RETINOMORPHIC VISION SYSTEMS I: PIXEL DESIGN

Kwabena Boahen
Physics of Computation Laboratory
California Institute of Technology
MS 136-93, Pasadena, CA 91125, USA
buster@pcmp.caltech.edu

ABSTRACT

I present and analyze test results from circuits that perform all four major operations performed by biological retinae using neurobiological principles: (1) continuous sensing for detection, (2) local automatic gain control for amplification, (3) spatiotemporal bandpass filtering for preprocessing, and (4) adaptive sampling for quantization. In the retinomorphic system that I describe, all these operations are performed at the pixel level, to eliminate redundancy, to reduce power dissipation, and to make efficient use of the capacity of the output channel.

1. SYSTEM CONCEPT

The primary difference between retinomorphic imagers and conventional ones is that retinomorphic imagers perform all operations at the pixel level [1]. The migration of more sophisticated signal processing down to the pixel level is driven by shrinking feature sizes in CMOS technology, allowing higher levels of integration to be achieved [2]. The retinomorphic approach uses the architecture and neurocircuitry of the nervous system as a blueprint for building low-level vision systems—systems that are *retinomorphic* in a literal sense [1]. This approach results in integrated systems that offer enriched functionality, by performing several functions within the same structure, and enhanced system-level performance using minimal-area devices ($3L \times 3L$), by distributing computation across several pixels.

The retinomorphic system described in this paper consists of two chips: a focal-plane image processor and a post-processor with a two-dimensional array of integrators. Both chips are fully functional; specifications and die photos are shown in Table 1 and in Figure 1. I describe the pixel design in Section 2, and the circuits that perform spatiotemporal bandpass filtering, local AGC, temporal integration, and adaptive quantization in Sections 3 through 6. My concluding remarks are in Section 7. The communication channel used to transmit asynchronous pulse streams between these two chips is described in the companion paper [3].

2. RETINOMORPHIC PIXEL DESIGN

The circuitry in each pixel of the retinomorphic processor is shown in Figure 2. In general terms, the principles of operation are as follows: A CMOS-compatible, vertical bipolar phototransistor performs **continuous sensing**; its emmiter current is proportional to the incident light intensity [4]. Two current spreading networks [5, 6, 7] diffuse the photocurrent signals over time and space; the first layer (node V0) excites the second layer (node W0), which reciprocates by inhibiting the first layer. The result is a **spatiotemporal bandpass filter** [8, 9, 10]. The second layer computes a measure of the light intensity, and feeds this information back to the input layer, where it is used to control light sensitivity. The result is **local automatic gain control** (AGC) [5]. A pulse generator converts analog currents from the excitatory layer into pulse-frequency. The diode-capacitor integrator computes a current that is proportional to the short-term average of the pulse frequency and this current is subtracted form the pulse generator's input. Hence, the more rapidly the input changes, the more rapidly the pulse generator fires. Adding a fixed charge quantum to the integrating capacitor produces a multiplicative change in current—due to the exponential current-voltage dependence in subthreshold. Hence, the larger the current level, the larger the step size. The result is **adaptive quantization** in amplitude *and* time. The diode-capacitor integrator is also used in the postprocessor to integrate the pulses and reconstruct the current level encoded.

	Imager	Postprocessor
Technology	2μm 2-poly 2-metal pwell	
Number of Pixels	64×64	
Pixel Size (L^2)	53×49	31.5×23
Transistors/pixel	32	8
Die Size (mm^2)	8.1×7.4	5.1×4.0
Supply	5 V	
Dissipation (0.2 MHz)	230 mW (total)	
Throughput	2 MHz	

Table 1. Specifications of two-chip retinomorphic system. L is the minimum feature size which is 2μm for the CMOS process used.

