VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function

Thesis by

Misha Mahowald

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1992

(Defended 12 May 1992)

Acknowledgments

my heart has no words to weave
the tapestry of time
of love and friendship
that are mine
as gifts

To those people who have known me from the beginning, so that all that i will ever do have a part of them in it: all my family, Mom and Dad, Sheila, Grandpa, and Aunt Tiny; my best friends, to Naomi and to Brian;

To Carver Mead who has opened the world for me;

To those who shared it: John Lazzaro, Tobi Delbrück, Dave Gillespie, Lloyd Watts, Kwabena Boahen;

To those who formed it: Massimo Sivilotti, John Wawrzynek, John Tanner, Steve DeWeerth, MaryAnn Maher, Andy Moore, Jim Campbell, Helen Derevan, William Ceasarotti;

To those who showed me to myself: Amro Umrani, John Allman, Bela Julesz, Lounette Dyer, Tom Tucker, David Feinstein;

To those who saved me from myself: Madelyn Sobel, Holly Campbell, Carlotta Glackin, Liz Smith, Carol Polansky, Thomas White, Pamela Abshire, Andy Dowsett, Todd Murray, Mark O'Dell, Tim Rentsch;

To those who welcomed me into the far out future: Christof Koch, Kevan Martin, and Rodney Douglas;

I am grateful for the weaving together of our lives, and for the strength and joy that the weaving has given me.

Abstract

This thesis describes the development and testing of a simple visual system fabricated using complementary metal-oxide-semiconductor (CMOS) very large scale integration (VLSI) technology. This visual system is composed of three subsystems. A silicon retina, fabricated on a single chip, transduces light and performs signal processing in a manner similar to a simple vertebrate retina. A stereocorrespondence chip uses bilateral retinal input to estimate the location of objects in depth. A silicon optic nerve allows communication between chips by a method that preserves the idiom of action potential transmission in the nervous system. Each of these subsystems illuminates various aspects of the relationship between VLSI analogs and their neurobiological counterparts. The overall synthetic visual system demonstrates that analog VLSI can capture a significant portion of the function of neural structures at a systems level, and concomitantly, that incorporating neural architectures leads to new engineering approaches to computation in VLSI. The relationship between neural systems and VLSI is rooted in the shared limitations imposed by computing in similar physical media. The systems discussed in this text support the belief that the physical limitations imposed by the computational medium significantly affect the evolving algorithm. Since circuits are essentially physical structures, I advocate the use of analog VLSI as powerful medium of abstraction, suitable for understanding and expressing the function of real neural systems. The working chip elevates the circuit description to a kind of synthetic formalism. The behaving physical circuit provides a formal test of theories of function that can be expressed in the language of circuits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

I have been exploring mechanisms of computation in VLSI analogs of biological visual systems. By mechanism of computation I mean a physical structure that executes an implicit algorithm mapping input to output. I argue that analog circuit design is an experimental tool for the investigation of neurobiological systems and that the analog circuits are an efficient language for expressing neural function. Furthermore, these VLSI circuits are the building blocks of entirely new kinds of computers whose algorithms are wed to their structure, and hence may be orders of magnitude more efficient at specific tasks than general-purpose computers.

The study of biological systems brings us up against the issue of the relationship between form and function, the objective and the subjective. Unlike man-made structures, whose purposes we know and whose histories we have documented, biological systems come to us without cosmic blueprint; all that we can see are the traces of the forces and constraints that gave rise to the modern organism. It is as if we are presented with the living artifacts of a magnificent civilization and are left to winnow out their purposes and functions. If we do not admit a purpose, we are reduced to a description of the brain as an organic soup of lipids, sugars and proteins. An alternative to objective description is to invoke natural selection as the process by which biological organisms have evolved. Of all the organs, the brain is unique because it is explicitly subjective. It generates abstractions to represent long-term and instantaneous information, which is used to facilitate survival. The task of perception is specifically not one of instrumentation. The goal is not to represent exactly what is out there but in some other way; instead, the brain must represent reality in the

most simplified way possible while maintaining the distinctions necessary for appropriate behavior. For example, it is critical to recognize a hungry tiger whether it is moving or standing still, in light or in shadow. The solution to this problem is not reserved for a homunculus in the higher centers of the brain who sits and watches the movie of reality through the eyes. The solution starts in the retina itself.

