Retinomorphic Vision Systems

Kwabena Boahen
Physics of Computation Laboratory
California Institute of Technology
MS 139-74, Pasadena, CA 91125, USA
E-mail: buster@pcmp.caltech.edu

Abstract

The new generation of silicon retinae has two defining characteristics. First, these synthetic retinae are morphologically equivalent to their biological counterparts—at an appropriate level of abstraction. Second, they accomplish 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. I introduce the term* retinomorphic *to refer to this subclass of the neuromorphic electronic systems [30]. I compare and contrast their design principles with the standard practice in imager design. I argue that neurobiological principles are best suited to perceptive systems [43] that go beyond reproducing the dynamic scene, like a conventional video camera does, to extracting salient information in real time [3]. I shall present results from a fully operational retinomorphic vision system and discuss the trade-offs involved in its design.*

1: Why build Retinomorphic Systems?

The retina is an exquisitely evolved piece of neuronal wetware. It contains about one 100 million black-and-white photoreceptors, complemented by 3 to 4 million color receptors. Its output—about one million axonal fibers that make up the optic nerve—conveys visual information to the rest of brain using an all-or-none pulse code. Compared to a state-of-the-art charge-coupled device (CCD) camera, it accomplishes many amazing feats.

Parallel processing of visual information begins in the retina with the presence of several channels specialized for nocturnal vision, color vision, spatial vision, and motion, among others. Under ideal conditions, these channels allow us to detect reliably the absorption of 10 photons in a pool of 5000 rods; to perceive color in wavelengths of light ranging from 400 nm to 670 nm; to detect 0.5 percent contrast; to resolve two lines substending an angle of 1/60 of a degree; and to tell the order of onset of two lines flashed 3 to 5 milliseconds apart. In addition, we can see well in both dim starlight and bright sunlight—a dynamic range of over ten decades! In contrast, an 8-bit CCD camera's 0.4 percent full-scale amplitude resolution comes close to matching the retina's contrast sensitivity but the electronic camera's 1/5 degree angular resolution and its 30 ms temporal resolution are an order of magnitude worse while its 50 dB dynamic range is six orders of magnitude short. We can therefore advance the state of the art in focal-plane image processing by studying the expanding body of knowledge gathered by neurobiologists about how the retina operates [22].

I outline the general design principles of the retina and contrast the retinomorphic approach with standard engineering practice in Section 2. Next, I look at the technology scaling trends in Section 3, and conclude that we already have the transistors required to perform all the retinal operations in the pixel. Then I describe a design for a retinomorphic pixel that accomplishes this goal in Section 4. Finally, I deal with the problem of reading out asynchronous bit-streams from the pixels in Section 5, and describe how a retinomorphic chip is interfaced with another neuromorphic chip in a complete system. My concluding remarks are in Section 6.

2: What are the Design Principles?

The design principles of the retina, which are borrowed by retinomorphic systems, are outlined in Table 1. The design principles employed by standard imager technology also are listed for comparison. These principles are elaborated in this section.

1086-1947/96 $5.00 © 1996 IEEE
Proceedings of MicroNeuro '96

Operation	Standard	Retinal
Detection	Integrate/reset	Continuous
Amplification	Global AGC	Local AGC
Preprocessing	Absent	Bandpass filter
Quantization	Fixed	Adaptive

Table 1. Retinal design principles. AGC - Automatic gain control.

Integrating detectors (e.g.,CCDs [18] and photogates [13]) suffer from blooming at high intensity levels and require a destructive readout (reset) operation. Continuous sensing detectors (e.g., photodiodes or phototransistors) do not bloom, and can therefore operate over a much larger dynamic range [28]. In addition, redundant readout operations can be eliminated with considerable power savings, because it is not necessary to reset the detector. Continuous-sensing detectors have been shunned because they suffer from gain and offset mismatches that give rise to salt-and-pepper noise in the image. However, preliminary results indicate that the learning capability of image-recognition systems can easily compensate for this fixed pattern noise [9].

Imagers that use global automatic gain control (AGC) can operate only under uniform lighting, because the variation of intensity across a scene exceeds their 8-bit dynamic range when shadows are present [1]. A charge-coupled device or photogate can achieve 12 bits (four decades) [13] and a photodiode or phototransistor can achieve 20 bits (six decades) [28, 12]—but the phototransistor's performance in the lowest two decades is plagued by slow temporal response. The system's dynamic range, however, is limited by the cost of precision analog read-out electronics and A/D converters, and by video standards. Introducing AGC at the pixel level increases dynamic range and increases resolution in the darker parts of the image without increasing the number of bits per sample. Following retinal principles, the gain is set to be inversely proportional to the local intensity, discounting gradual changes in intensity and producing an output that is proportional to contrast [5]. This adaptation greatly extends the dynamic range because lighting intensity varies by six decades from high noon to twilight, whereas contrast varies by at most a factor of 20 [35].

The intensity pattern that falls on the imager is highly redundant in space and time; that is, differences between adjacent samples in space or time are rare [14]. Bandpass spatiotemporal filtering is an optimal strategy for removing redundancy in the presence of white noise [4, 15, 41]. This preprocessing results in a sparse output representation in space and time (whitening), with the minimum amount of redundancy required to protect the signal from noise introduced by the signal source or by the circuit elements. Coupled with adaptive quantization, this efficient image representation requires much less bandwidth for transmission. It also enhances features at finer spatial and temporal scales, making recognition easier [9] and providing a spatial or temporal reference for motion computation.

Converters that automatically adapt their quantization in time and amplitude to the rate of change and to the amplitude probability distribution of the input signal, respectively, maximize the information that is transmitted through the output channel. In contrast, the quantization of traditional A/D converters is set to match the maximum rate of change and the smallest amplitude, respectively. This encoding produces a lot of redundant samples, because changes in the signal are rare [14]. Also, the large amplitude codes are seldom used since these signal amplitudes rarely occur in natural scenes [35]. Reassigning these codes to more probable amplitudes will increase the overall number of signals that can be discriminated. Thus information is maximized when all codes are equiprobable [36]. If conversion occurs in parallel at the pixel level, each converter can adapt its quantization independently. Corruption of analog signals by the switching noise produced by high-speed multiplexing is also avoided.

