c) The $e-$ output. White/black $= -15/66$ Hz. (d) The $o-$ output. White/black $= -15/132$ Hz.

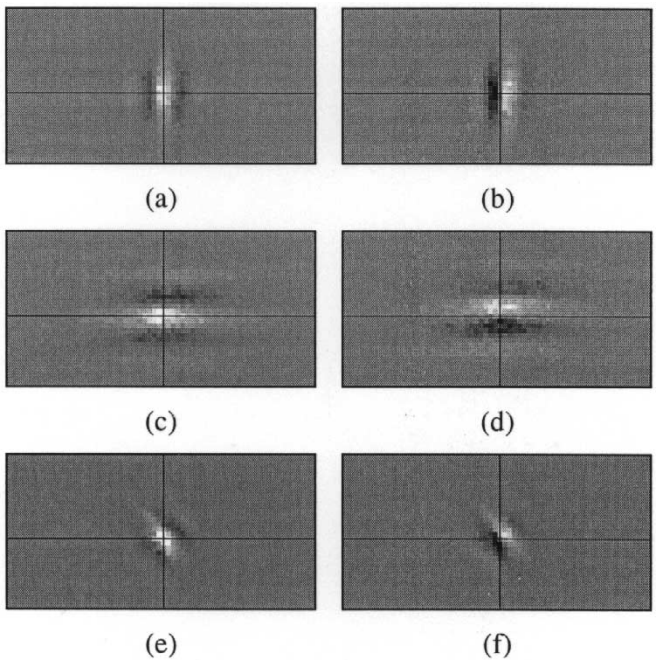


Fig. 9. (a) The ed output for vertical tuning. White/black $= \pm 188$ Hz. (b) The od output for vertical tuning. White/black $= \pm 132$ Hz. (c) The ed output for horizontal tuning. White/black $= \pm 121$ Hz. (d) The od output for horizontal tuning White/black $= \pm 116$ Hz. (e) The ed output for diagonal tuning. White/black $= \pm 721$ Hz. (f) The od output for diagonal tuning. White/black $= \pm 759$ Hz. The larger spike rate for diagonal tunings is primarily due to a difference in the tuning of the input gain (V_{int}) to the input integrator.

When the Gabor-type filter circuits turned on but with no input applied, the background spike rate and its variance increase due to the quiescent output current of the ON–OFF circuit. For the array tuned to vertical orientations, the average quiescent spike rate computed across the array was 15.1 Hz, with a standard deviation of 9.0 Hz. We computed these statistics tuning using a total of 392 714 spikes collected from the merge output during a 3.2 second time window. A similar increase is observed for other filter parameters.

To test the spatial impulse response of the array, we excited the "$e+$" input of pixel (17, 32) with a 50-kHz spike train from a pattern generator. All other inputs were silent. A logic analyzer connected to the merge output collected the output spike train, which is digitally processed for analysis. Fig. 8 shows the four outputs of the array. We computed the spike statistics using 390680 spikes collected over a 3.0 second window.

To show the tunability of the array, Fig. 9 shows the differences between the ON and OFF outputs for a spatial impulse input, when the array is tuned to vertical, diagonal, and horizontal orientations and different spatial scales. For vertical tuning, filter parameters which fit the response predicted by (1) in the least squares sense were $(\Omega_m, \Omega_n) = (0.78, -0.012)$ radians/pixel and $(\Delta\Omega_m, \Delta\Omega_n) = (0.43, 0.31)$ radians/pixel. The signal-to-noise ratio, defined as the energy in the ideal filter output with the best fit parameters divided by the energy in the difference between the actual and ideal filter outputs was 11.2 dB. For the diagonal orientation tuning, best fit parameters were $(\Omega_m, \Omega_n) = (0.60, 0.51)$, corresponding to a spatial frequency $\Omega = 0.78$ and orientation $\theta = \pi/4.5$, and $(\Delta\Omega_m, \Delta\Omega_n) = (0.53, 0.44)$. The signal-to-noise ratio was 11.8 dB. For the horizontal orientation tuning, best fit parameters were $(\Omega_m, \Omega_n) = (0.0059, 0.64)$ and $(\Delta\Omega_m, \Delta\Omega_n) = (0.17, 0.26)$. The signal-to-noise ratio was 8.8 dB.

