Optic Nerve Signals in a Neuromorphic Chip II: Testing and Results

Kareem A. Zaghloul, *Member, IEEE,* and Kwabena Boahen*

***Abstract*—Seeking to match the brain's computational efficiency [14], we draw inspiration from its neural circuits. To model the four main output (ganglion) cell types found in the retina, we morphed outer and inner retina circuits into a 96×60-photoreceptor, $3.5 \times 3.3\ \mathrm{mm}^2$, $0.35\ \mu\mathrm{m}$-CMOS chip. Our retinomorphic chip produces spike trains for 3600 ganglion cells (GCs), and consumes 62.7 mW at 45 spikes/s/GC. This chip, which is the first silicon retina to successfully model inner retina circuitry, approaches the spatial density of the retina. We present experimental measurements showing that the chip's subthreshold current-mode circuits realize luminance adaptation, bandpass spatiotemporal filtering, temporal adaptation and contrast gain control. The four different GC outputs produced by our chip encode light onset or offset in a sustained or transient fashion, producing a quadrature-like representation. The retinomorphic chip's circuit design is described in a companion paper [Zaghloul and Boahen (2004)].**

***Index Terms*—Adaptive circuits, neural systems, neuromorphic engineering, prosthetics, vision.**

I. Retinomorphic Systems

THE RETINA, one of the best studied neural systems, is a complex piece of biological wetware designed to optimally signal the onset or offset of visual stimuli in a sustained or transient fashion [19]. To encode these signals into spike patterns for transmission to higher processing centers, the retina has evolved intricate neuronal circuits that capture information contained within natural scenes efficiently [21]. This visual preprocessing, realized by the retina, occurs in two stages, the outer and inner retina. Each local retinal microcircuit plays a specific role in the retina's function, and neurophysiologists have extracted a wealth of data characterizing how its constituent cell types contribute to visual processing. These physiological functions can be replicated in artificial systems by emulating their underlying synaptic interactions.

In the companion paper [23], we introduced our model for neurocircuits in the outer and inner retina. In our outer retina model, interaction between cone and horizontal cell (HC) networks creates a bandpass spatiotemporal response that exhibits luminance adaptation and that produces a contrast signal at the cone terminal. Bipolar cells (BCs) in our model rectify these signals into ON and OFF channels to replicate the retina's complementary signaling. In our inner retina model, modulation of narrow-field amacrine cell (NA) presynaptic inhibition by wide-field amacrine cells (WAs) realizes contrast gain control and time-constant adaptation, which allows our model to optimally encode signals by adjusting its temporal filter based on input frequency and contrast.

In this paper, we present data characterizing how our outer and inner retina models [23], implemented in silicon, process visual information. Because we want to demonstrate that our circuits are viable models for retinal processing, our experimental protocols are similar to those used in earlier — now classic — physiological studies. We present similar visual stimuli to our retinomorphic chip and record spike outputs from our array. And because we have control over a number of parameters in our model and circuit, we adjust these parameters to explore the system's range of behavior. Such a characterization is useful in helping us fine tune processing in our retinomorphic chip to match outputs from the mammalian retina and to better quantify our chip's outputs for robotic or prosthetic applications.

The remainder of this paper is organized as follows. In Section II, we describe our chip's architecture, which is based on the retina's functional architecture and replicates signal convergence in its pathways. In Section III, we present typical outputs from our ganglion cell (GC) array to demonstrate how these outputs are structured. In Section IV, we characterize our outer retina model's ability to adapt to mean light intensity by presenting a contrast reversing grating at different mean luminances and exploring how different circuit parameters affect the retinomorphic chip's GC responses. In Section V, we explore how our outer and inner retina models filter input signals in space and time, and determine how this filtering changes as we adjust different properties of the circuit. Finally, in Section VI, we demonstrate that our inner retina model realizes contrast gain control and compare the chip's GC responses with our analytical results. Section VII concludes the paper.

II. Chip Architecture

In the mammalian retina, cone signals converge on to BCs [19], which makes the receptive field center Gaussian-like [18]. To implement signal convergence in our model, chip BCs connect the outputs from a central phototransistor and its six nearest neighbors (hexagonally tiled) to one inner retina circuit, as shown in Fig. 1(a). Each of our outer retina circuits actually produces two output currents. A central photoreceptor drives

Manuscript received September 6, 2002; revised June 25, 2003 This work was supported in part by the National Institutes of Health (NIH) under Vision Training Grant T32-EY07035 and in part by the Whitaker Foundation under Grant 37005-00-00. The work of K. A. Zaghloul was completed while he was with the Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104 USA. *Asterisk indicates corresponding author.*

K. A. Zaghloul is with the Department of Neurosurgery, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: kazaghloul@uphs.upenn.edu).

*K. Boahen is with the Department of Bioengineering, 120 Hayden Hall, 3320 Smith Walk, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: boahen@seas.upenn.edu).

Digital Object Identifier 10.1109/TBME.2003.821040

0018-9294/04$20.00 © 2004 IEEE

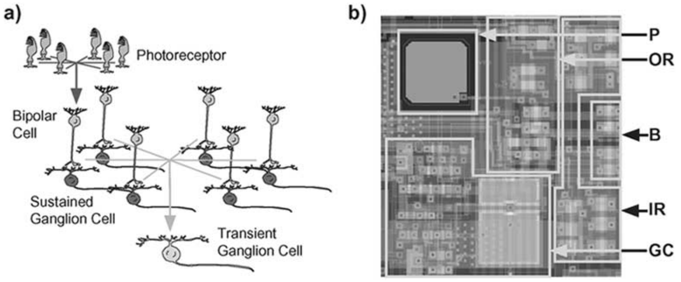


Fig. 1. Chip architecture and layout. (a) Signal convergence: Signals from a photoreceptor (not shown) and its six neighbors are pooled to provide synaptic input to each BC. Each BC generates a rectified output, either ON or OFF, that drives a local inner retina circuit. Sustained-type GCs receive input from a single local inner retina circuit. Signals from a central inner retina circuit (not shown) and its six neighbors are pooled to drive each transient-type GC. (b) Pixel layout: Each pixel, containing 38 transistors on average, has a photoreceptor (P), outer retina (OR) circuitry, BCs, and inner retina (IR) circuitry. Spike-generating GCs are found in five out of eight pixels; the remaining three contain a NA cell membrane capacitor. Each pixel is 34 × 40 $\mu\mathrm{m}^2$ in a 0.35 $\mu\mathrm{m}$ process.

the BC with both of these outputs while photoreceptors at the six vertices divide these outputs between their two nearest BCs. For symmetry, we implement a similar architecture for the reference current (see [23]).

Transient GCs in the mammalian retina pool their inputs from a larger region than sustained GCs. We maintain the same architecture in our chip. We pool inner retina signals in the same way that we pool outer retina signals, but inner retina circuits are tiled at one-quarter the density of the phototransistors that provide their input. Thus, each transient GC receives input from a central inner retina circuit and its six nearest neighbors, as shown in Fig. 1(a). This central circuit excites its GC with both copies of its transient output whereas those at the six vertices divide their outputs between their two nearest transient GCs. Thus, all GCs tile hexagonally, but the transient ones tile more sparsely than the sustained ones. This architecture gives transient GCs a larger receptive field, as found in cat Y-GCs. This pooling accounts for transient GCs' nonlinear subunits [9] since all the rectified BC signals can never sum to zero, at any one moment, in response to a sinusoidal grating. Conversely, if the cells were linear, a contrast-reversing grating exactly centered over the cell's receptive field would not modulate the cell's response since dark regions would exactly cancel out light regions.