3. SPATIOTEMPORAL BANDPASS FILTER

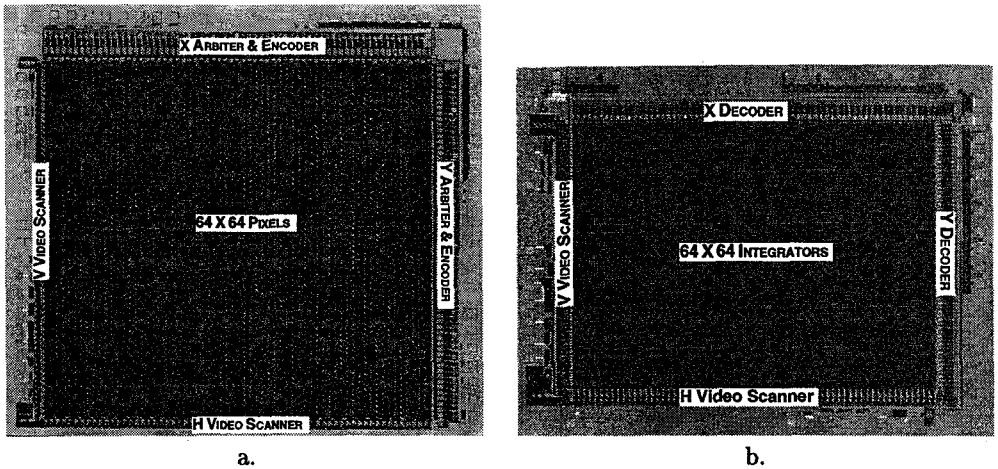
Using the small-signal equivalent model of the OPL circuit shown in Figure 3, we find that

$$I_o + \nabla^2 V_c / r_{cc} = g_{c0} V_c + c_{c0} \dot{V}_c + g_{ch} V_h, \quad (1)$$

$$g_{hc} V_c + \nabla^2 V_h / r_{hh} = g_{h0} V_h + c_{h0} \dot{V}_h, \quad (2)$$

in the continuum limit. Here, V_c is the voltage in the excitatory network, which models retinal cones; V_h is the voltage in the inhibitory network, which models retinal horizontal cells (HC); and I_o is the photocurrent [10]. These functions are now continuous functions of space, (x, y), and time, t; $\nabla^2 f$ is the Laplacian of f (i.e., $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2$) and $\dot{f}$ is the temporal derivative of f (i.e.,$\partial f/\partial t$). Models similar to this one were proposed and analyzed in [8, 9, 11].

0-7803-3073-0/96/$5.00



a. b.

Figure 1. Die photos of (a) Retinomorphic focal-plane processor and (b) Postprocessor.

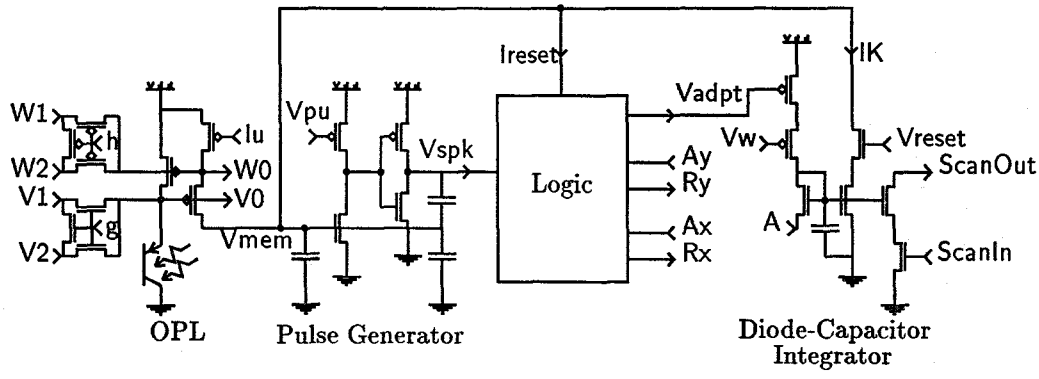


Figure 2. Pixel circuit for retinomorphic imager. The outer-plexiform–layer (OPL) circuit performs spatiotemporal bandpass filtering and local AGC. Nodes V0 and W0 are connected to their six nearest neighbors on a hexagonal grid by the delta-connected transistors. The logic circuit communicates the occurrence of a spike to the chip periphery, turns on Ireset, and takes Vadapt low. The pulse-generator and the diode-integrator capacitor form an adaptive neuron circuit. The remaining circuitry is used to scan out the integrator's output. Details of the logic circuit are revealed in the companion paper.

