10
Circuit Designs That Model the Properties of the Outer and Inner Retina

Kareem A. Zaghloul, MD, PhD and Kwabena Boahen, PhD

CONTENTS

SILICON MODELS

One goal of understanding neural systems is to develop prosthetic devices that can someday be used to replace lesioned neural tissue. For such prosthesis to be practical, the device must perform these computations as efficiently as, and at a physical scale comparable with the lesioned network, and should adapt its properties over time, independent of external control. The approach to design a successful prosthesis that faithfully replicates the computations performed by a neural circuit is based on a detailed understanding of that circuit's anatomic connections and functional computations.

The retina, one of the best studied neural systems is a complex piece of biological wetware designed to signal the onset or offset of visual stimuli in a sustained or transient fashion *(1)*. To encode its signals into spike patterns for transmission to higher processing centers, the retina has evolved intricate neuronal circuits that capture information, efficiently contained within natural scenes *(2)*. This visual preprocessing, realized by the retina, occurs in two stages, the outer retina and the 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.

From: *Ophthalmology Research: Visual Prosthesis and Ophthalmic Devices: New Hope in Sight*
Edited by: J. Tombran-Tink, C. Barnstable, and J. F. Rizzo © Humana Press Inc., Totowa, NJ

Present attempts to engineer a viable retinal prosthesis have focused on the significant problem of efficient electrical stimulation of neurons along the visual pathway *(3,4)*. Microelectrode arrays, implanted epiretinally or subretinally, evoke phosphenes in patients with visual loss (because of outer retinal degeneration) by relying on electrical stimulation of the remaining retinal cells to dictate firing patterns *(5,6)*. Whereas the epiretinal approach relies on an external camera to capture visual information and on an external processor to recreate retinal computation, subretinal devices use photodiodes embedded in the electrode array to locally transduce light into stimulating current. Cortical visual prostheses address disease processes affecting structures postsynaptic to the outer retina *(7,8)*. They are similar to epiretinal prostheses in that they also depend on external devices to capture and process visual information, but they must fully recreate thalamic function as well as retinal function.

Whereas the emphasis on electrical stimulation technology is important in addressing the difficult problem of interfacing with the nervous system, a fully implantable retinal prosthesis would ideally capture all of the functions performed by the mammalian retina in one autonomous device. These neural computations can be performed at an energy efficiency and physical scale comparable with biology by morphing neural circuits into electronic circuits *(9)*. Micron-sized transistors function as excitatory or inhibitory synapses or as gap junctions, thereby recreating the synaptic organization of the retina at a similar physical scale. The time-scale and energy dissipation can be matched as well by operating these transistors in the subthreshold region, where they conduct nanoamperes or even picoamperes, just like small populations of ion channels do. Furthermore, as millions of transistors can be fabricated on a thumb-nail-sized piece of silicon using very large scale integration technology, this neuromorphic approach offers a fully implantable solution for neural prostheses. This implementation efficiency is translated to a higher ability to explore model parameters to further understand the underlying biological system and, by communicating with other neuromorphic chips, a higher ability to replicate more complicated neural systems.

The first effort to morph the retina into silicon, though widely acclaimed, suffered from several shortcomings. First, only outer retina circuitry was morphed: the cones, horizontal cells (HC), and bipolar cells (BC) *(10)*. Second, a logarithmic photoreceptor (cf., cone) was used to capture a wide intensity range, but this degraded the signal-to-noise ratio by attenuating large amplitudes (i.e., signal) whereas leaving small amplitudes (i.e., noise) unchanged. Third, the spatiotemporal average (cf., HC) was subtracted to obtain contrast (cf., bipolar signal) or more precisely, the logarithm of contrast, but this made the signal-to-noise ratio even worse. Subsequent efforts *(11)* overcame these limitations by modulating synaptic strengths locally to control sensitivity, and by including the cone-to-cone gap junctions to attenuate noise. But they still omitted the inner retina, which contains upwards of 44 cell types *(12)*.

The most recent effort to model retinal processing in silicon incorporated outer retina circuitry as well as bipolar and amacrine cell interactions in the inner retina *(13)*. This outer retina circuit took the difference between the photoreceptor signal and its spatiotemporal average, computed by a network of coupled lateral elements HCs, through negative (inhibitory) feedback. Cone-coupling in this model attenuated high-frequency noise to realize a spatial bandpass filter and dynamic range was extended by implementing local automatic gain control. Furthermore, this model used HC activity to boost cone to HC excitation *(13)*, which eliminated the luminance-dependent receptive-field expansion and temporal instability that plagued previous efforts to modulate synaptic strengths locally *(14)*. This gain boosting mechanism has some physiological basis as glutamate release from cones is modulated by HC hemichannels *(15)*, and may be enhanced further by HC autofeedback *(16)*. The outer retina design adopts this approach.

The bipolar and amacrine cell interactions, introduced by this earlier design, represented a model for inner retina processing that much like the circuit discussed here, attempted to capture temporal adaptation through adjustment of amacrine cell feedback inhibition. However, this earlier design did not include the retina's complementary push-pull architecture *(1)*. Hence, at low frequency stimulation, baseline direct current (DC) levels tended to rise, and thus, modulation of the feedback loop gain was not realized as intended *(13)*. Furthermore, this previous implementation did not replicate synaptic interactions at the ganglion cell level, and thus, did not faithfully capture the behavior of the retina's major output pathways.

A silicon retina has been developed modeled on neural circuitry in both the outer and inner retina that addresses the design flaws in earlier designs *(13)* and extends them to include ganglion-cell level synaptic interactions. The model is based on the functional architecture of the mammalian retina and on physiological studies that have characterized the computations performed by the retina. By capturing both outer and inner retina circuitry using single-transistor synapses, the silicon retina, which was built, passes only an intermediate range of frequencies. It attenuates redundant low spatiotemporal frequencies and rejects noisy high frequencies, much like the retina does. And by modulating the strengths of its single-transistor synapses locally, the device adapts to luminance and to contrast. It responds faster, but more transiently as contrast increases, much like the retina does. This contrast gain control arises from a new inner retina circuit design that assigns specific roles to anatomically identified microcircuits in the mammalian retina.

This silicon retina outputs spike trains that capture the behavior of ON- and OFF-center *(17)* versions of wide-field transient and narrow-field sustained ganglion cells (GC) *(18)*, which provide ninety percent of the primate retina's optic nerve fibers *(19)*. And, more significant for a prosthetic application, these are the four major types that project through thalamus, to primary visual cortex. Furthermore, the silicon circuit is constructed at a scale comparable with the human retina and uses under a tenth of a watt, thereby satisfying the requirements of a fully implantable prosthesis.

This chapter describes the silicon implementation of the outer and inner retina circuits. It is organized as follows: In outer retina section, the outer retina model and its silicon implementation are presented. In inner retina section, the novel model for the inner retina and its silicon implementation are presented. In silicon chip layout section, the architecture of the silicon chip layout is presented. In silicon retina response section, the responses of the silicon retina to physiological stimuli are described and are compared with those of the mammalian retina. Finally, conclusion section concludes the article.

OUTER RETINA

The model for outer retinal circuitry is based on identified synaptic interactions and local microcircuits previously described in the literature, obtained using histological and physiological techniques. After constructing this model of the outer retina's synaptic connections, it is used as a blueprint with which the silicon retina is assembled.

Modeling the Outer Retina

The model for the outer retina (Fig. 1A) is designed to realize luminance adaptation by adjusting synaptic strengths locally. Photocurrents from the cone outer segments (CO) in the model drive a network of cone terminals, which subsequently, excite a network of HCs. Because they are coupled through gap junctions, these HCs compute the local average intensity in the model. This signal is used to modulate cone-to-cone coupling strength as well as the cone's membrane conductance (shunting inhibition). Using the local intensity signal to adjust these two synaptic strengths makes the cone terminal's sensitivity inversely proportional to luminance, whereas preventing the changes in spatial frequency tuning that plagued previous attempts at light adaptation *(11)*. To compensate for the resulting signal attenuation at the cone terminal, the HC's local intensity signal is also used to modulate the cone-to-HC synaptic strength. This autofeedback mechanism, whereby the HC regulates its own input, is similar to that found in the retina *(15,16)*.

From the synaptic interactions, the block diagram was derived for the outer retina shown in Fig. 1B by modeling both the cone and HC networks as spatial lowpass filters. The system level equations can be derived that describe how CO activity determines HC and cone terminal activity, represented by i_{hc} and i_{ct}, respectively, from this block diagram. These equations, reproduced from ref. *13* are as follows:

$$i_{\mathrm{hc}}(\rho) = \left(\frac{A}{\left(l_{\mathrm{c}}^2\rho^2 + 1\right)\left(l_{\mathrm{h}}^2\rho^2 + 1\right) + \frac{A}{B}} \right) \frac{i_{\mathrm{co}}}{B}$$

$$i_{\mathrm{ct}}(\rho) = \left(\frac{\left(l_{\mathrm{h}}^2\rho^2 + 1\right)}{\left(l_{\mathrm{c}}^2\rho^2 + 1\right)\left(l_{\mathrm{h}}^2\rho^2 + 1\right) + \frac{A}{B}} \right) \frac{i_{\mathrm{co}}}{B}$$

where i_{co} represents CO activity, excited by photocurrents, l_{c} and l_{h} are the cone- and horizontal-network space-constants, respectively, B is the attenuation from the CO to the cone terminal, A is the amplification from cone terminal to HC, and ρ is spatial frequency. HCs have stronger coupling in the model (i.e., l_{h} is larger than l_{c}), causing their spatial lowpass filter to attenuate lower spatial frequencies. Thus, HCs lowpass filter the signal whereas cone terminals bandpass filter it, as shown in Fig. 1C, contributing to the GCs' spatial frequency tuning *(20)*.