The evolutionary process endows the material with special significance. Man-made computational systems are created by a designer who regards the system from the outside. The creator analyzes the computation in terms of external constraints, and designs an algorithm to perform the computation which then can be implemented on a machine. Implementation is specifically excluded from algorithmic considerations. This was the approach expounded by David Marr in his classic text, *Vision*. Natural selection, however, seems unconcerned with the distinctions between objective constraint, algorithm, and implementation. The constraints on the computation are just as much a function of the material with which the machine is made as they are of the external task.

The CMOS VLSI medium is related to the neural medium at the base level. The CMOS transistor is analogous to the ligand- or voltage-sensitive channel that is the basis of neural computation. The current flow through a transistor is a monotonic function of the voltage difference across it and and a steep (exponential or square-law) function of the gate voltage. In a population of neuronal channels, the current through the membrane is a function of difference between the reversal potential and the intracellular potential, and exponentially gated by the intracellular potential or a transmitter concentration. In addition to these gain elements, threshold and saturating nonlinearities are computational primitives of both media. For example, current-limited transistors might be analogous to finite channel density in the nerve membrane. In addition, elementary arithmetic functions arise from Kirchoff's laws and systems dynamics are an inevitable consequence of conductances and capacitances. At a systems level, both neural and CMOS VLSI systems are made of large numbers of mismatched elements. This property necessitates mechanisms of self-calibration. Furthermore, limitations in the amount of DNA and in the technical difficulty of a circuit design both favor systems that are specified algorithmically. The computations performed by both of these systems are limited by the physical medium. Wiring density, bandwidth limitations, and the high cost of energy must be taken into account.

If we could succeed in defining algorithms that are consistent with the properties of the VLSI medium, we could create new machines that are far more computationally efficient than the general-purpose digital computer in use today. The digital representation allows the computation to proceed in spite of element mismatch but strips the transistor of much of its intrinsic computational power. The digital encoding of a number is susceptible to device failure, since single failures are just as likely to affect the most-significant as the least-significant bit. The fight against the intrinsic capacitances of the material leads to large power consumption as devices are forced to switch as quickly as possible. Biological systems have evolved algorithms and representations that are efficient and robust and which take advantage of natural dynamics. To the extent that neural systems and VLSI systems share the same fundamental hardware limitations and attempt to perform similar tasks, the algorithms and representations of neural systems will be appropriate for incorporation into VLSI circuits.

If the computational requirements of the algorithm are tailored more and more closely to the requirements of the medium, the specification of the algorithm may be most conveniently rendered by a description of the material that embodies it. Certainly in one sense, the best description of an organism is the organism itself. However, in contrast to description, understanding specifically requires the creation of an abstraction that is not a replica of the organism. This abstraction must capture essential qualities and relationships, which can be generalized to similar situations. To understand highly evolved biological systems, it is likely that the abstractions that preserve the essential qualities of the form of the organism will most efficiently communicate its function.

Analog circuits are abstractions capable of representing many of the essential qualities of neural systems. They are a more natural basis for representing neural circuits than words or equations because the patterns of interrelationship that are the essence of the collective function are implicit in the form of the circuit itself. Although a circuit may be physically instantiated, it also has an ideal existence. Like the circuits in the brain, analog circuits cannot physically be taken out of context because then they are not circuits anymore–their circular nature is interrupted. However, they can be distinguished and thought of as entities in much the same way as a word in a language. The attributes of this ideal circuit remain physical in nature. The circuit as abstraction is well situated between undifferentiated ideal

and uninformed matter.

The fundamental difference between the biological system and the analog is that the analog is explicitly designed to perform a particular function. Its circuits are defined by their functional significance, rather than an arbitrary morphological demarkation. The functional approach to understanding neural systems was introduced by Gordon Shepherd. He points out that the obvious physical characteristics of neural systems, like the cell boundaries of the neuron, are not the appropriate conceptual units of the nervous system. He states, "the neuron can no longer be regarded as the basic functional unit of the nervous system; rather, the nervous system is organized on the basis of functional units whose identity in many cases is independent of neuronal boundaries." The functional units can only be discerned by their participation in the computation. Because the analog is created for a purpose, the functional boundaries of the circuits are clear.