B. Temporal Response

We measured the temporal response of the arrays by applying a step change in the spike rate applied to the $e+$ input of pixel (17, 32) from 0 Hz to 25 kHz (stimulus onset) and vice versa (stimulus offset). Using a fast input spike rate better indicates the response of the current-mode processing array, since it minimizes temporal ripple in the output current of the current-mode integrator. More than 10 input spikes are integrated per output spike, so the output spike response is not influenced significantly by the temporal characteristics of the input spike train.

These experiments revealed a temporal asymmetry between the response to stimulus onset and offset. Fig. 10 shoes the $e+$ output at pixel (17, 32). The response to stimulus onset is essentially instantaneous. The steady state response is a spike rate of 1.6 kHz. This corresponds to an average interspike interval of 0.625 ms, which is the approximately the delay before the first spike. On the other hand, at stimulus offset the response took about 1 ms to die away. Fig. 11 shows a similar asymmetry in the response of the $o+$ neuron at pixel (17, 33). In this case the steady state firing rate to the stimulus is 360 Hz, corresponding to an interspike interval of 2.8 ms. The temporal asymmetry is primarily due to the nonlinearity introduced by the elementwise multiplication in (3), which slows down the network at low input levels, but speeds it up at high input levels. The dynamics of the current-mode integrated has similar characteristics [41], so part of the asymmetry can be attributed to this stage.

Despite the asymmetry, the settling times for onset and offset are both on the order a few milliseconds, which implies that for intended applications, the temporal dynamics of the array are negligible. For reference, consider that each frame in a video sequence occupies 30–40 ms or that the temporal bandwidth of cortical neurons is on the order of 10s of hertz.

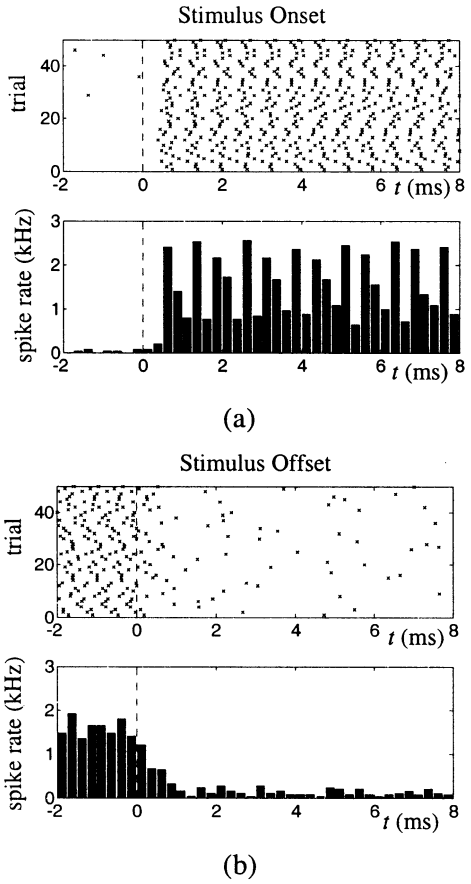


Fig. 10. (a) Temporal response at the $e+$ output at (17, 32) to stimulus onset at time zero. (b) Temporal response at the $e+$ output at (17, 32) to stimulus offset at time zero. The upper figures show spike rasters from 50 trials. The bottom figures show peri-stimulus time histograms computed over 100 trials with a 0.25-ms bin size.

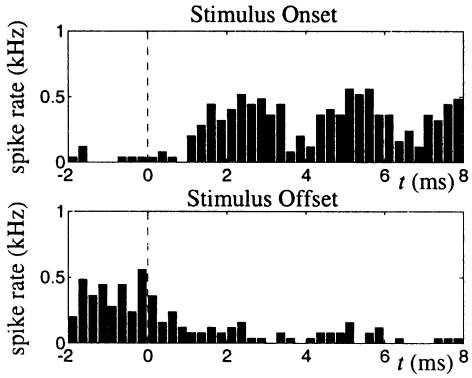


Fig. 11. Peri-stimulus time histograms of the response at the $o+$ output at (17, 33). Histograms were computed over 100 trials with a 0.25-ms bin size.

C. Power Dissipation

The power consumption is dominated by the activity of the communication circuits, rather than the processing circuits. We measured the power dissipation of the chip while stimulating pixel (16, 32) with spike trains ranging in frequency from 0 Hz to 100 kHz and plot the results in Fig. 12 as a function of average output activity per neuron, which is much higher than the input activity. The power increases linearly with the output activity.