To realize local automatic gain control in the model, HC activity is set proportional to intensity and this activity is used to modulate CO to cone terminal attenuation,

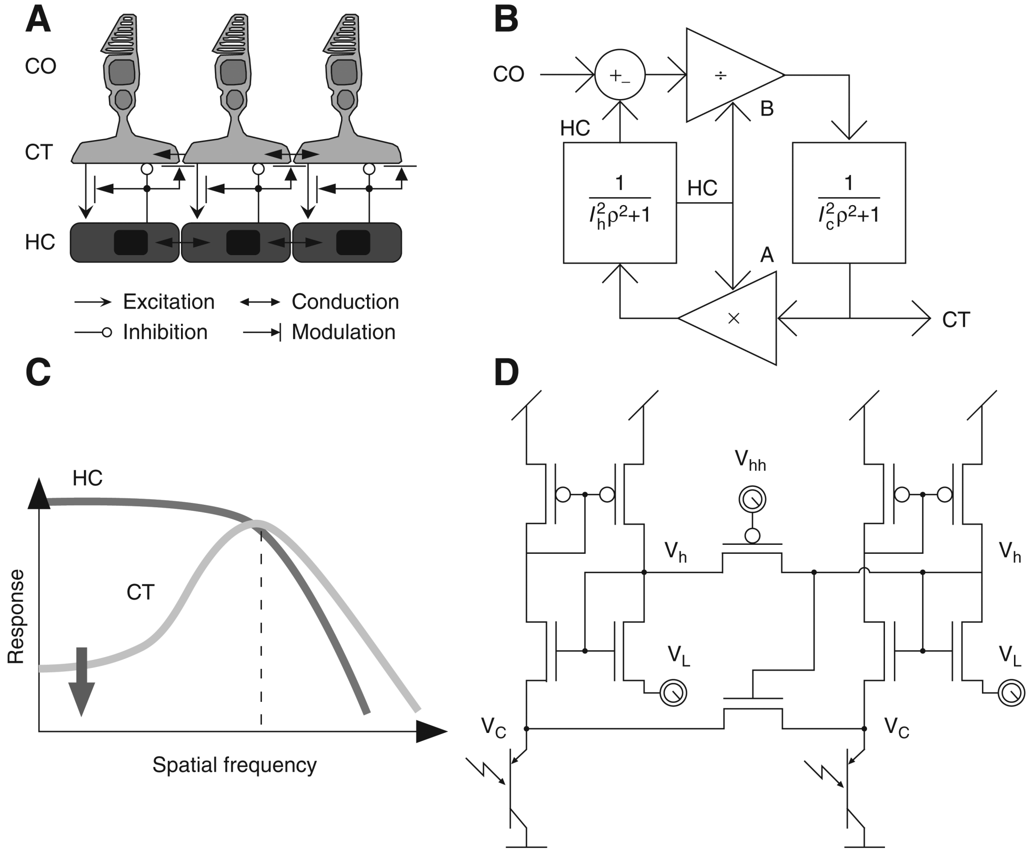

Fig. 1. Modeling the outer retina. **(A)** Neural circuit: cone terminals (CT) receive a signal that is proportional to incident light intensity from the cone outer segment (CO) and from neighboring cones, through gap junctions, and excite horizontal cells (HC). HCs spread their input laterally through gap junctions, provide shunting inhibition onto CT, and modulate cone coupling and cone excitation. **(B)** System diagram: signals travel from the CO to the cone terminal (CT) and on to the HC network, which provides negative feedback. Both cone terminals and HCs form networks, connected through gap junctions that are governed by their respective space constants, l_c and l_h. Excitation of HCs by cone terminals is modulated by the HCs, which also modulate the attenuation from the CO to the cone terminal together with modulation of cone gap junctions to keep l_c constant. These compensatory interactions realize local automatic gain control in the cone terminal and keep receptive field size invariant. **(C)** Frequency responses: both HCs and cone terminals (CT) lowpass filter input signals, but because of the HC network's larger space constant, HC inhibition eliminates low frequency signals, yielding a spatially bandpass response in the cone terminal. **(D)** Outer retina circuitry. A phototransistor draws current through an nMOS transistor whose source is tied to V_c, decreases in which represent increased cone terminal activity, and whose gate is tied to V_h, increases in which represent increased HC activity. This transistor passes a current proportional to the product of cone terminal and HC activity, thus, modeling shunting inhibition from HCs to cones. In addition, this current, mirrored through pMOS transistors, dumps charge on the HC node, V_h, modeling cone terminal excitation of the HCs, and its modulation by HCs. V_L, a global bias set externally sets the mean level of V_c. V_{hh}, another external, bias sets the strength of the HC gap junctions. The strength of those between cones are modulated locally by V_h *(30)*.

B, by changing cone-to-cone conductance. This modulation adapts cone sensitivity to different light intensities *(14)*. Furthermore, to overcome the expansion in receptive-field size that this modulation caused in earlier designs, HC modulation of cone gap-junctions is complemented with HC modulation of cone leakage conductance, through shunting inhibition, making l_c independent of luminance. The change in loop-gain with HC modulation of cone excitation was also compensated by keeping A, the amplification from cone terminal to HC, proportional to B, thus, fixing the peak spatial frequency of cone terminal response *(21)*.

Cone terminal activity in the model is primarily determined by spatial contrast, saturates when increasing signal fluctuations cause this ratio to become large, and is entirely independent of absolute luminance. From the outer retina model's system equations, how the silicon model's cone terminal activity depends on its cone and horizontal space constants and on contrast *(21)* was derived:

$$i_{\mathrm{CT}} = \frac{2r}{r + 2 - 1/r + 2c} c$$

where i_{CT} represents cone terminal activity and c represents stimulus contrast; $r = l_h/l_c$ is the ratio of the HC space constant to the cone space constant. This behavior is remarkably similar to the mammalian retina. Physiologists have found that cone responses as a function of contrast—intensity of light stimulation relative to a background—are described by a simple equation:

$$V = \frac{IV_m}{I + \sigma}$$

where V is the peak amplitude of the cone response produced by a given level of stimulating light intensity, I. V_m is the maximum response and σ is the background intensity. The response reaches half of the maximum when $I = \sigma$ *(22)*. This adaptive behavior is preserved across five decades of background intensities. Cone responses obtained from the model, where c corresponds to I/σ, behave similarly when l_h is comparable with l_c *(21)*.

Morphing the Outer Retina Into Silicon

The retinal models were morphed into a silicon chip by replacing each synapse or gap-junction in the model with a transistor. One of its terminals is connected to the presynaptic node, another to the postsynaptic node, and a third to the modulatory node. By permuting these assignments, excitation, inhibition, and conduction are realized, all of which are under modulatory control.

Modulation was implemented by exploiting the exponential I(V) relationship of the MOS (metal-oxide-semiconductor) transistor. In the subthreshold regime, the current from the drain terminal to the source terminal is the superposition of a forward component that decreases exponentially with the source voltage (V_s) and a reverse component that decreases similarly with the drain voltage (V_d); both components increase exponentially with the gate voltage (V_g). That is, $I_{\mathrm{ds}} = I_{\mathrm{o}} \mathrm{e}^{\kappa V_g}(e^{-V_s} - e^{-V_d})$ where $\kappa \approx 0.7$ is a nonideality factor; voltages are in units of $U_T = 25$ mV, at 25°C. This equation describes the n-type device; voltage and current signs are reversed for a p-type *(23)*. Hence, the transistor converts voltage to current exponentially and converts current back to voltage logarithmically. By adding a voltage offset representing the log of the modulation factor, the current can be multiplied by a constant, thereby modulating the synaptic strength. Synaptic strengths are modulated locally, within the chip, except for a small number of biases globally controlled by the user. Current-mirrors are added to reverse the direction of current when necessary.

Morphing the model for the outer retina yielded the electronic circuit shown in Fig. 1D. Omitting adaptation in the phototransduction cascade, photocurrents linearly proportional to luminance discharge the cone terminal node, V_c, which is defined as an increase in cone terminal activity. This drop in V_c produces a current that excites the HC network through an nMOS transistor followed by a pMOS current-mirror. This excitatory current is modulated by HC activity, represented by V_h, and increases as V_h increases to realize autofeedback. But this increased current also releases more charge on to V_c, thereby realizing horizontal-cell inhibition of cone terminal activity. Thus, a single transistor implements two distinct synaptic interactions, one excitatory, the other inhibitory. Cone nodes are electrically coupled to their six nearest neighbors through nMOS transistors whose gates are controlled locally by HCs, implementing the model of cone gap-junction modulation. HCs also communicate with one another, through pMOS transistors, but this coupling is controlled by an externally applied voltage (V_{hh}).

Cone terminals in the model drive two types of BCs that rectify the cone signal into ON and OFF channels, thus, reproducing the complementary signaling scheme found in the mammalian retina *(17)*. the cone signal using push–pull pMOS current mirrors is rectified, not shown here, but described in detail in *(21)*. Briefly, the current generated by cone terminal activity is compared with a reference current set at the mean. Subtracting these currents by mirroring them on to one another removes the baseline; the difference drives the ON bipolar cell if it is positive and drives the OFF bipolar cell if it is negative *(21)*. Thus, the bipolar circuitry divides signals into ON and OFF channels.

INNER RETINA

Adopting the same approach, that was taken with the outer retina, the model for the inner retina is based on identified synaptic interactions and local microcircuits previously described in the literature, obtained using histological and physiological techniques. After constructing this model of the inner retina's synaptic connections, it is used as a blueprint with which the silicon retina is assembled.

