

Implementing Neural Architectures Using Analog VLSI Circuits

MARY ANN C. MAHER, STEPHEN P. DEWEERTH, MISHA A. MAHOWALD,
AND CARVER A. MEAD

Abstract—Biological systems routinely perform computations, such as speech recognition and the calculation of visual motion, that baffle our most powerful computers. Analog very large-scale integrated (VLSI) technology allows us not only to study and simulate biological systems, but also to emulate them in designing artificial sensory systems. A methodology for building these systems in CMOS VLSI technology has been developed using analog micropower circuit elements that can be hierarchically combined. Using this methodology, experimental VLSI chips of visual and motor subsystems have been designed and fabricated. These chips exhibit behavior similar to that of biological systems, and perform computations useful for artificial sensory systems.

I. INTRODUCTION

CALCULATION of visual motion and speech recognition are two highly complex computations which biological systems perform routinely, but which are beyond the capability of our most powerful computers. Analog very large-scale integrated (VLSI) technology allows us to construct hardware models to study and simulate biological systems. We can also derive inspiration from biological models in building artificial sensory systems. Although they use imprecise and unreliable elements, biological systems obtain robustness to noisy input data and element failure through the use of highly-redundant, distributed architectures. Analog VLSI circuits provide an attractive medium for implementing such architectures in terms of density and speed. Large, regular structures which underlie peripheral sensory systems are natural and easy to implement in VLSI technology. Analog circuits in parallel architectures provide real-time computation for operating on sensory input. Processors and sensors can be integrated on the same chip, alleviating many of the problems (such as temporal aliasing due to sampling) inherent in designs that separate computation from sensing.

Modeling biological systems presents many challenges to the analog circuit designer. Neural computation is often an emergent property of the system, derived from the way the component elements are organized, and may not be evident in any single element. It is often difficult to separate a neural structure into functional units [1]. Major areas are richly interconnected and computation is intertwined, as a single neural structure subserves a multitude of functions simultaneously [2]. As a result, computational

strategies for building collective systems require the development of new architectures and a new design methodology. Mead [3] presents such a methodology for implementing biological inspired architectures. We shall illustrate this methodology and describe some of the neural organizing principles on which it is based. We also present two system designs: the silicon retina—a system in which the same physical structure that is used to compute gain-control also computes contrast ratio, time derivatives, and enhances edges, and the Tracker—a simple sensorimotor integration system that is able to actively track a bright spot of light.

II. SYSTEM PROPERTIES

Many parallels exist between biological “wetware” and analog silicon hardware [4]. Both use analog electrical signals, with current and differences in electrical potential as signal representations. Elementary computational primitives are a direct consequence of physical laws. Functions such as exponentials, due to the physics of energy barriers, are performed by the primitive devices—bilayer membranes and transistors. Time is an essential element for computation in both systems. The time constants of the processing elements are matched to the events in the inputs [5]. Builders of analog VLSI systems face many of the same resource constraints as do neural systems—limited wiring space and a high cost of communication imposed by the physical placement of computation elements.

Biological systems have evolved architectures that make ingenious and efficient use of these limited resources, and their study is insightful for VLSI implementations. Many “place encodings” exist in biology where information is encoded in an element’s spatial location and this location is part of the computation. For example, in vision systems, the location of a photoreceptor in the retina indicates the position of the incoming light in space forming a spatial map. The encoding is carried to the next level in cortex via a conformal mapping that maintains a spatial locality of the signals. At higher levels of cortex maps of features are made so that related features are stored together. In a spatial map, neurons representing similar information are arrayed as closely as possible so that processing may be shared. Local averages are computed by local signal aggregation with a minimum of wire length, a precious resource for the VLSI designer. These average values are the local operating points of the system and are used by an automatic gain control subsystem.

Manuscript received July 10, 1988; revised January 23, 1989. This paper was recommended by Guest Editors R. W. Newcomb and N. El-Leithy.

The authors are with the California Institute of Technology, Pasadena, CA 91125.

IEEE Log Number 8826919.

Some important differences exist between the two technologies, however. Neurons have a fan-out and fan-in of several thousand—much higher than is currently possible in VLSI technology. Also, the brain has more layers available for wiring; it has a $2 + \epsilon$ dimensional cortex, versus the two dimensions available to chip designers. Analog VLSI technology, however, has a speed advantage; it uses nanosecond logic, as compared to the millisecond logic used for typical neural computations. The integration levels of neural systems are much larger than is currently available in VLSI technology [3], [4]. We will need to use multi-chip systems or wafer-scale integration to achieve the density needed to implement complex neural functions. Neural computing devices are less precise and are less well matched than are the analog electronic devices used in VLSI circuits [4]. The fault tolerance of neural systems suggests that neural organizing principles can be adapted to produce wafer-scale architectures. We can use neurobiological systems as inspiration, but there will be some differences in implementation as we exploit the advantages of our technology. For example, we could use our speed advantage to time-multiplex signals on a single wire to make up for lack of wiring space.

III. THE CIRCUIT BUILDING BLOCKS

To manage the complexity of building large-scale neuro-morphic analog systems, we have developed a structured, hierarchical design methodology. At the most basic level, we have a transistor model which is simple, but is adequate for predicting relevant behavior at the circuit and subsystem levels. At the next level of abstraction, we have designed a set of elementary, yet powerful analog circuit building blocks. The basic building blocks are combined hierarchically into larger designs using composition rules. We must match signal types, and notice that each signal type lends itself to certain computations. We use voltages for distributing information and currents for doing summation via Kirchhoff's current law.