The layout of one pixel is shown in Fig. 1(b). It contains a phototransistor, outer retina circuitry, BCs, and one-quarter of the inner retina circuit. Hence, 2×2 and 4×4 adjacent pixels are needed to generate a complementary pair of sustained and transient outputs, respectively. Because transient GCs occur at a quarter resolution, not every pixel contains GC circuitry. Three out of every eight pixels instead contain a large capacitor that gives the NA its long time-constant. A spike generating circuit in the remaining five pixels converts GC inputs into spikes that are sent off chip.

Because analog signals cannot be relayed over long distances, retinal GCs use spikes to communicate with higher cortical structures. Similarly, each GC in the chip array converts the current it receives from the inner retina circuit to spikes, as shown in Fig. 2(a). Our spike generating circuit exhibits spike-rate adaptation through Ca^{++} activated K^{+} channel analogs, modeled with a current-mirror integrator [3].

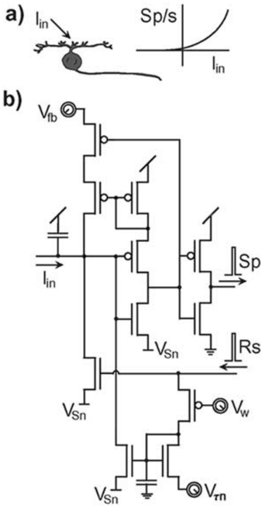


Fig. 2. Spike generation. (a) Input current to a retinal GC produces a spike that is conveyed down the optic nerve. Spike rate is a function of input current. (b) A CMOS circuit that transforms input current to spikes. I_{in} from the inner retina charges up a GC membrane capacitor. When the membrane voltage crosses threshold, the circuit produces a spike (Sp) that is relayed off chip by digital circuitry. This circuitry acknowledges receipt of the spike by sending a reset pulse (Rst) that discharges the membrane and dumps charge on a current-mirror-integrator that implements Ca^{++} spike-rate adaptation.

The CMOS circuit that transforms current into spikes is shown in Fig. 2(b). Briefly, input current charges up a GC membrane capacitor. As the membrane voltage approaches threshold, a positive feedback loop, modulated by V_{fb}, accelerates the voltage's rate of change [5]. Once threshold is passed, the circuit generates a pulse (or spike) that is relayed to digital circuitry. The digital circuitry acknowledges receipt of the spike by sending a reset pulse which discharges the membrane. The reset pulse, RST, also dumps a quanta of charge on to a current-mirror integrator through a pMOS transistor gated by V_{w}. Charge accumulating on the integrator models the build-up of Ca^{++} within the cell after it spikes. This charge, which leaks away with a time constant determined by $V_{\tau\mathrm{n}}$, draws current away from the membrane potential, modeling Ca^{++} mediated K^{+} channels. The source voltage for the neuron circuit, V_{Sn}, is set to be the same as the source voltage for the inner retina circuit, V_{S} [23].

Due to wiring limitations, we can not communicate each GC output off chip directly. Instead, we use an asynchronous, arbitered, multiplexer to read spikes out from the neurons [4]. Each GC interfaces with digital circuitry that communicates the spikes to an arbiter at the end of each row and each column of neurons, as shown in Fig. 3. The arbiter multiplexes spikes and outputs the location, or address, of each spiking neuron as they occur. Row and column addresses for each GC are communicated serially off chip. We can represent the spike activity of all 3600 GCs with just seven bits using this address-event representation. By noting the address of each event generated by the chip, we can decode GC type and location in the array.

III. CHIP OUTPUT

We designed and fabricated a 96×60 photoreceptor $3.5 \times 3.3\ \mathrm{mm}^2$ chip in 0.35 $\mu\mathrm{m}$ CMOS technology. Chip phototransistors are roughly square with a width of 10 $\mu\mathrm{m}$

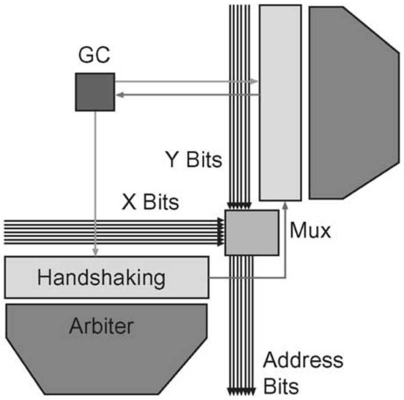


Fig. 3. Spike arbitration. A GC communicates spikes to peripheral digital handshaking and arbitration circuitry using row and column request lines. The arbiter chooses between spikes by selecting a row and column, encodes each incoming spike into a pair of seven-bit addresses, and communicates these addresses off chip. Row and column addresses are sent serially on the same address bus; a multiplexer toggles between the row and column encoders. The handshaking circuits relay reset signals back to the spiking GC.

and were tiled triangularly every 40 $\mu\mathrm{m}$ [23]. Our silicon chip generates spike train outputs for $2 \times 48 \times 30$ sustained-type and $2 \times 24 \times 15$ transient-type GCs in both ON and OFF channels, producing a total of 3600 spike outputs [23]. The chip's light response to a drifting vertical sinusoidal grating is shown in Fig. 4(a).

Spike trains from identical GCs in a single column of the chip array differ significantly due to variability between nominally identical transistors (e.g., ON transient GC spike rate $\mathrm{CV}\ (\mathrm{coefficient\ of\ variation}) = 57\%$, $\mathrm{ON\ sustained\ GC} = 162\%$). To further quantify this variability, we measured the mean firing rates for all GCs in the array in response to a 50% contrast 7.5 Hz 0.1096 cyc/deg drifting sinusoid. The distribution of these firing rates is shown in Fig. 4(b). From the figure, we see a Gaussian-like distribution of firing rates, when plotted on a log scale. Most cells fall within this log normal distribution, but there are some outliers, especially at low firing rates. This variability is most likely due to normally distributed threshold mismatch between transistors, which translates to log-normally distributed currents in subthreshold operation.

Despite this heterogeneity, we were able to get a robust measure of GC activity by summing responses from all GCs in a given column — much as physiologists average several trials from the same cell — and analyzing the spike histogram. The histograms shown in Fig. 4(a) demonstrate phase differences between the four GC types [23]: complementary ON and OFF channels respond out of phase with one another while transient cells lead sustained cells, exhibiting both earlier onset and shorter duration of firing.

