CHAPTER

15

SILICON RETINA

M. A. Mahowald Carver Mead

The retina is a thin sheet of neural tissue that partially lines the orb of the eye. This tiny outpost of the central nervous system is responsible for collecting all the visual information that reaches the brain. Signals from the retina must carry reliable information about properties of objects in the world over many orders of magnitude of illumination.

The high degree to which a perceived image is independent of the absolute illumination level is, in large part, a result of the initial analog stages of retinal processing, from the photoreceptors through the outer-plexiform layer. This processing relies on lateral inhibition to adapt the system to a wide range of viewing conditions, and to produce an output that is independent of the absolute illumination level. A byproduct of the mechanism of lateral inhibition is the enhancement of spatial edges in the image; in signal-processing terms, the operation performed by the chip resembles a Laplacian filter.

We have built a silicon retina that is modeled on the distal portion of the vertebrate retina. This chip generates, in real time, outputs that correspond directly to signals observed in the corresponding levels of biological retinas. The chip design uses the principles of signal aggregation discussed in Chapter 7. It demonstrates a tolerance for device imperfections that is characteristic of a collective system.

Portions reprinted with permission from *Neural Networks*, Vol. 1. Mead, C. and Mahowald, M.A. A Silicon Model of Early Visual Processing. © 1988, Pergamon Press.

RETINAL STRUCTURE

A thorough review of the biological literature up to 1973 is given in *The Vertebrate Retina* [Rodieck, 1973], and one of more recent work is presented in *The Retina: An Approachable Part of the Brain* [Dowling, 1987]. Although the details of each animal's anatomy are unique, the gross structure of the retina has been conserved throughout the vertebrates.

The major divisions of the retina are shown in cross-section in Figure 15.1. Light is transduced into an electrical potential by the photoreceptors at the top. The primary signal pathway proceeds from the photoreceptors through the **triad synapses** to the **bipolar cells**, and thence to the **retinal ganglion cells**, the output cells of the retina. This pathway penetrates two dense layers of neural processes and associated synapses. The **horizontal cells** are located just below the photoreceptors, in the **outer-plexiform layer**. The **inner-plexiform layer**, just above the ganglion cell bodies, contains amacrine cells. The horizontal and **amacrine cells** spread across a large area of the retina, in layers transverse to the primary signal flow.

Each cell type in the retina has unique characteristics. The horizontal cells, along with the photoreceptors and bipolar cells, represent information with smoothly-varying analog signals. The amacrine cells, which are concerned with extracting motion events, have active channels in their processes that propagate temporally sharp signals in response to broad inputs. The amacrine and horizontal cells are examples of *axonless* neurons. They receive inputs and generate outputs along the same neuronal processes. In contrast, the retinal ganglion cell possesses distinct dendrites and an axon that produces action potentials that are quasidigital (digital in amplitude but analog in time). The two-dimensional sheet of the retina is a marvelous canvas painted with the rich palate of biophysics. The diversity of cell types and computations demonstrates the power of combining a small number of physical elements in a hierarchical structure. Modeling the retina in silicon, we hope to develop a repertoire of computations based on the physics of the medium.

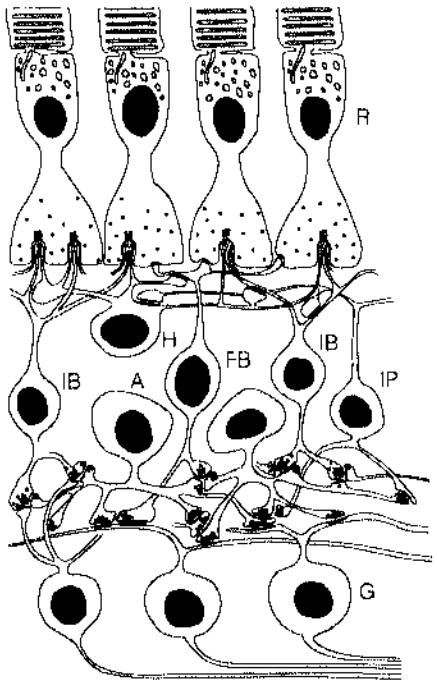


FIGURE 15.1 Artist's conception of a cross-section of a primate retina, indicating the primary cell types and signal pathways. The outer-plexiform layer is beneath the foot of the photoreceptors. The invagination into the foot of the photoreceptor is the site of the triad synapse. In the center of the invagination is a bipolar-cell process, flanked by two horizontal cell processes. R: photoreceptor, H: horizontal cell, IB: invaginating bipolar cell, FB: flat bipolar cell, A: amacrine cell, IP: interplexiform cell, G: ganglion cell. (*Source:* Adapted from [Dowling, 1987, p. 18].)

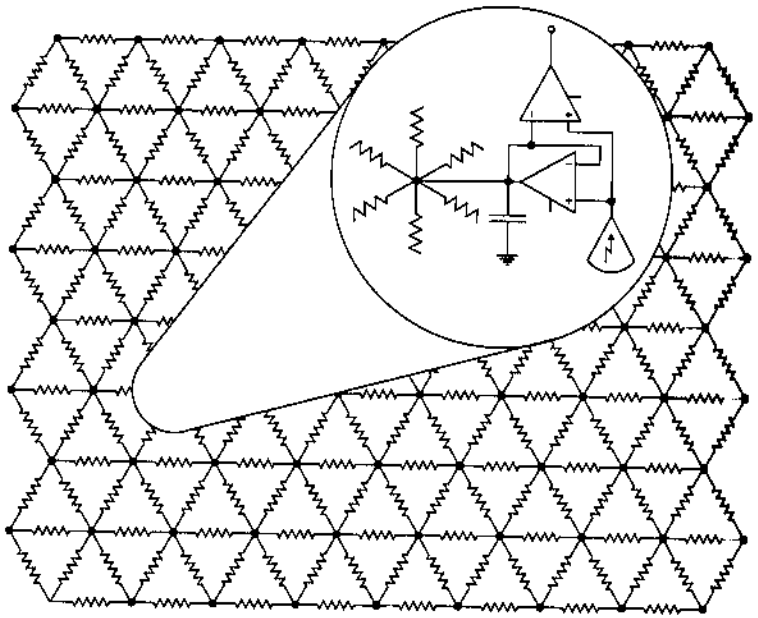


FIGURE 15.2 Diagram of the silicon retina showing the resistive network; a single pixel element is illustrated in the circular window. The silicon model of the triad synapse consists of a follower-connected transconductance amplifier by which the photoreceptor drives the resistive network, and an amplifier that takes the difference between the photoreceptor output and the voltage stored on the capacitance of the resistive network. These pixels are tiled in a hexagonal array. The resistive network results from a hexagonal tiling of pixels.

We begin with a simple model of the analog processing that occurs in the distal portion of the retina. Because our model of retinal processing is implemented on a physical substrate, it has a straightforward structural relationship to the vertebrate retina. A simplified plan of the silicon retina is shown in Figure 15.2. This view emphasizes the lateral spread of the resistive network, corresponding to the horizontal cell layer. The primary signal pathway proceeds through the photoreceptor and the circuitry representing the bipolar cell shown in the inset. The image signal is processed in parallel at each node of the network.