The circuit building blocks must be able to encompass the data representations used by neurons in different parts of the brain [7]. In a neural system, computation is often a series of transformations from one representation to another, as the most important outputs are passed on to the next level. In the motor system, muscles are innervated by neurons where the contraction of the muscle is proportional to the firing rate of the neuron. A neuron monotonically encodes a single variable. In contrast, in the visual system, neurons are tuned to respond to multiple properties of a stimulus. A neuron in visual cortex may respond to stimulus location, orientation and direction of motion [8]. A stimulus activates a number of neurons that respond over a limited range of inputs, but whose regions of sensitivity overlap. As a result, spatio-temporal patterns of activity of a cell population are used to represent data. For example, in the visual system, color is calculated from the analog ratios of the values of three different types of cone cells that have overlapping spectral sensitivities [9].

With these principles in mind, we can turn circuits into systems. In the following sections, we shall describe the

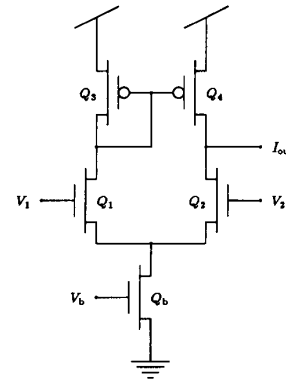


Fig. 1. Schematic of the transconductance amplifier circuit. The circuit consists of a bias transistor Q_b , a differential pair Q_1 - Q_2 , and a current mirror Q_3 - Q_4 .

transistor model and the functional analog building blocks. We shall then describe subsystems for global and local averaging.

3.1. The Transistor Model

We build analog circuits in which many of the MOS transistors operate in the weak-inversion (subthreshold) regime. In this regime, the MOS transistor behaves much like a bipolar transistor; the dominant conduction mode is the result of diffusion current. The drain current is exponential in the gate voltage. A simple model for the subthreshold transistor in terms of its gate voltage V_g , source voltage V_s and drain-source voltage V_{ds} is given by

$$I = I_0 e^{(\kappa/V_T)V_g} e^{-(1/V_T)V_s} \left(1 - e^{-(1/V_T)V_{ds}} + \frac{V_{ds}}{V_0} \right) \quad (1)$$

where κ measures the effectiveness of the gate voltage in determining the surface potential, I_0 is the zero-bias current, and V_0 is a measure of the drain resistance, also known as the Early voltage [10]. For our purposes, these parameters are considered to be constants of the fabrication process. The parameter V_T is a characteristic voltage equal to kT/q in which k is Boltzmann's constant, T is absolute temperature, and q is the charge on an electron.

Operation in the subthreshold region has several advantages for the construction of large analog systems [11]. Typical currents for a minimum-sized device are in the range of 10^{-12} to 10^{-7} A, so power consumption is low. Also, the transistor saturates after a few V_T of drain voltage. As a result, the drain voltage of a subthreshold transistor can be operated much closer to the source voltage than can the drain voltage of an above-threshold transistor. Finally, the transistor computes an exponential, a function we will use often.

3.2. The Transconductance Amplifier

One of our most important building blocks is the transconductance amplifier shown in Fig. 1. The circuit acts like an operational amplifier with high open-loop voltage gain (>1000), however, the transconductance (the gain from differential input voltage to output current) is

controllable. The bias transistor, Q_b , acts as a current source, setting the current through the differential-amplifier stage, controlling the transconductance G and the saturated output current I_b .

By using the simple model for the transistor, we can deduce the output current as a function of differential input voltage in the ideal case:

$$I = I_b \tanh\left(\frac{\kappa}{2V_T}(V_1 - V_2)\right). \quad (2)$$

The basic transconductance amplifier circuit uses only five transistors, but has a limited voltage gain and a limited range of output voltages over which it operates correctly. We use a wide-range transconductance amplifier [3] which has nine transistors when the circuit must operate over a large output voltage range.

The transconductance amplifier performs several interesting computations in different regions of operation. For small differential voltages, the amplifier is roughly linear, with transconductance

$$G = \frac{\partial I}{\partial (V_1 - V_2)} = I_b \frac{\kappa}{2V_T}. \quad (3)$$

For large differential input voltages, $|V_2 - V_1| \gg V_T/\kappa$, the circuit behaves like a threshold function with asymptotes $\pm I_b$.

3.3. Arithmetic Building Blocks

In addition to the transconductance amplifier, we have developed a set of primitive circuit elements that are sufficiently rich to encompass many different kinds of neural architectures. Basic transistor properties lead to simple square root and logarithm circuits that perform data compression. Due to the large dynamic range of sensory inputs, neural signals are often represented with a logarithmic encoding. Physiological recordings show that a biological photoreceptor's electrical response is logarithmic in light intensity over the central part of its range [12].

We can devise several of our functional building blocks as extensions of the transconductance amplifier. We make a unity-gain follower by connecting the open-circuited output of the transconductance amplifier back to its negative input. We design a half-way rectifier by adding a p-channel current mirror at the output of the transconductance amplifier. The p-channel current mirror copies only positive currents. A full-wave rectifier is composed of two half-wave rectifiers. Many of the circuits used in analog bipolar design, such as the Gilbert transconductance multiplier and other translinear circuits [13], [14] have been adopted in our subthreshold MOS implementations. We also use traditional analog micropower circuits [11].