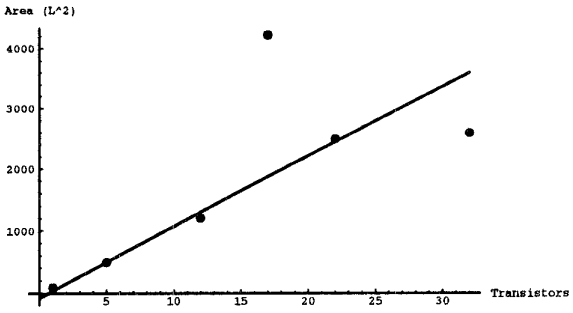

Figure 1. Scaling of pixel size with number of transistors. Area is measured in units of the minimum gate length (L) squared. The linear-regression fit indicates that each additional transistor costs $115L^2$. Data from Table 2.

[1] I am assuming a linear encoding which is the standard practice. This limits the dynamic range to 2^b for a b-bit encoding.

Pixel-Level Operations	No. of Pixels	Pixel Area (μm^2)	$L(\mu m)$	Size (L^2)	Transistors
Detection (CCD [18])	962 × 654	5.05 × 5.55	—	—	1
Amplification (CMD [33])	660 × 492	7.3 × 7.6	0.4*	85.5	1
Amplification (APS [13])	256 × 256	20 × 20	0.9	492.8	5
Filtering and LAGC [1]	230 × 210	39.6 × 43.8	1.2	1204.5	12
Conversion (Σ-Δ [17])	64 × 64	60 × 60	1.2	2500	22
Conversion (PFM [32])	10 × 10	104 × 104	1.6	4225	17
Retinomorphic	64 × 64	106 × 98	2.0	2597	32

Table 2. Trends in imager design. L is the minimum gate length. CCD - charge-coupled device; CMD - charge modulation device; APS - active pixel sensing; LAGC - local automatic gain control; Σ-Δ - sigma-delta; PFM - pulse-frequency modulation. *Estimated.

3: Do we have Enough Transistors?

The primary difference between retinomorphic imagers and conventional ones is that retinomorphic imagers perform all four operations listed in Table 1 at the pixel level. 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. The representative examples listed in Table 2 illustrate this trend. The pixel area, normalized by the square of the minimum gate length, is plotted versus the number of transistors in Figure 1.

If this trend continues, it will be possible to design pixels with 10 transistors that are 13 μm on a side and pixels with 35 transistors that are 25 μm on a side in a 0.4 μm process. As the size of the active devices becomes small compared to the sensor area—which is typically about 5 μm on a side—it will become cost effective to shrink the detector area and to use lenselet arrays to focus the light [42], freeing up that area for additional image-processing functions. Hence, it should be feasible to build a 730 × 730 pixel imager with 10 transistors per pixel, or a 380×380 pixel imager with 35 transistors per pixel, on a 1 cm square die in today's state-of-the-art 0.4 μm CMOS process. In comparison, the human fovea has only about 500 × 500 cones, the density is much higher, however: these cones occupy an area of just 1.5 mm × 1.5 mm!

It has been clear for over 20 years that this technology scaling trend is going to give us many more transistors than we know what to do with [20, 31]. The work described in this paper addresses this problem by taking inspiration from biological vision systems [27]. In particular, the retinomorphic approach uses the system architecture and neurocircuitry of the nervous system as a blueprint for building integrated, low-level, vision systems—systems that are *retinomorphic* in a literal sense. 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 postprocessor with a two-dimensional array of integrators. The system concept is shown in Figure 2. Both chips are fully functional; specifications and die photos are shown in Table 3 and in Figure 3. I describe the retinomorphic pixel design in Section 4.

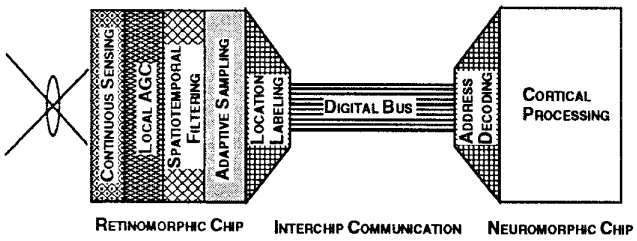

Figure 2. System concept. The retinomorphic chip acquires, conditions, prefilters, and quantizes the image. All these operations are performed at the pixel level. The interchip communication channel reads out digital pulses from the pixels by transmitting the location of pulses as they occur. A second neuromorphic decodes these address events and recreates the pulses.

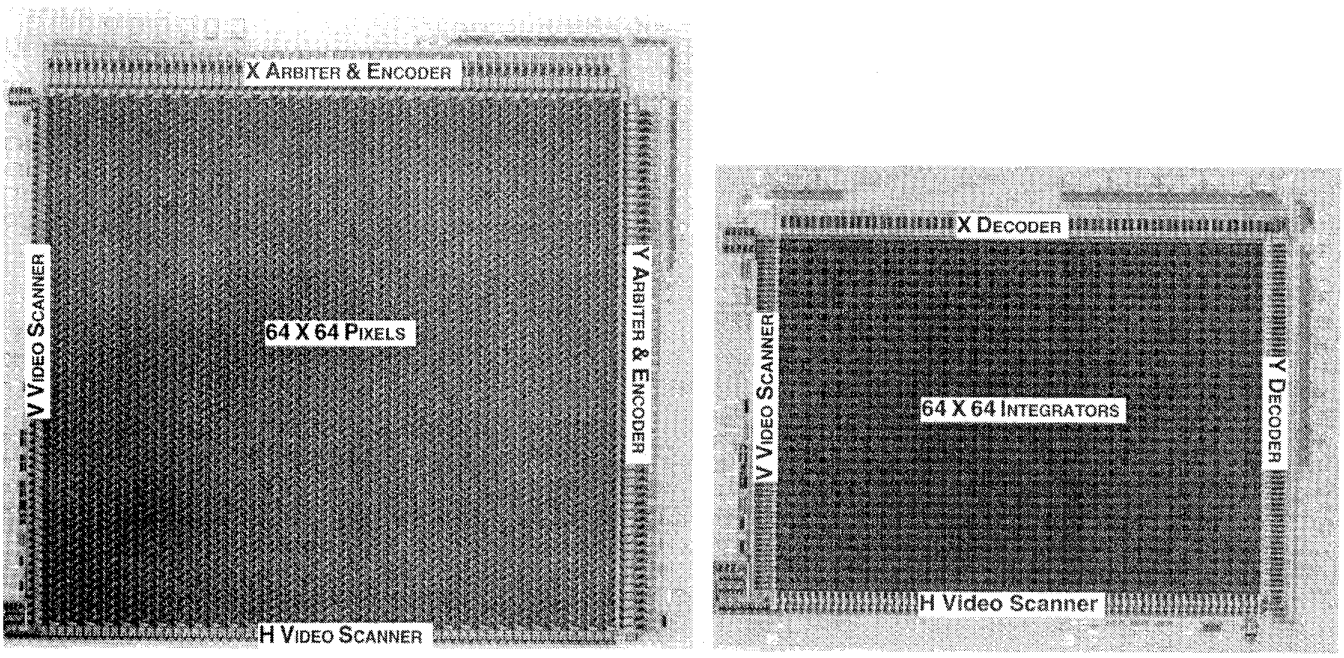

Figure 3. Die photos of (a) Retinomorphic focal-plane processor and (b) Postprocessor.