The key processing element in the outer-plexiform layer is the *triad synapse*, which is found in the base of the photoreceptor. The triad synapse is the point of contact among the photoreceptor, the horizontal cells, and the bipolar cells. We can describe our model of the computation performed at the triad synapse in terms of the synapse's three elements:

1. The photoreceptor takes the logarithm of the intensity
2. The horizontal cells form a resistive network that spatially and temporally averages the photoreceptor output
3. The bipolar cell's output is proportional to the difference between the photoreceptor signal and the horizontal cell signal

We will describe these elements in detail in the following sections.

Photoreceptor Circuit

The primary function of the **photoreceptor** is to transduce light into an electrical signal. For intermediate levels of illumination, this signal is proportional to the logarithm of the incoming light intensity. The logarithmic nature of the output of the biological photoreceptor is supported by psychophysical and electrophysiological evidence. Psychophysical investigations of human visual-sensitivity thresholds show that the threshold increment of illumination for detection of a stimulus is proportional to the background illumination over several orders of magnitude [Shapley et al., 1984]. Physiological recordings show that the photoreceptors' electrical response is logarithmic in light intensity over the central part of the photoreceptors' range, as are the responses of other cells in the distal retina [Rodieck, 1973]. The logarithmic nature of the response has two important system-level consequences:

1. An intensity range of many orders of magnitude is compressed into a manageable excursion in signal level.
2. The voltage difference between two points is proportional to the **contrast ratio** between the two corresponding points in the image. In a natural image, the contrast ratio is the ratio between the reflectances of two adjacent objects, reflectances which are independent of the illumination level.

The silicon photoreceptor circuit consists of a **photodetector**, which transduces light falling onto the retina into an electrical photocurrent, and a logarithmic element, which converts the photocurrent into an electrical potential proportional to the logarithm of the local light intensity. Our photodetector is a **vertical bipolar transistor**, which occurs as a natural byproduct in the CMOS process described in Appendix A. The base of the transistor is an isolated section of well, the emitter is a diffused area in the well, and the collector is the substrate. Photons with energies greater than the band gap of silicon create electron–hole pairs as they are absorbed. Electrons are collected by the n-type base of the pnp phototransistor, thereby lowering the energy barrier from emitter to base, and

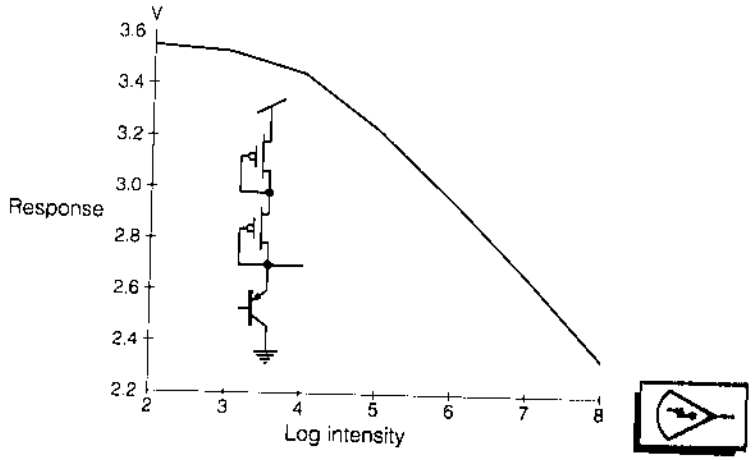


FIGURE 15.3 Measured response of a logarithmic photoreceptor. Photocurrent is proportional to incident-light intensity. Response is logarithmic over more than four orders of magnitude in intensity. Direct exposure of the chip to room illumination resulted in an output voltage of 2.1 volts. The symbol for the photoreceptor circuit is shown in the inset

increasing the flow of holes from emitter to collector. The gain of this process is determined by the number of holes that can cross the base before one hole recombines with an electron in the base. The photodetector in our silicon photoreceptor produces several hundred electrons for every photon absorbed by the structure.

The current from the photodetector is fed into two diode-connected MOS transistors in series. The photocurrent biases these transistors in the subthreshold region. This arrangement was described in Chapter 6; it produces a voltage proportional to the logarithm of the current, and therefore to the logarithm of the incoming intensity. We use two transistors to ensure that, under normal illumination conditions, the output voltage will be within the limited allowable voltage range of the resistive network. Even so, at very low light levels, the output voltage of the photoreceptor may be close enough to V_{DD} that the resistor bias circuit described in Chapter 7 cannot adequately bias the horizontal resistive connections.

The voltage out of this photoreceptor circuit is logarithmic over four to five orders of magnitude of incoming light intensity, as shown in Figure 15.3. The lowest photocurrent is about 10^{-14} amps, which translates to a light level of 10^5 photons per second. This level corresponds approximately to a moonlit scene focused on the chip through a standard camera lens, which is about the lowest illumination level visible to the cones in a vertebrate retina.

Horizontal Resistive Layer

The retina provides an excellent example of the computation that can be performed using a resistive network. The horizontal cells in most species are connected to one another by gap junctions to form an electrically continuous

network in which signals propagate by electrotonic spread [Dowling, 1987]. The lateral spread of information at the outer-plexiform layer is thus mediated by the resistive network formed by the horizontal cells. The voltage at every point in the network represents a spatially weighted average of the photoreceptor inputs. The farther away an input is from a point in the network, the less weight it is given. The horizontal cells usually are modeled as passive cables, in which the weighting function decreases exponentially with distance.

The properties of passive resistive networks were described in Chapter 7; additional details are given in Appendix C. Our silicon retina includes one such network, patterned after the horizontal cells of the retina. Each photoreceptor in the network is linked to its six neighbors with resistive elements, to form the hexagonal array shown in Figure 15.2. Each node of the array has a single bias circuit to control the strength of the six associated resistive connections. The photoreceptors act as voltage inputs that drive the horizontal network through conductances. This method of providing input to a resistive network is shown in Figure 7.12 (p. 120). By using a wide-range amplifier in place of a bidirectional conductance, we have turned the photoreceptor into an effective voltage source. No current can be drawn from the output node of the photoreceptor, because the amplifier input is connected to only the gate of a transistor.

The horizontal network computes a spatially weighted average of photoreceptor inputs. The spatial scale of the weighting function is determined by the product of the lateral resistance and the conductance coupling the photoreceptors into the network. Varying the conductance of the wide-range amplifier or the strength of the resistors changes the space constant of the network, and thus changes the effective area over which signals are averaged.

Both biological and silicon resistive networks have associated parasitic capacitances. The fine unmyelinated processes of the horizontal cells have a large surface-to-volume ratio, so their membrane capacitance to the extracellular fluid will average input signals over time as well as over space. Our integrated resistive elements have an unavoidable capacitance to the silicon substrate, so they provide the same kind of time integration as do their biological counterparts. The effects of delays due to electrotonic propagation in the network are most apparent when the input image changes suddenly.

Triad Synapse Computation

The receptive field of the bipolar cell shows an antagonistic center-surround response [Werblin, 1974]. The center of the bipolar cell receptive field is excited by the photoreceptors, whereas the antagonistic surround is due to the horizontal cells. The triad synapse is thus the obvious anatomical substrate for this computation. In our model, the center-surround computation is a result of the interaction of the photoreceptors, the horizontal cells, and the bipolar cells in the triad synapse.