Other functional units have biological correlates. The winner-take-all circuit [15] computes the maximum of the currents flowing into a set of input channels. The circuit causes the voltage on the output of the maximum channel to go high, while all other outputs are held low. This computation is similar to the nonspecific-inhibition

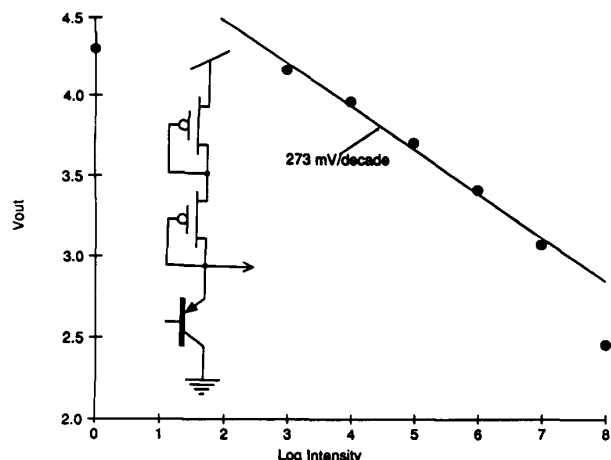


Fig. 2. Photoreceptor circuit schematic and measured response of the circuit. The photocurrent is proportional to the incident light intensity. The output voltage is logarithmic over more than four orders of magnitude in intensity. Data taken from Mead [3].

schemes found in biology, where the presence of a signal on one channel is used to inhibit other channels [16].

The time scales needed for sensory processing using neural organizations require large values of resistance. Resistor values in a typical CMOS process are quite low. Using a special process with undoped polysilicon, we could make the required resistors, but instead we have chosen to design active resistors with variable resistance using only the basic CMOS process [3]. We set the value of resistance with a subthreshold bias transistor. The resistor is monotonic with $I-V$ characteristics that pass through the origin. The resistor saturates when there is a large voltage differences across it, as does the transconductance amplifier. We will use this property to advantage at the system level.

3.4. The Photoreceptor

To process sensory data, we must have a set of primitives for transducing sensory inputs, such as light or sound, onto the silicon. We use the photoreceptor circuit shown in Fig. 2 to transduce light into an electrical signal [17]. The photoreceptor consists of a photodetector and a logarithmic element. The photodetector is a vertical bipolar transistor, which is a parasitic element of the basic CMOS process. The base is the well, the emitter is the source-drain diffusion, and the collector is the substrate. As incident photons create electron-hole pairs, electrons are collected by the base, transducing light into photocurrent with a gain of several hundred. The logarithmic element consists of two diode-connected p-channel transistors acting as a load for the photodetector. The transistors are biased into the weak-inversion regime by the photocurrent; they create an output voltage that is proportional to the logarithm of the current and hence to the logarithm of the incident light intensity. The voltage out of this photoreceptor is logarithmic over four to five orders of magnitude in light intensity, as shown in Fig. 2.

The logarithmic response compresses the intensity range of several orders of magnitude into a few hundred milli-

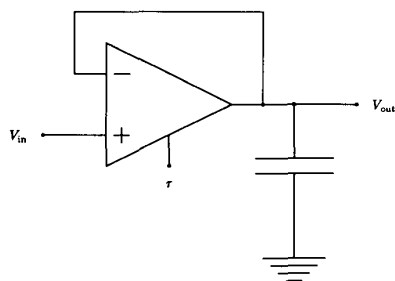


Fig. 3. Schematic of the follower integrator circuit. The capacitor current is proportional to the difference between V_{in} and V_{out} for small differences. The time constant integration, τ , is the capacitance multiplied by the conductance of the amplifier.

volts of output voltage range. As a result of this transformation, voltage differences between two points in a uniformly illuminated image correspond to the ratio between the reflectances of the two objects—the voltage difference represents a contrast ratio that is independent of illumination level.

3.5. Time-Varying Building Blocks

Most of the information that the sensory system must process concerns time-varying signals: visual motion, sound, etc. To process these signals, we have designed a set of functional building blocks for time-varying inputs. As shown in Fig. 3, we have created a first-order low-pass filter, called a follower integrator, by adding a capacitor to the basic follower circuit. We use an MOS transistor operated above threshold as a capacitor. For small signals, in the linear region of the amplifier, the follower integrator has the transfer function

$$\frac{V_{out}}{V_{in}} = \frac{1}{\tau s + 1} \quad (4)$$

where the time constant τ is equal to C/G , C is the value of capacitance, and G is the transconductance of the amplifier. For low frequencies, the circuit acts as a unity-gain follower; for high frequencies it acts as an integrator. We combined two cascaded follower integrator first-order sections with a positive feedback amplifier to build a second-order section. We set the poles of this second-order system by changing the conductance of the feedback amplifier.

Many computations emphasize temporal changes in the pattern of input signals. A derivative circuit has this property. We designed a differentiator by subtracting a signal from a time-integrated version of itself, which is computed by the follower integrator. We do this subtraction with a transconductance amplifier. We also have built circuits that emphasize temporal derivatives above a certain signal level.

So far, we have dealt with signals that are analog in both time and amplitude. Biological systems also use signals that are digital in amplitude and analog in time. These fully restored signals are used for transmitting data over long distances. The data are represented by the arrival

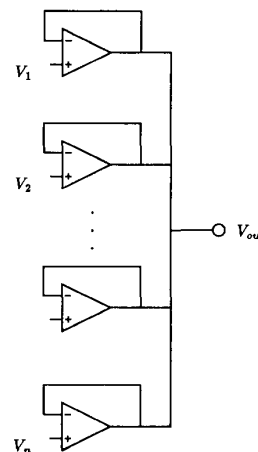


Fig. 4. Schematic of the follower aggregation circuit. Each follower supplies a current to the output node that is proportional to the difference between its input voltage and the output voltage.

time of nerve pulses [18]. We have designed a circuit, the Neuron, that integrates a current input and produces pulse outputs when the input voltage is above a certain threshold. The Neuron's output frequency is dependent on the input current, and the pulsewidth is controllable. This circuit is useful in encoding frequency coded data.