	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 3. Specifications of two-chip retinomorphic system.

4: How do we Build the Pixels?

The circuitry in each pixel of the retinomorphic processor is shown in Figure 4. In general terms, the principles of operation are as follows: The transducer is a vertical bipolar transistor; its emmiter current is proportional to the incident light intensity [28]. Two current spreading networks [5, 2, 44, 10] 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 spatiotemporally bandpass-filtered image [11, 34, 8]. 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 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. The diode-capacitor integrator is also used in the postprocessor to integrate the pulses and reconstruct the current level encoded.

4.1: Spatiotemporal Bandpass Filter

Using the small-signal equivalent model of the OPL circuit shown in Figure 5, 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 [8]. These functions are now continuous functions of space, (x, y), and time, t; $\nabla^2 f$ is the Laplacian of f

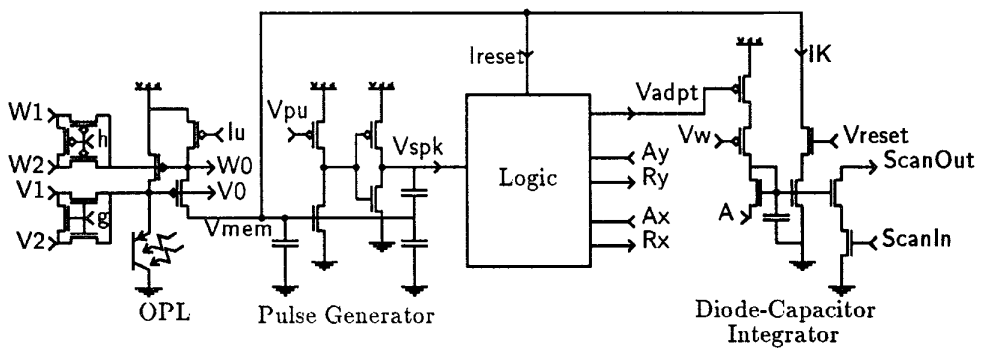

Figure 4. Pixel circuit for retinomorphic imager. The outer-plexiform–layer (OPL) circuit performs spatiotemporal bandpass filtering and local AGC. Nodes V0 and W0 are connected to their six nearest neighbors on a hexagonal grid by the delta-connected transistors. The logic circuit communicates the occurrence of a spike to the chip periphery, turns on Ireset, and takes Vadapt low. The remaining circuitry is used to scan out the integrator's output. Details of the logic circuit are revealed in Figure 9 (upper-left corner).

(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 [11, 34, 38].

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) [8]. 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 6. 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 [8]. The same behavior is observed in physiological data measured from cats [16] and psychophysical data measured from humans [21].

I gained a key the following insights from this analysis. There are tradeoffs among small low-frequency response, large dynamic range, and high sensitivity. A high-gain cone-to-HC synapse (i.e. small ϵ_h) is required to attenuate the cone's response to low spatial and temporal frequencies since $\tilde{H}_c(0,0) = \epsilon_h/g_{ch}$. However, increasing the gain of the cone-to-HC synapse decreases the dynamic range of the cone, (i.e. $V_c < \epsilon_h V_{\text{lin}}$, where V_{lin} is the linear range.) It also makes the circuit ring since $Q = (\epsilon_c\sqrt{\frac{\tau_h}{\tau_c}} + \epsilon_h\sqrt{\frac{\tau_c}{\tau_h}})^{-1}$. Smith and Sterling realized this constraint on the loop gain and proposed that feedforward inhibition to second-order cells (bipolar cells) may be used to attenuate the DC response [38]. Alternatively, we can decrease the gain of the HC-to-cone feedback synapse ($1/\epsilon_c$) or reduce the HC's time constant (τ_h) to maintain stability. Unfortunately, both changes reduce the peak sensitivity of the cone $\tilde{V}(0,\hat{\omega}) = Q\sqrt{(\tau_h(\epsilon_c\epsilon_h + 1)/\tau_c)}$. The circuit implementation shown in Figure 4 has high gain from the excitatory cone node (V0) to the inhibitory

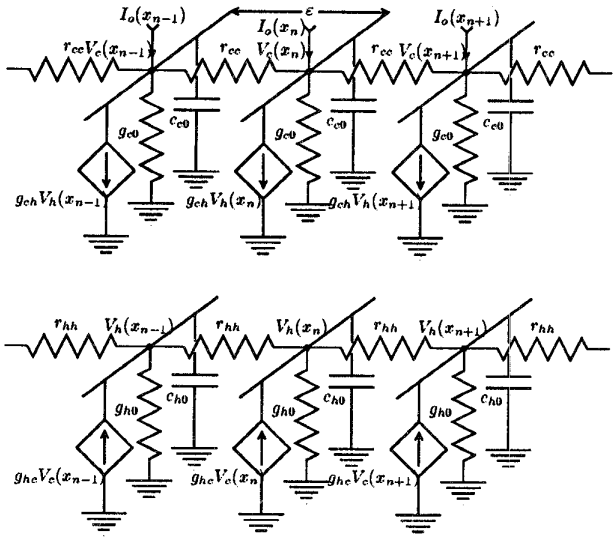

Figure 5. Linear circuit model of the retina's outer plexiform layer (OPL). Two resistive networks model the inter-cone and the inter-horizontal cell electrical synapses (gap junctions) and transconductances model the reciprocal chemical synapses between cones and horizontal cells. The circuit is analyzed in the continuum limit where $\epsilon \to 0$.