Assuming infinite spatial extent and homogeneous initial conditions, we can take Fourier transforms in space and time. Transforming the equations and solving, we find that

$$\tilde{H}_c = \frac{1}{g_{ch}} \frac{\ell_h^2\rho^2 + i\tau_h\omega + \epsilon_h}{(\ell_c^2\rho^2 + i\tau_c\omega + \epsilon_c)(\ell_h^2\rho^2 + i\tau_h\omega + \epsilon_h) + 1},$$

where $\tilde{H}_c(\rho,\omega) \equiv \tilde{V}_c/\tilde{I}_o$. $\tilde{f}(\rho_x,\rho_y,\omega)$ denotes the Fourier transform of $f(x,y,t)$; $\rho = \sqrt{(\rho_x^2+\rho_y^2)}$ is spatial frequency, and ω is temporal frequency (both in radians) [10]. Here, $\tau_c = c_{c0}/g_{ch}$ and $\tau_h = c_{h0}/g_{hc}$ are the time constants associated with the HC-to-cone coupling and the cone-to-HC coupling, respectively; $\ell_c = (r_{cc}g_{ch})^{-1/2}$ and $\ell_h = (r_{hh}g_{hc})^{-1/2}$ are the space constants of the decoupled networks, with transconductances replaced by conductances to ground; and $\epsilon_c = g_{c0}/g_{ch}$ and $\epsilon_h = g_{h0}/g_{hc}$ are the ratios of leakage conductance to the transconductance. The reciprocals of ϵ_c and ϵ_h are the open-loop voltage gains from the HC to the cone, and vice versa.

The spatiotemporal frequency response of the excitatory cone network obtained from this analysis is plotted in Figure 4. The set of parameters values used was: $\ell_c = 0.05°$, $\ell_h = 0.2°$, $\tau_c = 30$ms, $\tau_h = 200$ms, $\epsilon_c = 0.3$, $\epsilon_h = 0.1$, $g_{ch} = 0.2$pA/mV. Observe that the temporal frequency response is bandpass at low spatial frequencies (flicker sensitivity), and the spatial frequency response is bandpass at low temporal frequencies (grating sensitivity). However, the overall response is not linearly separable; that is, it is not simply the composition of a bandpass spatial filter and a bandpass temporal filter. The spatial tuning becomes lowpass at high temporal frequencies and the temporal tuning becomes lowpass at high spatial frequencies [10].

4. LOCAL AGC

We achieve local AGC by making the intercone conductance $(1/r_{cc})$ proportional to the local average of the photocurrent, since

$$V_c(x) = r_{cc}I_oLe^{-|x|/L}\sin(|x|/L - \pi/4)/(2\sqrt{2}),$$

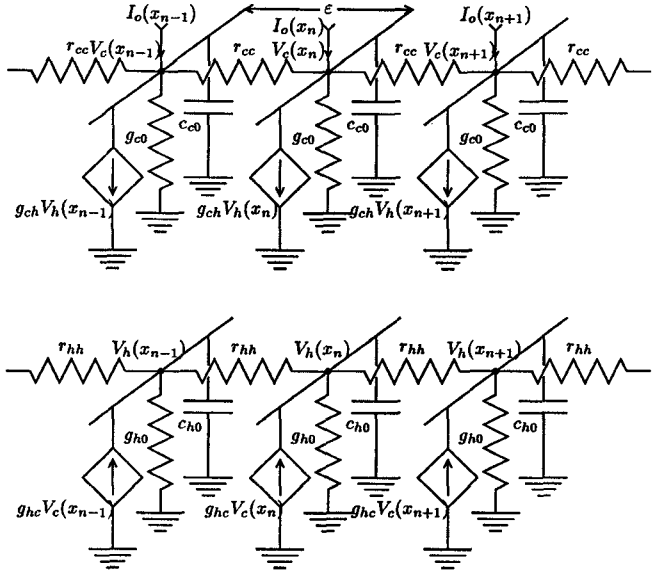


Figure 3. Linear circuit model of the retina's outer plexiform layer (OPL). Two resistive networks model the inter-cone and the inter-horizontal cell electrical synapses (gap junctions) and transconductances model the reciprocal chemical synapses between cones and horizontal cells. The circuit is analyzed in the continuum limit where $\epsilon \to 0$.