Modeling the Inner Retina

The model of the inner retina, which realizes lowpass and highpass temporal filtering, adapts temporal dynamics to input frequency and to contrast, and drives ganglion cell responses, is shown in Fig. 2A. BCs from the outer retina excite amacrine cells with either narrow or wide fields. To ensure that only the ON or OFF channel is active at any time, bipolar and amacrine cells receive inhibition from the complementary channel in the model, similar to vertical inhibition found between the inner-plexiform

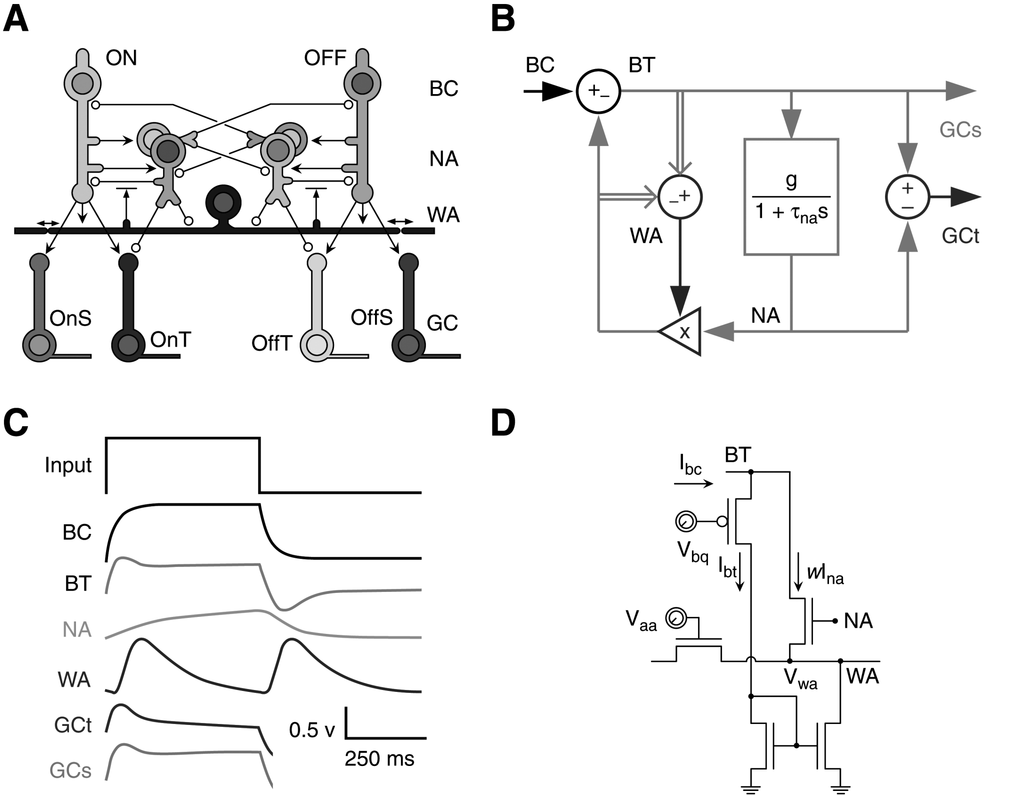

Fig. 2. Inner retina model and circuit. **(A)** Inner retina synaptic interactions: On and Off bipolar cells (BC) relay cone signals to ganglion cells (GC), and excite narrow- and wide-field amacrine cells (NA, WA). Narrow-field amacrine cells inhibit BC, wide-field amacrine cells, and transient GCs; their inhibition onto wide-field amacrine cells is shunting. Wide-field amacrine cells modulate narrow-field amacrine cell presynaptic inhibition and spread their signals laterally through gap junctions. BC also excite local interneurons that inhibit complementary bipolar cell and narrow-field amacrine cells. Four types of GCs signal the onset or offset of visual stimuli in a sustained (OnS, OffS) or transient (OnT, OffT) fashion. **(B)** System diagram: narrow-field amacrine cell (NA) signals represent a lowpass filtered version of bipolar terminal (BT) signals and provide negative feedback on to the bipolar cell (BC). The wide-field amacrine cell (WA) network modulates the gain of NA feedback (X). WA receives full-wave rectified excitation from BT and full-wave rectified inhibition from modulated NA (double arrows). BT drives sustained GCs directly whereas the difference between BT and NA drives transient ganglion cells (GCt). **(C)** Numerical solution to inner retina model with a unit step input of 1V. Traces show 1 s of On cell responses for BC, BT, NA, GCt, and WA. Outer retina time constant, τ_o, is 96 ms; τ_{na} is 1 s. **(D)** Wide field amacrine cell modulation: bipolar terminal (BT) activity (I_{bt}) excites a network of wide-field amacrine cells (WA) through a current mirror; it also excites the narrow-field amacrine cell (NA, excitation circuitry not shown). Wide-field amacrine cell activity modulates the strength of narrow-field amacrine cell feedback inhibition on to the bipolar terminal, subtracting a current wI_{na} from the bipolar cell's excitatory input I_{bc}. The same current is also subtracted from the wide-field amacrine cell's excitatory input, I_{bt}, thereby inhibiting it. V_{bq} controls the quiescent current supplied to the inner retina by the bipolar terminal and V_{aa} controls the extent of gap-junction coupling in the wide-field amacrine cell network *(30)*.

layer's ON or OFF laminae *(24)*. The signal is made at the bipolar cell's terminal more transient (highpass filtered) than its cone input by applying sustained (lowpass filtered) inhibition from the narrow-field amacrine-cell *(25)*. The bipolar cell terminal in the model also excites two types of GCs, which is called transient and sustained. In transient GCs, feedforward inhibition from the narrow-field amacrine cells cancels residual sustained excitation from the bipolar terminal, similar to the synaptic complex found in mammalian retina *(26)*.

Contrast gain control in the inner retina model is determined by the modulatory effects of wide-field amacrine cell activity. Wide-field amacrine cell activity represents the local measure of contrast, weighed on neighboring spatial locations. These cells are excited by both ON and OFF BCs and inhibited by both ON and OFF narrow-field amacrine cells in the model, similar to ON–OFF amacrine cells found in the retina *(27)*, and are coupled together through gap-junctions to form a wide-field amacrine cell network. By modulating their own inhibitory inputs as well as narrow-field amacrine cell inhibition at bipolar terminals, the wide-field amacrine cells compute temporal contrast. That is, their activity reflects the ratio between contrast fluctuations (highpass signal) and average contrast (lowpass signal). As this temporal contrast increases, their modulatory activity increases, the net effect of which is to make the GCs respond more quickly and more transiently. There is also an overall decrease in sensitivity because of the less sustained nature of the response. This adaptation captures properties of contrast gain control in the mammalian retina *(21,28)*.

From the synaptic interactions of Fig. 2A, the block diagram was derived for the inner retina, shown in Fig. 2B, by modeling the narrow-field amacrine cell as a lowpass filter. Responses of the different inner retina cell types in this model to a step input are shown in Fig. 2C. Bipolar cell activity is a low-pass filtered version of light input to the outer retina. Increase in bipolar cell activity causes an increase at the bipolar terminal and slower increase in the narrow-field amacrine cell. The difference between bipolar terminal activity and gain-modulated narrow-field amacrine cell activity determines wide-field amacrine cell activity, which, in turn, sets the gain of narrow-field amacrine cell feedback inhibition on to the bipolar terminal and on to the wide-field amacrine cell. Thus, after a step input, bipolar terminal activity initially rises, but narrow-field amacrine cell inhibition, setting in later, attenuates this rise until bipolar terminal activity is equaled by gain-modulated narrow-field amacrine cell activity. Wide-field amacrine cell activity rises above its baseline value of unity at both onset and offset, driven by full-wave rectified input from ON and OFF bipolar terminals. The bipolar terminal drives the sustained ganglion cell, which responds for the duration of the step, whereas the difference between bipolar terminal and narrow-field amacrine cell activity drives the transient ganglion cell, which decays to zero.

From the block diagram in Fig. 2B, the system level equations can be derived for narrow-field amacrine and bipolar cell acitivity with the help of the Laplace transform:

$$i_{\mathrm{na}} = \frac{g\varepsilon}{\tau_A s + 1} i_{\mathrm{bc}}, i_{\mathrm{bt}} = \frac{\tau_A s + \varepsilon}{\tau_A s + 1} i_{\mathrm{bc}}$$

where g is the gain of the excitation from the bipolar terminal to the narrow-field amacrine cell, and where

$$\tau_A \equiv \varepsilon\tau_{\mathrm{na}}, \varepsilon \equiv \frac{1}{1+wg}$$

τ_{na} is the time constant of the narrow-field amacrine cell and w represents wide-field amacrine cell activity, which determines feedback strength. From the equations, it can be seen that the bipolar terminal highpass filters and the narrow-field amacrine cell lowpass filters the bipolar cell signal; they have the same corner frequency, $1/\tau_A$. This closed-loop time-constant, τ_A, depends on the loop gain wg, and therefore on wide-field amacrine cell activity. For example, stimulating the inner retina with a high frequency would provide more bipolar terminal excitation (highpass response) than narrow-field amacrine cell inhibition (lowpass response) to the wide-field amacrine cell network. Wide-field amacrine cell activity, and hence w, would subsequently rise, reducing the closed-loop time-constant τ_A, until the corner frequency $1/\tau_A$ reaches a point where bipolar terminal excitation equaled narrow-field amacrine cell inhibition. This drop in τ_A, accompanied by a similar drop in ε, will also reduce overall sensitivity and advance the phase of the response.