The output of our silicon retina is analogous to the output of a bipolar cell in a vertebrate retina. Our triad synapse consists of two elements (Figure 15.2):

1. A wide-range amplifier provides a conductance through which the resistive network is driven toward the photoreceptor output potential
2. A second amplifier senses the voltage difference across the conductance, and generates an output proportional to the difference between the photoreceptor output and the network potential at that location

The output of our bipolar cell thus represents the difference between a center intensity and a weighted average of the intensities of surrounding points in the image.

IMPLEMENTATION

The floorplan for the retina is shown in Figure 15.4. The chip consists of an array of pixels, and a scanning arrangement for reading the results of retinal processing. The output of any pixel can be accessed through the scanner, which is made up of a vertical scan register along the left side of the chip and a horizontal scan register along the bottom of the chip. Each scan-register stage has 1-bit of

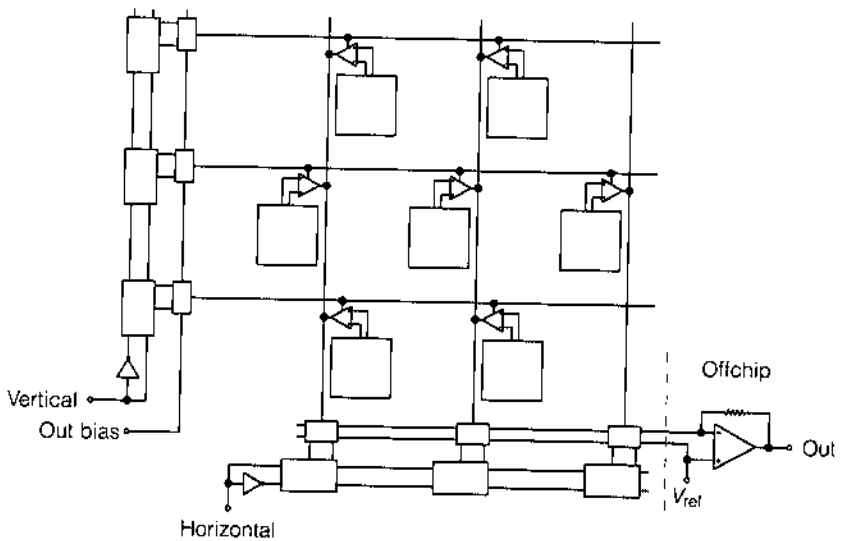


FIGURE 15.4 Layout of the retina chip. The main pixel array is made up of alternating rows of rectangular tiles, arranged to form a hexagonal array. The scanner along the left side allows any row of pixels to be selected. The scanner along the bottom allows the output current of any selected pixel to be gated onto the output line, where it is sensed by the offchip current-sensing amplifier. A schematic illustration of the shift-register cell used in the scanners is shown in Figure 15.6. The horizontal and vertical switching circuits are shown in Figure 15.7. The complete schematic of the pixel is shown in Figure 15.5. In the completed chip, the array was 48 by 48 pixels in extent. A fabrication layout of a small version of this chip is presented on the inside of the front cover.

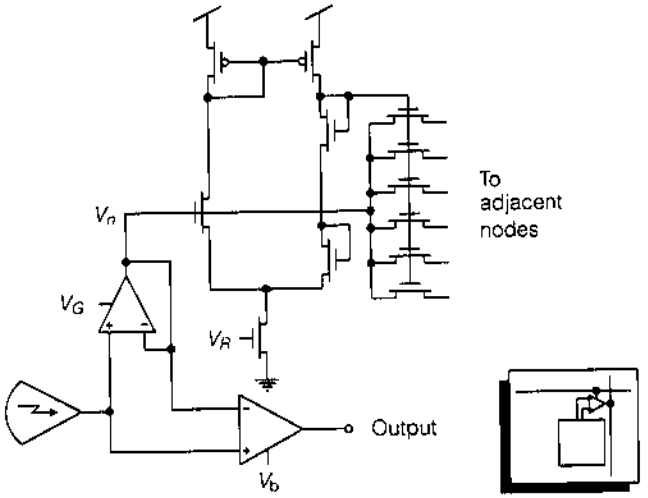


FIGURE 15.5 Detailed schematic illustration of all circuitry within an individual pixel of the silicon retina. The logarithmic photoreceptor circuit is shown in the lower-left corner. The follower-connected transconductance amplifier forms the G conductance for that node of the resistive network. The six pass transistors form this end of the resistive connections to the six neighboring pixels. The single bias circuit, described in Figure 7.10 (p. 118), is shared among all six pass transistors. The output amplifier does not need to have a wide-range design, because the output line can be run at a constant potential, near V_{DD}. In the design described in this chapter, each pixel was 109 λ wide by 97 λ high. The symbol for the pixel circuit is shown in the inset.

shift register, with the associated signal-selection circuits. Each register normally is operated with a binary 1 in the selected stage, and binary 0s in all other stages. The selected stage of the vertical register connects the *out-bias* voltage to the horizontal scan line running through all pixels in the corresponding row of the array. The deselected stages force the voltage on their horizontal scan lines to ground. Each horizontal scan line is connected to the bias control (V_b) of the output amplifiers of all pixels in the row. The output of each pixel in a selected row is represented by a current; that current is enabled onto the vertical scan line by the V_b bias on the horizontal scan line. The current scale for all outputs is set by the out-bias voltage, which is supplied from offchip. A schematic diagram of all circuits in the pixel is shown in Figure 15.5.

The shift-register stage is made with complementary set–reset logic (CSRL) [Mead et al., 1985], and is shown in Figure 15.6. Signals are two-rail; the data value is represented on the top rail, and its complement is represented on the bottom rail. As ϕ_2 rises, the power supply to the second pair of cross-coupled inverters is limited by the upper p-channel power-down transistor. By the time ϕ_2 reaches the threshold of the two pass transistors, the current to the second pair of cross-coupled inverters is limited to about one-half of the maximum that can be supplied when the clock is low. For this reason, the first stage, which is

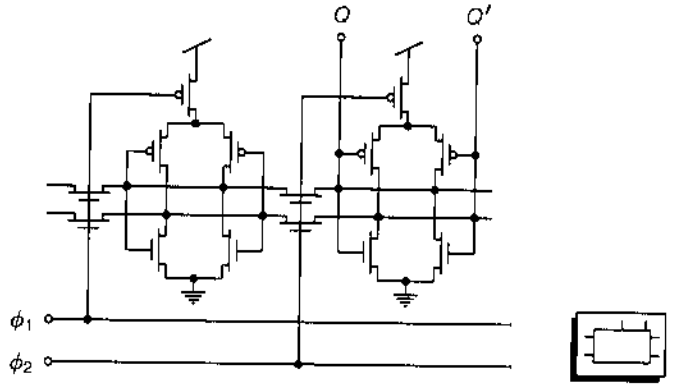


FIGURE 15.6 Shift-register stage used in the horizontal and vertical scanner. Each stage consists of a pair of cross-coupled inverters, which have a p-channel transistor in series with their common power supply. Because the output of each inverter feeds the input of the other, the cross-coupled stage has two stable states: (1) a binary 1 is stored on the top rail and a 0 on the bottom rail; or (2) a binary 0 is stored on the top rail and a 1 on the bottom rail.