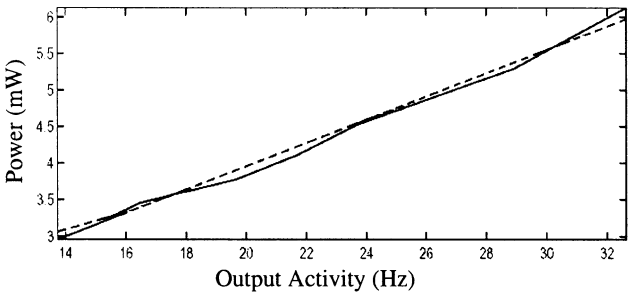


Fig. 12. The solid line plots total power consumption versus the average output activity per neuron. The dotted line is a linear least squares fit to the data, which has slope 0.16 mW/Hz and vertical offset 0.77 mW.

The quiescent power consumption with no input, but an average output activity of 14 Hz, is about 3 mW. The buffer circuits that drive the pads account for around 75% of the total power consumption. The digital spike communication circuits account for around 24%. The analog circuits consume less than 1%.

D. Comparison With Single-Ended Architecture

An implementation of the same filter kernels using a single-ended representation, described in [43], requires half as many transistors for implementing the analog filtering network. If connected to an AER interface, each pixel would require half as many integrators and spiking neurons. However, the ON–OFF implementation described here has several advantages which outweigh the additional hardware cost.

First, the filtering network has reduced quiescent (zero input) power dissipation. The single-ended implementation encodes positive and negative signals as variations around a quiescent bias current I_{bias} which dissipates power, even if the output of the filter is zero. In the ON–OFF implementation, the analogous bias current is the quiescent output currents in the ON–OFF circuit. To compare the power dissipation across a range of operating conditions, we assume that the maximum absolute signal current $I_{\max}$ in the two cases is the same. Since the bias current limits the maximum negative signal excursion, the power dissipation of the single-ended implementation is $P_1 = cI_{\max}$ where c is a constant of proportionality depending upon the supply voltage and the array tuning.

An upper bound on the quiescent power dissipation of the ON–OFF implementation is

$$P_{\text{ON-OFF}} = 2c\left(2^{\frac{1}{\kappa+1}} I_{bq}\right)$$

where the factor 2 arises because the ON–OFF implementation has twice as many paths from V_{DD} to ground. We estimate this by assuming that the output currents of the ON–OFF circuit are $2^{1/(\kappa+1)} I_{bq}$ and computing the total current flowing from V_{DD}, ignoring the reduction in the output current due to the feedback which supplies current to the input. Since M_3 and M_4 should be in saturation, the source voltages of M_5 and M_6 must be a several multiples (r) of U_T below V_{DD}. Thus, the maximum output current of the ON–OFF circuit is $I_{\max} = I_0 \exp(\kappa(V_{\text{DD}} - V_{bq})/U_T - r)$. Combining this with

(13) gives $I_{bq} = I_0(I_{\max}/I_0)^{(\kappa/(\kappa+1))}e^{(\kappa r)/(\kappa+1)}$, which implies

$$P_1 = \left(2^{-(\frac{\kappa+2}{\kappa+1})}e^{-(\frac{\kappa r}{\kappa+1})}(I_{\max}/I_0)^{\frac{1}{\kappa+1}}\right)P_{\text{ON-OFF}}.$$

Typical parameters for the TSMC0.25 um process are $I_0 = 3.1$ pA and $\kappa = 0.68$. Choosing $I_{\max} = 10$ nA and $r = 4$, we obtain $P_1 = 8P_{\text{ON-OFF}}$.

However, there is a tradeoff between latency and power. In the ON–OFF implementation, weak signals are processed slower than fast signals because the elementwise product with the state slows the dynamics of the array when the current levels are smaller. The signal gain is independent of signal strength. Since this chip takes input from a contrast sensitive silicon retina, weak signals correspond to areas with little contrast. The slower response improves signal-to-noise ratio for weak signals by increasing temporal smoothing. Biological systems exploit the same strategy. In the retina, rods, which are sensitive for dim light, respond slower than cones, which are sensitive to bright light [44]. The response of cat retinal ganglion cells speeds up for higher contrast signals [45].