The chip actually produced useful images despite all this variability, as illustrated by presenting it a face. As shown in Fig. 4(c) *top*, edges are enhanced by sustained GC activity in the static image. To confirm that the chip captures the visual information, we reconstructed the natural stimulus from the sustained GC spike activity (see Fig. 4(c) *bottom*). We did this by convolving ON and OFF sustained GC spike output with a simple difference-of-Gaussian model, whose excitatory and inhibitory standard deviations were determined by fitting GC spatial frequency responses with a one-dimensional difference-of-Gaussian. The ratio of excitatory to inhibitory standard deviations was 0.15. We matched temporal filtering as well by convolving with a temporal low-pass filter with a time constant of 22.7 ms, computing a new frame every 20 ms. We took the difference between images obtained from ON and OFF spikes and displayed it on a gray-scale, with ON and OFF activity corresponding to bright and dark pixels, respectively. Activity from transient GCs did not enhance the resolution of the reconstructed image and was not included. Our reconstruction produces an image that is easily recognizable, even with only 30×48 pixels and just 0.4 spikes/cell/frame, suggesting that cortical structures, as well as visual prostheses or robots, can extract useful visual information from the retina's neural code through simple linear filtering.

IV. LUMINANCE ADAPTATION

Our outer retina circuit's nonlinear behavior was designed to generate cone terminal (CT) signals that are entirely proportional to contrast. Because of this local automatic gain control, we expected the circuit's output current to be independent of luminance. The mammalian retina exhibits this behavior, where GC responses depend on contrast but not on luminance [20], and we hypothesize that this luminance adaptation takes place in the outer retina.

To characterize luminance adaptation, we measured GC responses to a 7.5 Hz drifting sinusoid (50% contrast), of various spatial frequencies, as we changed mean intensity. We found that intensity had a slight effect on the highest sensitivity achieved; this effect was larger for ON cells. OFF transient (OffT) GC peak responses, which were 408 sp/s, 432 sp/s and 218 sp/s as we decreased mean intensity from 196 $\mathrm{cd/m^2}$ to 33 $\mathrm{cd/m^2}$ to 3.3 $\mathrm{cd/m^2}$, are relatively unchanged until mean intensity drops to 3.3 $\mathrm{cd/m^2}$. For OnT GCs, the peak response dropped from 771 sp/s to 485 sp/s to 117 sp/s as we decreased mean intensity from 196 $\mathrm{cd/m^2}$ to 33 $\mathrm{cd/m^2}$ to 3.3 $\mathrm{cd/m^2}$. Similarly, OFF sustained (OffS) GC peak responses were 504 sp/s, 520 sp/s and 346 sp/s while OnS GC peak responses were 425 sp/s, 225 sp/s and 63 sp/s at these three mean intensities, respectively.

The overall reduction in sensitivity for both channels most likely arises from stray photocurrents interfering with GC spike-rate adaptation. Mean spike activity is affected by these leakage currents, which determine the spike-rate adaptation time-constant. As intensity drops and, therefore, as these photocurrents decrease, this time-constant increases, causing a drop in quiescent spike rate and overall sensitivity. The asymmetry between ON and OFF pathways is explained by mismatch between drain voltages in the outer retina circuit, which distorts the rectification, causing ON sensitivity to decrease more than OFF sensitivity.

Stray photocurrents hampers our outer retina circuit's ability to adapt to luminances through their effect on sensitivity. However, we believe that our design will work if these are eliminated. To prove this, we compensate for the change in sensitivity by manually increasing lateral current spread in the HC network

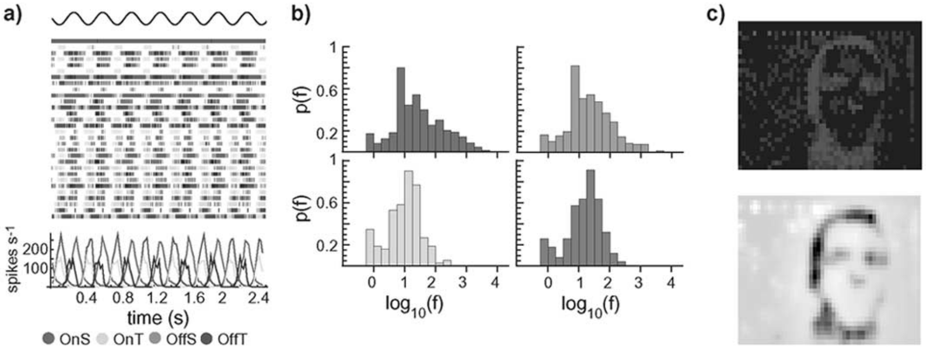


Fig. 4. Chip Output. (a) A raster plot of the spikes (top) and histogram (bottom, $\text{bin width} = 20$ ms) recorded from a single column of the chip array. The stimulus was a 3 Hz 50%-contrast drifting sinusoidal grating (0.14 cyc/deg) whose luminance varied horizontally across the screen and was constant in the vertical direction. We use a 50% contrast stimulus in all responses presented here unless otherwise noted. The four GC type outputs are color coded (see legend). We computed the amplitude of the fundamental Fourier component of these histograms, which is plotted in all frequency responses presented here, unless otherwise noted. (b) The distribution of firing rates for the four types of GC outputs, demonstrating the amount of variability in our chip. Log of the firing rate, f, is plotted on the abscissa, and the probability density, $p(f)$, is plotted on the ordinate. Histograms are computed from all active cells of a given type in the array (151 out of 360 OnT, 202/360 OffT, 890/1440 OnS, and 792/1440 OffS cells in the array exhibited no activity). (c) In response to a static natural image, edges are enhanced by sustained type GC outputs of the chip (top). Reconstruction of the image from sustained GC activity (bottom) demonstrates fidelity of retinal encoding despite the variability.

(i.e., decreased V_{hh} [23]), which boosts CT activity [23]. Because $l_h \propto e^{-\kappa V_{\mathrm{hh}}/2\mathrm{U_T}}$, decreasing V_{hh} causes an increase in l_h. When the space constants of the cone and HC networks, l_c and l_h, are similar, increasing l_h increases CT sensitivity, but the effect saturates when $l_h \gg l_c$.

As we did not know the exact value of the subthreshold slope coefficient, κ, we determined how much we should change V_{hh} to compensate for the light-related change in spike-rate adaptation empirically by recording the GC response to a 7.5 Hz 0.1096 cyc/deg grating. We adjusted V_{hh} to keep this response fixed at different mean intensities. Because we measured the GC response, while adjusting V_{hh}, at only one spatiotemporal frequency — and not across the entire spatiotemporal spectrum — this technique is only approximate, as it does not account for shifts in the peak. For every decade reduction in photocurrent, we had to decrease V_{hh} by 85 mV to maintain the same response at this spatiotemporal frequency.

In addition to decreasing V_{hh} with mean intensity, we also increased NA cell membrane leakage, I_τ, to compensate for smaller stray photocurrents at lower light intensities [23]. The inner retina's open loop time-constant, τ_{na}, is governed by the size of the NA cell capacitor, and by I_τ, which drains these capacitors. To conserve space, we restricted the size of our NA capacitors to 1 pF each. This leaves little room for the magnitude of I_τ. In testing the chip, we found that we had to set V_τ, the gate voltage of the transistor that produces I_τ, at 50 mV to attain reasonable responses. However, this makes I_τ susceptible to stray leakage currents generated in the substrate by incident photons. Therefore, in addition to decreasing V_{hh} to compensate for nonlinearities [23], we also had to increase V_τ to maintain the same level of I_τ at lower light levels.