The system behavior governed by these equations is remarkably similar to the contrast gain control model proposed by Victor *(29)*, which accounts for the compression of retinal responses in amplitude and in time with increasing contrast. Victor proposed a model for the inner retina whose highpass filter's time-constant, T_S, is determined by a "neural measure of contrast," c. The governing equation is:

$$T_s = \frac{T_O}{1+\dfrac{c}{c_{1/2}}}$$

This model's time-constant depends on contrast in the same way that the model's time constant depends on wide-field amacrine cell activity, where Victor's T_O is similar to the τ_{na} and where Victor's ratio $c/c_{1/2}$ is represented by how much wide-field amacrine cell activity increases above its value at DC in the model. As this activity is sensitive to temporal contrast *(21)*, it is proposed that wide-field amacrine cells are the anatomical substrate that computes Victor's neural measure of contrast.

In the model, the loop gain wg is set by the local *temporal contrast*. Wide-field amacrine cell activity reflects inputs from bipolar terminals and narrow-field amacrine cells weighted across spatial locations, so these pooled excitatory and inhibitory inputs should balance when the system is properly adapted:

$$\begin{aligned} &w \updownarrow i_{\mathrm{na}} \updownarrow = \updownarrow i_{\mathrm{bt}} \updownarrow + i_{\mathrm{surr}} \\ &\Rightarrow w = (\updownarrow i_{\mathrm{bt}} \updownarrow + i_{\mathrm{surr}}) / \updownarrow i_{\mathrm{na}} \updownarrow \end{aligned}$$

where i_{surr} is defined as the gap-junction current, resulting from spatial differences in wide-field amacrine cell activity w. $\updownarrow i_{\mathrm{bt}} \updownarrow$ and $\updownarrow i_{\mathrm{na}} \updownarrow$ are full-wave rectified versions of i_{bt} and i_{na}, computed by summing On and Off signals. If all different phases are pooled spatially, these full-wave rectified signals will not fluctuate and will be proportional to amplitude. And if i_{surr}, is ignored w will simply be $|i_{\text{bt}}|/|i_{\text{na}}|$, yielding a temporal measure of contrast, because it is the ratio of a temporal difference (highpass signal, i_{bt}) and a temporal average (lowpass signal, i_{na}). At DC, this ratio is equal to *1/g*; hence the loop gain is unity and the DC gain from the bipolar cell to its terminal is $\varepsilon = 1/2$.

Ganglion cell responses in the inner retina model are derived from the bipolar terminal and narrow-field amacrine cell signals. Specifically, bipolar terminal signals directly excite both sustained and transient types of GCs, but transient cells receive feedforward narrow-field amacrine cell inhibition as well. The system equations determining sustained and transient ganglion cell responses (GCs and GCt), derived from the equations above, as a function of the input to the inner retina, i_{bc}, are as follows:

$$i_{\text{GCs}} = \frac{j\tau_A\omega + \varepsilon}{j\tau_A\omega + 1} i_{\text{bc}}$$

$$i_{\text{GCt}} = \frac{j\tau_A\omega + \varepsilon(1-g)}{j\tau_A\omega + 1} i_{\text{bc}}$$

where ω is temporal frequency (and $j = \sqrt{-1}$). When bipolar terminal to narrow-field amacrine cell excitation has unity gain ($g = 1$), feedforward inhibition causes a purely high-pass (transient) response in the transient ganglion cell whereas the sustained ganglion cell retains a low-pass (sustained) component. With a small loop-gain, wg, the DC gain, ε, approaches 1/2 and the bipolar terminal and sustained ganglion cell response becomes all-pass. However, as the loop gain increases, ε decreases and the response becomes highpass. The change in ε with loop gain is matched in both bipolar terminals and narrow-field amacrine cells, and so taking the difference between these two signals always cancels the sustained component in the transient ganglion cell *(21)*. Thus, the transient ganglion cell produces a purely highpass response irrespective of the system's loop gain.

Morphing the Inner Retina Into Silicon

The model for the inner retina is implemented in part by the electronic circuit shown in Fig. 2D. Wide-field amacrine cell modulation of narrow-field amacrine cell inhibition is realized by applying voltages representing wide- and narrow-field amacrine activity to a transistor's source and gate terminals, respectively. This transistor drains current from the node that represents bipolar terminal activity, implementing presynaptic inhibition by the narrow-field amacrine cell. It also sources current onto the node that represents wide-field amacrine activity, charging up that voltage, V_{wa}. This increase corresponds to inhibition of wide-field amacrine activity since, as V_{wa} increases, the strength of narrow-field amacrine inhibition *(w)* decreases. Conversely, as V_{wa} decreases, the strength of this inhibition increases. A pMOS transistor and a current-mirror realize excitation of the wide-field amacrine by the bipolar terminal. Wide-field amacrine cell nodes are coupled to one another through nMOS transistors gated by V_{aa}.

The circuit diagram of Fig. 2D represents only one of the inner retina circuit's complementary halves and does not show bipolar terminal to narrow-field amacrine excitation, lowpass filtering in the narrow-field amacrine cell, or push-pull inhibition. In the complete circuit, ON and OFF signals from bipolar cell circuitry drive either half of the inner retina circuit. Excitatory current from the bipolar terminal excites the narrow-field amacrine cell node on one side of the circuit whereas also inhibiting the narrow-field amacrine cell node on the complementary side, thus, realizing push-pull inhibition. To realize linear lowpass filtering, these inputs are divided by the sum of ON and OFF narrow-field amacrine cell activity, which compensates for the exponential voltage-current relationship in the subthreshold regime *(30)*. ON and OFF bipolar terminal and narrow-field amacrine cell signals converge on to the wide-field amacrine cell network, implementing full-wave rectification. Finally, signals from the bipolar terminal and narrow-field amacrine cells are used to drive ON and OFF transient and sustained GCs.

Because analog signals cannot be relayed over long distances, retinal GCs use spikes to communicate with central structures. Similarly, each ganglion cell in the chip array converts the current it receives from the inner retina circuit to spikes. The spike generating circuit (not shown) exhibits spike-rate adaptation through Ca^{++} activated K^{+} channel analogs, modeled with a current-mirror integrator *(31)*.

SILICON CHIP LAYOUT

In the mammalian retina, cone signals converge on to BCs *(1)*, which makes the receptive field center Gaussian-like *(32)*. To implement signal convergence in the 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. 3A. Each of the outer retina circuits actually produces two output currents. A central photoreceptor drives the bipolar cell with both of these outputs while photoreceptors at the six vertices divide these outputs between their two nearest BCs. For symmetry, a similar architecture is implemented for the reference current (*see* ref. *21*).

Transient GCs in the mammalian retina pool their inputs from a larger region than sustained GCs. This difference is maintained in the chip by pooling inner retina signals in the same way that outer retina signals are pooled. Thus, each transient ganglion cell receives input from a central inner retina circuit and its six nearest neighbors, as shown in Fig. 3A. This central circuit excites its ganglion cell with both copies of its transient output whereas those at the six vertices divide their outputs between their two nearest transient GCs. As inner retina circuits are tiled at one-quarter the density of the phototransistors that provide their input, this results in a larger receptive field for each ganglion cell than for each bipolar cell. 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. It also accounts for transient GCs' nonlinear subunits *(20)* since all the rectified bipolar cell 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 on the cell's receptive field would not modulate the cell's response as decreases in dark regions would exactly cancel out increases in light regions.

The chip design was fabricated in a 0.35 µm minimum feature-size process with its cell mosaics tiled at a scale similar to the mammalian retina (Fig. 3B). Phototransistors are tiled triangularly 40 µm apart; this spacing is only about two and a half times that of human cones at 5 mm nasal eccentricity *(33)*. The phototransistors are only 10 µm

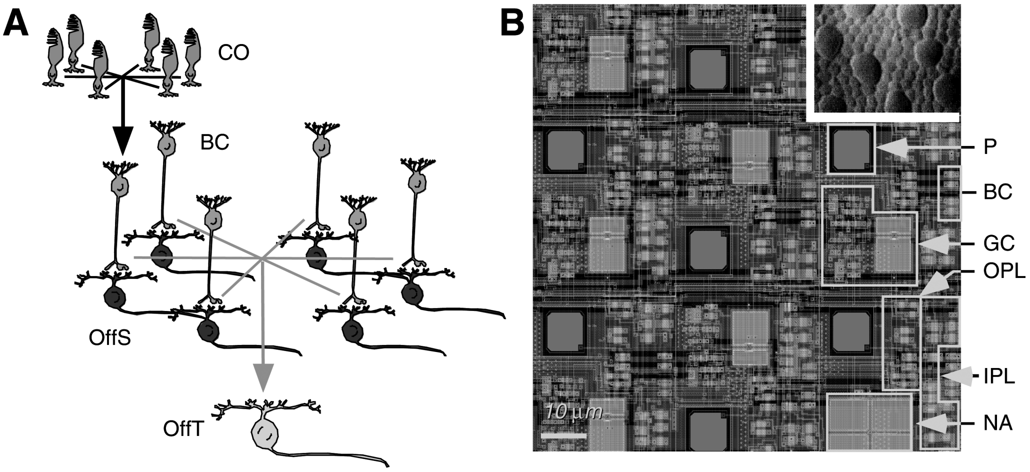

Fig. 3. Chip layout. **(A)** Functional architecture: signals from a central photoreceptor (not shown) and its six neighbors (CO) are pooled to provide synaptic input to each bipolar cell (BC). Each bipolar cell generates a rectified output, either On or Off that drives a local inner plexiform layer (IPL) circuit. Sustained GCs (OffS), which have a dendritic field diameter of 80 µm, receive input from a single local IPL circuit. Transient GCs (OffT), however, receive signals from a central IPL circuit (not shown) and its six neighbors, and hence, their dendritic field is 240 µm wide. **(B)** Chip design and human photoreceptor mosaic: each pixel with 38 transistors on average has a phototransistor (P), outer plexiform (synaptic) layer circuitry, bipolar cells, and IPL circuitry. Spike-generating GC are found in five out of eight pixels; the remaining three contain a narrow field amacrine (NA) cell membrane capacitor. Inset: Tangential view of human cone (large) and rod (small) mosaic at 5 mm eccentricity, plotted at the same scale (reproduced from refs. *30, 33*).

on a side, leaving ample space for postsynaptic circuitry, which is interspersed between them. This spacing is necessary because, unlike neural tissue, silicon microfabrication technology cannot produce three-dimensional structures.

