MAY 1991
$3.95

SCIENTIFIC AMERICAN

Exploring the genetic heritage of racehorses.

Can anyons explain high-temperature superconductivity?

The impact of Kuwait's burning oil wells.

***Silicon sees** a cat. This retina-on-a-chip mimics the functions of cells in the human eye.*

The Silicon Retina

A chip based on the neural architecture of the eye proves a new, more powerful way of doing computations

by Misha A. Mahowald and Carver Mead

The eye is the window through which the mind perceives the world around it. It is also a window through which to discern the workings of the brain. The retina, a thin sheet of tissue that lines the orb of the eye, converts raw light into the nerve signals that the brain interprets as visual images. This tiny outpost of the central nervous system must extract all the essential features of the visual scene rapidly and reliably under lighting conditions that range from the dark of a moonless night to the stark glare of the noontime sun.

The retina's ability to perform these tasks outstrips that of the most powerful supercomputers. Yet individual neurons in the retina are about a million times slower than electronic devices and consume one ten-millionth as much power. They also operate with far less precision than do digital computers. Understanding how the retina manages this feat will undoubtedly yield profound insights into the computational principles of other, less accessible regions of the brain.

Clearly, biological computation must be very different from its digital counterpart. To elucidate this difference, we decided to build a silicon chip inspired by the neural architecture and function of the retina. Our artificial retina generates, in real time, outputs that mimic signals observed in real retinas. Our success persuades us that this approach not only clarifies the nature of biological computation but also demonstrates that the principles of neural information processing offer a powerful new engineering paradigm.

MISHA A. MAHOWALD and CARVER MEAD work on analog very large scale integrated circuits at the California Institute of Technology. Mahowald, a doctoral student, designs neuronally inspired vision systems. She received her B.Sc. in biology from Caltech in 1985. Mead is Gordon and Betty Moore Professor of Computer Science at Caltech, where he has taught for more than 30 years. He played a major role in the development of design methods for digital VLSI and is co-author of the standard textbook in the field. He is now working to model in silicon biological structures such as the cochlea and the retina.

Conventional electronic image-processing systems bear little resemblance to the human retina. Typically they consist of a photosensitive array that delivers signals corresponding to the absolute value of the illumination at each point in an image, backed by a formidable computer that attempts to extract geometric features from the resulting digital data.

The retina, in contrast, contains five layers of cells, through which information flows both vertically (from one layer to the next) and horizontally (among neighboring cells in the same layer). The sensing of photons and the processing of the information they contain are inextricably combined. We believe that this architecture is crucial to the formation of visual images.

The top three layers of the retina—photoreceptors, horizontal cells and bipolar cells—are the best understood. These outer layers are the ones whose organization we have chosen to emulate in the silicon retina.

The first layer consists of rod and cone cells that convert incoming light to electrical signals. Horizontal cells—the second layer—make connections to both photoreceptors and bipolar cells through the triad synapse. Each horizontal cell is also connected to its neighbors by gap junctions through which ions diffuse. The potential of any given horizontal cell is thus determined by the spatially weighted average of the potentials of cells around it. Nearby cells make the strongest contribution; distant ones, relatively less.

Each bipolar cell receives inputs from a photoreceptor and a horizontal cell and then produces a signal proportional to the difference between the two. Information from the bipolar cell passes through the amacrine cell layer to the ganglion cells and thence toward the optic nerve.

The most crucial function of these first three layers is adaptation. The photoreceptors, horizontal cells and bipolar cells take widely varying amounts of incoming light and adapt their response to produce a signal with a much narrower dynamic range that nonetheless captures the important information in a scene. Adaptation is necessary if the system is to respond sensitively to small local changes in the image against a background whose intensity may vary by a factor of a million from midnight to high noon.

The retina copes with this tremendous input range in several stages. The first biological trick is to use two different kinds of receptors: rods are sensitive to low light levels and cones to higher ones. Furthermore, the cones themselves can alter the range of light intensities to which they respond, depending on the average long-term brightness in a scene. (These adaptive mechanisms explain why people stepping into bright sunlight from semidarkness experience the scene as washed out and overexposed.)