in one-dimensional space, when $g_{c0} = g_{h0} = 0$ [12]. This can be shown by taking the inverse Fourier transform of $\tilde{H}_c$. The effective space constant of the coupled, dual-layer network is therefore $L = \sqrt{\ell_c \ell_h} = (r_{cc} g_{ch} r_{hh} g_{hc})^{-1/4}$. Gain adaptation is realized in the circuit simply by the fact that (Vdd – V0) equals the sum of the gate–source voltages of two devices. The currents passed by these devices represent the activity in the inhibitory network, I_h, which is equal to the local average of the intensity, and the activity of the excitatory network, I_c, which is equal to the Laplacian of the smoothed intensity profile (see Equation 2). Hence, by the translinear principle [13, 7], the current that spreads in the excitatory network is proportional to the product, $I_c I_h$, of these currents. Since I_h scales with the intensity, the internode conductance in the excitatory cone network will scale accordingly [5].

The receptive field contracts as we increase r_{cc} to increase the gain, since the space constant L also depends on the intercone conductance. This undesirable side-effect is evident in the images produced by this OPL circuit that are shown in Figure 5; this data is from the chip described in [5]. Images of the same scenes acquired with a CCD camera are included for comparison [14]. The retinomorphic front-end pulls out information in the shadows whereas the output of the CCD camera has hit its lower limit, demonstrating that local AGC indeed increases the dynamic range. The spatiotemporal bandpass filtering also removes gradual changes in intensity and enhances edges and curved surfaces. Unfortunately, the retinomorphic chip's output is more noisy in the darker parts of the image. When the space constant decreases, salt-and-pepper noise is no longer attenuated because the cutoff frequency shifts upwards. The dominant noise source is the poor matching

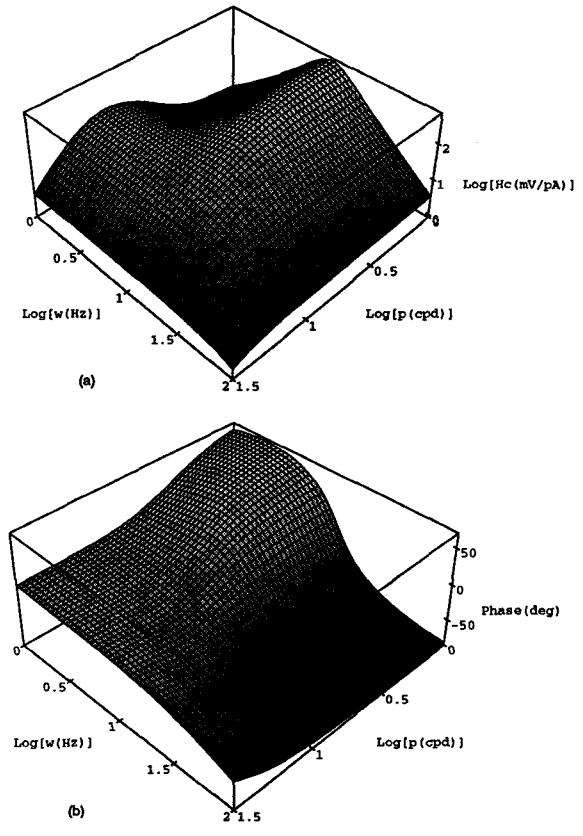


Figure 4. Spatiotemporal sensitivity of linear OPL model. Three-dimensional plots showing (a) magnitude and (b) phase versus spatial frequency (p) and temporal frequency (w).

among the small ($4L \times 3.5L$) transistors used which operate in subthreshold [15]—not shot noise in the photon flux. Nevertheless, when it replaced the CCD as the front-end of a face-recognition system, the OPL chip reduced the error rates by 50% [14].

5. DIODE-CAPACITOR INTEGRATOR

This integrator is based on the well-known current mirror circuit. A large capacitor at the input of the mirror integrates charge, and the diode-connected transistor leaks charge away. In subthreshold, the current has an exponential dependence on the gate voltage and therefore the small-signal conductance of the diode-connected transistor is proportional the current. Hence, the time-constant will change as the current level changes. This circuit's temporal behavior is described by a nonlinear differential equation

$$Q_T \frac{dI_{out}}{dt} = I_{out}(t)(I_{in}(t) - \tfrac{1}{A} I_{out}(t)),$$

where $U_T \equiv kT/q$ is the thermal voltage, $A = \exp(V_A/U_T)$ is the current gain of the mirror, and $Q_T \equiv CU_T/\kappa$ is the charge required to e-fold the current [4, 16].