One pixel—the basic element that is tiled to create the silicon retina—contains a phototransistor, outer retina circuitry, a pair of BCs, and one-quarter of the inner retina circuit. Hence, 2 × 2- and 4 × 4-pixel blocks 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 ganglion cell circuitry. Three out of every eight pixels instead contain a large capacitor that gives the narrow-field amacrine cell its long time-constant. A spike generating circuit in the remaining five pixels converts GC inputs into spikes that are sent off chip. The 3.5×3.3 mm^2 silicon die has 5760 phototransistors at a density of 722/mm^2 and 3600 GCs at a density of 461/mm^2—tiled in $2 \times 48 \times 30$ and $2 \times 24 \times 15$ mosaics of sustained and transient On and Off GCs.

Because of wiring limitations, each ganglion cell output off chip cannot be communicated directly. Instead, an asynchronous, arbitered, multiplexer are used to read spikes out from the silicon neurons *(34)*. Each ganglion cell interfaces with digital circuitry that communicates the occurrence of a spike to the row and each column arbiters. The arbiters select one row and one column at a time, and an encoder outputs their addresses. Row and column addresses are communicated serially off chip. The spike activity of all 3600 GCs can be represented with just seven bits using this address-event representation. By noting the address of each event generated by the chip, Ganglion cell location and type can be decoded in the array. In a future prosthetic application, such a scheme would be unnecessary as each ganglion cell's local spiking circuit could drive a local micro-electrode to stimulate adjacent neural tissue.

SILICON RETINA RESPONSES

To gauge the validity of the outer and inner retina models and to verify their ability to recreate retinal function, the responses of the silicon retina are compared with those of the mammalian retina. Specifically, the silicon retina's responses were recorded to different spatial and temporal frequencies, to different luminances, and to different contrasts and these responses were compared with physiological measurements available in the literature.

Spatiotemporal Filtering

The silicon retina's GCs respond to a restricted band of spatiotemporal frequencies, with transient cells displaying nonlinear spatial summation. In response to a drifting sinusoidal grating, spike trains from GCs of the same type differ significantly because of the cumulative effect of variability between transistors (CV = 20–25% for currents in identically sized and biased transistors). It was possible to obtain results that match physiological data by averaging responses from all cells in a given column, much as physiologists average several trials from the same cell.

The results reveal that at low temporal frequencies, both low and high spatial frequencies are attenuated in our silicon chip. However, sustained cells respond to a higher range of frequencies, as expected from their smaller receptive fields (Fig. 4A,B). Spatial frequency responses of ganglion cell activity (Fig. 4b) were fit with a balanced difference-of-Gaussians, an empirical model for retinal GCs *(35)*. Briefly, the frequency profile of a one-dimensional zero-mean inhibitory Gaussian was computed, with standard deviation σ_{Inh} and unit area, subtracted from a one-dimensional zero-mean excitatory Gaussian, with standard deviation σ_{Exc} and unit area. This frequency response was compared with the data and was optimized σ_{Exc} and σ_{Inh} to give the best fit.

When the phase of a contrast-reversing sinusoidal grating was varied, frequency doubling in transient cells was observed (Fig. 4C,D). This nonlinear summation is the fundamental distinction between narrow- and wide-field mammalian GCs *(20,36,37)* and arises because the bipolar cell signals are rectified before they are summed *(20)*.

Sustained cells in the silicon retina retain bandpass spatial filtering at all temporal frequencies (Fig. 5A,B). This pattern of spatiotemporal filtering matches the mammalian retina, except for a resonance found at very high temporal frequencies *(38)*. In the silicon retina, fast wide-field amacrine-cell modulation augments slow horizontal-cell inhibition to suppress low spatial frequencies, irrespective of whether they are presented at high or low temporal frequencies. And the optics and the cone–cone gap-junctions blur high spatial frequencies, also irrespective of temporal frequency. As a result, the sustained cells pass a restricted band of spatial frequencies at all temporal frequencies.

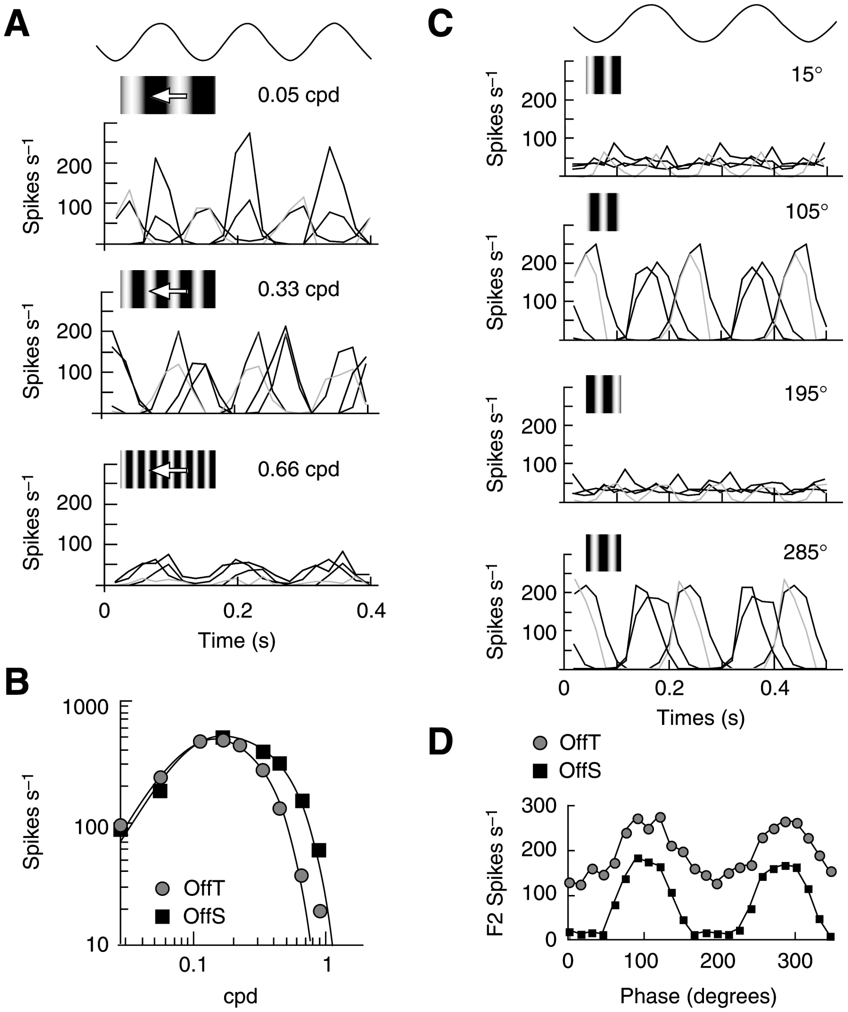

Fig. 4. Spatial filtering and nonlinear summation: **(A)** Varying spatial frequency: responses to 7.5 Hz horizontally drifting sinusoids with three different spatial frequencies. The responses are the strongest at an intermediate frequency, except for OnT cells, which showed an anomalous preference for low frequencies (*see* Fig. 2A for color code). **(B)** Spatial frequency tuning: OffT and OffS amplitudes are plotted for all spatial frequencies tested; they both peaked at 0.164 cpd, but OffS cells pass a higher range of frequencies. Solid lines are the best fit of a balanced difference-of-Gaussian model (OffT: $\sigma_{Exc}/\sigma_{Inh} = 0.20$; OffS: $\sigma_{Exc}/\sigma_{Inh} = 0.15$) *(35)*. **(C)** Varying spatial phase: responses to a 5 Hz 0.33 cpd contrast-reversing grating at four different spatial phases. Transient cells (yellow and blue) show frequency doubled responses at 15° and 195°. **(D)** Null Test: amplitudes of the second Fourier component (F2) of OffT and OffS responses are plotted for all phases tested. The sustained cells' F2 response disappeared at certain phases, but it could not be nulled in the transient cells. Fluctuations in F2 amplitude arise from uneven spatial sampling in the silicon retina *(30)*.