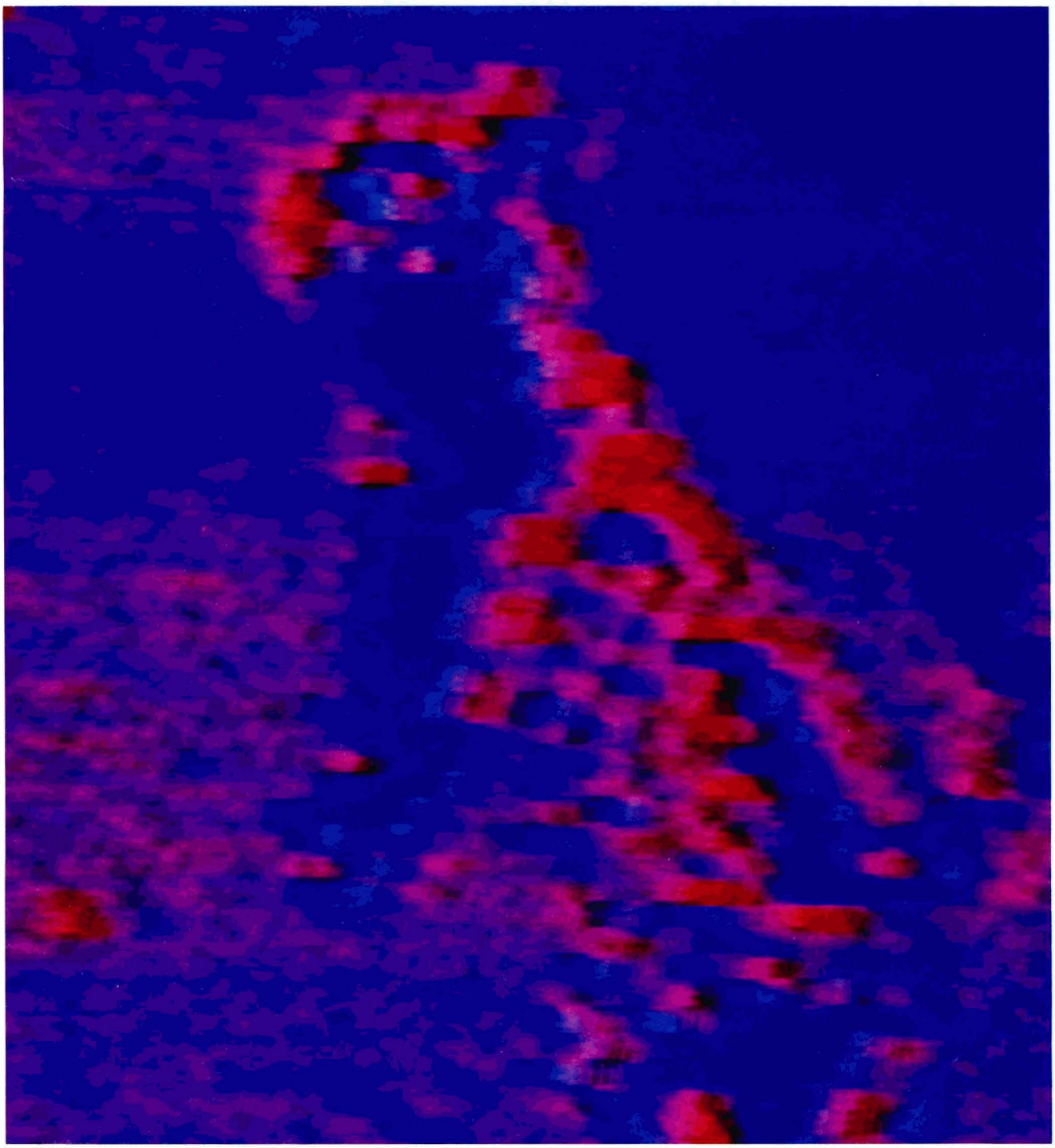
The bipolar cells have a narrower dynamic range than either the rods or the cones. The crucial element in enhancing their response to the important elements in an image is the triad synapse. The triad synapse mediates feedback between the horizontal cells and the cones. As a result, the bipolar cell does not have to respond to the absolute brightness of the scene; it responds only to the difference between the photoreceptor signal and the local average signal as computed by the horizontal cell network.

In addition, both the photoreceptors and the horizontal cells produce logarithmic signals, so that the output of the bipolar cell—the difference between the two—actually corresponds to the ratio of local light intensity to back-

ground intensity, irrespective of the absolute light level. Performing further visual processing in terms of the intensity ratio enables the retina to see detail in shaded and bright areas within the same scene.