Incorporating the changes in V_{hh} and I_τ to maintain response sensitivity, we verified that our outer retina circuit adapts to mean luminance and encodes stimulus contrast. We recorded the output of chip ON transient GCs [23] to a 0.22 cyc/deg 3 Hz reversing grating whose contrast varied between 3.25 and 50% at four different mean luminances. The data is shown in Fig. 5. Chip responses maintain contrast sensitivity over at least one and a half decades of mean luminance (our experimental setup was limited to 200 $\mathrm{cd/m^2}$). To verify that our outer retina model accounts for the behavior shown in Fig. 5, we fit the chip's responses to different contrasts with [23, Equation 4], allowing only the ratio l_h/l_c to increase as intensity decreased. We found that as intensity decreased from 192 $\mathrm{cd/m^2}$ to 19 $\mathrm{cd/m^2}$, for example, the best fits for the data were attained when l_h/l_c increased by 6% (from 0.4617 to 0.4893).

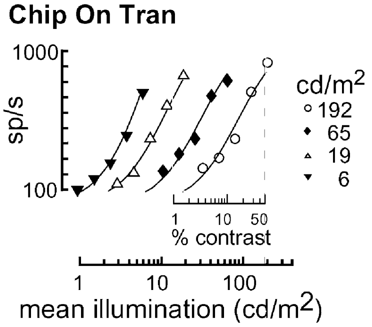


Fig. 5. Luminance adaptation. Chip ON-transient cell responses to a sinusoidal grating (0.22 cyc/deg) whose contrast varied between 3.25% and 50% and reversed at 3 Hz, for four different mean luminances. Response versus contrast (small x-axis) curves are shifted horizontally according to mean luminance (large x-axis) such that the 50% contrast response is aligned with that particular mean luminance. Solid lines represent the best fits of [23, Equation 4] to the data, where we only allowed the ratio of horizontal space constant to cone space constant, l_h/l_c, to vary for different intensities. Values for l_h/l_c were 0.4617, 0.4668, 0.4893, and 0.6932 as intensity dropped from 192 $\mathrm{cd/m^2}$ to 6 $\mathrm{cd/m^2}$.

To capture this data, we had reduced V_{hh} by 85 mV for this tenfold reduction in intensity. This change in V_{hh} would correspond to a three-fold increase in l_h, if we only consider l_h's dependence on V_{hh}. However, l_h is also proportional to $e^{(1-\kappa)V_h/2\mathrm{U_T}}$, as seen in our outer retina analysis [23]. Because the HC's leakage current, I_h ($\propto e^{\kappa V_{\mathrm{hh}}/\mathrm{U_T}}$), decreases more slowly than the lateral current, I_{hh} ($\propto e^{V_h}$) because $\kappa < 1$ [23]. A decrease in light intensity, thus, results in a *smaller* HC space constant, l_h, which is offset by the increase that results from manually decreasing V_{hh}. Indeed, our fits in Fig. 5 suggest that

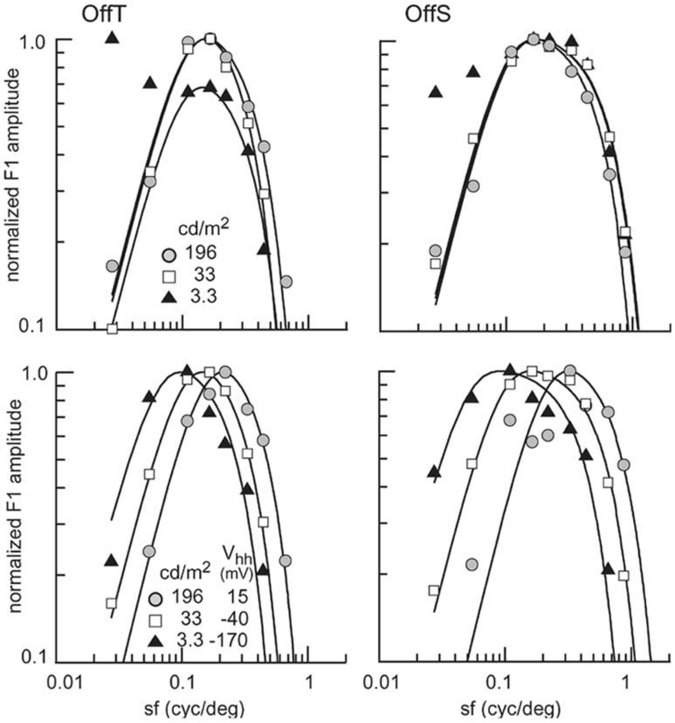


Fig. 6. Spatial filtering at different mean intensities. Responses of chip OffS and OffT cells to 7.5-Hz horizontally drifting sinusoids (50% contrast) with different spatial frequencies. Normalized responses recorded at three different mean luminances *without* changing $V_{\rm hh}$ to compensate for the outer retina nonideality are shown on top. Solid lines are the best fit of a balanced difference-of-Gaussian model (OffT: $\sigma_{\rm Exc}/\sigma_{\rm Inh} = 0.21$, 0.27, and 0.21 for mean luminances of 196, 33, and 3.3 $\rm cd/m^2$ respectively; OffS: $\sigma_{\rm Exc}/\sigma_{\rm Inh} = 0.14$, 0.11, and 0.11). We ignored the lowest two spatial frequencies for the purposes of our fits. Normalized responses recorded while changing both mean luminance and $V_{\rm hh}$ are shown on bottom. $V_{\rm hh}$ values are in units of mV. Solid lines are the best fit of a balanced difference-of-Gaussian model (OffT: $\sigma_{\rm Exc}/\sigma_{\rm Inh} = 0.26, 0.23$, and 0.18 for mean luminances of 196, 33, and 3.3 $\rm cd/m^2$ respectively; OffS: $\sigma_{\rm Exc}/\sigma_{\rm Inh} = 0.19, 0.11$, and 0.09).

decreasing light intensity while manually decreasing $V_{\rm hh}$ reults in only a small increase in the HC space constant, l_h.

V. Spatiotemporal Filtering

Our outer retina model is designed to bandpass filter signals in both space and time, as described in the accompanying paper [23]. We found the spatial frequency at which sensitivity peaked to be essentially independent of intensity, as expected, for both OffS and OffT GC responses [Fig. 6 (top)]. Compensating for the light dependent spike-rate adaptation time-constant by decreasing $V_{\rm hh}$, however, has the effect of expanding the receptive field size, or lowering the peak spatial frequency, ρ_A, similar to the expansion observed in mammalian retina at lower light intensities [12].

In Fig. 6 (bottom), we measured GC spatial responses at different mean intensities while compensating for sensitivity by changing $V_{\rm hh}$, as described above. Since ρ_A is inversely proportional to $\sqrt{l_c l_h}$, its dependence on $V_{\rm hh}$ is described by $\rho_A \propto e^{\kappa V_{\rm hh}/4}$. For $\kappa = 0.7$, decreasing $V_{\rm hh}$ from 15 mV to -40 mV, for example, should cause a 28% reduction in ρ_A, ignoring the negligible change due to intensity. We found that the peak spatial frequency of the transient OFF GC in fact decreased by 25% (from 0.2192 to 0.1644 cyc/deg). The change in spatial profile for both OFF transient and sustained GC responses for further reductions in $V_{\rm hh}$ at different mean intensities is shown in Fig. 6 (bottom). As expected, the peak spatial frequency continues to decrease, further expanding the GCs receptive field.