These building blocks for time-varying signals can be combined into subsystems, and are important in the design of auditory processing elements. We cascaded first-order sections into an analog delay line. The silicon cochlea chip uses second-order sections in a frequency-selective analog delay line [19]. We composed neuron circuits into a delay line that propagates fully restored pulses. A variation on this design, based on principles taken from biological axons, allows bidirectional pulse propagation. The pulse representation used by the axon delay lines makes computations such as correlations particularly easy to perform.

IV. LOCAL AND GLOBAL AGGREGATION

An essential computation in neural systems is the calculation of averages. The follower integrator circuit computes a temporal average—a reference against which the temporal variation of signals can be compared. We shall now describe subsystems that perform spatial averages. The follower aggregation circuit, shown in Fig. 4, consists of follower stages with their outputs connected. Assuming operation in the linear regime, each amplifier contributes a current $G_i(V_i - V_{out})$ to the output node. Using Kirchhoff's current law at the output node yields

$$\sum_{i=1}^n G_i(V_i - V_{out}) = 0 \quad (5)$$

$$V_{out} = \frac{\sum_{i=1}^n G_i V_i}{\sum_{i=1}^n G_i} \quad (6)$$

Thus the circuit takes the average of the inputs V_i weighted by the transconductances G_i . If any input voltage is significantly different from the average, however, the transconductance-amplifier current saturates and the contribution

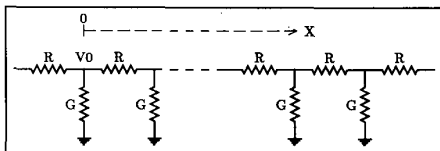


Fig. 5. A resistive ladder network. A signal injected at node V_0 will decay exponentially with distance at a scale set by the product RG .

by that data point is limited. This property gives the circuit a robustness against bad data points. If all values are scattered by many V_T/κ , then the circuit performs a weighted median calculation with all amplifiers saturated.

The follower-aggregation circuit performs a global average. Neural-type circuits also perform local aggregation useful for computing local averages. Data can be averaged over a local neighborhood, with data spatially distant from the point of aggregation contributing less to the average. Local averages can be computed by a resistive network. A one-dimensional resistive-ladder network is shown in Fig. 5. A signal injected into this network decays exponentially with distance from the source. For a uniform, continuous network, the voltage along the network as a function of distance x , has the form

$$V = V_0 e^{-x/L} \quad (7)$$

where

$$\frac{1}{L} = \sqrt{RG}. \quad (8)$$

Here, R is the resistance per unit length and G is the conductance to ground. L is a measure of the neighborhood over which the average is taken and $\alpha = 1/L$ is called the *space constant*. For a discrete network, the result is

$$V_n = \gamma^n V_0 \quad (9)$$

where

$$\gamma = 1 + \frac{1}{2L^2} - \frac{1}{L} \sqrt{1 + \frac{1}{4L^2}} \quad (10)$$

and R and G are given per section. For large values of L , the continuous approximation to the discrete network is quite good. If the conductance to ground is small when compared with the conductance to the network, the signal will propagate for a large distance before it dies out, and L will be larger.

Inputs to the network can be provided by voltages or by currents. The voltage source is placed between the conductance G and ground. Multiple inputs cause the network to perform a weighted average at each node by superposition; the farther away the inputs are from a node, the less weight they are given. The voltage at a given node due to multiple input currents I_n is

$$V_k = \frac{1}{2G_0} \sum_n \gamma^{n-k} I_n \quad (11)$$

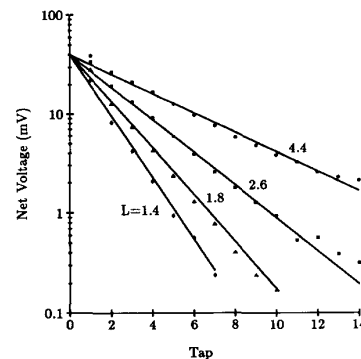


Fig. 6. Voltage versus distance for a one-dimensional discrete resistive ladder network for different values of L . The solid curves were computed from theory for each value of L . The dots are data taken from the output of the discrete line. Data taken from Mead [3].

where G_0 is the effective conductance of a semi-infinite network. For a continuous network, G_0 is given by

$$G_0 = \sqrt{\frac{R}{G}}. \quad (12)$$

For the discrete case, the result is

$$G_0 = \sqrt{\frac{R}{G}} \sqrt{1 + \frac{1}{4L^2}}. \quad (13)$$

We design a silicon implementation of the resistive network by replacing the resistor by our active resistor circuit and the conductance by a follower. The transconductance of the amplifier corresponds to G in Fig. 5. Inputs to the follower correspond to voltage sources. Data from an experimental network for several different values of L are shown in Fig. 6.

This resistive averaging network is useful for a smoothing operation. The superposition principle also applies in two dimensions, but the weighting function is more complex [3]. The two-dimensional network computes a smooth fit to each point of data included in a region of diameter L . Because the resistors saturate, when there is too much voltage drop across a single resistor, a discontinuity will occur. This saturating property of the resistor is useful for image segmentation [3]. The network segments an image into regions over which the image is smooth.

V. THE SILICON RETINA

The silicon retina [20] is a system built from our analog functional building blocks. It illustrates many of the properties of neural systems. The model for the retina of each type of animal is different, but we have conserved the gross structure of vertebrate retina in our design of the silicon retina. The chip generates, in real time, outputs that correspond to signals observed in biological retinas, and exhibits a tolerance for device imperfections.