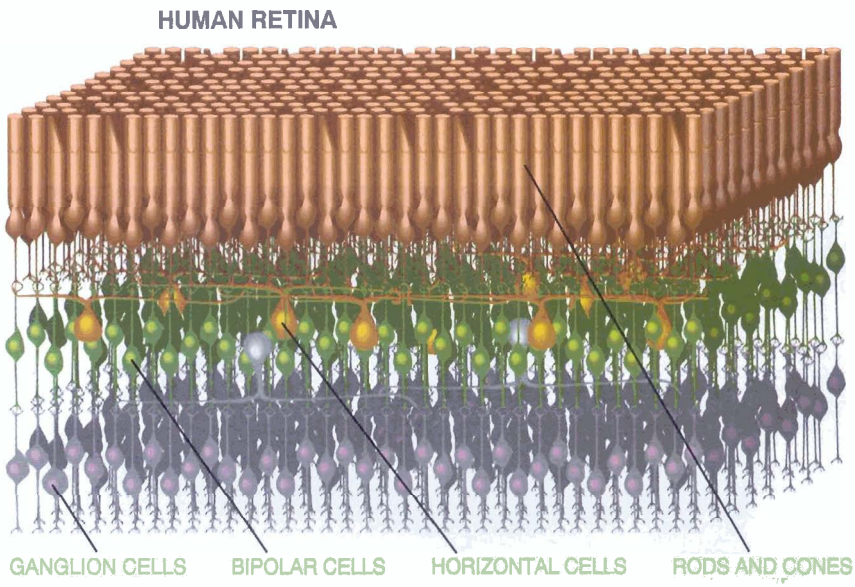
This local adaptation does not just ensure reliable signaling of small changes in image brightness. It also suppresses features of images that are not of interest while enhancing those that are. Large, uniform areas produce only weak visual signals because the impulses from any single photoreceptor are essentially canceled by the spatial average signal from the horizontal cell network. Edges, in contrast, produce strong signals because receptors on both sides of the edge sense light levels significantly different from the local average.

The relatively slow temporal response of the horizontal cell network also enhances the visual system's response to moving images. Photoreceptors produce signals from the image of a moving object while the horizontal cell signal against which they are compared is still reporting the previous intensity level. Unlike a camera, which produces a single snapshot of an image, the ret-

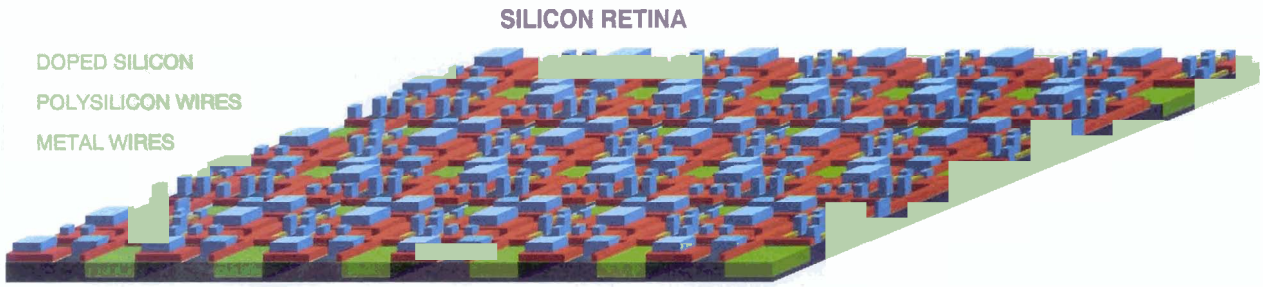


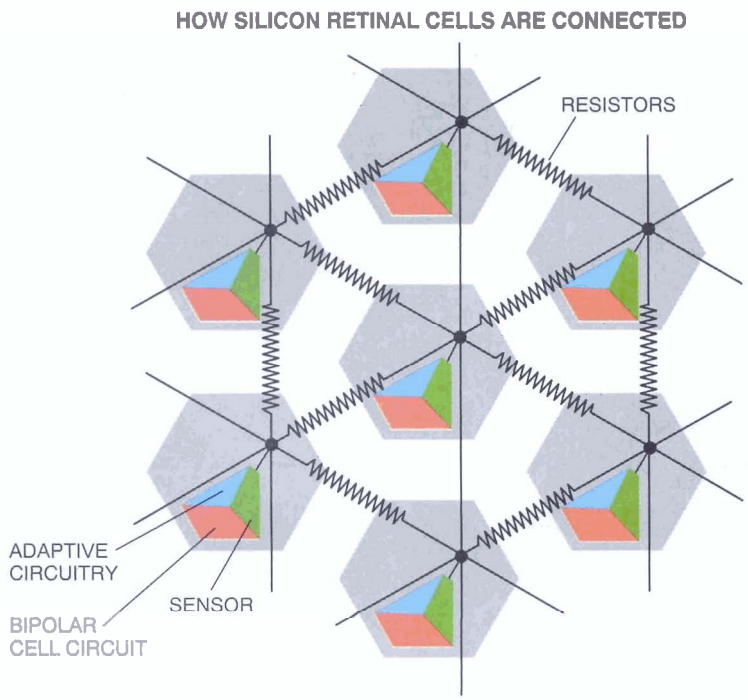
MOVING CAT as seen by silicon retina shows initial stages of biological image processing. (Areas of the image that are darker than their surroundings appear blue; those that are lighter appear red.) The retina responds most strongly to moving images: the cat's head and forelegs appear in sharp relief while stationary parts of its body fade into the background.