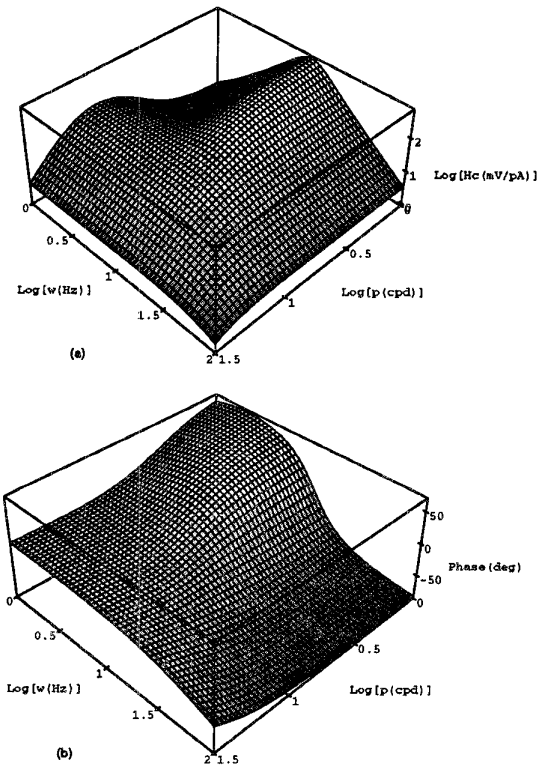

Figure 6. Spatiotemporal sensitivity of linear OPL model. Three-dimensional plots showing (a) magnitude and (b) phase versus spatial frequency (p) and temporal frequency (w).

HC node (W0), giving it small DC response and high sensitivity but poor temporal stability.

4.2: Local AGC

We achieve local AGC by making the intercone conductance ($1/r_{cc}$) proportional to the local average of the photocurrent. This 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 [19, 2], 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 coupling in the excitatory cone network will scale accordingly [5].

It remains to show that the response of the excitatory cone network is proportional to the intercone resistance [7]. We can obtain closed-form solutions for the impulse response in one-dimensional space:

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

where $L = \sqrt{\ell_c \ell_h} = (r_{cc} g_{ch} r_{hh} g_{hc})^{-1/4}$ is the effective space constant of the dual-layer network. These solutions are valid for the case $g_{c0} = g_{h0} = 0$, which is a fairly good approximation of the actual circuit. Linear-system theory thus predicts that the gain of the cone is equal to the product of the space constant and the intercone coupling resistance.

This analysis reveals a compromise made by this approach to local AGC. The space constant also depends on the intercone conductance: $L = (r_{cc} g_{ch} r_{hh} g_{hc})^{-1/4}$ Thus, as we increase r_{cc} to increase the gain, the receptive field contracts. This effect is evident in the images produced by this OPL circuit that are shown in Fig-

Figure 7. 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.

ure 7; this data is from the chip described in [5]. Images of the same scenes acquired with a CCD camera are included for comparison [9]. 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 among the small ($4L \times 3.5L$) transistors used—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% [9].

4.3: 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

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

where $\mathrm{U_T} \equiv \mathrm{kT/q}$ is the thermal voltage, $\mathrm{A} = \exp(V_{\mathrm{A}}/\mathrm{U_T})$ is the current gain of the mirror, and $\mathrm{Q_T} \equiv \mathrm{CU_T}/\kappa$ is the charge required to e-fold the current [28, 6].

The output produced by a periodic sequence of current pulses is

$$I_{\mathrm{out}}(t) = \frac{I_{\mathrm{out}}(t_0 + nT)}{\frac{I_{\mathrm{out}}(t_0+nT)}{\mathrm{AQ_T}}(t - (t_0 + nT)) + 1},$$

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

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

during the interspike interval, $t_0 + nT < t < t_0 + (n+1)T$, where $\hat{I}_T \equiv \alpha \mathrm{AQ_T}/T$, and $\alpha \equiv (\exp(q_\alpha/\mathrm{Q_T}) - 1)$ is the percentage by which the output current is incremented by each spike [6]. The fixed quantity of charge q_α supplied by each current pulse multiplies the current by $\exp(q_\alpha/\mathrm{Q_T})$, since it takes $\mathrm{Q_T}$ to e-fold. Hence the incremental change in the output current caused by a spike is not fixed; it is proportional to the output current level at the time that the spike occurs. The peak output current levels attained immediately after each spike converge to $\hat{I}_T \equiv \alpha \mathrm{AQ_T}/T$ when $(1+\alpha)^{-n} \ll 1$. Therefore, the equilibrium output current level is proportional to the pulse frequency.

4.4: Adaptive Neuron Circuit

We build an adaptive neuron circuit 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) [29]. (See Figure 4). The high-gain amplifier serves as a thresholding device and the capacitive divider provides hysteresis. In addition, there is a reset current (Ireset) produced by the logic circuit that terminates the spike. Other designs for adaptive neurons are described in [26, 23].

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

$$\mathrm{C_{mem}}\frac{dV_{\mathrm{mem}}}{dt} = I_{\mathrm{in}} - I_{\mathrm{K}} - \mathrm{Q_{th}}\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}), \quad (3)$$

$$\mathrm{Q_T}\frac{dI_{\mathrm{K}}}{dt} = I_{\mathrm{K}}(\mathrm{q_\alpha}\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}) - \frac{1}{\mathrm{A}}I_{\mathrm{K}}), \quad (4)$$

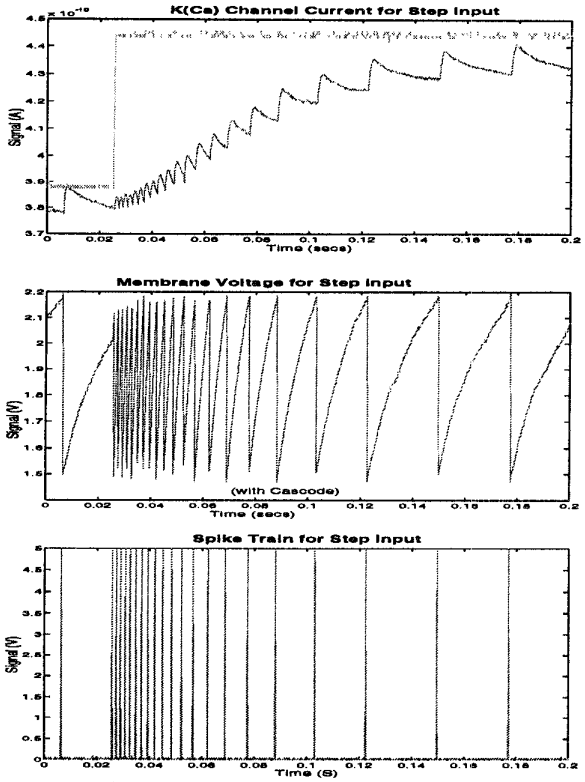

Figure 8. Adaptive neuron's step response. Top: The neuron's input current and the integrator's output current. Middle: Input voltage ramping up between the reset and threshold levels. Bottom: The spike train.