On the rising edge of ϕ_2, the power to the second cross-coupled stage is cut off, and the data from the previous stage are able to pass through the pass transistors. When ϕ_2 falls, power is restored, and the data are held statically. A similar transfer occurs to the following stage during ϕ_1.

fully powered up, can force its state through the pass transistors into the second stage on the rising edge of ϕ_2. Similarly, the second stage can transfer its contents into the following stage on the rising edge of ϕ_1. Unidirectionality of the transfer is guaranteed because the clock signals are used both for the pass transistors and for the power-down transistor of the receiving stage. The clocks must be nonoverlapping.

We can envision the dynamics of a transfer by considering the interesting case, when the new datum is a 0 (upper rail low) and the previously stored datum is a 1 (upper rail high). While ϕ_2 is high, the upper rail discharges toward ground and the lower rail charges toward V_{DD}. The clock rates are limited by the capacitances of the nodes and the resistances of the pass transistors. After some time, the two voltages cross over. The clock ϕ_2 may return to zero any time after the crossover occurs, and a successful transfer will result. There is no need to wait for the two signal rails to pass any absolute threshold. Each cross-coupled stage can be viewed as a sense amplifier with very high differential gain. The positive-feedback action will fully restore both datum and complement, as long as the *relative* values of these two signals are of correct sign when the clock falls. If we attempt to bring the clock low before the crossover time at which the two signals become equal, the signals return to their previous values, and the transfer fails. The crossover time thus represents the minimum time during which the clock must be high, and thereby limits the maximum frequency of operation. At all times, at least one clock is low, so the CSRL shift register is fully static.

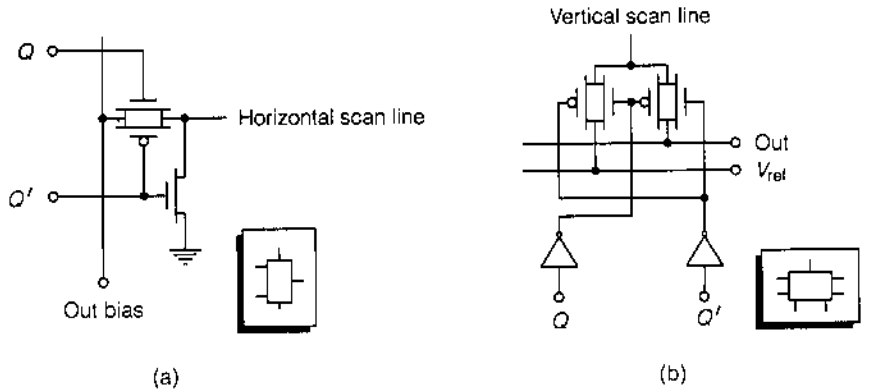


FIGURE 15.7 (a) Schematic diagram of the driver for a horizontal scan line. (b) Multiplexor for a vertical scan line. A binary 1 in the vertical scan register gates the *out-bias* voltage onto the selected row, while the scan lines to the other rows are held at ground. A 1 in the horizontal scan register gates the output of the corresponding column onto the output line; all other output lines are connected to the reference line.

The circuits associated with driving a horizontal scan line and selecting data from a vertical scan line are shown in Figure 15.7. The current in a vertical scan line is connected to one of two output lines through a pair of complementary pass-transistor analog switches. If a binary 1 is stored in the corresponding stage of the horizontal shift-register, the vertical scan line is connected to the line labeled *out*. If a binary 0 is stored in the stage, the vertical scan line is connected to the line labeled V_{ref}. The current from the selected column thus flows in the *out* line, and the current from all unselected columns flows in the V_{ref} line. The chip is designed to be used with the off-chip current-sense amplifier shown to the right of the broken line in Figure 15.4. The *out* line is held at the V_{ref} potential by negative feedback from the amplifier output through the resistor. The principal advantage of this arrangement is that all vertical scan lines—selected and unselected—are held at the same potential. Thus, no transient is introduced as the vertical scan line is selected. In addition, capacitive transients due to the charge in the pass-transistor channels are minimized by the complementary nature of the analog switches [Sivilotti et al., 1987].

The scanners can be operated in one of two modes: static probe or serial access. In static-probe mode, a single row and column are selected, and the output of a single pixel is observed as a function of time, as the stimulus incident on the chip is changed. In serial-access mode, both vertical and horizontal shift registers are clocked at regular intervals to provide a sequential scan of the processed image for display on a television monitor. A binary 1 is applied at *horizontal*, and is clocked through the horizontal shift register in the time required by a single scan line in the television display. A binary 1 is applied at *vertical*, and is clocked through the vertical shift register in the time required by one frame of the television display. The vertical scan lines are accessed in sequential order via a single binary 1 being clocked through the horizontal shift register. After all pixels in a given row have been accessed, the single binary 1 in the vertical shift register is advanced to the next position, and the horizontal scan is repeated. The horizontal scan can be fast because it involves current steering and does not require voltage changes on the capacitance of a long scan wire. The vertical selection, which involves the settling of the output bias on the selected amplifiers, has the entire horizontal flyback time of the television display to settle, before it must be stable for the next horizontal scan.

The core of the chip is made up of rectangular tiles with height-to-width ratios of $\sqrt{3}$ to 2. Each tile contains the circuitry for a single pixel, as shown in Figure 15.5, with the wiring necessary to connect the pixel to its nearest neighbors. Each tile also contains the sections of global wiring necessary to form signal nets for V_{DD}, the bias controls for the resistive network, and the horizontal and vertical scan lines. The photoreceptors are located near the vertical scan line, such that alternating rows of left- and right-facing cells form a hexagonal array. This arrangement allows the vertical scan wire to be shared between adjacent rows, being accessed from the left by the odd rows, and from the right by even rows. To protect the processing circuitry from the effects of stray minority carriers, we have covered the entire chip with a solid sheet of second-layer metal, with openings directly over the photoreceptors; this layer is used for distributing ground to the pixels. We designed several versions of this chip over 3 years. A small section of the array of the most recent version is shown on the front endpaper of this book.

PERFORMANCE

Neurophysiologists have undertaken a tremendous variety of experiments in an attempt to understand how the retina performs computations, and they have come up with many explanations for retinal operation. Different investigators emphasize different aspects of retinal function, such as spatial-frequency filtering, adaptation and gain control, edge enhancement, and statistical optimization [Srinivasan et al., 1982]. It is entirely in the nature of biological systems that the results of several experiments designed to demonstrate one or another of these points of view can be explained by the properties of the single underlying structure. A highly evolved mechanism is able to subserve a multitude of purposes simultaneously.

Our experiments on the silicon retina have yielded results remarkably similar to those obtained from biological systems. From an engineering point of view, the primary function of the computation performed by silicon retina is to provide an automatic gain control that extends the useful operating range of the system. It is essential that a sensory system be sensitive to changes in its input, no matter what the viewing conditions. The structure executing this gain-control operation can perform many other functions as well, such as computing the contrast ratio or enhancing edges in the image. Thus, the mechanisms responsible for keeping

the system operating over an enormous range of image intensity and contrast have important consequences with regard to the representation of data.