From Fig. 6 (top), we also find that spatial filtering in ON and OFF pathways is not identical. OffT peak response lies at 0.1644 cyc/deg whereas the corresponding OnT peak response lies at 0.1096 cyc/deg (not shown). Similarly, OffS begins to roll off at 0.3288 cyc/deg while OnS begins its rolloff at 0.2192 cyc/deg (not shown). This implies that the OFF channel has a smaller effective space constant than the ON channel. Both channels, however, are driven by the same outer retina circuitry, so this difference most likely arises from asymmetric rectification in the bipolar circuit [23].

We believe that saturation in the ON channel is responsible for the asymmetric spatial filtering. Currents diverted to the ON channel in the bipolar circuit saturate, whereas currents in the OFF channel do not. Both ON and OFF GCs inherit a Mexican-hat-like receptive field — a narrow excitatory center and a broader inhibitory surround — from the outer retina. The width of this Mexican hat determines the system's peak spatial frequency, ρ_A. Saturation flattens the ON channel's excitatory center which leads to a relative increase in its width. This leads to a decrease in the ON channel's corner spatial frequency, ρ_A, which is what we observe in the data.

We expect sustained GC to be spatially bandpass for all temporal frequencies. Sustained GCs' responses represent an all-pass version of BC signals and are, therefore, dominated by the outer retina's temporal low-pass filter [23]. To verify that we realized invariant spatial bandpass filtering, we measured sustained GC activity in response to a horizontally drifting 50% contrast vertical sinusoidal grating at different temporal and spatial frequencies. The spatiotemporal profiles of sustained GC responses, which reflect activity at the bipolar terminal [23], is shown in Fig. 7. From the figure, we see that sustained GCs' bandpass spatial filtering is largely invariant with temporal frequency.

In contrast, theoretical studies of the outer retina reveal a transition to low-pass spatial filtering at high temporal frequencies [3]. This transformation occurs because HC inhibition is ineffective at high temporal frequencies because of its long time-constant. However, WA cells, which respond much faster (in our chip, as well as in the retina), suppress low spatial frequencies at the bipolar terminal. This suppression could account for the attenuated response at low-spatial-high-temporal frequencies seen in chip sustained GCs and in the mammalian retina [11]. In fact, chip sustained GCs capture the overall suppression seen in mammalian cells at low spatial frequencies. Because of this suppression, spatial tuning remains bandpass at all but the highest temporal frequencies — except for a resonance at very high temporal frequencies seen in the cat data [11].

On the other hand, we expect transient GC responses to be temporally bandpass for all spatial frequencies. To verify that we realized this behavior, we measured GC activity in response to the same horizontally drifting 50% contrast sinusoidal grating. The spatiotemporal profiles of transient GC responses are shown in Fig. 8. From the figure, we see that transient GCs exhibit bandpass temporal filtering at all spatial frequencies.

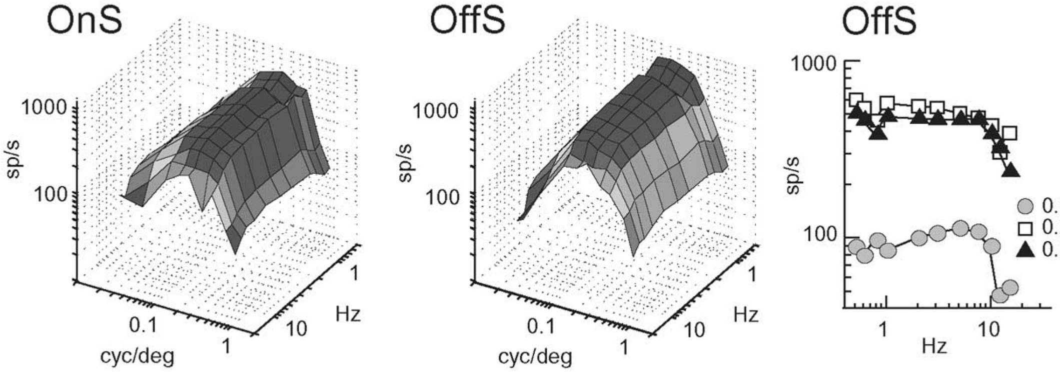


Fig. 7. Sustained cell spatiotemporal response. Three dimensional plots of chip sustained GC responses, which reflect activity at the bipolar terminal, to horizontally drifting sinusoids of different spatial and temporal frequencies. OFF sustained GC temporal frequency responses are shown on the right, in two dimensions, for three different spatial frequencies. Chip sustained cells are bandpass in space for all temproal frequencies.

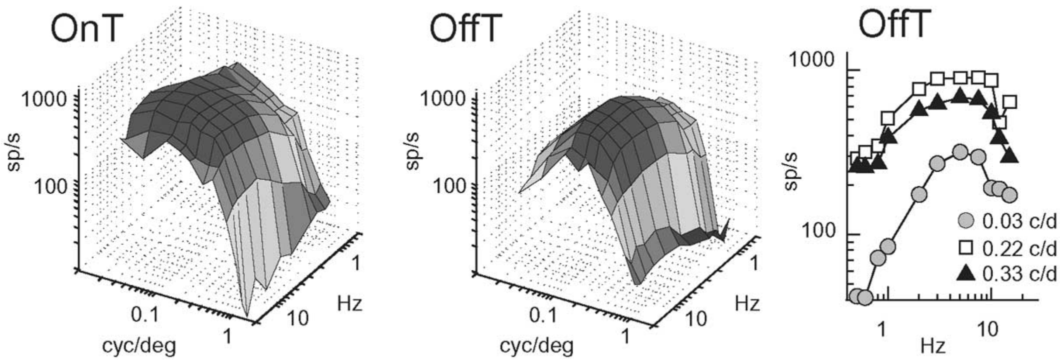


Fig. 8. Transient cell spatiotemporal response. Responses of chip transient GCs, which incorporate additional processing in the inner retina, to horizontally drifting sinusoids of different spatial and temporal frequencies. ON transient GC temporal frequency responses are shown on the right, in two dimensions, for three different spatial frequencies. Chip transient cells are bandpass in time for all spatial frequencies. They also spatially bandpass filter input signals at all temporal frequencies.

Unlike their sustained counterparts [23], which are largely low-pass temporally. In addition, transient cells' peak spatial frequency is lower than the corresponding sustained cells due to their larger receptive fields.

In contrast, theoretical studies of the outer retina reveal a transition to low-pass temporal filtering at high spatial frequencies [3]. This transformation occurs because HC inhibition is also ineffective at high spatial frequencies because most of their excitatory input is lost to neighboring HCs through gap-junctions [3]. This low-pass filtered temporal signal is maintained at the bipolar terminal. However, feedforward inhibition eliminates these low frequencies, leaving a bandpass temporal response in the transient GCs [23].