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; $Q_{\text{T}} = C_{\text{Ca}}U_{\text{T}}/\kappa$; 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

$$Q_{\text{th}} = I_{\text{in}}\Delta_n - AQ_{\text{T}} \ln\left(\frac{I_{\text{K}n}}{AQ_{\text{T}}}\Delta_n + 1\right),$$

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 [6]. 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.

When adaptation is complete, the interspike intervals become equal and we have $I_{\text{K}n} = \alpha AQ_{\text{T}}/\Delta_n$. Hence,

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

(remember that $q_\alpha = Q_{\text{T}} \ln(1+\alpha)$). 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 firing-rate adaptation reduces the firing rate by a factor of $\gamma \equiv 1 + Aq_\alpha/Q_{\text{th}}$.

It is preferable to have $I_{\text{K}}(t) < I_{\text{in}}(t)$ for all t, because V_{mem} stays close to the threshold, making the latency shorter and less variable and keeping the integrator's output device in saturation. The circuit operates in this regime if $\gamma < 2/\alpha$ [6]. A tradeoff is imposed by the desire to operate in this regime. If we want a large adaptation attenuation factor γ, we must use a small charge quantum q_α, making the number of spikes required to adapt large. The response of the adaptive neuron circuit to a 14 percent change in its input current is shown in Figure 8; this data also demonstrates the integration of pulse trains by the diode-capacitor integrator and the adaptive step-size.

We need to quantize at the pixel level in order to use adaptive quantization. In the next section, I describe a communication channel that supports asynchronous pixel-level A/D conversion. The design is optimized for activity in the array that is sparse in space but clustered in time.

5: How do we Transmit the Pulses?

The address-event (AE) communication protocol [37, 25, 24] is a random-access, time-division multiplexing (RA-TDM) communication protocol. RA-TDM is an alternative to the more common sequential-access, time-division multiplexing (SA-TDM) protocol. SA-TDM sequentially polls all the users, allocating a fixed fraction of the channel capacity to each user. Hence, its efficiency degrades as the fraction of active users decreases, because bandwidth is tied up by polling inactive users. In contrast, RA-TDM services only the active users, hence the channel capacity is dynamically allocated. However, this enhancement comes at the cost of sending $\log_2 N$-bit addresses to identify one out of N users, instead of just one bit to indicate whether a user is active or not. We also need

to introduce mechanisms that dynamically allocate the channel capacity and, in particular, that arbitrate contention for channel access. I discuss these issues in the next two subsections and introduce a simple activity model to quantify the trade-offs involved. And finally, I describe an implementation for a RA-TDM channel.

5.1: Activity Model

We are given a desired Nyquist sampling rate f_{Nyq}, determined by the bandwidth of the signal to be quantized. We use an adaptive quantizer that samples at f_{Nyq} when the signal is changing, and samples at f_{Nyq}/γ when the signal is not changing. Let the probability that a pixel samples at f_{Nyq} be a, that is a is the fraction of pixels that are active at any time. RA-TDM achieves a sampling rate of f_{N} at the bit rate

$$f_{\text{bits}} = f_{\text{Nyq}}(a + (1-a)/\gamma)\log_2(N-1),$$

per pixel, since a percent of the time it samples at f_{N}; the remaining $(1-a)$ percent of the time it samples at f_{N}/γ; and, each time it samples, $\log_2(N-1)$ bits are sent to encode pixel identity. On the other hand, SA-TDM achieves a sampling rate of f_{Nyq} at the bit rate f_{Nyq} per neuron, since there is no distinction between active and passive neurons and a single bit is read from the pixel. Therefore, RA-TDM will be more efficient if

$$a < \frac{\gamma}{\gamma-1}\left(\frac{1}{\log_2(N-1)} - \frac{1}{\gamma}\right).$$

For example, in a 64×64 array with sampling rate attenuation γ of 40, the active fraction a must be less than 6.1 percent. In a retinomorphic system, the adaptive neuron performs sampling rate attenuation and the spatiotemporal bandpass filter makes the output activity sparse.

Given a certain fixed channel capacity (F_{chan}), in samples per second, we can ask what Nyquist rate f_{Nyq} each channel can achieve. For RA-TDM, channel capacity is allocated dynamically in a ratio $a : (1-a)/\gamma$ between active pixels and passive pixels, hence

$$f_{\text{Nyq}} = f_{\text{chan}}/(a + (1-a)/\gamma)$$

where $f_{\text{chan}} \equiv F_{\text{chan}}/N$ is the capacity per pixel. In contrast, SA-TDM achieves only f_{chan}. For instance, if $f_{\text{chan}} = 100$ Hz, RA-TDM achieves $f_{\text{Nyq}} = 1.36$ KHz with an active fraction of 5 percent and a sampling rate attenuation factor of 40.

5.2: To arbitrate or not to arbitrate?

The original implementation of a RA-TDM protocol included arbitration and queuing mechanisms [37, 25] to allow for graceful degradation of information in the face of the heavy, but sporadic, demands on bandwidth due to synchronous firing triggered by events occurring in the scene. Some temporal dispersion of the burst of spikes is incurred when the channel capacity is exceeded, but this scales linearly with the load. Compared with the exponential increase in information that is lost due to collisions when no arbitration occurs [32]. However, arbitration increases the length of the communication cycle, reducing the channel capacity which is defined as the reciprocal of the cycle time.

This trade-off may be quantified using the following well-known result for the collision probability [40]

$$p_{\text{coll}} = 1 - e^{-2/b},$$

where b is the ratio of the channel capacity to the sampling rate, i. e. $1/b$ is the probability that a spike occurs during a period equal to the communication cycle time. If we arbitrate, we will achieve a certain cycle time, and a corresponding channel capacity, in a given VLSI technology. This channel can operate at 100 percent capacity (i.e., $b = 1$) because the 0.86 collision probability is not a problem—users just wait their turn. When spikes occur in bursts that last for T_{burst}, they are dispersed over an interval no larger than $(\hat{F}_{\text{burst}} T_{\text{burst}}/F_{\text{channel}})$, where $\hat{F}_{\text{burst}}$ is the peak rate reached and F_{chan} is the throughput. Now, we do not arbitrate, we will achieve a shorter cycle time, and a proportionate increase in capacity. Let us assume a factor of 10 improvement, which is optimistic. With $b = 10$, we find $p_{\text{coll}} = 18\%$. Thus, the simple-nonarbitered channel offers more throughput if collision rates higher than 18% are tolerable. For lower collision rates, the complex-arbitered channel offers more throughput, even though its cycle time is an order of magnitude longer. More reasonable failure probabilities of 5 percent require the nonarbitered channel to operate at only 2.5 percent of its capacity.