Second, the ON–OFF network exhibits reduced fixed pattern noise in the output. The primary source of fixed pattern noise in the single-ended architecture is mismatch in the transistors supplying the bias current, which adds spatial noise to the filter input. By reducing the quiescent bias current, the ON–OFF network reduces the fixed pattern noise. In [43], the standard deviation of the fixed pattern noise was 26–38% of the bias current. In this network, the standard deviation of the quiescent spike rate in Fig. 8 with no input (9.1 Hz) was 1.2% of the peak spike rate (752 Hz).

Third, the ON–OFF signal representation includes half-wave rectification of the filter output, while the single-ended architecture does not. Although this could be added as a separate circuit to the single-ended architecture, its design is complicated by the large fixed pattern noise in the output, which means that the reference point around which to rectify varies from pixel to pixel.

Fourth, the ON–OFF output representation conserves bandwidth on the AER bus. The low quiescent output currents map to near zero quiescent spike rates at the output of the spiking neuron circuit. For Fig. 8, the average quiescent spike rate (15.1 Hz) was 2.0% of the peak spike rate (752 Hz). If the output current of the single-ended architecture is fed into a spiking neuron circuit, quiescent spike rate must be 50% of the maximum spike rate, assuming that the maximum positive and negative signal excursions are identical. Given that power dissipation is dominated by the communication circuits, this would significantly increase power consumption as well.

Finally, the ON–OFF circuit design is more symmetric. First, all current gains α_{2m} and α_{2n} in the ON–OFF architecture are positive, and are implemented using nMOS current mirrors. The single-ended architecture requires positive and negative current gains. Negative gains require an extra mirroring step through a pair of pMOS transistors, which increases mismatch. Second, the positive and negative signal excursions have the same limit in the ON–OFF architecture, both being limited by V_{bq}. For the single-ended architecture, the maximum negative signal is limited by the bias current while the maximum positive signal is limited by the largest current before the transistors leave weak inversion.

VII. Conclusion

Inspired by the functionality of visual cortical neurons, we have designed an orientation selective image filtering chip that uses an ON–OFF signal representation. The resulting circuit architecture has compelling engineering advantages over previous single-ended feedback circuit architectures for orientation selective filtering.

Our current work seeks to incorporate this chip into a multichip functional model of the primary visual cortex. Each chip contains an array of neurons, all selective for the same orientation but different image locations. Sets of chips implement hypercolumns of neurons selective for different orientations. Because both input and output are AER encoded spike trains, this network will be able to include feedback interactions, such as competition between orientations to enhance orientation selectivity [46]. The orientation selective neurons may also be used in building neurons selective along other stimulus dimensions, such as binocular disparity and direction of motion. Because the network includes a rectifying nonlinearity, it may also be useful in modeling responses to second-order stimuli using a "filter-rectify-filter" model [47].

Acknowledgment

The authors would like to thank P. Merolla and J. Arthur for their work in designing the transmitter and receiver circuits for the AER interface, and K. Hynna and B. Taba for their assistance in generating the chip layout.

References

[1] D. G. Albrecht and W. S. Geisler, "Motion selectivity and the contrast response function of simple cells in the visual cortex," *Vis. Neurosci.*, vol. 7, pp. 531–546, 1991.

[2] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, pp. 181–197, 1992.

[3] ——, "Half-squaring in responses of cat striate cells," *Vis. Neurosci.*, vol. 9, pp. 427–443, 1992.

[4] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[5] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A*, vol. 2, pp. 284–299, 1985.

[6] A. B. Watson and J. A. J. Ahumada, "Model of human visual-motion sensing," *J. Opt. Soc. Amer. A*, vol. 2, pp. 322–342, 1985.

[7] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors," *Science*, vol. 249, pp. 1037–1041, 1990.

[8] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vis. Res.*, vol. 20, pp. 847–856, 1980.

[9] S. Marceljia, "Mathematical description of the responses of simple cortical cells," *J. Opt. Soc. Amer.*, vol. 70, pp. 1297–1300, 1980.

[10] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurosci.*, vol. 58, no. 6, pp. 1233–1258, 1987.

[11] D. A. Pollen and S. Ronner, "Visual cortical neurons as localized spatial-frequency filters," *IEEE Trans. Syst., Man, Cybern.*, vol. 13, pp. 907–916, 1973.