Existing both in the domain of form and function, the analog system provides a framework in which to integrate information from disparate experimental and theoretical techniques. It incorporates the electrical characteristics that are the domain of electrophysiology, the behavioral characteristics that are the domain of psychophysics and the purposive characteristics that are the domain of engineering. At all of these levels of complexity, the analog can be compared to the neural system. This dialectical interaction between the analog and the real system depends simultaneously on the similarities and the differences between the two media.

Analog VLSI is, ultimately, a synthetic tool for understanding neural systems. Although the circuit exists in an abstract sense, its true power is revealed only when it is realized. The instantiated circuits exist as written words that express neural function and our understanding is tested by the performance of the systems that we build.

1.1 Overview

I describe the components of a primitive analog CMOS vision system for doing real-time stereopsis. These components include a silicon retina, an optic nerve, and a stereoscopic matching array. The choice of representation of information and the computations performed using that representation are suggested by the properties of the analog medium in

which the system is implemented.

1.1.1 The Silicon Retina

Chapter 1 describes the silicon retina. The silicon retina shares several features with its biological counterpart. Transducers and computational elements (silicon neurons) are arrayed in a thin layer over a two-dimensional surface. The lens focuses the image directly on the transducers. Image processing occurs in parallel at each node of the array. The computation proceeds in real-time. Like the biological retina, the silicon retina acts to reduce the bandwidth needed to communicate reliable information. The need for data compression arises because ambient light intensity varies over many orders of magnitude, yet the local intensity variation in a single image is usually small. In the presence of noise, communication of absolute image intensity would require a large dynamic range to encode reliably small differences in image intensity over the full possible illumination range. The retina reduces the bandwidth by subtracting average intensity levels from the image and reporting only spatial and temporal changes. Adaptation at the photoreceptor level prevents amplification of static component mismatch.

The goal of keeping retinal output confined to a reasonable bandwidth creates an abstract representation of an image. For example, by reporting only contrast edges the retinal output provides sensory invariance. The contrast of white square on a black background is invariant under changes in illumination even though the photon flux from the black background in bright illumination may be larger than the photon flux from the white square in dim illumination. The abstraction created by the silicon retina results in output patterns that are reminiscent of several visual illusions, such as simultaneous contrast, the Herring grid illusion and the Mach band illusion. The process of adaptation results in the formation of after-images, since the receptors cannot tell the difference between internal miscalibration and a persistent pattern of illumination.

The advantages of using a network of resistors and conductances to perform this computation are discussed. The concept of conductance, typically seen as a passive discrete element, is generalized to include the action of nonlinear feedback circuits. A relationship between feedback inhibition and the conductance properties of light sensitive channels in vertebrate cones is postulated.

The abstraction created by the silicon retina has several ramifications for further visual processing in our analog CMOS system. The form of retinal output suggests a novel representation that allows data to be transferred reliably between chips.

1.1.2 Optic Nerve

Communication of high bandwidth information between chips is a major impediment to progress in building parallel, neural-network computers. I have designed an interchip communication protocol that takes advantage of the representation of the visual world created by the silicon retina. Experimental results are reported from a simple two chip system: a silicon retina and a receiver that copies the image from the retina. I call this system the silicon optic nerve.

The silicon optic nerve is based on a self-timed digital multiplexing technique which uses an *address-event representation*. The address-event representation has much in common with the action-potential representation used by real neurons. Like neuronal action potentials, events in this system are stereotyped digital amplitude events and the interval between events is analog. Information is encoded in the time between events. The action potential events of all the neurons are transmitted one at a time in an asynchronous fashion as the address of the neuron that issued the event. This encoding scheme reduces N wires to $(1 + \log_2 N)$ wires. The retinal encoding of visual information insures that only a few of the neurons in the retinal array will be firing with high spike rates in response to the image. This protocol devotes all of the bandwidth of the bus to transmitting accurate temporal information when the data rate is low. As the data rate increases, the protocol resembles more and more closely the traditional sequential scanning methods of data transfer.

Because the address-event representation preserves timing information, it is particularly suited for signaling dynamic events. These dynamic events are an integral part of real-time sensorimotor processing. If this data transmission protocol is widely adopted, it will allow many types of chips to easily be interfaced to each other.