From our analysis in the accompanying paper [23], we expect that reducing our NA cell's time-constant, $\tau_{\rm na}$, will shift the transient GC's (GCt) temporal profile to higher frequencies, but leave the sustained GC's (GCs) response unchanged. To verify this prediction, we measured GC responses to a 0.2192 cyc/deg drifting sinusoidal grating at different temporal frequencies and recorded the temporal profile for different levels of V_τ, the bias voltage that sets $\tau_{\rm na}$. As we increased V_τ from 10 mV to 50 mV to 90 mV, the peak ON GCt response was 386 sp/s, 297 sp/s, and 418 sp/s while the peak ON GCs response was 319 sp/s, 271 sp/s, and 310 sp/s respectively. The peak values were relatively constant, suggesting that we successfully compensated for the effect of V_τ on the bipolar terminal (BT)-to-NA gain, g, by adjusting a second bias, $V_{\tau \rm s}$, to keep g constant [23]. So we normalized these responses and focused on changes in frequency. As shown in Fig. 9, increasing V_τ caused low-frequency GCt

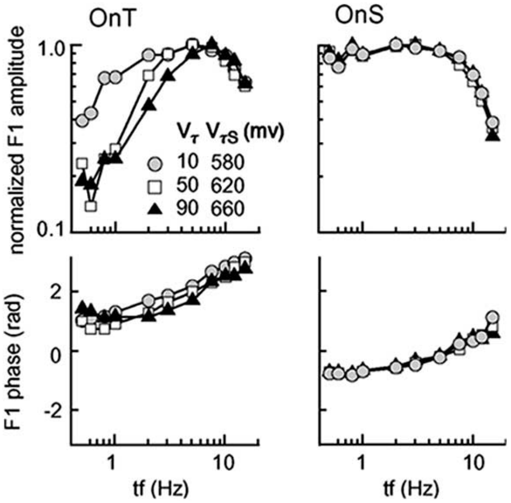


Fig. 9. Changing the open loop time constant. Temporal frequency responses of chip OnT and OnS GCs to a 0.2192 cyc/deg drifting sinusoidal grating. Profiles are shown for three different values of V_τ, which determines the open loop time constant, $\tau_{\rm na}$. Increasing V_τ causes a decrease in $\tau_{\rm na}$, thus increasing the system's corner frequency. To compensate for changes in dc loop gain, we also changed $V_{\tau \rm s}$ (values shown). Phase data are shown on the bottom.

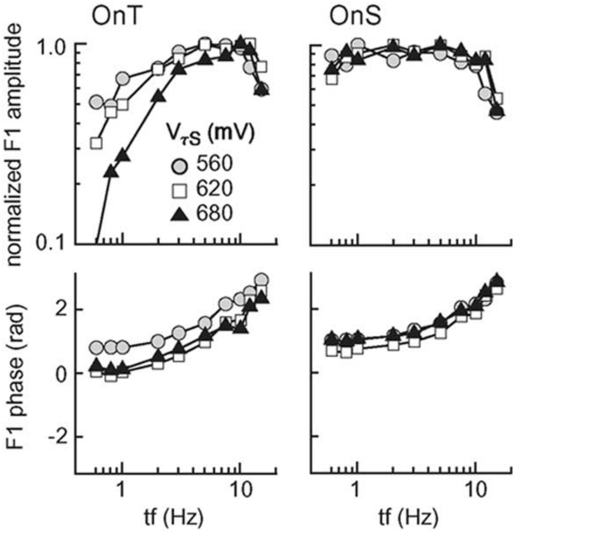


Fig. 10. Changing the open loop gain. Temporal frequency responses of chip OnT and OnS GCs to a 0.2192 cyc/deg drifting sinusoidal grating. Profiles are shown for three different values of $V_{\tau \mathrm{s}}$, which determines the open loop gain, g. Increasing $V_{\tau \mathrm{s}}$ above the dc unity value of 620 mV increases the open-loop gain, g, increasing the system's corner frequency. Phase data are shown on the bottom for the three different $V_{\tau \mathrm{s}}$ conditions.

responses to be attenuated, as the system's corner frequency increased. Changing V_τ has no effect on GCs' temporal frequency profile.

Increasing $V_{\tau \mathrm{s}}$ without increasing V_τ increases the gain of BT to NA excitation, g, and thereby modifies the loop-gain. From our analysis of loop-gain in the inner retina [23], we expect GCs' responses to remain unchanged while GCt's low-frequency responses are attenuated. Conversely, lowering $V_{\tau \mathrm{s}}$ would make the low-frequency roll-off less severe. To verify that g had this effect, we measured the GC temporal frequency response using a 50% contrast 0.2192 cyc/deg drifting sinusoidal grating for different levels of $V_{\tau \mathrm{s}}$. In this case, we empirically determined that $V_{\tau \mathrm{s}} = 620\ \mathrm{mV}$ corresponds to $g = 1$ for $V_\tau = 50\ \mathrm{mV}$. At 50% contrast, as we increased $V_{\tau \mathrm{s}}$ from 560 mV to 620 mV to 680 mV, the peak ON GCt response dropped from 1600 sp/s to 568 sp/s to 117 sp/s while the peak ON GC's response remained relatively unchanged and was 261 sp/s, 377 sp/s, and 338 sp/s respectively. Intuitively, one can understand the drop in GCt sensitivity by recognizing that increasing g boosts NA activity, which provides more feedforward inhibition on to GCt. To focus on g's effect on the system's temporal dynamics, however, we plotted the normalized responses, shown in Fig. 10. GCt's low-frequency responses were attenuated as we increased $V_{\tau \mathrm{s}}$, as expected. There was little effect on the high-frequency responses. And as expected, there was also little effect on GCs' responses.

VI. Contrast Gain Control

As discussed in the accompanying paper [23], modulation of presynaptic inhibition at the bipolar terminal by WA cells realizes contrast gain control. This mechanism is frequency dependent, as WA computes a temporal measure of contrast. The differential effect of frequency can be measured by simultaneously stimulating the retina with the sum of several sinusoids, approximating a white noise stimulus. Therefore, we measured the chip's temporal frequency sensitivity in response to a 0.14 cyc/deg contrast reversing sinusoidal grating whose temporal modulation signal was the sum of eight sinusoids. The temporal frequencies of the input were 0.214 Hz, 0.458 Hz, 0.946 Hz, 1.923 Hz, 3.876 Hz, 7.782 Hz, 15.594 Hz, and 31.219 Hz. These frequencies, chosen to minimize higher order interactions, are identical to those Victor and Shapley used to demonstrate contrast gain control [17].

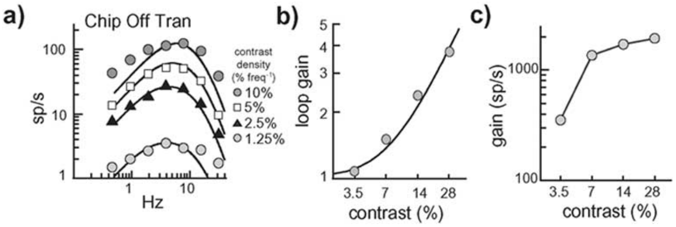


Fig. 11. Contrast gain control. (a) Chip Off-Transient cell response to a 0.14 cyc/deg contrast reversing sinusoidal grating whose temporal modulation signal was a sum of eight sinusoids. The amplitude of the fundamental Fourier component at seven of the eight frequencies used, for four different modulation contrasts, are shown. Solid lines are the best fit of an analytical model of the chip circuitry. (b) The loop gain that best fit Equation 1 increases as stimulus contrast increases. The solid line represents our prediction for this change in loop gain [23]. (c) As stimulus contrast increases, the system gain that best fits the data saturates, suggesting that contrast gain control causes a reduction in GC sensitivity.