[12] D. L. Ringach, "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex," *J. Neurophysiol.*, vol. 88, pp. 455–463, 2002.

[13] L. Raffo, "Resistive network implementing maps of Gabor functions of any phase," *Electron. Lett.*, vol. 31, no. 22, pp. 1913–1914, 1995.
[14] P. Venier, A. Mortara, X. Arreguit, and E. A. Vittoz, "An integrated cortical layer for orientation enhancement," *IEEE J. Solid-State Circuits*, vol. 32, pp. 177–186, Feb. 1997.
[15] G. Cauwenberghs and J. Waskiewicz, "Focal-plane analog VLSI cellular implementation of the boundary contour system," *IEEE Trans. Circuits Syst. I*, vol. 46, pp. 327–334, Feb. 1999.
[16] R. Etienne-Cummings, Z. K. Kalayjian, and D. Cai, "A programmable focal-plane MIMD image processing chip," *IEEE Journal Solid-State Circuits*, vol. 36, pp. 64–73, Jan. 2001.
[17] T. Serrano-Gotarredona, A. G. Andreou, and B. Linares-Barranco, "AER image filtering architecture for vision-processing systems," *IEEE Trans. Circuits Syst. I*, vol. 46, pp. 1064–1071, Sept. 1999.
[18] S. C. Liu, J. Kramer, G. Indiveri, T. Delbruck, T. Burg, and R. Douglas, "Orientation selective a VLSI spiking neurons," *Neural Netw.*, vol. 14, pp. 629–643, 2001.
[19] B. E. Shi, T. Y. W. Choi, and K. Boahen, "ON–OFF differential current-mode circuits for Gabor-type spatial filtering," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. II, Phoenix, AZ, 2002, pp. 724–727.
[20] T. Y. W. Choi, B. E. Shi, and K. Boahen, "An orientation selective 2-D AER transceiver," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. IV, Bangkok, Thailand, 2003, pp. 800–803.
[21] B. E. Shi, "Cortically inspired visual processing with a four-layer cellular neural network," in *Proc. Int. Joint Conf. Neural Networks*, Portland, OR, 2003.
[22] D. Ferster and K. D. Miller, "Neural mechanisms of orientation selectivity in the visual cortex," *Ann. Rev. Neurosci.*, vol. 23, pp. 441–471, 2000.
[23] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1257–1272, 1988.
[24] B. E. Shi, "Focal plane implementation of 2D steerable and scalable cortical filters," *J. VLSI Signal Processing*, vol. 23, no. 2/3, pp. 319–334, 1999.
[25] D. G. Kelley, "Stability in contractive nonlinear neural networks," *IEEE Trans. Biomed. Eng.*, vol. 37, pp. 231–242, Mar. 1990.
[26] B. Shi, "Oriented spatial pattern formation in a four-layer CMOS cellular neural network," *Int. J. Bifurcation Chaos*, to be published.
[27] T. Poggio and C. Koch, "Ill-posed problems in early vision: From computational theory to analogue networks," *Proc. R. Soc. Lond. B*, vol. 226, pp. 303–323, 1985.
[28] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 314–319, 1985.
[29] J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Oxford, U.K.: Clarendon, 1873, vol. 1.
[30] W. Millar, "Some general theorems for nonlinear systems possessing resistance," *Philosoph. Mag.*, ser. 7, vol. 42, pp. 1150–1160, 1951.
[31] D. L. Standley and J. L. Wyatt Jr., "Stability criterion for lateral inhibition and related networks that is robust in the presence of integrated circuit parasitics," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 675–681, May 1989.
[32] J. L. Wyatt, "Little-known properties of resistive grids that are useful in analog vision chip designs," in *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, C. Koch and H. Li, Eds. Los Alamitos, CA: IEEE Computer Society Press, 1995, pp. 72–104.
[33] B. Shi, "Pseudoresistive networks and the pseudovoltage-based cocontent," *IEEE Trans. Circuits Syst. I*, vol. 50, pp. 56–64, Jan. 2003.
[34] K. A. Zaghloul, "A silicon implementation of a novel model for retinal processing," Ph.D. dissertation, Dept. Neurosci., Univ. Pennsylvania, Phladelphia, 2001.
[35] M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex*. New York: Cambridge Univ. Press, 1991.
[36] V. Braitenberg and A. Schuz, *Anatomy of the Cortex: Statistics and Geometry*. Berlin, Germany: Springer-Verlag, 1991.
[37] J. Bullier, J. M. Hupe, A. C. Hames, and P. Girard, "The role of feedback connections in shaping the responses of visual cortical neurons," in *Vision: From Neurons to Cognition, Progress in Brain Research*, C. Casanova and M. Ptito, Eds. Amsterdam, The Netherlands: Elsevier, 2001, vol. 134, ch. 13, pp. 193–204.
[38] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Trans. Circuits Syst. II*, vol. 47, pp. 416–434, May 2000.
[39] A. G. Andreou and K. A. Boahen, "Neural information processing II," in *Analog VLSI: Signal and Information Processing*, M. Ismail and T. Fiez, Eds. New York: McGraw-Hill, 1994.
[40] E. Vittoz and X. Arreguit, "Linear networks based on transistors," *Electron. Lett.*, vol. 29, pp. 297–299, 1993.
[41] K. A. Boahen, "The retinomorphic approach: Pixel-parallel adaptive amplification, filtering, and quantization," *Analog Integr. Circuits Signal Processing*, vol. 13, pp. 53–68, 1997.
[42] E. Culurciello, R. Etienne-Cummings, and K. Boahen, "Arbitrated address event representation digital image sensor," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 2001, pp. 92–93.
[43] B. E. Shi, "A low power orientation selective vision sensor," *IEEE Trans. Circuits Syst. II*, vol. 47, pp. 435–440, May 2000.
[44] J. E. Dowling, *The Retinal: An Approachable Part of the Brain*. Cambridge, MA: Belknap, 1987.
[45] R. Shapley and J. Victor, "Nonlinear spatial summation and the contrast gain control of cat retinal ganglion cells," *J. Physiol.*, vol. 290, pp. 141–161, 1979.
[46] B. E. Shi and K. Boahen, "Competitively coupled orientation selective cellular neural networks," *IEEE Trans. Circuits Syst. I*, vol. 49, pp. 388–394, Mar. 2002.
[47] C. L. Baker Jr. and I. Mareschal, "Processing of second-order stimuli in the visual cortex," in *Vision: From Neurons to Cognition, Progress in Brain Research*, C. Casanova and M. Ptito, Eds. Amsterdam, The Netherlands: Elsevier, 2001, vol. 134, ch. 12.