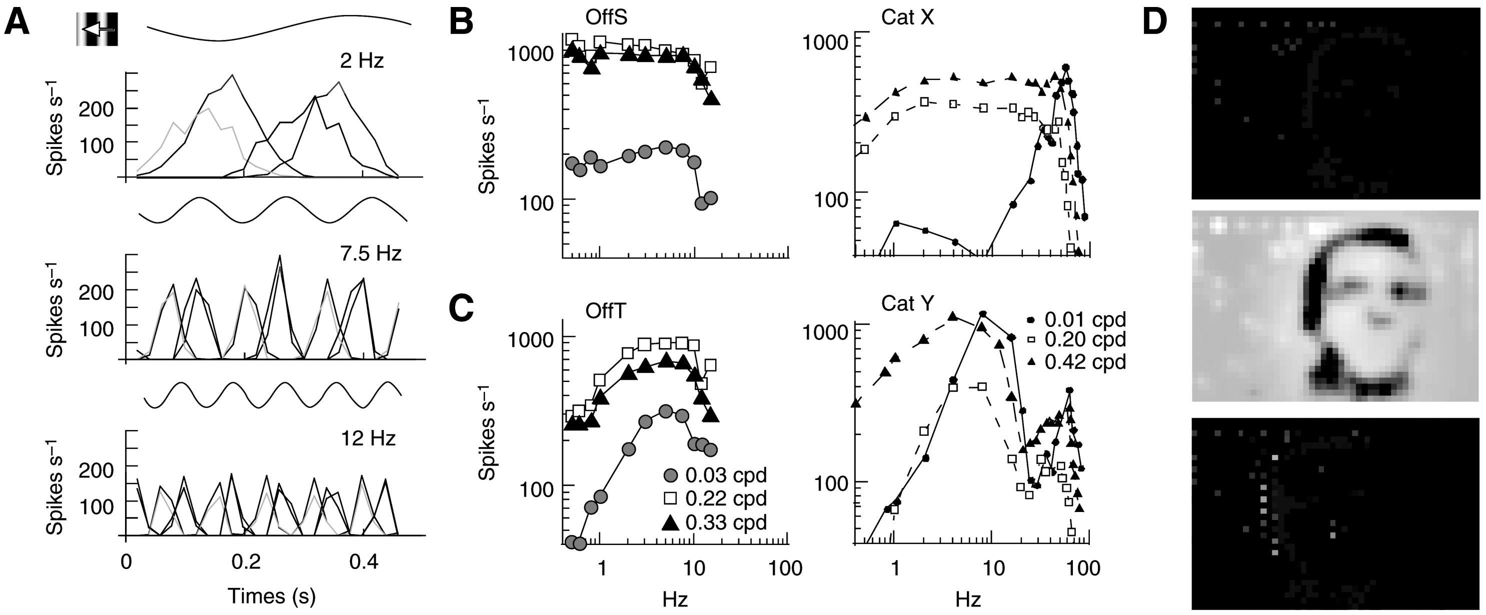

Fig. 5. Temporal filtering: **(A)** Varying temporal frequency: responses to 0.22 cpd horizontally drifting sinusoids at three different temporal frequencies. The response is the strongest at an intermediate frequency for transient cells (yellow and blue), whereas sustained cell (red and green) responses decline monotonically. **(B)** Sustained-cell temporal frequency tuning: responses of OffS and cat ON-center X-cells to low, medium, and high spatial-frequency sinusoidal gratings drifting horizontally at different temporal frequencies (*see* legend in C). Both pass all temporal frequencies less than 10 Hz, except at low spatial frequencies. However, the cat data display a high-frequency resonance (Cat data are reproduced from ref. *38*). The ordinate here and in **c** represents responsivity, which is the amplitude of the fundamental Fourier component divided by the stimulus contrast. **(C)** Transient-cell temporal frequency tuning: same as in B but for OffT cells and cat ON-center Y-cells. Both pass a restricted band of temporal frequencies at all spatial frequencies less than 0.33 cpd. **(D)** Responses to a face: in the static image (top), only sustained GCs respond. Reconstruction of the image from their activity (middle) demonstrates fidelity of retinal encoding. In the moving image (bottom), transient GCs respond as well, highlighting moving edges. The velocity of the image was approx 26.96°/s. The mean spike rate was 19 spikes/cell/s *(30)*.

On the other hand, the silicon retina's transient cells retain bandpass temporal filtering at all spatial frequencies (Fig. 5A,C). This pattern also matches the mammalian retina, except for the high-frequency resonance *(38)*. In the silicon retina, focussed narrow-field amacrine-cell inhibition augments diffuse horizontal-cell inhibition to suppress low temporal frequencies, irrespective of whether they are presented at high or low spatial frequencies. And the cone membrane's capacitance smears high temporal frequencies, also irrespective of spatial frequency. As a result, the transient cells pass a restricted band of temporal frequencies at all spatial frequencies.

The overall effect of spatiotemporal filtering is the best illustrated by natural stimuli (Fig. 5D). Bandpass spatial filtering in sustained GCs enhances edges in the static image. During rapid motion, bandpass temporal filtering in transient GCs captures this information with surprisingly little blurring. To confirm that the chip encodes visual information, the natural stimulus was reconstructed from the sustained ganglion cell spike activity. The visual image was reconstructed from the spikes by convolving ON and OFF sustained ganglion cell spike output with the same difference-of-Gaussian model, whose excitatory and inhibitory standard deviations were determined by the fit to the ganglion cell spatial frequency responses (σ_{Exc} and σ_{Inh}, Fig. 4B). The outputs of this convolution were passed through a temporal low-pass filter with a time constant of 22.7 ms, computing a new frame every 20 ms. The difference between images obtained from ON and OFF spikes was taken, and was displayed 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. Passing spike output through this simple spatiotemporal filter produces an image that is easily recognizable, even with only 30 × 48 pixels and just 0.4 spikes/cell/frame. This result suggests that cortical structures receiving input from such a visual prostheses can extract useful visual information from the silicon retina's neural code through simple linear filtering.

Light and Contrast Adaptation

The silicon retina's GCs adapt to mean luminance and encode stimulus contrast (Fig. 6). They maintain contrast sensitivity during at least one and a half decades of mean luminance. This intensity range was limited on the low end by leakage currents; the transistors pass a few pico-amps even when their gate voltage is zero. And it was limited on the high end by the projector (could not exceed 200 cd/m^2) in the experimental set-up and by stray photocurrents (light-induced leakage currents) in the silicon chip. To obtain the results presented here, the effect of these photocurrents was compensated by changing two externally applied voltages that would otherwise require no adjustment.

From the outer retina model, discussed in outer retina section, how the silicon model's cone activity depends on contrast was derived. The analytical result is fitted to the measured luminance adaptation curves (Fig. 6C), allowing *r* to increase with decreasing intensities as that data were obtained by exploiting the dependence of contrast sensitivity on horizontal-cell coupling (controlled globally in the circuit by V_{hh}), thereby the effect of stray photocurrents was being compensated. These photocurrents, which set an upper limit for the membrane time-constants that can be realized, reduce the silicon retina's sensitivity by speeding up ganglion-cell spike-frequency adaptation

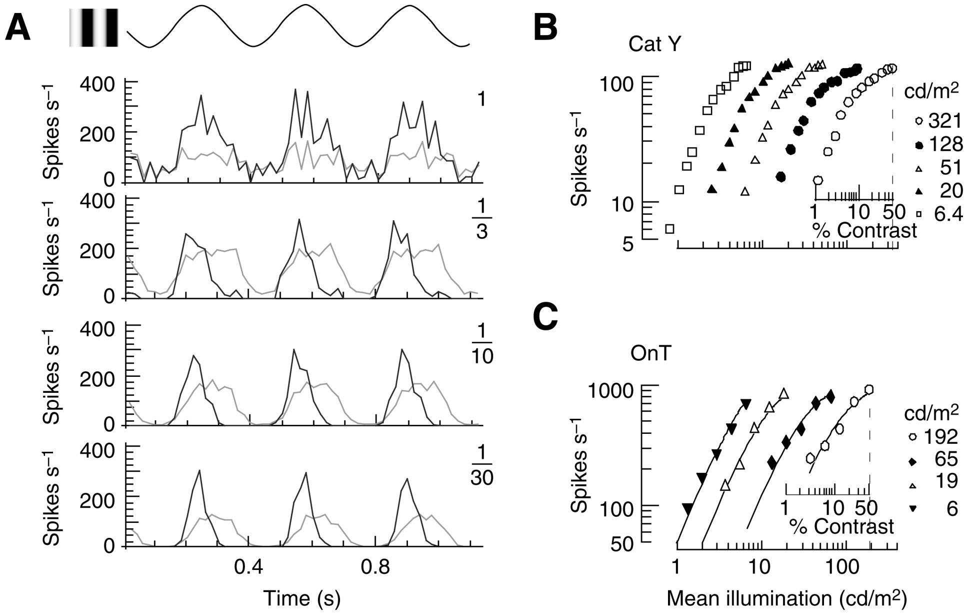

Fig. 6. Luminance adaptation: **(A)** Varying intensity: OnT and OnS responses to a sinusoidal grating (0.22 cpd) whose mean intensity was attenuated by amounts listed, using neutral density filters. Because of increases in sensitivity, the response amplitude hardly changes. The noisier responses at high intensity are due increased background activity, which tends to invoke synchronous firing because of cross-talk (i.e., ephatic interactions) in the silicon chip. **(B)** Cat On-center Y-cell intensity curves: the sinusoidal grating's (0.2 cpd) contrast varied from 1 to 50% and reversed at 2 Hz, for five mean luminances *(39)*. Here, and in C, response vs contrast (small *x*-axis) curves are shifted to align the 50% contrast response with that particular mean luminance (large *x*-axis). Mean luminance is converted from trolands to cd per m^2 based on a 5 mm diameter pupil (adapted from ref. *39*). **(C)** OnT intensity curves: the sinusoidal grating's (0.22 cpd) contrast varied from 3.25 to 50% and reversed at 3 Hz, for four different mean luminances. Solid lines represent the best fit of the equation given in the text for cone terminal activity; the fit indicates that the ratio of the horizontal cell and cone terminal space constants increased from 0.46 to 0.69, with a corresponding drop in peak spatial frequency from 0.22 to 0.16 cpd *(30)*.

and narrow-field amacrine-cell presynaptic inhibition as intensity increases. The latter effect was compensated by adjusting a second externally applied voltage that sets the narrow-field amacrine-cell's (baseline) membrane leakage.