The output produced by a periodic sequence of current

Figure 5. CCD Camera (top row) versus OPL imager chip (bottom row) under variable lighting. The CCD camera performs global AGC whereas the OPL chip performs local AGC and bandpass-filtering.

pulses is

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(t_0 + nT)}{\frac{I_{\text{out}}(t_0+nT)}{\text{AQ}_\text{T}}(t - (t_0 + nT)) + 1},$$

immediately after the $(n+1)$th pulse, and decays like

$$I_{\text{out}}(t_0 + nT) = \{1/\hat{I}_T + (1/I_{\text{out}}(t_0) - 1/\hat{I}_T)(1+\alpha)^{-n}\}^{-1}$$

during the interspike interval, $t_0 + nT < t < t_0 + (n+1)T$, where $\hat{I}_T \equiv \alpha \text{AQ}_\text{T}/T$, and $\alpha \equiv (\exp(q_\alpha/\text{Q}_\text{T}) - 1)$ is the percentage by which the output current is incremented by each spike [16]. The fixed quantity of charge q_α supplied by each current pulse multiplies the current by $\exp(q_\alpha/\text{Q}_\text{T})$, since it takes Q_T to e-fold. The peak output current levels attained immediately after each spike converge to $\hat{I}_T \equiv \alpha \text{AQ}_\text{T}/T$ when $(1+\alpha)^{-n} \ll 1$. Therefore, the equilibrium output current level is proportional to the pulse frequency.

6. ADAPTIVE NEURON CIRCUIT

We build an adaptive neuron circuit by taking a pulse generator and placing a diode-capacitor integrator around it, in a negative feedback configuration. The pulse-generation circuit has a high-gain amplifier (two digital inverters) with positive feedback around it (capacitive divider) [17]. (See Figure 2). The high-gain amplifier serves as a thresholding device and the positive feedback provides hysteresis. Positive feedback also increases the slew rate of the input and reduces the rise and fall times of the output pulse. This is especially important when the neuron has to drive global column and row lines to communicate the occurrence of a pulse [3], since the 0.1V/ms rate at which its input charges (determined by the desired firing rate), translates to a 10mV/μs rate at the output (with a voltage gain of 100 from two inverters). Transmission speed is limited by this slow slew rate—not by drive capacity—which lengthens the communication cycle period. The neuron also has a reset current (Ireset), produced by the logic circuit, that terminates the spike. Other designs for adaptive neurons are described in [18, 19].

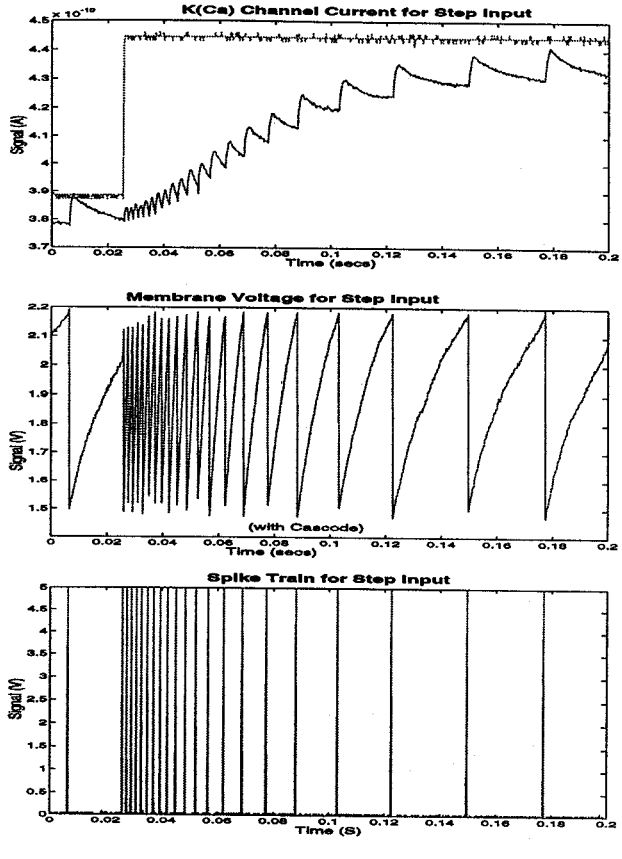


Figure 6. Adaptive neuron's step response. Top: The neuron's input current and the integrator's output current. Middle: Input voltage ramping up between the reset and threshold levels. Bottom: The spike train.