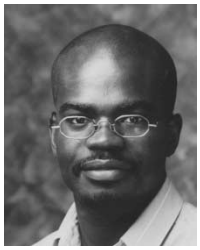
Thomas Yu Wing Choi (S'03) was born in Hong Kong, SAR, China, in 1975. He received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, in 1997, 1999, and 2003, respectively.

His research interests include analog VLSI design and neural networks.

Bertram E. Shi (S'93–M'95–SM'00–F'01) received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, and the Ph.D. degree in electrical engineering from the University of California at Berkeley.

He is currently an Associate Professor in the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China. His research interests are in analog VLSI and cellular neural networks, bio-inspired and neuromorphic engineering, machine vision, and image processing.

Dr. Shi was the Student Activities Chair of the IEEE Hong Kong Section, Secretary and Chair for the IEEE Circuits and Systems Society Technical Committee on Cellular Neural Networks and Array Computing and Distinguished Lecturer for the IEEE Circuits and Systems Society. He served as the Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS.



Kwabena A. Boahen received the B.S. and M.S.E. degrees in electrical and computer engineering from Johns Hopkins University, Baltimore, MD, in the concurrent masters-bachelors program, both in 1989, and the Ph.D. degree in computation and neural systems from the California Institute of Technology, Pasadena, in 1997.

He is an Associate Professor in the Bioengineering Department at the University of Pennsylvania, Philadelphia, where he holds a secondary appointment in electrical engineering. His current research interests include mixed-mode multichip VLSI models of biological sensory and perceptual systems, and their epigenetic development, and asynchronous digital interfaces for interchip connectivity.

Dr. Boahen was awarded a Packard Fellowship in 1999, a National Science Foundation CAREER Grant in 2001, and an Office of Naval Research YIP Grant in 2002. He is a member of Tau Beta Kappa and has held a Sloan Fellowship for Theoretical Neurobiology at the California Institute of Technology.