Modeling Neural Structures in Silicon



The human retina consists of cells that conduct neural signals both within layers and from one layer to another. The silicon retina models the functions of the outermost three layers—photoreceptors (rods and cones), horizontal cells and bipolar cells. The rods and cones transform light into electrical signals; the horizontal cells, meanwhile, respond to the average light intensity in their neighborhood. Bipolar cells transmit a signal corresponding to the ratio of the signals from rods and horizontal cells through the ganglion cells, where it is further processed before being delivered to the brain.





Each silicon photoreceptor mimics a cone cell. It contains both a photosensor and adaptive circuitry that adjusts its response to cope with changing light levels. A network of variable resistors mimics the horizontal cell layer, supplying feedback based on the average amount of light striking nearby photoreceptors. And bipolar cell circuitry amplifies the difference between the signal from the photoreceptor and the local average. The physical layout of the chip (*above*) contains circuitry in staggered blocks. Silicon areas doped with impurities (*green*) are the basis for transistors and photosensors, polysilicon (*red*) forms wires and resistors, and metal lines (*blue*) act as low-resistance wires. The functional diagram at the left shows the arrangement of receptor circuitry and the hexagonal grid of variable resistors that makes up the horizontal cell network. The response of the retinal circuit closely approximates the behavior of the human retina.

ina devotes itself largely to reporting changes.

By the mid-1980s neuroscientists had learned enough about the operation of nerves and synapses to know there is no mystery to what they do. In no single instance is there a function done by a neural element that cannot, from the point of view of a systems designer, be duplicated by electronic devices. Our goal in building a silicon retina was not to reproduce the biology to the last detail but rather to create a simplified version that contains the minimum structure needed to mimic the biological function.

Each pixel of our model retina consists of three parts: a photoreceptor, horizontal cell connections and a bipolar cell. The photoreceptor includes both a photosensitive element and a feedback loop that mimics the slow adaptive mechanism of cones in the biological retina. The photosensor, a bipolar transistor, produces a current proportional to the number of photons it absorbs. The feedback loop amplifies the difference between the instantaneous photocurrent and its long-term average level. The output voltage of this circuit is proportional to the logarithm of the light intensity.

At its utmost sensitivity, the photoreceptor can form images from light fluxes of about 100,000 photons per second—about the intensity of light from a moonlit scene focused on the chip through a standard camera lens. (That is also near the low end of the operating range of vertebrate retina cones.) Large changes in intensity saturate the photoreceptor response until it has adapted to the new light level.

To imitate the horizontal cells, we built a simple hexagonal network of resistors and capacitors. Each node in the network is linked to a single photoreceptor and, through identical variable resistors, to its six neighboring nodes. The capacitors correspond to the charge storage capacity of horizontal cell membranes, whose fine branchings present a large surface for storing ionic charge from the extracellular fluid. The resistors, meanwhile, model the gap junctions that couple adjacent horizontal cells in the vertebrate retina.

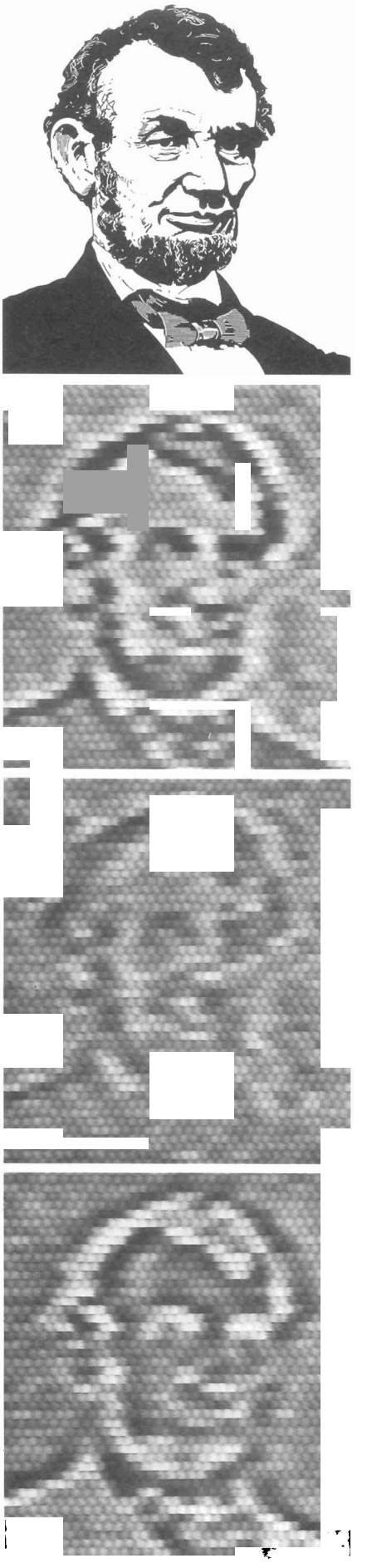
The voltage at each node in the horizontal cell network therefore presents a spatially weighted average of the photoreceptor inputs to the network. By varying the value of the resistor, we can modulate the effective area over which signals are averaged—the greater the resistance, the smaller the area over which the signals can spread. The horizontal cells also feed back to the photoreceptors and reduce their response to areas of uniform intensity.