The complete adaptive neuron circuit is described by two coupled differential equations:

$$\text{C}_\text{mem}\frac{dV_\text{mem}}{dt} = I_\text{in} - I_\text{K} - \text{Q}_\text{th}\delta(V_\text{mem} - \text{V}_\text{th}), \quad (3)$$

$$\text{Q}_\text{T}\frac{dI_\text{K}}{dt} = I_\text{K}(\text{q}_\alpha\delta(V_\text{mem} - \text{V}_\text{th}) - \frac{1}{\text{A}}I_\text{K}), \quad (4)$$

where I_in is the current supplied to node Vmem by the OPL circuit, and C_mem is the total capacitance connected to that node; I_K is the current subtracted from node Vmem by the integrator; C_Ca is the integrator's capacitance; $\text{Q}_\text{T} = \text{C}_\text{Ca}\text{U}_\text{T}/\kappa$; and Q_th is the repolarization charge; that is, the charge we must supply to Vmem to bring it from the reset level to the threshold level (V_th). For a constant input current, we can integrate these equations and obtain

$$\text{Q}_\text{th} = I_\text{in}\Delta_n - \text{AQ}_\text{T}\ln(I_{\text{K}n}\Delta_n/\text{AQ}_\text{T} + 1),$$

where $\Delta_n \equiv t_{n+1} - t_n$ is the interspike interval. This analysis ignores the parasitic coupling capacitance between Vmem and the integrator's input node, and that capacitance can have a large influence on the circuit's behavior [16]. In this particular design, the cascode device between the integrator's output and the pulse generator's input (tied to Vreset) eliminates virtually all coupling.

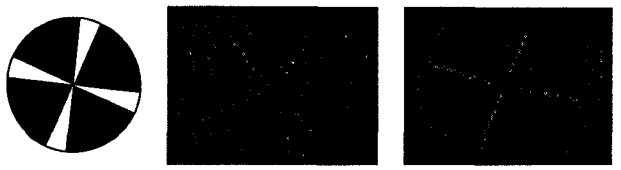


Figure 7. Video frames from AE postprocessor chip showing real-time temporal integration of pulses. The stimulus is a windmill pattern (left) that rotates counterclockwise slowly (middle) and quickly (right).

When adaptation is complete, the interspike intervals become equal, and the integrator's output current converges to $I_{Kn} = \alpha A Q_T / \Delta_n$. Hence,

$$\Delta_n = (Q_{th} + Aq_\alpha)/I_{in} = ZQ_{th}/I_{in}$$

(remember that $q_\alpha = Q_T \ln(1+\alpha)$). This result is understood as follows. During the interspike interval, Δ_n, the input current must supply the charge Q_{th} to the capacitors tied to Vmem and supply the charge Aq_α removed by the integrator, where q_α is the quantity of charge added to the integration capacitor by each spike. Notice that adaptation reduces the firing rate by a factor of $Z \equiv 1 + Aq_\alpha/Q_{th}$. The response of the adaptive neuron circuit to a 14 percent change in its input current is shown in Figure 6; this data also demonstrates the integration of pulse trains by the diode-capacitor integrator and the adaptive step-size.

7. CONCLUSIONS

The output of the postprocessor that integrates the spike trains—after transduction, local AGC, bandpass filtering, adaptive quantization, and interchip communication—is shown in Figure 7. The sparseness of the output representation is evident. When the windmill moves, neurons at locations where the intensity is increasing (white region invades black) fire more rapidly; hence, the leading edges of the white vanes are more prominent. The mean spike rate was 30Hz per pixel, and the two-chip system dissipated 190 mW at this spike rate.