Sensitivity Curves

The computation performed in the distal portion of the retina prevents the output from saturating over an incredible range of illumination levels. By logarithmically compressing the input signal, the photoreceptor takes the first step toward increasing the retina's dynamic range. The next step is a level normalization, implemented by means of the resistive network. The horizontal cells of the retina provide a spatially averaged version of the photoreceptor outputs, with which the local photoreceptor potential can be compared. The triad synapse senses the difference between the photoreceptor output and the potential of the horizontal cells, and generates a bipolar-cell output from this difference. The maximum response occurs when the photoreceptor potential is different from the space–time averaged outputs of many photoreceptors in the local neighborhood. This situation occurs when the image is changing rapidly in either space or time.

Figure 15.8 shows the shift in operating point of the bipolar-cell output of both a biological and a silicon retina, as a function of surround illumination. At a fixed surround illumination level, the output of the bipolar cell has a familiar tanh characteristic; it saturates to produce a constant output at very low or very

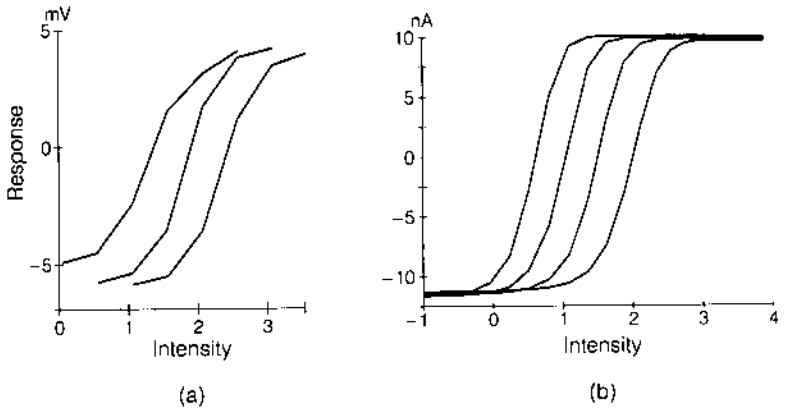


FIGURE 15.8 Curve shifting. Intensity-response curves shift to higher intensities at higher background illuminations. (a) Intensity-response curves for a depolarizing bipolar cell elicited by full-field flashes. The test flashes were substituted for constant background illuminations. These curves are plotted from the peaks of bipolar response to substituted test flashes. Peak responses are plotted, measured from the membrane potential just prior to response. (*Source:* Data from [Werblin, 1974]) (b) Intensity-response curves for a single pixel of the silicon retina. Curves are plotted for four different background intensities. The stimulus was a small disk centered on the receptive field of the pixel. The steady-state response is plotted.

high center intensities, and it is sensitive to changes in input over the middle of its range. Using the potential of the resistive network as a reference centers the range over which the output responds on the signal level averaged over the local surround. The full gain of the triad synapse can thus be used to report features of the image without fear that the output will be driven into saturation in the absence of local image information.

The action of the horizontal cell layer is an example of lateral inhibition, a ubiquitous feature of peripheral sensory systems [von Békésy, 1967]. Lateral inhibition is used to provide a reference value with which to compare the signal. This reference value is the operating point of the system. In the retina, the operating point of the system is the local average of intensity as computed by the horizontal cells. Because it uses a local rather than a global average, the eye is able to see detail in both the light and dark areas of high-contrast scenes, a task that would overwhelm a television camera, which uses only global adaptation.

Time Response

Time is an intrinsic part of an analog computation. In analog perception systems, the time scale of the computation must be matched to the time scale of external events, and to other real-time parts of the system. Biological vision systems use an inherently dynamic processing strategy. As emphasized in Chapter 13, body and eye movements are an important part of the computation.

Figure 15.9 shows the response of a single output to a sudden increase in incident illumination. Output from a bipolar cell in a biological retina is provided for comparison. The initial peak represents the difference between the voltage at the photoreceptor caused by the step input and the old averaged voltage stored on the capacitance of the resistive network. As the resistive network equilibrates to the new input level, the output of the amplifier diminishes. The final plateau value is a function of the size of the stimulus, which changes the average value of the intensity of the image as computed by the resistive network. Having computed a new average value of intensity, the resistive network causes the output of the amplifier to overshoot when the stimulus is turned off. As the network decays to its former value, the output returns to the baseline.

The temporal response of the silicon retina depends on the properties of the horizontal network. The voltage stored on the capacitance of the resistive network is the temporally as well as spatially averaged output of the photoreceptors. The horizontal network is like the follower–integrator circuit discussed in Chapter 9, which weights its input by an amount that decreases exponentially into the past. The time constant of integration is set by the bias voltages of the wide-range amplifier and of the resistors. The time constant can be varied independently of the space constant, which depends on only the difference between these bias voltages, rather than on their absolute magnitude.

The form of time response of the system varies with the space constant of the network. When the resistance value is low, γ approaches one, and the network

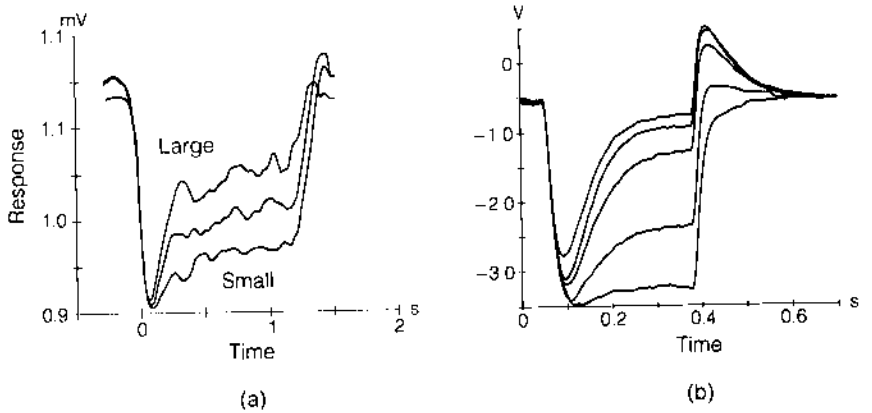


FIGURE 15.9 Temporal response to different-sized test flashes. (a) Response of a bipolar cell of the mud puppy, *Necturus maculosus*. (*Source:* Data from [Werblin, 1974]) (b) Output of a pixel in the silicon retina. Test flashes of the same intensity but of different diameters were centered on the receptive field of the unit. The space constant of the network was $\gamma = 0.3$. Larger flashes increased the excitation of the surround. The surround response was delayed due to the capacitance of the resistive network. Because the surround level is subtracted from the center response, the output shows a decrease for long times. This decrease is larger for larger flashes. The overshoot at stimulus offset decays as the surround returns to its resting level.

is computing the global average. A test flash of any limited size will produce a sustained output. Conversely, when the resistance value is high, γ approaches zero, and the triad synapse is just a diff1 circuit (Figure 10.5 (p. 167)), which has no sustained output. Because the rise time of the photoreceptor is finite, the space constant also can affect the initial peak of the time response. The dynamics of a small test flash are dominated by a pixel charging the capacitance of the surrounding area through the resistive network. In contrast, a pixel in the middle of a large test flash is charging mainly its own capacitance, because adjacent nodes of the network are being charged by their associated photoreceptors. The peak value of the output is thus larger for a small test flash than it is for larger test flashes.