The final output of each pixel in the silicon retina comes from an amplifier that senses the voltage difference between the output of a photoreceptor unit and the corresponding node in the horizontal cell network. The behavior of this amplifier resembles that of the vertebrate bipolar cell.

The result is a semiconductor chip containing roughly 2,500 pixels—photoreceptors and their associated image-processing circuitry—in a 50-by-50 array. The retina chip also incorporates wiring and amplifier circuits that enable us either to study the output of each pixel individually or to scan the outputs of all the pixels and feed them to a television monitor, which displays the image processed by the entire array. (The retina has gone through about 20 iterations, each requiring a few months for the chip's design and fabrication. It continues to evolve and to generate new, special-purpose designs to test particular hypotheses about image formation.)

The behavior of the adaptive retina is remarkably similar to that of biological systems. We first examined how the output of a single pixel responds to changes in light intensity when the surrounding cells are at a fixed background illumination. The shape of the response curve is similar to that of bipolar cells in the vertebrate retina. In addition, changes in the background illumination alter the potential of the horizontal cell network so that the response curve of the silicon retina shifts in the same manner as in biological retinas.

The silicon retina also has a temporal response that closely resembles that of bipolar cells. When the intensity of light is suddenly increased, there is a large jump in output voltage, equal to the difference between the new input and the previous average voltage stored in the resistive network. The response then settles down to a plateau as the



LINCOLN PORTRAIT (*top*) eventually disappears as the silicon retina adapts itself to an immobile picture. Once the retina has "adapted the image away," substitution of a blank sheet of paper yields a negative afterimage—just as the human visual system perceives afterimages when the eye looks away from bright objects. The bright band around Lincoln's head in the first image arises because the retina enhances the contrast of borders between light and dark areas.

network computes a new average voltage. When the light is suddenly decreased to its original intensity, the output voltage plunges below its original value because the network now has a larger average potential than it had originally. Finally, as the network returns to the original average value, the output also returns to its former state. In a biological retina the slow response of the horizontal cells ensures that rapid full-field changes in intensity—which might correspond to the shadow of a predator passing over an animal—pass through the bipolar cells without attenuation.

In subsequent tests, we found our silicon retina to be subject to many of the same optical illusions that humans perceive. The most obvious illusion is that of simultaneous contrast: a gray square appears darker when placed against a white background than when placed against a black background. Other illusions include the Mach bands (apparent bright and dark bands adjacent to transitions from dark to light) and the Herring grid, in which gray spots appear at the intersection of a grid of white lines [*see box on opposite page*].

Such optical illusions provide important insight into the biological retina's role in reducing the bandwidth of visual information and extracting only the essential features of the image. The illusions are created because the retina selectively encodes visual information. That our retinal model also sometimes generates an illusory output gives us additional confidence in our interpretation of the principles by which the biological retina operates.

The behavior of the artificial retina demonstrates the remarkable power of the analog computing paradigm embodied in neural circuits. The digital paradigm dominating computation today assumes that information must be digitized to guard against noise and degradation. In a digital device, voltages within a certain range are translated into bits having a value of, say, one, whereas voltages within a different range are translated into zeros. Each device along the computational pathway restores the voltages to their proper range. Digitization imposes precision on an inherently imprecise physical system.

A neuron, in contrast, is an analog device: its computations are based on smoothly varying ion currents rather than on bits representing discrete ones and zeros. Yet neural systems are superbly efficient information processors. One reason is that neural systems work with basic physics rather than trying constantly to work against it.

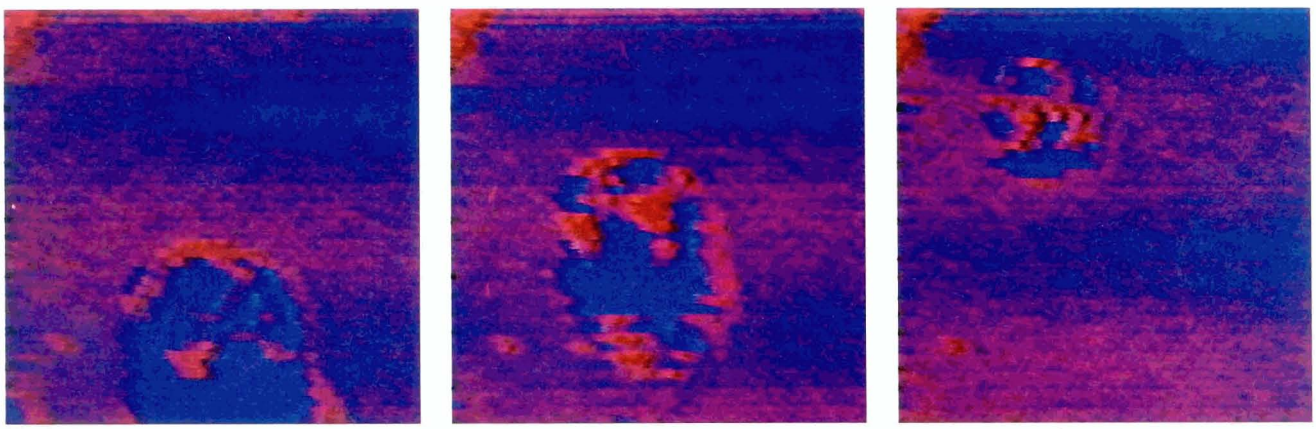
Although nature knows nothing of bits, Boolean algebra or linear systems theory, a vast array of physical phenomena implement important mathematical functions. The conservation of charge, for example, dictates that electric currents will add and subtract. Thermodynamic properties of ions cause the current flowing into a cell to be an exponential function of the voltage across the membrane.