We presented the white-noise like stimulus at four input contrast: 1.25%, 2.5%, 5%, and 10%, defined as the ratio of the peak deviation of each component over their common mean. Thus, we found that our OffT cells shift their sensitivity profile to higher temporal frequencies as the contrast per unit frequency increases from 1.25% to 10%, as shown in Fig. 11(a). This figure also demonstrates saturation in the response with increasing contrast.

To verify that the contrast changes we observed in the transient GC responses were consistent with our model, we fit the curves in Fig. 11(a) with the inner retina system equations derived in the accompanying paper [23]. We introduced sinusoidal inputs of contrast per unit frequency d. The outer retina is approximated by a low-pass temporal filter with time constant τ_o, whose output drives a transient GC response that is the difference between BT and NA. Thus, the GC response, in spikes/s, is given by

$$\mathrm{GCt} = S\left| d\frac{j\tau_A\omega + \epsilon(1-g)}{j\tau_A\omega + 1}\right| \left|\frac{1}{j\tau_o\omega + 1}\right|^2 \left|\frac{1}{j\tau_p\omega + 1}\right|$$

where $\tau_A \equiv \epsilon\tau_{\mathrm{na}}$, $\epsilon = 1/(1+wg)$, and w is the WA-modulated strength of NA inhibition onto BT. We also included a term that models the low-pass filtering behavior of the chip's photoreceptors whose time-constant is τ_p. We fit the four data sets by allowing the system gain term, S, and the loop gain, wg to vary across different stimulus contrasts, and fixed the remaining parameters. The best fits of this model to the four input contrast densities are shown as the solid lines in Fig. 11(a). Fitted parameter values were $\tau_p = 33$ ms, $\tau_o = 77$ ms, $\tau_{\mathrm{na}} = 1.0382\ \mathrm{s}$, and $g = 1.07$.

The system's loop gain increased with stimulus contrast, as expected. The best fits for loop gain, wg, in the four contrast conditions are shown in Fig. 11(b). As input contrast, c, increased by a factor of eight, the loop gain increased by a factor of 4, just as predicted by our loop-gain analysis in the accompanying paper [23]. Fitting the loop-gain equation [23] to the best estimates of wg [solid line in Fig. 11(b)] yields a value of $b_0 = 0.0784$ for the residual bipolar terminal activity.

However, we found that our contrast gain control mechanism cannot by itself account for the reduction in gain we see in the GC response. Although our predictions for wg were right, we had to introduce a variable system gain, S, that saturates with increasing input contrast, to account for sensitivity, as shown in Fig. 11(c). Other nonlinearities present in our CMOS circuit may account for this. For example, at the first synapse in the outer retina circuit, cone activity, which is determined by light input [23], gates an nMOS transistor to drive BC circuitry. Because $\kappa < 1$, the signal passed to the BC is compressed, causing saturation in the GC response in addition to that due to contrast gain control. Aside from this static nonlinearity, our prediction for how the chip adapts to contrast and temporal frequency was borne out, validating the theory, and suggesting that we have implemented a valid model for contrast gain control. However, as contrast gain control occurs at the bipolar terminal, we expected to observe its effects in sustained cells as well. However, it was not as dramatic in these cells, suggesting that NA cell feed-forward inhibition enhances contrast gain control.

VII. Conclusion

Due to the retina's complexity, we adopted a strategy that captured the major computations realized by the retina in a simplified model. Our model, and the silicon implementation of that model, produces multiple representations of the visual scene analogous to the retina's four major output pathways, and incorporates linear spatiotemporal filtering as well as nonlinear operations, including luminance adaptation, contrast gain control, and nonlinear spatial summation [23]. Our test results demonstrate that we succeeding in implementing our model, based on the neurocircuitry of the retina's cone pathway, in silicon. Furthermore, we have realized these computations at a photoreceptor density that is only 2.5 times as sparse as the human cone density at 5 mm eccentricity [6].

We implemented these functions in CMOS circuits by remaining faithful to the underlying retinal neuroanatomy — we morphed neural circuits into silicon circuits. Based on metabolic rates for rabbit retina [1], we estimate that our chip uses a thousand times as much energy per GC, consuming 17 μW per GC (62.7 mW total), at an average spike rate of 45 spikes/s. Ongoing advances in chip fabrication technology will allow us to improve our chip's energy efficiency as well as its spatial resolution and dynamic range.

We plan to close the performance gap between our chip and the mammalian retina by redesigning some of our CMOS circuits. For instance, asymmetric rectification in our model bipolar cells gives rise to differences in sensitivity between On and Off channels. The mammalian retina also exhibits an asymmetry at the BC, but in this case, the asymmetry lies in the quiescent levels of activity in On and Off pathways. Our asymmetry, which arises from saturation in the On channel, can be avoided by not taking the reciprocal of the cone signal.

Turning to the inner retina, while chip transient GCs exhibit contrast gain control, this effect is significantly less pronounced in chip sustained cells. We hypothesize that this difference arises from the presence or absence of feedforward amacrine cell inhibition. Sustained GCs in the mammalian retina receive some feedforward inhibition from amacrine cells, although the amount of inhibition is less than received by transient GCs. In our model, however, sustained cells receive no feedforward inhibition [10], [13]. Thus, an additional design issue to address would be to incorporate some feedforward inhibition, so as to potentiate the effects of contrast gain control, while maintaining the sustained behavior of our narrow-field sustained-type GCs.

While these design limitations and energy efficiency issues are important and still need to be addressed, our goal in these two companion papers has been to create a model based on the functional anatomy of the retina and to implement that model in silicon. We have demonstrated here that our chip performs bandpass spatiotemporal filtering, luminance adaptation, and contrast gain control. We will present comparisons between our chip's behavior and physiological measurements from the mammalian retina in a forthcoming paper.

Extensions of this work can be used to gain a deeper understanding of the computations in the retina. With a real-time model of retinal processing whose parameters can be easily adjusted, we can explore how certain components of our model affect GC response. Furthermore, because we have access to an entire array of GC outputs, we can also explore firing patterns among populations of cells. Physiologists have recently begun exploring these questions by using multi-electrode arrays to record behavior across several cells in the mammalian retina simultaneously [2], [15]. They have observed correlations in firing patterns between GCs that may play a role in conveying information to higher cortical structures [8], [7]. And they have revealed the presence of waves of retinal activity that may be important in development [22]. With our chip's array of spike outputs, researchers can begin to investigate these questions in a large-scale, real-time model.

In addition, our chip can be used to facilitate the design and fabrication of more complicated neural systems in silicon. With our approach, neuromorphic systems can be designed and implemented that replicate processing in the thalamus and in higher cortical structures. These higher level systems rely on sensory input, and our retinomorphic chip can serve as the front-end for these systems. By capturing the neural code of the mammalian retina, our chip can provide researchers with realistic retinal input with which they can design and test subsequent neuromorphic circuits.