Edge Response

We can view the suppression of spatially and temporally smooth image information as a filtering operation designed to enhance edges in the image. The outputs of the bipolar cells directly drive the sustained X-type retinal-ganglion cells of the mud puppy, *Necturus maculosus*. Consequently, the receptive-field properties of this type of ganglion cell can be traced to those of the bipolar cells [Werblin et al., 1969]. Although the formation of the receptive field of the X-type ganglion cells of the cat is somewhat more complex [Dowling, 1987], the end result is qualitatively similar. The receptive fields of these cells are described as

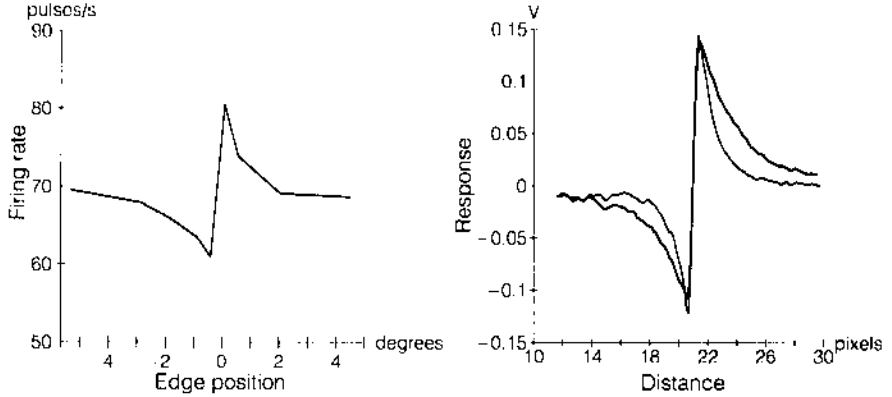


FIGURE 15.10 Spatial-derivative response of a retinal ganglion cell and of a pixel to a contrast edge. The vertical edge was held stationary at different distances from the receptive-field center. Contrast of the edge was 0.2 in both experiments. (a) On-center X-type ganglion cell of the cat. The contrast edge was turned alternately on and off. The average pulse density over the period 10 to 20 seconds after the introduction of the edge was measured for each edge position (*Source:* [Enroth-Cugell et al., 1966]). (b) Pixel output measured at steady state as the edge was moved in increments of 0.01 centimeters at the image plane. Interpixel spacing corresponded to 0.11 centimeters at the image plane. Response is shown for two different space constants. The rate of decay of the response is determined by the space constant of the resistive network.

antagonistic center-surround fields. Activation of the center of the receptive field stimulates the cell's response, and activation of the surround produces inhibition. Cells with this organization are strongly affected by discontinuities in intensity. The response of a sustained X-type ganglion cell to a contrast edge placed at different positions relative to its receptive field is shown in Figure 15.10(a). The spatial pattern of activity found in the cat is similar to the response of our silicon retina to a spatial-intensity step, as shown in Figure 15.10(b). The way the second spatial derivative is computed is illustrated in Figure 15.11. The surround value computed by the resistive network reflects the average intensity over a restricted region of the image. As the sharp edge passes over the receptive-field center, the output undergoes a sharp transition from lower than the average to above the average. Sharp edges thus generate large output, whereas smooth areas of the image produce no output, because the local center intensity matches the average intensity.

The center-surround computation sometimes is referred to as *a difference of Gaussians*. Laplacian filters, which have been used widely in computer vision systems, can be approximated by a difference of Gaussians [Marr, 1982]. These filters have been used to help computers localize objects; they work because discontinuities in intensity frequently correspond to object edges. Both of these mathematical forms express, in an analytically tractable way, the computation

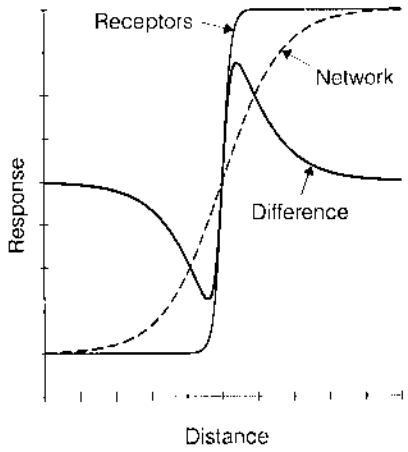


FIGURE 15.11 Model illustrating the mechanism of the generation of pixel response to spatial edge in intensity. The solid line, labeled *receptors*, represents the voltage outputs of the photoreceptors along a cross-section perpendicular to the edge. The resistive network computes a weighted local average of the photoreceptor intensity, shown by the dashed line. The average intensity differs from the actual intensity at the stimulus edge, because the photoreceptors on one side of the edge pull the network on the other side toward their potential. The difference between the photoreceptor output and the resistive network is the predicted pixel output, shown in the trace labeled *difference*. This mechanism results in increased output at places in the image where the first derivative of the intensity is changing.

that occurs as a natural result of an efficient physical implementation of local level normalization.

Space Constant of the Resistive Network

The resistive net is an economical way to generate a center-surround type of receptive field because the wiring is shared among many elements. Furthermore, we can vary the extent of the surround by changing the space constant of the network. In several species, the space constant of the horizontal-cell network is modulated by the release of dopamine [Dowling, 1987].

Figure 15.12 shows the exponential nature of the spatial decay of the response on one side of an edge for different space constants. The edge stimulus, being uniform in one dimension, generates current flow in only the transverse direction. The one-dimensional network therefore is a good approximation to the response of the two-dimensional network to an edge. As we noted in Chapter 7, the continuum approximation to the solution for a one-dimensional network is

$$V = V_0 e^{-\frac{1}{L}|x|} \tag{7.1}$$

where

$$\frac{1}{L} = \sqrt{RG} \tag{7.2}$$

As discussed in Appendix C, when we choose the unit of length to be $\sqrt{3}/2$ times the spacing of points in the horizontal lattice, the continuum approximation is very good, even for values of L as low as 1. The value of L is determined by the product of the conductance G and the resistance R. Both G and R are exponential functions of their respective bias controls:

$$G \propto e^{V_G} \qquad \text{and} \qquad R \propto e^{-V_R} \tag{15.1}$$

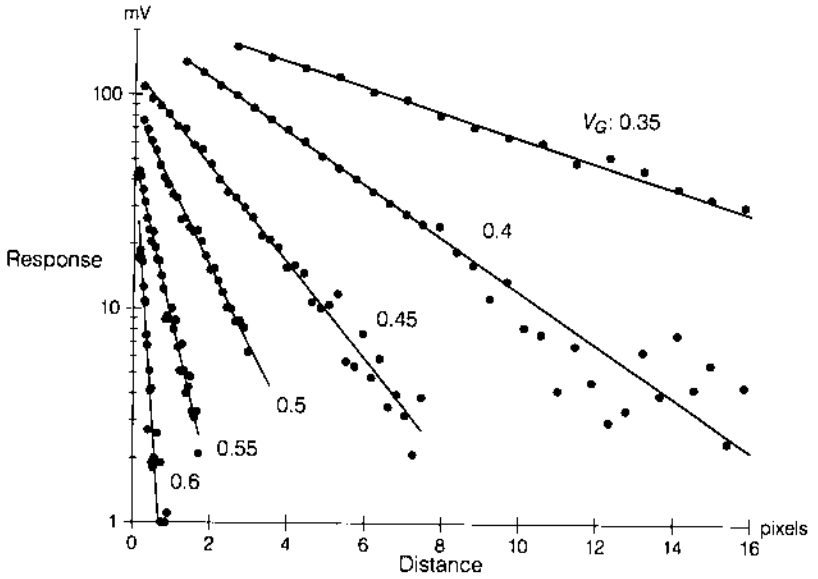


FIGURE 15.12 Exponential decay of one side of the response to an edge, as shown in Figure 15.10(b). Each curve was taken with the setting of the V_G control shown. For all curves, V_R was 0.55 volt. The slope of the decay corresponds to the space constant of the network.