Working with physics helps to explain why the most efficient digital integrated circuits envisioned will consume about 10^{-9} joule per operation, whereas neurons expend only 10^{-16} joule. In digital systems, data and computational operations must be converted into binary code, a process that requires about 10,000 digital voltage changes per operation. Analog devices carry out the same operation in one step and so decrease the power consumption of silicon circuits by a factor of about 10,000.

Even more important, however, the capacity of analog neural circuits to operate in unpredictable environments depends on their ability to represent information in context. They respond to differences in signal amplitude rather than to absolute signal levels, thus largely eliminating the need for precise calibration. The context for a neural signal may be the local average light intensity—as it is when a photoreceptor signal is balanced against the signal from the horizontal cell network at a triad synapse. Or it may be the previous behavior of a neural circuit itself, as in the long-term adaptation of a photoreceptor to changing light levels. The context of a signal may also be some more complex collection of neural patterns, including those that constitute learning.

The interplay of context and adaptation is a fundamental principle of the neural paradigm. It also imposes some interesting constraints on neurally inspired circuits. Because only changes and differences convey information, constant change is a necessity for neural systems—rather than a source of difficulty, as it is for digital systems. When showing an image to the digital retina, for example, we must constantly keep it in motion, or the retina will adapt and no longer perceive it. This requirement for change firmly situates a neural circuit in the world that it observes, in contrast to digital circuits, whose design implicitly assumes separation between the system and the outside world.

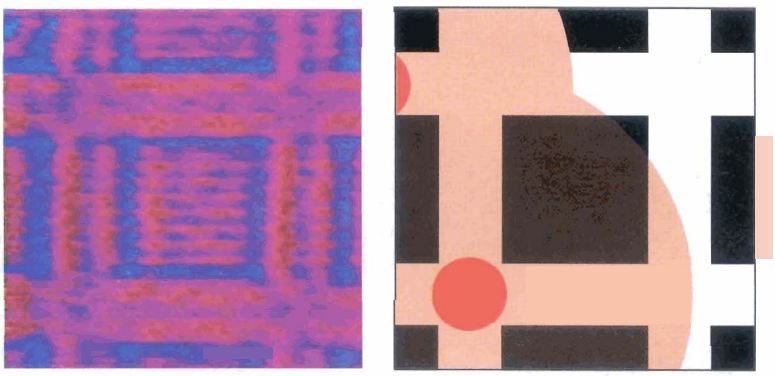
We have taken the first step in simulating the computations done by the brain to process a visual image. How readily can this strategy be extended to other types of brain computations? It may seem that the essentially two-dimensional nature of today's integrated circuits would severely limit efforts to model neural



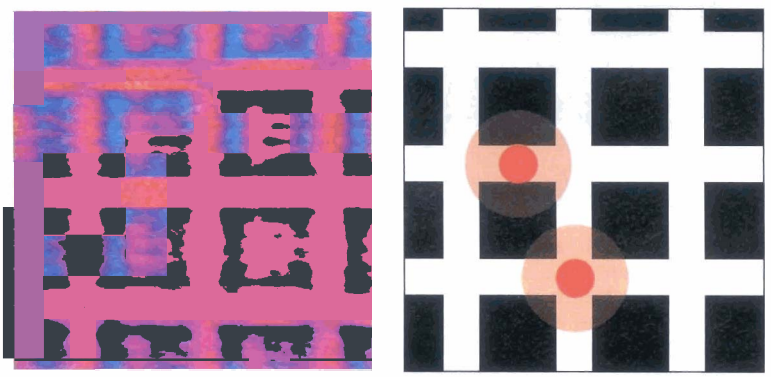
SOCCERBALL in motion shows how the delayed response of the horizontal cell network affects the retina's perception. The ball leaves behind a trail of excitation: bright where the dark spots have just passed; dark where bright parts have been.