I address the short-comings of arbitered RA-TDM channels in the implementation described here by introducing pipelining [39]. I achieved additional reductions in the average cycle time by exploiting locality in the arbiter tree and sending spikes from all active pixels in a selected row without redoing the arbitration between rows. This design is also more robust than Sivilotti and Mahowald's original implementation [37, 25], because it makes fewer timing assumptions and uses static state-holding elements instead of dynamic ones. Lazzaro has made different improvements to the original design that make it also more robust [24].

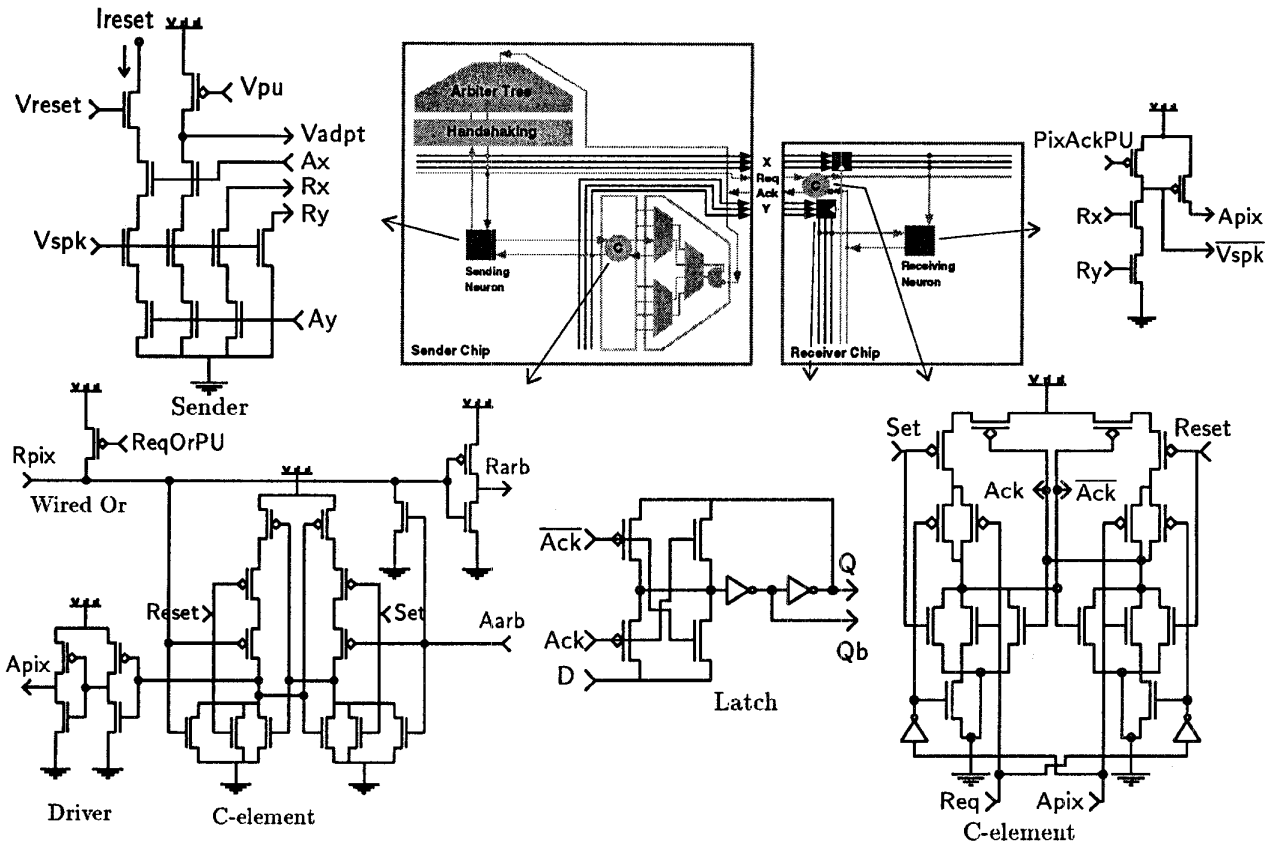

Figure 9. Pipelined address-event interchip-communication channel.

5.3: Circuits

The new design is shown in Figure 9. Like the original design, this interface is completely self-timed, thus every communication must be acknowledged by a feedback signal. These acknowledge signals are also used to implement the queue: you make a pixel wait just by refusing to acknowledge it. At the beginning of a communication cycle, the request and acknowledge signals are both low. The communication cycle proceeds as follows.

On the sender side, spiking neurons first make requests to the Y-arbiter, which selects only one row at a time. It does so in a hierachical fashion, using a decision tree that is built out of two-input arbiter cells [37, 25, 24]. All spiking neurons in the selected row then make requests to the X-arbiter. The address encoders drive the addresses of the selected row and column on to the bus, and Req goes high. When Ack goes high, the select signals are disabled by the AND gate at the top of the arbiter tree and Req is taken low. (The arbiter is described in detail in [25, 24]). The C-elements [39] between the arbiters and the rows/columns will delay this, if necessary, until the pixel has withdrawn its row/column requests, confirming that it has reset.

On the receiver side, as soon as Req goes high, the address bits are latched and Ack goes high immediately. So while the sender chip is deactivating its internal request/select signals, the receiver decodes the addresses and selects the corresponding pixel. When the sender takes Req low, the receiver responds by taking Ack low. The receiver's C-element will delay this, if necessary, until the pixel activates the wired-OR circuit, confirming that it got the spike.