The analytical fits to the measured luminance adaptation curves support the conclusion that the silicon retina's local modulation of synaptic strengths was largely responsible for adaptation. As mean luminance increased from 6 to 192 cd/m^2, the value assigned to the ratio between the HC and cone space constants, *r*, in the fitted equation decreased monotonically from 0.69 to 0.46 (unexpectedly, the best fits were found with *r* less than one; the corresponding reduction in peak spatial frequency response was from 0.22 to 0.16 cpd). Thus, even though we intervened by manually adjusting V$_{hh}$, the resulting change in *r*, a drop of only 36%, fell far short of what is required to reduce the sensitivity by a decade and a half. Overall, the silicon retina's ganglion-cell activity remained weakly correlated with absolute light intensity because of the residual effect

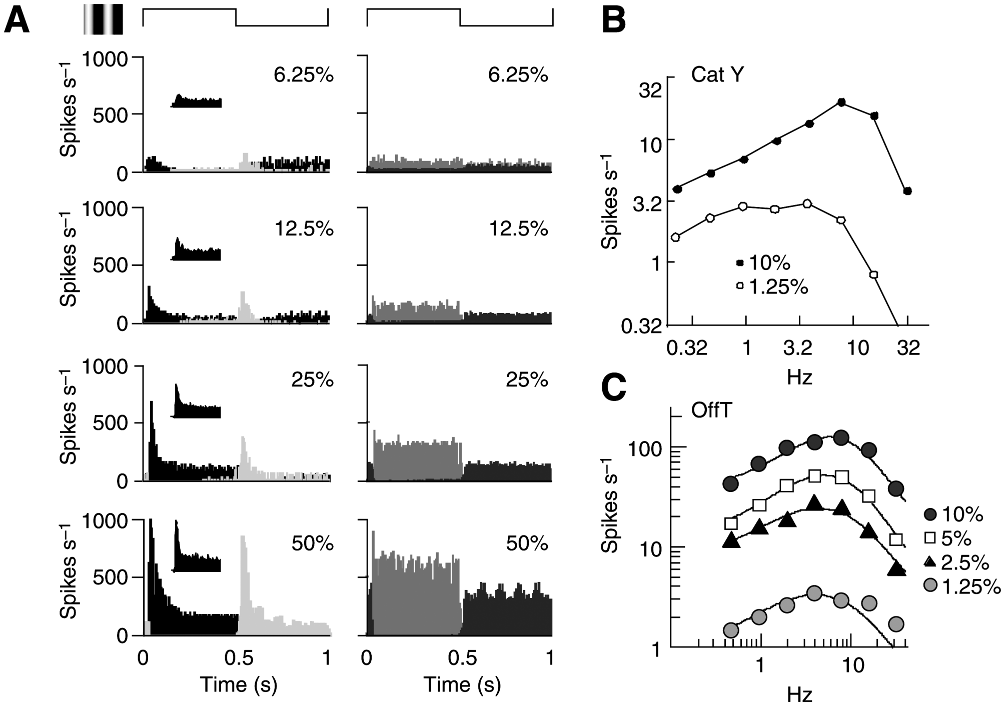

Fig. 7. Contrast gain control: **(A)** Varying contrast: responses to a 1 Hz square-wave contrast reversal of a sinusoidal grating (0.22 cpd) at four different peak stimulus contrasts. Bin width is 4 ms. Responses increase sublinearly and change more rapidly with increasing contrast; these effects are more pronounced in the transient cells. Their responses decayed with a time-constant that decreased from 28 to 22 ms as contrast increased from 6.25 to 50%. Inset: Response of ON-center cat X cell to half a cycle of the same stimulus *(29)*. **(B)** ON-center cat Y cell contrast-dependent temporal filtering: a stationary sinusoidal grating (0.25 cpd) whose contrast was determined by the sum of eight sinusoids was used. All eight sinusoids had the same amplitude whose value relative to the background is stated. The amplitude of the fundamental Fourier component at these eight frequencies was plotted for two different amplitudes (reproduced from ref. *28*). The peak sensitivity shifted from 3.9 to 7.8 Hz as the contrast increased from 1.25 to 10%. **(C)** OffT contrast-dependent temporal filtering: the stimulus was the same except that the spatial frequencywas reduced to 0.14 cpd and four different contrasts were tested. The peak sensitivity shifted from 3.9 to 7.8 Hz as the contrast increased from 1.25 to 10%. Solid lines are the best fit of an analytical model, which indicated that an increase in the strength of narrow-field amacrine-cell feedback inhibition from $w = 1.07$ to $w = 3.76$ could account for this change in temporal dynamics *(30)*.

of the stray photocurrents. Thus, as found in mammalian ganglion cell behavior *(39)*, responses at low contrasts are weaker at lower light intensities.

The silicon retina's GCs also adapt to temporal contrast. When presented with contrast-reversing gratings, the transient GCs respond more quickly, but more transiently with increasing contrast (Fig. 7A). And as the responses became even more transient, the peak-firing rate tended to saturate at the highest contrast levels. This adaptation is similar to the contrast gain control observed in mammalian narrow-field sustained *(29)* and wide-field transient GCs. However, it was not as dramatic in the silicon retina's sustained cells, whose responses did not decay nor saturate as much; they did, however, display a more rapid onset with increasing contrast. This difference between the silicon retina's sustained and transient GCs suggests that narrow-field amacrine cell feed-forward inhibition enhances contrast gain control by making the response more transient.

To better quantify the effect of contrast gain control *(28)*, the silicon retina's temporal frequency was measured, tuning at different contrasts, using a mixture of sinusoids with a flat spectrum. The OFF-Transient cells' peak response shifted to higher frequencies with increasing contrast, moving by an amount similar to that observed in the mammalian retina (Fig. 7B,C). But while this shift in tuning was accompanied by an overall strengthening of responses at all frequencies in the data, it was accompanied by preferential strengthening of high frequency responses in the cat data.

From the inner retina model, discussed in inner retina section, how the silicon model's ganglion cell responses depend on input contrast for a flat spectrum (i.e., white noise) was derived *(21)*. This stimulus was characterized by contrast per unit frequency *d*. The transient ganglion cell response, i_{Gt}, in spikes per second, is given by

$$i_{\mathrm{Gt}} = S\left|\left(d\,\frac{j\tau_A\omega + \varepsilon(1-g)}{j\tau_A\omega + 1}\right)\left(\frac{1}{j\tau_o\omega + 1}\right)^2\left(\frac{1}{j\tau_P\omega + 1}\right)\right|$$

where $\tau_A = \varepsilon\tau_{\mathrm{na}}$, $\varepsilon = 1/1(1 + \mathrm{wg})$. τ_{na} is the narrow-field amacrine cell's time constant, *g* is the gain from the bipolar terminal to the narrow-field amacrine cell, *w* is the wide-field amacrine cell-modulated strength of narrow-field amacrine cell inhibition onto the bipolar terminal, and *S* is a gain constant. The sustained ganglion cell response can be obtained by setting *g* to zero. The outer retina was approximated, using a lowpass temporal filter with time constant τ_o. The lowpass filtering behavior of the chip's photoreceptors was also included whose time-constant is τ_p. The four data sets are fitted (Fig. 7C) by allowing the system gain term, *S*, and the inhibition modulation *(w)* to vary across different stimulus contrasts, and fixed the remaining parameters.

The best fits of this model to the four input contrast densities (solid lines in Fig. 7C) support the conclusion that the silicon retina's wide-field-amacrine-mediated modulation of presynaptic inhibition was responsible for contrast gain control. The parameter values for these fits were $\tau_p = 33$ ms, $\tau_o = 77$ ms, $\tau_{\mathrm{na}} = 1.0382$ s, and $g = 1.07$. The system gain, *S*, saturated over the four contrasts (352–1358 to 1711–1929), whereas the inhibition strength *(w)* increased exponentially (1.07–1.51 to 2.38–3.76). Thus, the inhibition strength increased with contrast as it was expected, but the system gain was not constant because of a static nonlinearity *(40)*, canceling part of the expected decrease in sensitivity.

CONCLUSIONS

By autonomously extracting the same visual information encoded by the mammalian retina at a similar physical scale and energy efficiency, the silicon retina satisfies the criteria for a fully implantable ocular prosthesis. The device approximates the behavior of the mammalian retina in both linear response and nonlinear adaptation. This success validates the neuromorphic modeling approach for advanced prosthetic applications. In addition, this real-time silicon model may be useful in both further testing specific hypotheses about the retina and in serving as a realistic retinal front-end to other downstream processes like cortical models, other artificial neural systems, or robots.

The approach in constructing this silicon retina was to model synaptic connections found in the mammalian retina and to implement them using transistor primitives. For example, reciprocal inhibition between bipolar and amacrine cells in complementary On and Off channels in the model mimics vertical inhibition between On and Off laminae *(24)* and serial inhibition found between amacrine cells *(41)* in the mammalian retina. Furthermore, the model proposes an anatomical substrate for computing Victor and Shapley's "neural measure of contrast" *(28,29)*, suggesting that wide-field amacrine cells play this role in the mammalian retina. Whereas this hypothesis remains to be tested experimentally, there are instances in which the model oversimplifies the functional architecture of the mammalian retina. For instance, dopaminergic amacrine cells *(42)*, and light sensing GCs *(43)*, are likely important in modulating mammalian retinal cone and horizontal gap-junction conductance. These dopaminergic cells are not included in the model, and instead the local horizontal-cell signal is relied on to modulate cone coupling.

Capturing the mammalian retina's synaptic organization proved sufficient to recreate its computations in the silicon retina, although there were some specific quantitative differences of note, which arose from technological limitations. The device's sensitivity dropped when luminance was high and its background firing rates increased. These deficiencies can be addressed by reducing photo-induced leakage currents, which tend to speed up inhibition and firing rates in the silicon retina GCs, and by reducing crosstalk between silicon GCs, which produced high background activity when bright or large stimuli were presented. The device's sustained GCs also displayed little or no contrast gain control. A faster photodetector should rectify this situation, which appears to be related to the absence of any significant transient component in the sustained GCs' responses. This speed-up will lead to a transient component at the bipolar terminal, and hence in the sustained GCs.