Optical Illusions and the Silicon Retina

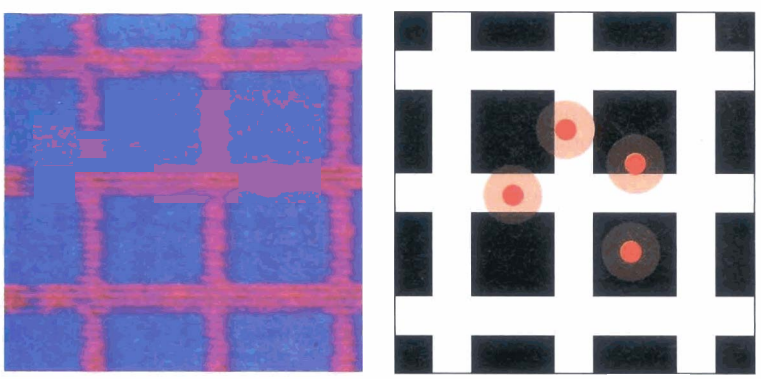
That the silicon retina is subject to some of the same misperceptions as is the human visual system suggests it has captured some essential biological principles. The Herring grid is one well-studied illusion: gray patches appear at the intersections of a grid of black squares on a white background. These patches occur because the retina's response at a given point in the visual field depends on the light intensity at nearby points. (This is the so-called center-surround effect.) The neighborhood of the intersections contains more white space and so reduces the apparent brightness of the intersection itself. A simpler example of the same effect is the illusion of simultaneous contrast (*bottom*), in which a gray square appears darker or lighter depending on the brightness of its background.



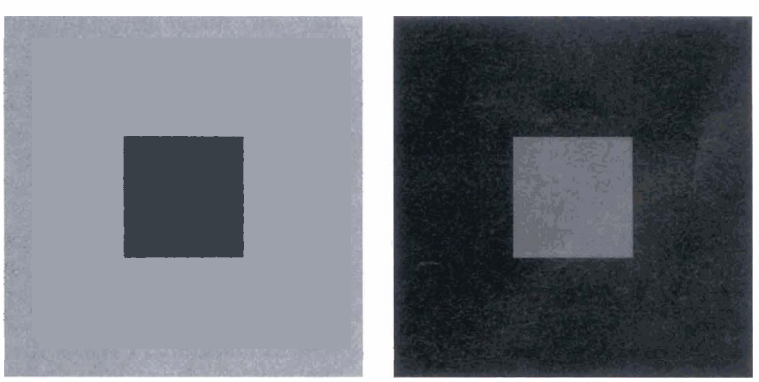
Close-up of the grid reveals no illusory brightness change because both the center and the surround of the receptive field are smaller than the space between the squares.



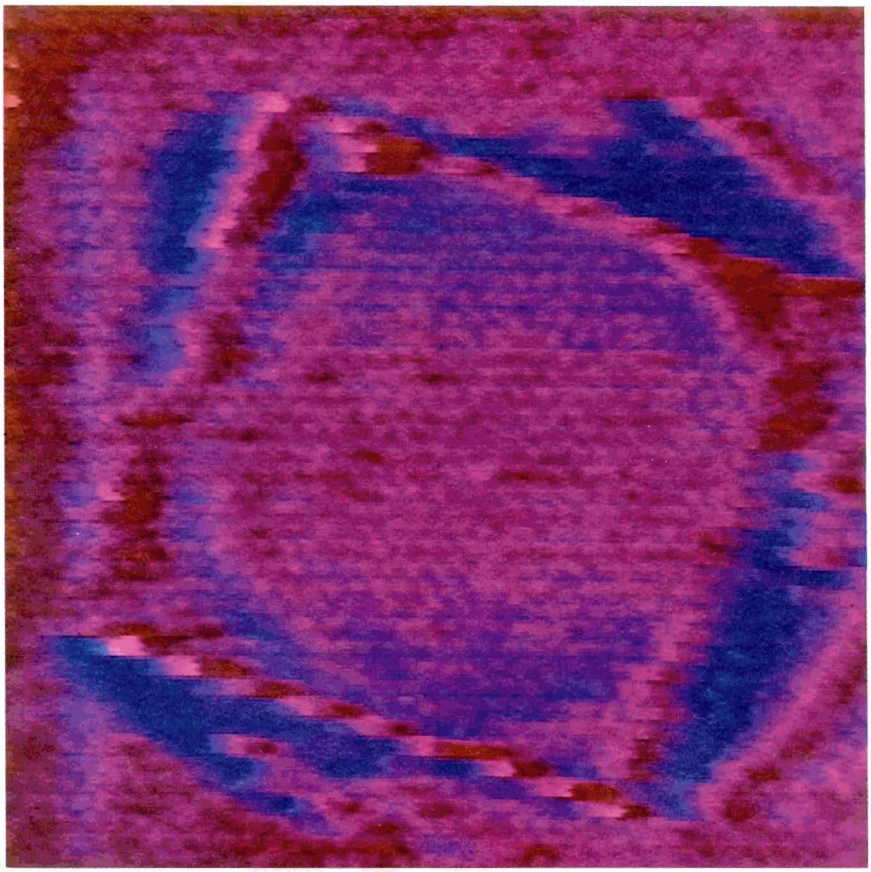
When the size of the center receptive field is comparable to the space between the squares, the illusion appears.