1.1.3 Stereopsis

I have designed a chip for fusing data from two one-dimensional retinal regions into a single depth image. The chip is designed to use the communications framework described

above to receive data from two retina chips. Because the retinal output using the address-event representation has not been perfected yet, artificial data was provided by a hardware interface to a digital computer.

The algorithm for stereo matching embedded in the stereo chip is novel. It evolved from an earlier attempt to do stereo matching on an analog VLSI chip that was based on Marr and Poggio's cooperative stereo correspondence algorithm. Marr and Poggio's algorithm uses a place valued encoding of disparity; it requires an array of correlators, each one tuned to a different disparity. Positive feedback between correlators in the array helps disambiguate matching, so that fusion can be achieved even in dense arrays of identical targets in which there are many possible false matches. The pattern of interaction between correlators limits the images which can be correctly fused to those that are in a fronto-parallel plane with respect to the observer. We have extended their algorithm so that it performs correctly on images that are tilted in depth. The innovation in this chip is the transformation of place encoding into an analog value encoding of disparity. This is convenient because the analog domain provides a natural representation for surface interpolation. The analog encoding is used to guide the stereo matching process that takes place in the correlator array.

The performance of this chip can explain the human performance on dot patterns described by Mitchison and McKee. These patterns are constructed with combinations of ambiguous and unambiguous targets and attempt to reveal the matching strategy used by the visual system. The results obtained with these patterns have not previously been explained by any computational model. The correspondence between the performance of the circuit and humans on these patterns, suggests that the model may be capturing something fundamental about the way stereoscopic fusion is achieved by the visual system.

The nodes of the analog circuit mimic the response of the major disparity cell types observed in the macaque monkey primary visual cortex. There is no generally accepted biological explanation of how the tuning characteristics arise, or what their computational function is. However, since the circuit was designed to perform a specific computation, the computational function and origin of these tuning characteristics are known in this case. The tuning characteristics of these nodes cannot be understood by analysis of inputs. Instead the responses are a result of the neuron's embedding in a nonlinear network. There is no proof that the analog circuit is performing the same computation that is performed

in the monkey; however, the circuit does provide a number of hypotheses of brain function that are testable. For example, the analogy between a particular class of nodes in the circuit and disparity flat cells suggests that the disparity flat cells are smooth inhibitory cells. This hypothesis can be tested using currently available electrophysiological techniques.

Chapter 2

The Silicon Retina

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 encoding of visual information in the retina is generated, in large part, by the initial analog stages of retinal processing, from the photoreceptors through the outer-plexiform layer (OPL). Processing in the OPL relies on lateral inhibition to adapt the system to a wide range of viewing conditions, and to produce an output that is mostly independent of the absolute illumination level. A byproduct of lateral inhibition is the enhancement of spatial and temporal changes in the image.

In collaboration with Carver Mead, I have designed two versions of silicon retinas modeled on the outer-plexiform layer (OPL) of the vertebrate retina. These chips generate, in real time, outputs that correspond directly to signals observed in the corresponding levels of biological retinas. In this chapter, I describe the silicon retinas and compare and contrast their performances. I interpret some of the biophysical mechanisms of signal processing in the OPL of the vertebrate retina in light of these silicon circuits.

2.1 Vertebrate Retina

The retina has been the subject of a tremendous number of investigations (see Dowling [8] for a review). Although the details of each animal's retina are unique, the gross structure

of the retina has been conserved throughout the vertebrates.

2.1.1 Basic Anatomy

The major divisions of the retina are shown in cross-section in Figure 2.1. 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 (OPL). The inner-plexiform layer (IPL), 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. The OPL and IPL are the site of interaction between the various cell types of the retina.

2.1.2 Function of the Outer-Plexiform Layer

The most salient feature of the OPL is its ability to adapt to prevailing light conditions. The photoreceptors, horizontal cells and bipolar cells take widely varying amounts of incoming light and produce a signal with much narrower dynamic range that nonetheless captures the important information in an image. The outer-plexiform layer allows a system with limited output range and finite analog resolution to communicate small local changes in image intensity when the background intensities may vary by a factor of one million.