The cells in the first layers of the retina are shown in Fig. 7 [21]. Light is transduced into an electrical signal via the photoreceptors at the top. The primary pathway proceeds vertically from the photoreceptors through the triad synapse to the bipolar cells and then to the ganglion cells.

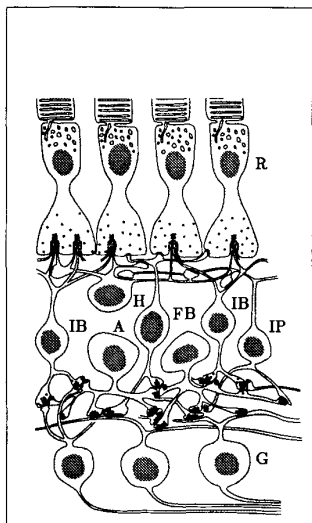


Fig. 7. An artist's conception of a cross section of a primate retina. *R*: photoreceptor, *H*: horizontal cell, *IB*: invaginating bipolar cell, *FB*: flat bipolar cell, *A*: amacrine cell, *IP*: inner plexiform cell, *G*: ganglion cell. Adapted from Dowling [21].

This pathway intersects two horizontal pathways: the horizontal cells of the outer-plexiform layer and the amacrine cells of the inner-plexiform layer. The triad synapse is the point of contact among the photoreceptor, the bipolar cell, and the horizontal network. In just a few layers of cells in the retina, a remarkable amount of computation is done: the image becomes independent of the absolute light level and as the retina adapts to a wide range of viewing conditions, it enhances edges and emphasizes time-derivatives.

A schematic drawing of the silicon retina is shown in Fig. 8. The horizontal network is modeled as a resistive network. We model the photoreceptor, the bipolar cell, and the triad synapse as shown in the inset. A wide-range amplifier provides a conductance through which the resistive network is driven toward the photoreceptor-output potential. The horizontal cells form a network that averages the photoreceptor output spatially and temporally. 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 bipolar cells' output is thus proportional to the difference between the photoreceptor signal and the horizontal-cell signal. Because the silicon model is implemented in a physical substrate, it has a straightforward structural relationship to the vertebrate retina and provides an example of a spatial mapping. Each photoreceptor in the network is linked to its six neighbors with resistive elements to form a hexagonal array. By using a wide-range amplifier in place of a bidirectional conductance, we make the photoreceptor an effective voltage source that provides input into the resistive network. The spatial scale of the weighting function α is determined by the product of the lateral resistance and the conductance

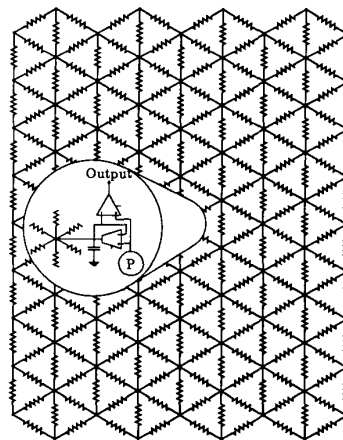


Fig. 8. Diagram of the silicon retina. The horizontal cell layer is represented by the resistive network. The pixels are tiled in a hexagonal array. The circuit schematic for a single pixel representing the triad synapse is shown in the inset with *P* representing the photoreceptor.

coupling the photoreceptors into the network as described in Section IV.

The chip consists of an array of pixels and a scanning arrangement for reading the results of the retinal processing. The output of any pixel can be accessed through the scanner, which is made up of a vertical scan register and a horizontal scan register [22]. Each scan-register stage has 1-bit of shift register with the associated signal-selection circuits. 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. 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. The core of the chip is made of rectangular tiles with a height-to-width ratio of $\sqrt{3}$ to 2 to approximate a hexagonal grid. Each tile contains the circuitry for a single pixel, along with the wiring necessary to connect the pixel to its nearest neighbors.

The photoreceptor, the horizontal cells, and the bipolar cells in the triad synapse interact in a *center-surround* organization. In this organization, the signal average from a central area is subtracted from the average over a larger surrounding area, and the difference is reported at the output. The center of the bipolar-cell receptive field is excited by the photoreceptors, whereas the antagonistic surround is due to the horizontal cells. The output of the bipolar cell represents the difference between a center intensity and a weighted average of the intensities of the surround. The horizontal network provides a smooth reference for local computation. If the visual system used a global average as a reference, details in very light or very dark areas would be invisible.

Fig. 9 shows the shift in operating point of the bipolar-cell output of both the biological and silicon retinas as a function of surround illumination. At a fixed surround

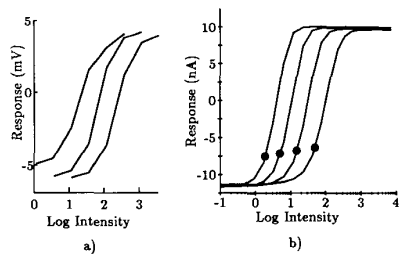


Fig. 9. (a) Intensity versus output peak response for a depolarizing bipolar cell responding to full-field flashes. Data from Werblin [23]. (b) Intensity versus steady-state output current for a single pixel of the silicon retina for four different background intensities. The curves shift to higher intensities at higher background illuminations.