The illusion disappears again when the grid is viewed from a distance, because the average intensity registered by the surround is roughly the same everywhere.



The small squares in both of these images are the same shade of gray. The retina, however, perceives brightness in relation to an object's background, and so the small square on the right appears lighter.



ROTATING SQUARE appears to leave a trail of dark (*blue*) pixels as it spins. The effect results from the slow decay of the voltage in the silicon retina's horizontal cell network: the bright square increases the potential of the network so that background pixels appear dark in comparison. Meanwhile the circular area in the center of the square appears in the background color; its intensity does not change over time, and so the retina adapts it away.

tissue. But many parts of the central nervous system are in fact thin sheets that carry two-dimensional representations of computationally relevant information. The retina is merely the most obvious example. Furthermore, in both neural and silicon systems, the active devices—be they synapses or transistors—occupy no more than 1 or 2 percent of the space; "wire" occupies the remaining area. One can be sure, therefore, that the limitation of connectivity has forced the design of many parts of the brain into a highly specific form.

Specialized wiring patterns are one clear adaptation to situations in which the number of processing elements is limited by the total amount of wire needed to accomplish a computation. The brain's wiring, for instance, ensures that closely related information is mapped onto neighboring groups of neurons. As an example, the cortical areas that perform the early processing of visual information preserve the spatial relations of the image. This map-like organization of the cortex allows most of the brain's wiring to be short and highly shared. Similarly, we designed the silicon retina so that the resistors of the horizontal cell network implement computations for the entire circuit, not just the immediately adjacent cells.

The future development of the silicon retina and similar neurally inspired chips leads along two potentially divergent paths. One is the development of improved machine vision. A single chip containing an array of relatively simple analog circuits, after all, can perform the same functions as a multiple-chip system containing an image sensor and many powerful microprocessors and large memory chips. Some work is already in progress toward binocular circuits—side-by-side silicon retinas that can determine the distance of objects in a scene.

Real vision (or something somewhat closer to it than what exists now) will probably require retina chips containing perhaps 100 times more pixels as well as additional circuits that mimic the movement-sensitive and edge-enhancing functions of the amacrine and ganglion cells. Ultimately such systems will also incorporate additional neural circuits to recognize the patterns that the retina generates.

Another path will take researchers toward a grander objective: understanding the brain. For years, biologists have tacitly assumed that when they have understood the operation of each molecule in a nerve membrane, they will understand the operation of the brain. But both the digital and the analog paradigms of computation make it clear that this assumption is wrong. After all, a computer is built from a completely known arrangement of devices whose operation is understood in minute detail. Yet it is often impossible to prove that even a simple computer program will calculate its desired result or, for that matter, whether the computation will even terminate.

No matter how well the brain's architecture is mapped out, such mapping alone will not lead to a global view of the principles and representations on which the nervous system is organized. The interactions of the computations are simply too complex. If, however, workers can build silicon systems according to a deliberate and well-defined biological metaphor, they may be able to test and advance researchers' understanding of the nervous system.

The success of this venture can create a bridge between neurobiology and the information sciences, and it will also greatly deepen the understanding of computation as a physical process. It will give rise to an entirely new view of information processing that harnesses the power of analog collective systems to solve problems that are intractable by conventional digital methods.

FURTHER READING

THE CONTROL OF SENSITIVITY IN THE RETINA. Frank S. Werblin in *Scientific American*, Vol. 228, No. 1, pages 71–79; January 1973.

THE RETINA: AN APPROACHABLE PART OF THE BRAIN. John E. Dowling. Belknap Press of Harvard University Press, 1987.

ADAPTIVE RETINA. Carver Mead in *Analog VLSI Implementation of Neural Systems*. Edited by Carver Mead and Mohammed Ismail. Kluwer Academic Publishers, 1989.

AN ELECTRONIC PHOTORECEPTOR SENSITIVE TO SMALL CHANGES IN INTENSITY. T. Delbrück and C. A. Mead in *Advances in Neural Information Processing Systems I*. Edited by David Touretzky. Morgan Kaufmann Publishers, 1989.

SILICON RETINA. M. A. Mahowald and Carver Mead in *Analog VLSI and Neural Systems*. Edited by Carver Mead. Addison-Wesley, 1989.