The goal of designing and fabricating a silicon retina was realized that can operate and adapt independent of external control to a large extent, but there remains some degree of manual intervention necessary to make the device work properly. Whereas the voltages applied to the biases that set mean cone terminal activity, mean bipolar terminal activity, mean ganglion cell activity, and coupling strength in wide-field amacrine cells remained fixed, the voltages applied to the biases had to be manually adjusted that set the coupling strength between HCs and the bias that sets the narrow-field amacrine cell leakage current to compensate for light-dependent leakage currents, a shortcoming of silicon microtechnology at these small length scales. However, this fine tuning was only required for light adaptation. Any bias voltages was not adjusted whatsoever during any of the other experiments. Hence, addressing the leakage issue will make it possible to hard-wire all of the silicon retina's external biases to specific voltages, as required by a final prosthetic application.

The artificial retina satisfies the requirements of a neural prosthesis by matching the biological retina in size and weight and using under a tenth of a watt. Rabbit retina uses 16.2 nW per ganglion cell (82 μmoles of ATP g per min *[44]*, or 88 mW g/min, times 70 mg average weight, divided by 380,000 GCs *[12]*). In contrast, the chip consumes

17 µW per ganglion cell (62.7 mW for the entire chip) at an average spike rate of 45 spikes s per ganglion cell. Although, this energy consumption is one thousand times less efficient than the mammalian retina, it still represents a 100-fold improvement over conventional microprocessors. A 1 GHz Pentium® processor operating at 10 W would dissipate 2.2 mW per ganglion cell to compute the response of a13 × 13 × 13 kernel (X × Y × T) updated at 100 times/s. With an upper limit on a proposed intraocular implant's power dissipation of 100 mW, the Pentium® could thus only compute the responses of under 40 GCs, or a 6 by 6 array, which is too small for functional vision *(5)*. However, with the same 100 mW limit, our neuromophic chip's energy efficiency allows it to compute the responses of 4000 GCs, roughly a 60 × 60 array. This energy efficiency is expected to improve further, together with spatial resolution and dynamic range, as microfabrication technology advances.

In conclusion, based on detailed knowledge of the retina's neuronal specializations, synaptic organization, and functional architecture *(45)*, thirteen neuronal types have been constructed in silicon and linked them together in two synaptic layers on a physical scale comparable with the human retina. Furthermore, a silicon retina has been created that modulates its synaptic strengths locally. The silicon retina realizes luminance adaptation, without using logarithmic compression, and contrast gain control independent of external control, thus, capturing properties of retinal neural adaptation for the first time. The success modeling neural adaptation using single-transistor primitives suggests that a similar approach could be used to morph other neural systems into silicon as well; this may eventually lead to fully implantable neural prostheses *(46,47)* that do not require external interfaces.

REFERENCES

1. Sterling P. Retina. In: Shepherd GM, ed. Synaptic organization of the brain*,* 4th ed., New York: Oxford University Press, NY, 1998.
2. van Hateren J. A theory of maximizing sensory information. Biol Cybernetics 1992;68:23–29.
3. Rizzo JFr, Wyatt J, Humayun M, et al. Retinal prosthesis: An encouraging first decade with major challenges ahead. Opthalmology 2001;108:13–14.

4 .Margalit E, Maia M, Weiland JD, et al. Retinal prosthesis for the blind. Surv of Ophthalmol 2002;47:335–356.

5. Humayun MS, de Juan EJ, Weiland JD, et al. Pattern electrical stimulation of the human retina. Vision Res 1999;39:2569–2576.
6. Chow AY, Pardue MT, Chow VY, et al. Implantation of silicon chip microphotodiode arrays in the cat subretinal space. IEEE Trans Neural Syst Rehabil Eng 2001;9:86–95.
7. Normann RA, Maynard EM, Rousche PJ, Warren DJ. A neural interface for a cortical vision prosthesis. Vision Res 1999;39:2577–2587.
8. Dobelle WH. Artificial vision for the blind by connecting a television camera to the visual cortex. ASAIO J 1999;46:3–9.
9. Mead CA. Analog VLSI and neural systems, Addison Wesley, Reading, MA, 1989.
10. Mahowald M, Mead C. A silicon model of early visual processing. Neural Networks 1988;1.
11. Baccus SA, Meister M. Fast and slow contrast adaptation in retinal circuitry. Neuron 1997;36:909–919.
12. Masland R. The fundamental plan of the retina. Nature Neurosci 2001;4:877–886.
13. Boahen K. A retinomorphic chip with parallel pathways: Encoding Incresing, On, Decreasing, and Off visual signals. J Analog Integrated Circtuits Signal Processing 2001;30(2).

14. Boahen KA, Andreou AG. A contrast sensitive silicon retina with reciprocal synapses. In: Moody JE, Hanson SJ, Lippmann RP, eds. Advances in neural information processing systems. Vol. 4, Morgan Kaufmann, San Mateo, CA, 764–772.
15. Kamermans M, Fahrenfort I, Schultz K, Janssen-Bienhold U, Sjoerdsma T, Weiler R. Hemichannel-mediated inhibition in the outer retina. Science 2001;292:1178–1180.
16. Kamermans M, Werblin F. GABA-mediated positive autofeedback loop controls horizontal cell kinetics in tiger salamander retina. J Neurosci 1992;12:2451–2463.
17. Kuffler SW. Discharge patterns and functional organization of mammalian retina. J Neurophysiol 1953;16:37–68.
18. Werblin FS, Dowling JE. Organization of the retina of the mudpuppy, Necturus maculosus. II. Intracellular recording. J Neurophyiol 1969;32:339–355.
19. Rodieck RW. The primate retina. Comp Primate Biol 1988;4:203–278.
20. Enroth-Cugell C, Freeman AW. The receptive field spatial structure of cat retinal Y cells. J Physiol 1987;384:49–79.
21. Zaghloul KA, Boahen K. Optic nerve signals in a neuromorphic chip I: Outer and inner retina models. IEEE Trans Biomed Eng 2004;51:657–666.
22. Normann RA, Perlman I. The effect of background illumination on the photoresponses of rod and green cones. J Physiol 1979;286:491–507.
23. Tsividis YP. Operation and modeling of the MOS transistor. New York:McGraw-Hill Book Company, 1987.
24. Roska B, Werblin F. Vertical interactions across ten parallel, stacked representations in the mammalian retina. Nature 2001;410:583–587.
25. Maguire G, Lukasiewicz P. Amacrine cell interactions underlying the response to change in tiger salamander retina. J Neurosci 1989;9:726–735.
26. Kolb H, Nelson R. OFF-alpha and OFF-beta ganglion cells in cat retina: II. Neural circuitry as revealed by electron microscopy of HRP stains. J Comp Neurol 1993;329:85–110.
27. Freed MA, Pflug R, Kolb H, Nelson R. On-Off amacrine cells in cat retina. J Comp Neur 1996;364:556–566.
28. Shapley R, Victor JD. The contrast gain control of the cat retina. Vision Res 1979;19:431–434.
29. Victor JD. The dynamics of cat retinal X cell centre. J Physiol 1987;386:219–246.
30. Zaghloul KA, Boahen K. An On-Off log domain circuit that recreates adaptive filtering in the retina. IEEE Trans Circuits Syst 2005;52:99–107.
31. Boahen KA. The retinomorphic approach: Pixel-parallel adaptive amplification, filtering, and quantization. J Analog Integrated Circtuits Signal Processing 1997;13:53–68.
32. Smith RG. Simulation of an anatomically defined local circuit—The cone-horizontal cell network in cat retina. Visual Neurosci 1995;12:545–561.
33. Curcio CA, Sloan KR, Kalina RE, Hendrickson AE. Human photoreceptor topography. J Comp Neur 1990;292:497–523.
34. Boahen KA. Point-to-point connectivity between neuromorphic chips using address-events. IEEE Trans Circuits Syst 1999;47:100–116.
35. Rodieck R. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision Res 1965;5:583–601.
36. Enroth-Cugell C, Robson JG. The contrast sensitivity of retinal ganglion cells of the cat. J Physiol 1966;187:517–552.
37. Demb JB, Zaghloul KA, Haarsma L, Sterling P. Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina. J Neurosci 2001; 21:7447–7454.
38. Frishman LJ, Freeman AW, Troy JB, Schweitzer-Tong DE, Enroth-Cugell C. Spatiotemporal frequency responses of cat retinal ganglion cells. J Gen Physiol 1987; 89:599–628.

39. Troy JB, Enroth-Cugell C. X and Y ganglion cells inform the cat's brain about contrast in the retinal image. Exp Brain Res 1993;93:383–390.
40. Zaghloul KA, Boahen K. Optic nerve signals in a neuromorphic chip II: Testing and results. IEEE Trans Biomed Eng 2004;51:667–675.
41. Dowling JE, Boycott BB. Organization of the primate retina: electron microscopy. Proc R Soc Lond B 1966;166:80–111.
42. Jensen RJ, Daw NW. Effects of dopamine and its agonists and antagonists on the receptive field properties of ganglion cells in the rabbit retina. Neuroscience 1986;17:837–855.
43. Berson DM, Dunn FA, Takao M. Phototransduction by retinal ganglion cells that set the circadian clock. Science 2002;295:1070–1073.
44. Ames A, Li YY, Heher EC, Kimble CR. Energy metabolism of rabbit retina as related to function: high cost of Na^+ transport. J Neurosci 1992;12:840–853.
45. Freed MA, Sterling P. The ON-alpha ganglion cell of the cat retina and its presynaptic cell types. J Neurosci 1988;8:2303–2320.
46. Craelius W. The bionic man: Restoring mobility. Science 2002;295:1018–1021.
47. Zrenner E. Will retinal implants restore vision. Science 2002;295:1022–1025.