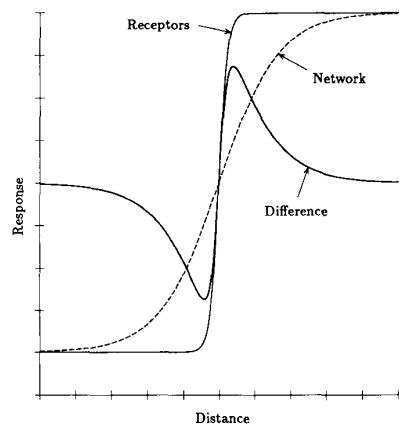


Fig. 10. Model illustrating pixel response to a spatial edge in intensity. The solid line represents the voltage outputs of the photoreceptors along a cross section perpendicular to the edge. The dashed line shows the resistive network output. The difference between the photoreceptor output and the resistive network is shown in the trace labeled difference.

illumination level, the output of the bipolar cell saturates to produce a constant output at very low or very 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, it centers the range over which the output responds on the signal level averaged over the local surround. The action of the horizontal-cell layer is an example of lateral inhibition. As a sharp edge passes over the receptive-field center, the output undergoes an abrupt transition from lower than average to above average. Sharp edges thus generate large output, whereas smooth areas produce no output, because the local center intensity matches the average intensity. Fig. 10 shows a model illustrating the mechanism of the generation of a pixel's response to a spatial edge in intensity. Fig. 11 shows the actual response of the silicon retina to an edge stimulus. The output response is large at the position of the edge. The center-surround computation is a good approximation to a Laplacian filter, which is used widely in computer vision systems. Other experiments on the silicon retina including its time properties are reported in [20].

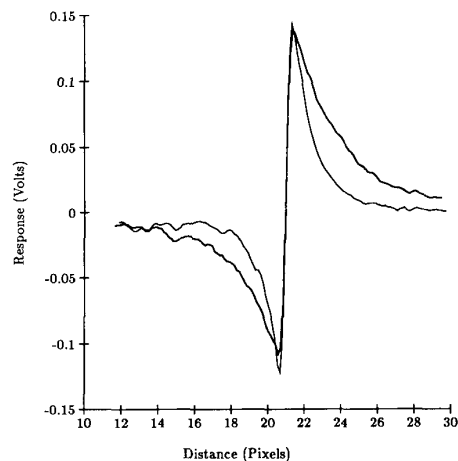


Fig. 11. Pixel output of the silicon retina in response to a contrast edge for two different space constants. Data from Mead [3].

VI. SENSORIMOTOR INTEGRATION

Neurons in the retina spatially encode information in their activity. Motor neurons, such as those that control the eye muscles in the ocular motor system, encode scalar information in their firing rates. In order to perform ocular motor functions such as saccades, place encodings from retinotopic maps must be covered into the frequency encoding used by the motor neurons [24]–[26]. We have designed and fabricated VLSI chips that convert the place encoding of a stimulus in an image into a frequency encoding for driving a motor system. These chips extract information from a visual map created using a two-dimensional array of photodetectors and local processing and then use servo techniques to create signals useful for driving motors. We shall describe a system that calculates a useful function of an image—its center of intensity and combines this information with a servo to perform simple sensory motor integration.

6.1. Center of Intensity

The silicon retina has an output for every pixel. We shall now describe a system that extracts information from the image and reduces the number of outputs that are passed on to subsequent stages of processing. The chip, the Tracker [27], calculates the center of intensity of a visual field. Its computation effectively determines the position of a bright spot in a visual image, provided that the background is sufficiently dim. The silicon retina design closely follows the biological metaphor; in contrast, the Tracker chip represents a more traditional engineering-oriented approach.

Fig. 12 shows a follower-aggregation network composed of a one-dimensional array of phototransistors and transconductance amplifiers. The network computes the weighted average of the phototransistor positions, where each position is weighted by the photocurrent in the corresponding phototransistor. We have modified the basic

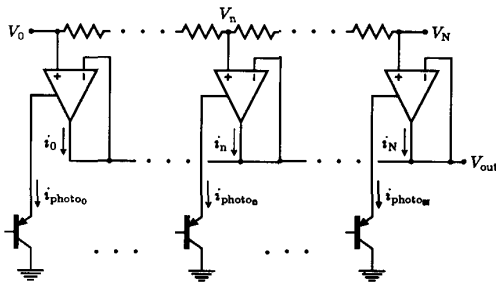


Fig. 12. Follower aggregation network with followers modified to use phototransistor currents as their bias currents. Inputs to the array correspond to photodetector positions encoded by a resistive line.

transconductance amplifier by replacing the bias transistor Q_b (Fig. 1) with a phototransistor. The amplifier bias current is supplied by the phototransistor, so the transfer function now is

$$I = I_{\text{photo}} \tanh\left(\frac{\kappa}{2V_T}(V_1 - V_2)\right) \quad \text{and} \quad G = \frac{I_{\text{photo}}\kappa}{2V_T}. \quad (14)$$

The position of each phototransistor in the array is encoded by a resistive line with end voltages V_0 and V_N . The resistive line is used as a voltage divider, which sets up a linear voltage gradient along the array. The network behaves like the one described in Section IV, with conductance set by the phototransistor and inputs set to equal the phototransistor positions.

For small $V_0 - V_N$, all amplifiers are in their linear regions and the output voltage represents the mean of the distribution of incoming light intensity. For large $V_0 - V_N$, the output voltage represents the median of the distribution, since most of the outputs will be independent of the position in the array. Detectors farther away from the bright spot contribute less to the output than they would in a weighted mean, and the localization is more accurate. The output signal levels are also much larger with a reduced sensitivity to offsets, so this mode of operation is preferred.