The logic in the sending neuron is shown in the upper-left corner of Figure 9; it is similar to that described in [25, 24]. The neuron takes Vspk high when it spikes, and pulls the row request line Ry low. The column request line Rx is pulled low when the row select line Ay goes high. Finally, Ireset is turned on when the column select line Ax goes high, and the neuron is

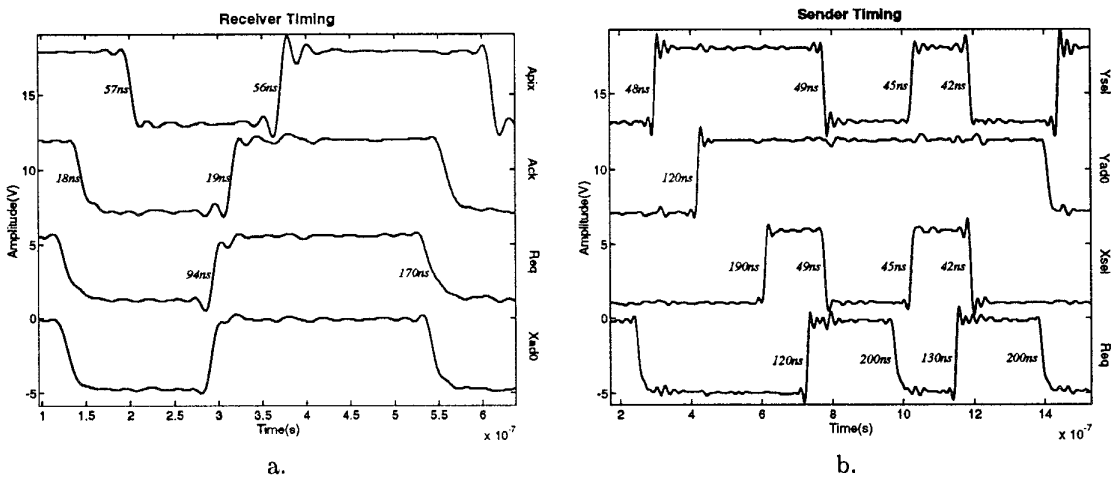

Figure 10. Measured AE channel timing. All the delays given are measured from the preceding transition: (a) Timing of Req and Ack signals relative to X-address bit (Xad0) and receiver pixel's acknowledge (Apix). (b) Timing of Req signal relative to the select signals fed into the top of the arbiter trees (Ysel and Xsel), and the Y-address bit (Yad0). Arbitration occurs in both the Y and X dimensions during the first cycle, but only in the X dimension during the second cycle.

reset. Vadpt is also pulled low to dump some charge on the integrator. A third transistor, driven by Vspk, was added to the reset chain to make the width of the reset pulse independent of the communication-cycle time.

The sender's C-element circuit is shown in the lower-left corner of Figure 9 (it is slightly different form the receiver's). The AND gate at the top of the arbiter tree disables the arbiter's outputs (Aarb) when Ack is high, preventing the arbiter from activating another row/column while Ack is high. There are two differences between this circuit and Lazzaro et.al.'s handshaking circuit [24]. First, Lazzaro et.al. reset all the arbiter's inputs to prevent the arbiter from granting another request while Ack is high. My design leaves the inputs undisturbed, preserving arbiter's state, and thereby reducing the time and power required to arbitrate. This approach also allows all spiking neurons in a selected row to be serviced without redoing the Y-arbitration. Second, Lazzaro et.al. assume that the pixel will withdraw its request before the receiver acknowledges, whereas I make no timing assumptions. Their assumption may not hold if the receiver is pipelined, and when the assumption fails, the row/column select lines may be cleared before the pixel has been reset.

The receiver's C-element is shown in the lower-right corner of Figure 9. This C-element's output signal drives the Ack signal, strobes the latches, and activates the address decoders. The address-bit latch and the logic inside the receiving pixel are shown also shown in the figure (middle of lower row and upper-right corner, respectively). The latch is opaque when Ack is high and transparent when it is low. The pixel logic produces an active low spike whose duration depends on the delay of the wired-OR and the decoder and the duration of the sender's reset phase. Circuits for the blocks that are not described here—namely the arbiter cell, the address encoder, and the address decoder–are given in [25, 24].

Timing measurements for the AE channel are shown in Figure 10. The cycle time is 730 ns if arbitration is performed in both dimensions and 420 ns when arbitration is performed in only the X dimension (i.e., the pixel sent is from the same row as the previous pixel). A timing analysis was performed on this design. The slow steps are propagating high-going selects, Xsel ↑, Ysel ↑, down the six-level arbiter tree, which takes 100ns, and pulling up or down the wired-OR column/row request lines, which takes 120ns.

The output of the postprocessor—after image acquisition, analog preprocessing, quantization, AE encoding, interchip communication, AE decoding, and integration of charge packets in the receiver's diode-capacitor integrators—is shown in Figure 11. The

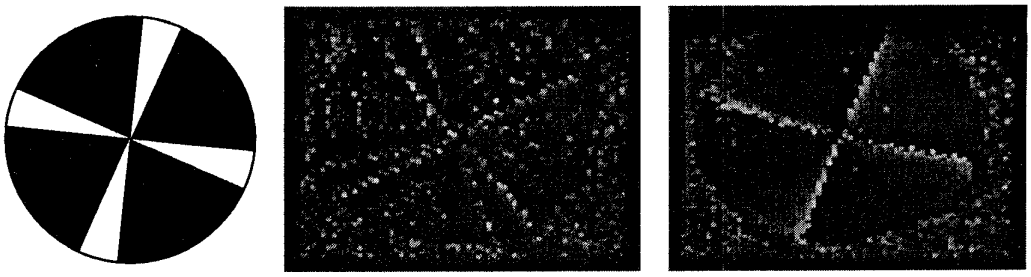

Figure 11. Video frames from AE postprocessor chip showing real-time temporal integration of pulses. The stimulus is a windmill pattern (a) that rotates counterclockwise slowly (b) and quickly (c).

sparseness of the output representation is evident. When the windmill moves, neurons at locations where the intensity is increasing (white region invades black) become active, hence the leading edges of the white vanes are more prominent. These neurons fire more rapidly as the speed increases because the temporal derivative increases. The time-constant of the receiver's diode-capacitor integrator is shorter than that of the sender, so temporal integration occurs only at high spike rates. This mismatch wipes out DC information, and results in an overall high-pass frequency response that enhances the response to motion. The mean spike rate was 30Hz per pixel, and the two-chip system dissipated 190 mW at this spike rate.

6: Any Questions?

I have described an approach to building machine vision systems that exploit regularities in natural scenes to optimize their information gathering capacity. This approach is inspired by the biological retina, and requires sophisticated signal processing at the pixel-level and efficient use of the capacity of the output channel. The retinomorphic approach also offers substantial savings in power dissipation. Specific implementations of all the circuit and system-level functions required were presented, and the design tradeoffs made were described to point the way towards more effective solutions.