Substituting Equation 15.1 into Equation 7.1 (p. 108), we obtain

$$\frac{1}{L} = \sqrt{RG} \propto e^{(V_G - V_R)/2} \tag{15.2}$$

The space constant thus should be a function of $V_G - V_R$, and should not be dependent on the absolute voltage level. The constant of proportionality in Equation 15.2 contains the width-to-length ratios for transistors in the horizontal resistor and in the resistor bias circuit, and those for transistors in the transconductance amplifier. Figure 15.13 shows the edge response of the silicon retina measured for several values of bias voltages, with a fixed difference between V_G and V_R, and thus a fixed ratio between the transconductance bias current and the resistor bias current. The form of the static response of the system is unchanged, as expected.

If we assume the continuum form of the decay, Equation 7.2 (p. 108) applies to the horizontal network over the range of L values involved, and we can compare the slopes of the decay curves in Figure 15.12 with the theoretical expression given in Equation 15.2, where all voltages are expressed in terms of $kT/(q\kappa)$. The comparison is shown in Figure 15.14; the voltage dependence of the decay constant is in excellent agreement with the theoretical prediction. The absolute value of the curve in Figure 15.14 was adjusted for the best fit to the data, and is higher, by a factor of about two, than the value deduced from the device geometries in

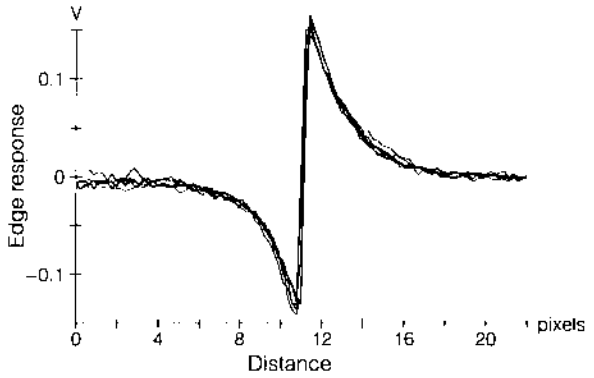
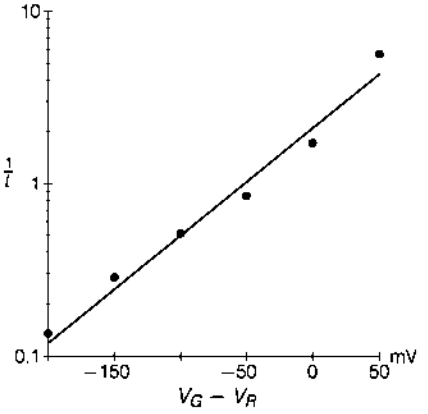


FIGURE 15.13 The response of a pixel to a 0.2 contrast edge measured for a fixed difference between the conductance bias voltage and the resistor bias voltage. (DC offsets in the response were subtracted out.) The space constant of the network depends on only the ratio of conductance bias current to resistor bias current. Resistor bias voltages were 100 millivolts greater than were the conductance bias voltages. The form of the response stayed essentially unchanged as bias voltages were swept over a 250-millivolt range, thereby changing the bias current by more than three orders of magnitude.

the resistive connections and in the transconductance amplifiers. A number of factors may be responsible for this discrepancy, including inaccurate calibration of the interpixel spacing, partial saturation of resistive connections due to voltage offsets, uncertainties in the channel lengths of short-channel devices, and so on. None of these factors should have a large effect on the voltage dependence of the decay, in keeping with our observations.

The space constant determines the peak amplitude of the response as well as the decay constant of the exponential. The decay length L is small when the conductance feeding the local input to the network is large relative to the

FIGURE 15.14 Space constant of the response data of Figure 15.12, plotted as a function of $V_G - V_R$. The straight line is the theoretical expression taken from Equation 15.2, using the measured value of $\kappa = 0.73$. The magnitude of the curve was adjusted for best fit to the data, and is about a factor of two higher than expected from the width-to-length ratios of transistors in the transconductance amplifier and in the resistor bias circuit.



lateral conductance. Under these conditions, the difference between the local photoreceptor and the network also is small, because the average is dominated by the local input. The decay length L is large when the conductance feeding the local input to the network is small relative to the lateral conductance. Under these conditions, the difference between the local photoreceptor and the network approaches the full difference between the local photoreceptor and the average over many photoreceptors, because the average is affected very little by the local input. This dependence of peak amplitude on space constant can be seen in the curves in Figure 15.12. The precise nature of this dependence cannot be determined from the continuum limit, because the input conductance is inherently tied to the discrete nature of the network. Feinstein discusses these matters in more detail [Feinstein, 1988].

Mach Bands

Retinal processing has important consequences for higher-level vision. Many of the most striking phenomena known from perceptual psychology are a result of the first levels of neural processing. In the visual systems of higher animals, the center-surround response to local stimuli is responsible for some of the strongest visual illusions. For example, Mach bands, the Hermann–Hering grid illusion, and the Craik–O'Brian–Cornsweet illusion may all be traced to simple inhibitory interactions among elements of the retina [Ratliff, 1965; Julesz, 1971].

The response of a pixel to a ramp stimulus is plotted in Figure 15.15. Because the retina performs a second-order filtering of the image, changes in the first derivative of intensity are enhanced. **Mach bands** are illusory bright and dark bands that appear at the edges of an intensity ramp. The positions of the illusory bands correspond to the positions where the retinal output is enhanced due to changes in the first derivative of the intensity.

The retina, as the first stage in the visual system, provides gain control and image enhancement, as well as transduction of light into electrical signals. The evolutionary advantage of this kind of preprocessing is evidenced by the ubiquitous occurrence of retina structures in the vertebrates, and even in invertebrates such as the octopus. From an engineering viewpoint, the retina greatly reduces the signal bandwidth required to transmit visual information to the brain, thereby greatly reducing the size of the optic nerve and allowing more effective computation at the next level. Thus, the retina is a prime example of a system that performs information processing by *selectively rejecting irrelevant information.* Any operation that discards information will, of necessity, create **ambiguities**, in which several distinct input images create an output that is indistinguishable by the next level of processing. To the visual researcher, these *optical illusions* provide valuable insight into the nature of information processing at various stages in the visual system. The fact that our retinal model generates an illusory output when exposed to the same stimulus that evokes a Mach-band illusion in humans gives us additional confidence that we have correctly interpreted the principles on which the biological system operates.