Taking inspiration from biology, I have described an approach to building machine vision systems that perform sophisticated signal processing at the pixel level. Such systems can be maximally adaptive to their inputs and thereby optimize their information gathering capacity. Specific implementations of all the circuit functions required were presented. The interchip communication system used is described in a companion paper [3].

8. ACKNOWLEDGMENTS

I thank my advisor, Carver Mead, for sharing his insights into the operation of the nervous system, and my former advisor, Andreas Andreou for setting me on this path. This work is supported in part by the Office of Naval Research, by ARPA, by the Beckman Foundation, and by the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation Engineering Research Center Program, and by the California Trade and Commerce Agency, Office of Strategic Technology.

REFERENCES

[1] Kwabena A Boahen. Retinomorphic vision systems. In *MicroNeuro'96: Fifth Int. Conf. on Neural Networks and Fuzzy Systems*, Los Alamitos, CA, February 1996. IEEE Computer Soc. Press.

[2] Carver A Mead. Scaling of mos technology to submicrometer feature sizes. *J. VLSI Signal Processing*, 8:9–25, 1994.

[3] Kwabena A Boahen. Retinomorphic vision systems ii: Communication channel design. In *IEEE Int. Symp. on Circuits and Systems*, Piscataway, NJ, July 1996. IEEE Circuits and Systems Soc., IEEE Press.

[4] Carver A Mead. A sensitive electronic photoreceptor. In H. Fuchs, editor, *1985 Chapel Hill Conference on VLSI*, pages 463–471. Computer Science Press, 1985.

[5] Kwabena Boahen and Andreas Andreou. A contrast-sensitive retina with reciprocal synapses. In J E Moody, editor, *Advances in neural information processing 4*, volume 4, San Mateo CA, 1991. Morgan Kaufman.

[6] Eric Vittoz and Xavier Arreguit. Linear networks based on transistors. *Electronics Letters*, 29:297–299, 1993.

[7] Andreas Andreou and Kwabena Boahen. Translinear circuits in subthreshold mos. *J. Analog Integrated Circ. Sig. Proc.*, March 1996.

[8] P C Chen and A W Freeman. A model for spatiotemporal frequency responses in the x cell pathway of cat's retina. *Vision Res.*, 29:271–291, 1989.

[9] S Ohshima, T Yagi, and Y Funashi. Computational studies on the interaction between red cone and h1 horizontal cell. *Vision Res.*, 35(1):149–160, 1994.

[10] Kwabena A Boahen. Spatiotemporal sensitivity of the retina: A physical model. Technical Report CNS-TR-91-06, California Institute of Technology, Pasadena CA, 1991.

[11] R G Smith. Simulation of an anatomically define local circuit — the cone-horizontal cell network in cat retina. *Visual Neurosci.*, 12(3):545–561, May-Jun 1995.

[12] Kwabena A Boahen. Towards a second generation silicon retina. Technical Report CNS-TR-90-06, California Institute of Technology, Pasadena CA, 1990.

[13] Barrie Gilbert. Translinear circuits: A proposed classification. *Electronics Letters*, 11(6):136, 1975.

[14] J Buhman, M Lades, and Eeckman F. Illumination-invariant face recognition with a contrast sensitive silicon retina. In J D Cowan, G Tesauro, and J Alspector, editors, *Advances in neural information processing 6*, volume 6, San Mateo CA, 1994. Morgan Kaufman.

[15] A Pavasović, A G Andreou, and Westgate C R. Characterization of subthreshold mos mismatch in transistors for vlsi systems. *J. Analog Integr. Circ. and Sig. Proc.*, 6:75–84, June 1994.

[16] Kwabena A Boahen. The adaptive neuron and the diode-capacitor integrator. *In preparation.*

[17] Carver A Mead. *Analog VLSI and Neural Systems.* Addison Wesley, Reading MA, 1989.

[18] M Mahowald and D Douglas. A silicon neuron. *Nature*, 354(6345):515–518, 1991.

[19] John Lazzaro. Temporal adaptation in a silicon auditory nerve. In D Tourestzky, editor, *Advances in Neural Information Processing 4*, volume 4. Morgan Kaufmannn Pub., 1992.