7: Acknowledgments

I thank my advisor, Carver Mead, for sharing his insights into the operation of the nervous system. Thanks to Misha Mahowald for making available layouts for the arbiter, the address encoders, and the address decoder, to John Lazzaro, Alain Martin, and Jose Tierno for helpful discussions on address-events and asynchronous VLSI, to Tobi Delbruck for help with the Mac address-event–interface, and to Jeff Dickson for help with PCBs design. This work was funded by ONR and ARPA and I am supported by a Sloan fellowship.

References

[1] A. Andreou and K. Boahen. A 48,000 pixel, 590,000 transistor silicon retina in current-mode subthreshold cmos. In *Proc. 37th Midwest Symposium on Circ. and Sys.*, pages 97–102, Lafayette, Louisiana, 1994.

[2] A. Andreou and K. Boahen. Translinear circuits in subthreshold mos. *J. Analog Integrated Circ. Sig. Proc.*, March 1996.

[3] A. G. Andreou. Low power analog vlsi systems for sensory information processing. In B. Sheu, E. Sanchez-Sinencio, and M. Ismail, editors, *Microsystems technologies for multimedia applications*. IEEE Press, Los Alamitos CA, 1995.

[4] J. Atick and N. Redlich. What does the retina know about natural scene. *Neural Computation*, 4(2):196–210, 1992.

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

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

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

[9] J. Buhman, M. Lades, and E. 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.

[10] K. Bult and G. J. Geelen. An inherently linear and compact most-only current division technique. *IEEE J. Solid-State Circ.*, 27(12):1730–1735, 1992.

[11] 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.

[12] T. Delbruck and C. Mead. Photoreceptor circuit with wide dynamic range. In *Proceedings of the International Circuits and Systems Meeting*, IEEE Circuits and Systems Society, London, England, 1994.

[13] A. Dickinson, B. Ackland, E. El-Sayed, D. Inglis, and E. R. Fossum. Standard cmos active pixel image sensors for multimedia applications. In W. Dally, editor, *Proceedings of the 16th Conference on Advanced Research in VLSI*, pages 214–224, Chapel Hill, North Carolina, 1995. IEEE Press, Los Alamitos CA.

[14] D. Dong and J. Atick. Statistics of natural time-varying scenes. *Network: Computation in Neural Systems*, 6(3):345–358, 1995.

[15] D. Dong and J. Atick. Temporal decorrelation - a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178, 1995.

[16] C. Enroth-Cugell, J. G. Robson, D. E. Schweitzer-Tong, and A. B. Watson. Spatiotemporal interactions in cat retinal ganglion cells showing linear spatial summation. *J. Physiol.*, 341:279–307, 1983.

[17] B. Fowler, A. E. Gamal, and D. Yang. A cmos area image sensor with pixel-level a/d conversion. In J. H. Wuorinen, editor, *Digest of Technical Papers*, volume 37 of *IEEE International Solid-State Circuits Conference*, pages 226–227, San Francisco, California, 1994.

[18] K. Fujikawa, I. Hirota, H. Mori, T. Matsuda, M. Sato, Y. Takamura, S. Kitayama, and J. Suzuki. A 1/3 inch 630k-pixel it-ccd image sensor with multi-function capability. In J. H. Wuorinen, editor, *Digest of Technical Papers*, volume 38 of *IEEE International Solid-State Circuits Conference*, pages 218–219, San Francisco, California, 1995.

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

[20] B. Hoeneisen and C. A. Mead. Fundamental limitations in microelectronics-i: Mos technology. *IEEE J. Solid-State Circ.*, 15:819–829, 1972.

[21] D. H. Kelly. Motion and vision ii. stabilized spatiotemporal threshold surface. *J. Opt. Soc. Am.*, 69(10):1340–1349, 1979.

[22] S. B. Laughlin. *Matching Coding, Circuits, Cells, and Molecules to Signals: General principles of Retinal Design in the Fly's Eye*, volume 13(1) of *Progress in Retinal and Eye Research*, chapter 7, pages 165–196. Pergamon Press, Oxford, England, 1994.

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

[24] J. Lazzaro, J. Wawrzynek, M. Mahowald, M. Sivilotti, and D. Gillespie. Silicon auditory processors as computer peripherals. *IEEE Trans. on Neural Networks*, 4(3):523–528, 1993.

[25] M. Mahowald. *An Analog VLSI Stereoscopic Vision System*. Kluwer Academic Pub., Boston, MA, 1994.

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

[27] M. Mahowald and C. Mead. The silicon retina. *Scientific American*, 264(5):76–82, 1991.

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

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

[30] C. A. Mead. Neuromorphic electronic systems. *Proc. IEEE*, 78(10):1629–1636, 1990.

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

[32] A. Mortara, E. Vittoz, and P. Venier. A communication scheme for analog vlsi perceptive systems. *IEEE Trans. Solid-State Circ.*, 30(6):660–669, 1995.

[33] M. Ogata, T. Nakamura, K. Matsumoto, R. Ohta, and R. Hyuga. A smart pixel cmd image sensor. *IEEE Trans. Electron Dev.*, 38(5):1005–1010, 1991.

[34] 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.

[35] W. A. Richards. A lightness scale for image intensity. *Appl. Opt.*, 21:2569–2582, 1982.

[36] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. Univ. Illinois Press, Urbana IL, 1949.

[37] M. Sivilotti. *Wiring considerations in Analog VLSI Systems, with application to Field-Programmable Networks*. PhD thesis, California Institute of Technology, Pasadena CA, 1991.

[38] 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.

[39] I. E. Sutherland. Micropipelines. *Communications of the ACM*, 32(6):720–738, 1989.

[40] A. S. Tanenbaum. *Computer Networks*. Prentice-Hall International, 2 edition, 1989.

[41] J. H. van Hateren. A theory of maximizing sensory information. *Biol. Cybern.*, 68:23–29, 1992.

[42] W. B. Veldkamp. Wireless focal planes: On the road to amacronic sensors. *IEEE J. Quantum Electronics*, 29(2):801–813, 1993.

[43] E. Vittoz. Analog vlsi signal processing: why, where, and how. *J. Analog Integrated Circ. Sig. Proc.*, 6:27–44, 1994.

[44] E. Vittoz and X. Arreguit. Linear networks based on transistors. *Electronics Letters*, 29:297–299, 1993.