This center-of-intensity calculation can be generalized to two dimensions, as shown in Fig. 13. Because the two dimensions are independent, the value of the intensity at any point along each axis is taken to be the sum of the currents from the receptors in a line perpendicular to that axis. The receptors are spatially alternated, so that the currents from adjacent receptors are added to opposing axes. A 200×200 pixel version of the chip was fabricated in a $2\text{-}\mu\text{m}$ CMOS process. We used the polysilicon layer to implement the resistive voltage dividers. We tested the chip using a light-emitting diode (LED), a precision motion table, and a uniformly reflective background. The results of moving the LED along one axis are shown in Fig. 14. The output voltage correctly encoded the position of the LED along the chip. As the stimulus moved off the ends of the chip, the output voltage returned to the value obtained for a uniformly illuminated background. We repeated the procedure along different positions and we calculated the

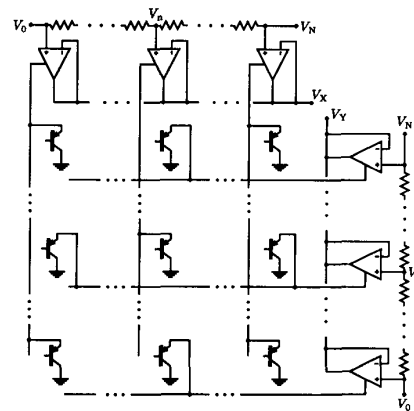


Fig. 13. Two-dimensional follower aggregation network. The phototransistor current outputs are summed onto wires running parallel to each axis. These current sums bias the transconductance amplifiers in the two follower aggregation networks located at the edges of the array. The receptors are spatially alternated, so that adjacent phototransistors contribute to opposing axes.

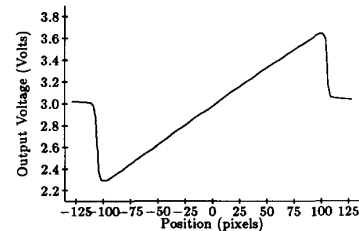


Fig. 14. Output voltage of the Tracker chip versus position of an LED along an axis. Data taken from DeWeerth and Mead [27].

relative and maximum errors to be less than 1 percent. We measured a repeatability of 0.05 percent, and a monotonicity of better than 0.1 percent. An advantage of the Tracker over commercially available schemes is that it can be extended to calculate multiple bright spots or to display selective attention to a given bright spot [28].

6.2. Sensorimotor Integration

Using traditional positional servo techniques, we developed a framework for converting simple sensory information into appropriate signals for driving actuators [29]. Sensory and servo information can be combined on the same chip, affording advantages in terms of lower pad count, higher speed, lower area, lower cost, and lower discrete-part count. We used the same follower-integrator, differential amplifier and derivative circuits as those we used to design the sensory systems. We used the Neuron circuit to convert the servo outputs into a pulse train for driving the actuators. The output pulsewidth determines the size of each elementary correction being applied to the actuator. The actuator is driven sufficiently hard by each pulse to overcome the static friction. Because the amount of drive is set by the duty cycle of the pulses and not by an analog current, power amplification is very easy to accomplish. We used a dual-rail pulse encoding for bidirectional operation of the actuators. As an example of this framework, we implemented a simple position-derivative (P-D)

servo. We combined this servo with the 2-D Tracker circuit on a single chip. The chip was able to actively track a bright spot of light, a useful engineering task.

We have also implemented a system combining a 1-dimensional version of this chip with a planar model of the oculomotor plant. The pulse outputs from the chip are used to drive a pair of motors representing the antagonistic muscles that control one axis of ocular rotation. The eyeball is modelled using a turntable with the chip and a lens mounted at its center. The system fixates on a stimulus presented to the visual field of the chip. The system is a simple model of sensorimotor integration that performs a transformation from a place encoding to a frequency encoding. We envision using more complex local processing at each pixel to extract other features of the image. We can also use this system as a basis for more realistic models of the ocular-motor system.

VII. CONCLUSION

We have demonstrated a methodology for building large analog integrated circuits that uses neural organizing principles. We presented experimental data from two systems designed using this methodology. In our laboratory, we have built other experimental sensory systems. The silicon retina enhances time derivatives and edges in an image. Other systems we have designed extract high-level visual information, such as uniform motion [30], depth from stereopsis [31], center of intensity, and edge orientation [32]. Resistive networks have been the basis for chips that interpolate, smooth, and enhance edges in surfaces that have been sparsely and noisily sampled [33]. We have designed an artificial cochlea [19] and used it with models of the binaural auditory localization from the owl in an auditory-localization system [34]. We have developed a chip that maps visual input to auditory output. We hope to assist blind people in making an internal model of their surroundings using this chip [35]. We also have built models of central pattern generators [36].

VLSI technology provides real-time, parallel architectures that allow us to implement neural systems efficiently. Building blocks are now available for translating algorithms into architecture; processing, sensing, and control can be integrated on the same chip. Biology provides examples of many intriguing engineering principles, such as the center-surround organization, local averaging, logarithmic encoding, and pulse encoding. In addition, building large-scale analog neuromorphic systems has led to novel uses of silicon technology. As we learn more about neural organizing principles, we will be able to model and emulate more complex systems. We believe our methodology will be useful for developing artificial sensory systems and prostheses, and will allow us to build analog VLSI chips that perform computations that are currently intractable on even the largest computers.