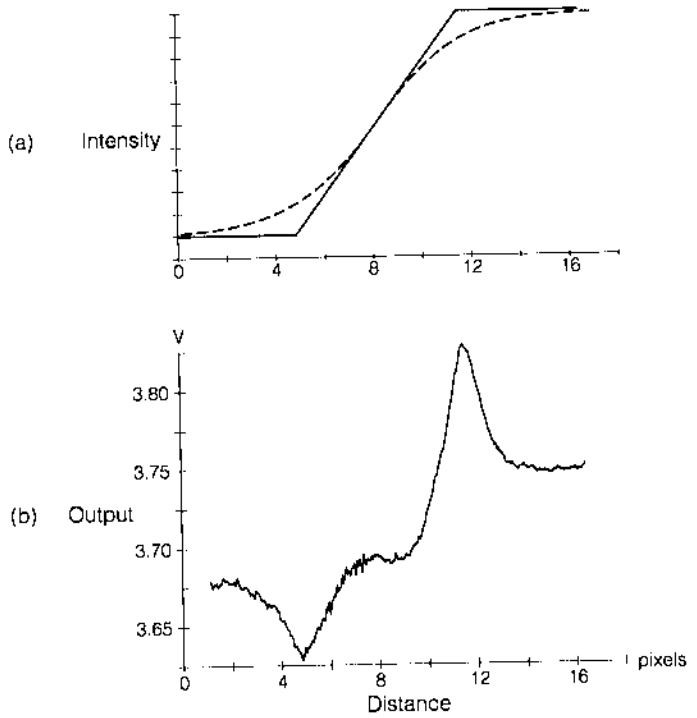


FIGURE 15.15 Mach bands are illusory bright and dark bands that appear at the edges of a ramp of intensity. The interaction between retinal output and higher-level neural processing is believed to explain the perception of Mach bands.

(a) Ramp stimulus illustrates the function of a second-order filter. The solid line indicates the intensity profile of an ideal Mach-band stimulus. The dashed line is the weighted local average of the intensity. The difference between the local average and the point intensity is the output of the retina. The magnitude of the difference is large at the point in the image where the first derivative is changing.

(b) Response of a pixel to ramp stimulus. This stimulus is a shadow cast by an opaque sheet between an extended light source and the image plane. The stimulus is moved over the retina in 50-micron steps. The enhanced response at the edges of the ramp is due to the second-order behavior of the retinal response. The shift in DC value across the response is due to intensity variation as the light source approaches the pixel.

SUMMARY

We have taken the first step in simulating the computations done by the brain to process a visual image. We have used a medium that has a structure in many ways similar to neurobiological structures. Following the biological metaphor has led us to develop a system that is nearly optimal from many points of view. The constraints on our silicon system are similar to those on neurobiological systems. As in the biological retina, density is limited by the total amount of wire required to accomplish the computation. The retina, like many other areas of the brain, minimizes wire by arranging the signal representation such that as much wire as possible can be shared. The resistive network is the ultimate example of shared wiring. By including a pixel's own input in the average, we can compute the weighted average over a neighborhood for every position in the image, using the same shared structure.

The principle of shared wire is found, in less extreme forms, throughout the brain. Computation is always done in the context of neighboring information. For a neighborhood to be meaningful, nearby areas in the neural structure must represent information that is more closely related than is that represented by areas farther away. Visual areas in the cortex that begin the processing sequence are mapped retinotopically. Higher-level areas represent more abstract information, but areas that are close together still represent similar information. It is this *map* property that organizes the cortex such that most wires can be short and highly shared; it is perhaps the single most important architectural principle in the brain.

REFERENCES

Dowling, J.E. *The Retina: An Approachable Part of the Brain.* Cambridge, MA: Belknap Press of Harvard University Press, 1987.

Enroth-Cugell, C. and Robson, J.G. The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology,* 187:517, 1966.

Feinstein, D. The hexagonal resistive network and the circular approximation. *Caltech Computer Science Technical Report,* Caltech-CS-TR-88-7, California Institute of Technology, Pasadena, CA, 1988.

Julesz, B. *Foundations of Cyclopean Perception.* Chicago, IL: The University of Chicago Press, 1971.

Marr, D. *Vision.* San Francisco, CA: W.H. Freeman, 1982.

Mead, C. and Wawrzynek, J. A new discipline for CMOS design. In Fuchs, H. (ed), *1985 Chapel Hill Conference on Very Large Scale Integration.* Chapel Hill, NC: Computer Science Press, 1985, p. 87.

Ratliff, F. *Mach Bands: Quantitative Studies on Neural Networks in the Retina.* San Francisco: Holden-Day, 1965.

Rodieck, R.W. *The Vertebrate Retina.* San Francisco, CA: W.H. Freeman, 1973.

Shapley, R. and Enroth-Cugell, C. Visual adaptation and retinal gain controls. In Osborne, N.N. and Chader, G.J. (eds), *Progress in Retinal Research,* vol 3. Oxford, England: Pergamon Press, 1984, p. 263.

Sivilotti, M.A., Mahowald, M.A., and Mead, C.A. Real-time visual computations using analog CMOS processing arrays. In Losleben, P. (ed), *Stanford Conference on Very Large Scale Integration.* Cambridge, MA: MIT Press, 1987, p. 295.

Srinivasan, M.V., Laughlin, S.B., and Dubs, A. Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London,* Series B, 216:427, 1982.

von Békésy, G. *Sensory Inhibition.* Princeton, NJ: Princeton University Press, 1967.

Werblin, F.S. and Dowling, J.E. Organization of the retina of the mudpuppy, *Necturus maculosus.* II. Intracellular recording. *Journal of Neurophysiology,* 32:339, 1969.

Werblin, F.S. Control of retinal sensitivity, II. Lateral interactions at the outer plexiform layer. *Journal of General Physiology,* 63:62, 1974.

CHAPTER

16

ELECTRONIC COCHLEA

Richard F. Lyon Carver Mead

When we understand how hearing works, we will be able to build amazing machines with brainlike abilities to interpret the world through sounds that is, to *hear.* As part of our endeavor to decipher the auditory nervous system, we can use models that incorporate current ideas of how that system works to engineer simple *electronic* systems that hear in simple ways. The relative success of these *engineered* systems then helps us to evaluate our knowledge about hearing, and helps to motivate further research.

As a first step in building machines that hear, we have implemented an analog electronic cochlea that incorporates much of the current state of knowledge about cochlear structure and function. The biological cochlea (inner ear) is a complex three-dimensional fluid-dynamic system, illustrated schematically in Figure 16.1. In the process of designing, building, and testing the electronic cochlea, we have had to put together a coherent view of the function of the biological cochlea from the diverse ideas in the literature. This view and the resulting design are the subjects of this chapter.

We hear through the sound-analyzing action of the cochlea and of the auditory centers of the brain. As does vision, hearing provides a representation of events and objects in the world that are relevant to survival.

Portions reprinted with permission from *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 36, No. 7, July 1988.