Finally, our chip can be used in prosthetic applications. By using much less power — and weight and space — than required by computer-based solutions, retinomorphic chips could eventually serve as an *in situ* replacement. The most successful current retinal prosthesis designs are based on an external camera and processor [16]. By addressing our aforementioned design and power issues, we hope to use our approach to develop a fully integrated retinal prosthesis that supersedes current prosthetic designs.

REFERENCES

[1] A. Ames, Y. Y. Li, E. C. Heher, and C. R. Kimble, “Energy metabolism of rabbit retina as related to function: High cost of Na+ transport,” *J. Neurosci.*, vol. 12, pp. 840–853, 1992.

[2] M. J. Berry, D. K. Warland, and M. Meister, “The structure and precision of retinal spike trains,” *Proc. Nat. Acad. Sci.*, vol. 94, pp. 5411–6, 1997.

[3] K. A. Boahen, “The retinomorphic approach: Pixel-parallel adaptive amplification, filtering, and quantization,” in *Anaglog. Integr. Circuits Signal Processing*, vol. 13, 1997, pp. 53–68.

[4] ——, “A throuput-on-demand address-event transmitter for neuromorphic chips,” in *Proc. 20th Conf. Advanced Research in VLSI (ARVLSI’99)*, 1999, pp. 76–86.

[5] E. Culurciello, R. Etienne-Cummings, and K. Boahen, “High dynamic range, arbitrated address event representation digital imager,” in *Proc. IEEE Inter Symp Circuits and Systems*, vol. 3, 2001, pp. 505–508.

[6] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, “Human photoreceptor topography,” *J. Comp. Neurol.*, vol. 292, pp. 497–523, 1990.

[7] S. H. DeVries, “Correlated firing in rabbit retinal ganglion cells,” *J. Neurophysiol.*, vol. 81, no. 2, pp. 908–20, 1999.

[8] S. H. DeVries and D. A. Baylor, “Mosaic arrangement of ganglion cell receptive fields in rabbit retina,” *J. Neurophysiol.*, vol. 78, no. 4, pp. 2048–60, 1997.

[9] C. Enroth-Cugell and A. W. Freeman, “The receptive field spatial structure of cat retinal Y cells,” *J Physiol.*, vol. 384, pp. 49–79, 1987.

[10] M. A. Freed and P. Sterling, “The ON-alpha ganglion cell of the cat retina and its presynaptic cell types,” *J. Neurosci.*, vol. 8, no. 7, pp. 2303–20, 1988.

[11] L. J. Frishman, A. W. Freeman, J. B. Troy, D. E. Schweitzer-Tong, and C. Enroth-Cugell, “Spatiotemporal frequency responses of cat retinal ganglion cells,” *J. Gen. Physiol.*, vol. 89, pp. 599–628, 1987.

[12] R. J. Jensen and N. W. Daw, “Effects of dopamine and its agonists and antagonists on the receptive field properties of ganglion cells in the rabbit retina,” *Neuroscience*, vol. 17, no. 3, pp. 837–855, 1986.

[13] H. Kolb and R. Nelson, “OFF-alpha and OFF-beta ganglion cells in cat retina: II. Neural circuitry as revealed by electron microscopy of HRP stains,” *J. Comp. Neurol.*, vol. 329, no. 1, pp. 85–110, 1993.

[14] C. Mead, “Neuromorphic electronic systems,” *Proc. IEEE*, vol. 78, pp. 1629–36, 1990.

[15] M. Meister and M. J. Berry, “The neural code of the retina,” *Neuron*, vol. 22, pp. 435–450, 1999.

[16] J. F. Rizzo, J. Wyatt, M. Humayun, E. de Juan, W. Liu, A. Chow, R. Eckmiller, E. Zrenner, T. Yagi, and G. Abrams, “Retinal prosthesis: An encouraging first decade with major challenges ahead,” *Opthalmology*, vol. 108, no. 1, pp. 13–4, 2001.

[17] R. Shapley and J. D. Victor, “The contrast gain control of the cat retina,” *Vis. Res.*, vol. 19, pp. 431–434, 1979.

[18] R. G. Smith, “Simulation of an anatomically defined local circuit – The cone-horizontal cell network in cat retina,” *Visual Neurosci.*, vol. 12, pp. 545–561, 1995.

[19] P. Sterling, “Retina,” in *The Synaptic Organization of the Brain*, Fourth ed, G.M. Shepherd, Ed. New York: Oxford University Press, 1998.

[20] J. B. Troy and C. Enroth-Cugell, “X and Y ganglion cells inform the cat’s brain about contrast in the retinal image,” *Exp. Brain Res.*, vol. 93, pp. 383–390, 1993.

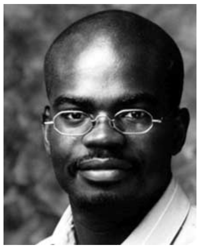
[21] J. H. van Hateren, “A theory of maximizing sensory information,” *Biological Cybern.*, vol. 68, pp. 23–29, 1992.

[22] R. O. Wong, M. Meister, and C. J. Shatz, “Transient period of correlated bursting activity during development of the mammalian retina,” *Neuron*, vol. 11, no. 5, pp. 923–38, 1993.

[23] K. A. Zaghloul and K. Boahen, “Optic nerve signals in a neuroomorphic chip I: Outer and inner retina models,” *IEEE Trans Biomedical Eng.*, vol. 51, pp. 657–666, Apr. 2004.

[24] ——, “A silicon model of the mammalian retina,” A silicon model of the mammalian retina*Neuron*, 2004, submitted for publication.

Kareem A. Zaghloul (S’00–M’03) received the B.S. degree from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, in the in 1995, where he made Tau Beta Kappa and Eta Kappa Nu. He recently completed a combined M.D./Ph.D. program at the University of Pennsylvania, Philadelphia. The Ph.D. degree was awarded in the Department of Neuroscience, where he worked on understanding information processing in the mammalian retina with K. A. Boahen. His work was supported by a Vision Training Grant from the National Institutes of Health and a Ben Franklin Fellowship from the University of Pennsylvania School of Medicine. He is currently a Resident Physician in the Department of Neurosurgery at the University of Pennsylvania.



Kwabena A. Boahen received the B.S. and M.S.E. degrees in electrical and computer engineering from the Johns Hopkins University, Baltimore MD, in the concurrent masters-bachelors program, both in 1989, where he made Tau Beta Kappa. He received the Ph.D. degree in computation and neural systems from the California Institute of Technology, Pasadena, in 1997, where he held a Sloan Fellowship for Theoretical Neurobiology.

He is an Associate Professor in the Bioengineering Department, University of Pennsylvania, Philadelphia, where he holds a secondary appointment in the electrical engineering. He was awarded a Packard Fellowship in 1999. His current research interests include mixed-mode multichip VLSI models of biological sensory and perceptual systems, and their epigenetic develoipment, and asynchronous digital interfaces for interchip connectivity.

Dr. Boahen received a National Science Foundation (NSF) CAREER award in 2001 and an Office of Naval Research (ONR) YIP award in 2002.