REFERENCES

- [1] G. Shepherd, "The neuron doctrine: A revision of functional concepts," *Yale J. Biol. Med.*, vol. 45, no. 6, pp. 584-599, Dec. 1972.
- [2] R. Shapley and C. Enroth Cugell, "Visual adaptation and retinal gain controls," in *Progress in Retinal Research*, N. N. Osborne and G. J. Chader, Eds., vol. 3, Oxford, England: Pergamon, 1984, p. 263.
- [3] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison Wesley, in press.
- [4] G. M. Shepherd, *Neurobiology*. New York: Oxford Univ. Press, 1983.
- [5] C. E. Carr and M. Konishi, "Axonal delay lines for time measurement in the owl's brain stem," in *Proc. Nat. Acad. of Sci.*, vol. 85, pp. 8311-8315, 1988.
- [6] M. V. Srivivasan, S. B. Laughlin and A. Dubs, "Predictive coding: A fresh view of inhibition in the retina," in *Proc. Royal Society of London*, series B, vol. 216, p. 427, 1982.
- [7] D. H. Ballard, "Cortical connections and parallel processing: Structure and function," *The Behavioral and Brain Sci.*, vol. 9, pp. 67-120, 1986.
- [8] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106-154, 1962.
- [9] H. B. Barlow and J. D. Mollon, *The Senses*. Cambridge, England: Cambridge Univ. Press, 1982.
- [10] J. M. Early, "Effects of space charge layer widening in junction transistors," *Proc. IRE*, vol. 40, pp. 1401-1406, 1952.
- [11] E. A. Vittoz, "Micropower techniques," in *Design of MOS VLSI Circuits for Telecommunications*, Y. Tsvidis and P. Antognetti, Eds., Englewood Cliffs, NJ: Prentice Hall, 1985.
- [12] R. W. Rodieck, *The Vertebrate Retina*. San Francisco, CA: Freeman, 1973.
- [13] B. A. Gilbert, "Precise four-quadrant multiplier with subnanosecond response," *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 365-373, 1968.
- [14] ———, "Translinear circuits: A proposed classification," *Electron. Lett.*, vol. 11, p. 126, 1975.
- [15] J. Lazzaro, S. Ryckebusch, M. Mahowald, and C. Mead, "Winner-take-all networks of order N complexity," in *Proc. 1988 IEEE Conf. on Neural Information Processing—Natural and Synthetic*, Denver, 1988.
- [16] G. von Bekesy, *Sensory Inhibition*. Princeton, NJ: Princeton University Press, 1967.
- [17] C. A. Mead, "A sensitive electronic photoreceptor," in *1985 Chapel Hill Conf. on Very Large Scale Integration*, H. Fuchs, Ed., Rockville, MD: Computer Sci. Press, 1985.
- [18] G. M. Shepherd, *The Synaptic Organization of the Brain*. New York: Oxford Univ. Press, 1979.
- [19] R. F. Lyon and C. A. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1119-1134, 1988.
- [20] C. A. Mead and M. A. Mahowald, "A silicon model of early visual processing," *Neural Networks*, vol. 1, no. 1, pp. 91-97, 1988.
- [21] J. Dowling, *An Approachable Part of the Brain*. Cambridge, MA: Harvard Univ. Press, 1987.
- [22] M. A. Sivilotti, M. A. Mahowald, and C. A. Mead, "Real-time visual computation using analog CMOS processing arrays," in *Advanced Research in VLSI: Proc. 1987 Stanford Conf.* Cambridge, MA: MIT Press, 1987.
- [23] F. S. Werblin, "Control of retinal sensitivity. II. Lateral interactions at the outer plexiform layer," *J. Gen. Physiol.*, vol. 63, p. 62, 1974.
- [24] D. A. Robinson, "Eye movements evoked by collicular stimulation in the alert monkey," *Vision Res.*, vol. 12, pp. 1795-1808, 1972.
- [25] M. F. Jay and D. L. Sparks, "Sensorimotor integration in the primate Superior Colliculus. I. Motor convergence," *J. Neurophysiol.*, vol. 57, no. 1, pp. 22-34, Jan. 1987.
- [26] ———, "Sensorimotor integration in the primate Superior Colliculus. II. Coordinates of auditory signals," *J. Neurophysiol.*, vol. 57, no. 1, pp. 35-55, Jan. 1987.
- [27] S. P. DeWeerth and C. A. Mead, "A two-dimensional visual tracking array," in *Advanced Research in VLSI: Proc. Fifth MIT Conf.* Cambridge, MA: MIT Press, 1988.
- [28] C. B. Umminger, "Locating bright spots in an image with analog VLSI circuitry," Senior thesis, California Inst. of Technol., 1988.
- [29] S. P. DeWeerth, "A VLSI framework for motor control," Masters thesis, California Inst. of Technol., 1987.
- [30] J. E. Tanner, *Integrated Optical Motion Detection*, Ph.D. thesis, California Inst. of Technol., 1986.
- [31] M. Mahowald and T. Delbruck, "An analog VLSI Implementation of the Marr-Poggio Stereo Correspondence Algorithm," *Neural Networks*, vol. 1 (supplement 1), p. 392, 1988.
- [32] T. Allen, C. Mead, F. Faggin, and G. Gribble, "An orientation-selective VLSI retina," in *Visual Communication and Image Processing '88, Proc. SPIE Conf.*, Nov. 9-11, pp. 1040-1046.
- [33] J. Hutchinson, C. Koch, J. Luo, and C. Mead, "Computing motion using analog and binary resistive networks," *Computer*, pp. 55-63, Mar. 1988.
- [34] J. Lazzaro and C. Mead, "Silicon models of auditory localization," *Neural Comput.*, vol. 1, no. 1, p. 1, 1989.

- [35] L. Nielsen, M. Mahowald, and C. A. Mead, "SeeHear," in *Proc. Fifth SCIA Conf.*, July 1987. **Stephen P. DeWeerth**, photograph and biography not available at time of publication.
- [36] S. Ryckebusch, C. Mead, and J. Bower, "Modeling a central pattern generator in analog CMOS VLSI," in *Proc. 1988 IEEE Conf. on Neural Information Processing—Natural and Synthetic*, Denver, CO, 1988. ✖

✖

M. A. Mahowald, photograph and biography not available at time of publication.

✖

Mary Ann C. Maher, photograph and biography not available at time of publication.

Carver A. Mead, photograph and biography not available at time of publication.