The initial stage of retinal processing is performed by the photoreceptors that transduce light into an analog electrical signal. In fact, all of the neurons in the OPL represent information with smoothly varying analog signals, rather than action potentials used by most neurons. The photoreceptor amplifies the photon-event with a second messenger cascade. The absorption of a single photon activates an enzyme that catalyzes the destruction of many molecules of cGMP. Lowering the cGMP concentration causes sodium-permeable channels to close and the cell becomes hyperpolarized.

Because the photoreceptor must respond over several orders of magnitude in photon flux, it must change its gain to be commensurate with the average number of incoming photons. Cones possess an intrinsic light-adaptation mechanism operating over a time

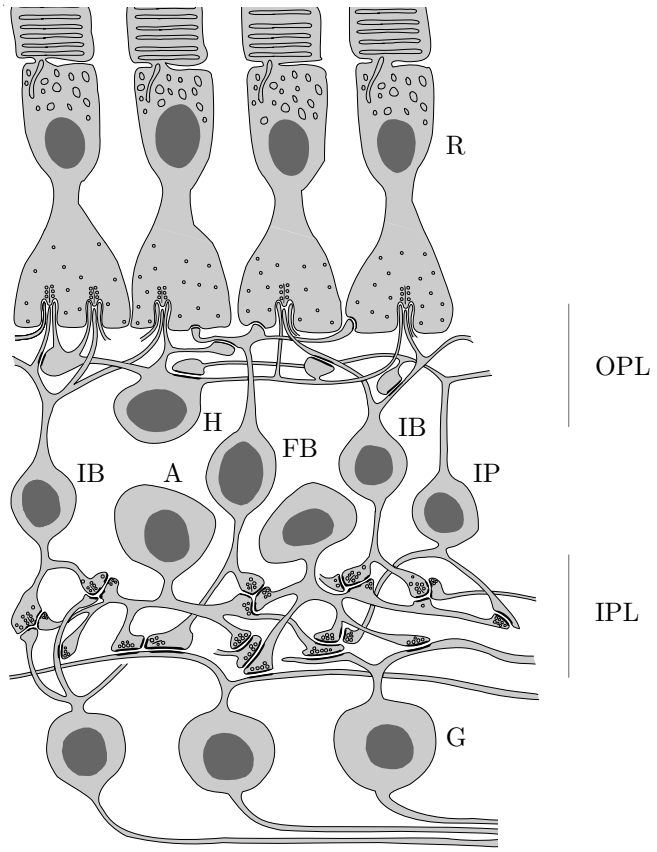

Figure 2.1: Artist's conception of a cross-section of a primate retina, indicating the primary cell types and signal pathways. The outer-plexiform layer is beneath the foot of the photoreceptors. The invagination into the foot of the photoreceptor is the site of the triad synapse. In the center of the 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. Adapted from Dowling [8]

course of a couple of seconds that is based on calcium regulation of cGMP synthesis (for a review see [26]). Calcium enters the cell through open sodium-permeable channels and is removed by a pump. When channels close in response to light stimulation, more calcium is pumped out than is flowing in so the intracellular concentration of calcium decreases to a lower equilibrium value. Low calcium concentration stimulates the synthesis of cGMP so the light-adapted cone recovers quickly from the absorption of light (the destruction of cGMP). Light-adaptation allows the cone to respond at high illumination without depleting its store of sodium-permeable channels and consequently saturating its voltage to maximum hyperpolarization. A more rapid adjustment of the electrical operating point of the cone is provided by interactions between the cones and the OPL network.

Some of the cellular interactions in the OPL are summarized briefly here. The cones make excitatory (glutaminergic) synapses onto both horizontal cells and bipolar cells [8]. The horizontal cells inhibit both the cones and the bipolar cells by electrogenic GABA release [15]. In addition, horizontal cells are electrically coupled to each other with gap junctions. The gap junctions couple the horizontal cells into a resistive sheet. The sheet has capacitance to the extra-cellular fluid due to the cell membranes. The horizontal cells thus compute a spatially and temporally smoothed version of the photoreceptor signal. This average is used as a reference point for the system.

The horizontal cells provide two forms of lateral inhibition, one by feedback inhibition to the cones, the other by feedforward inhibition of the bipolar cells. Feedback inhibition allows the cones to respond with high gain to small local changes in illumination and still span a large input range [31]. Feedforward inhibition to the bipolar cells gives these cells their classical center-surround receptive field structure; bipolar cells amplify the difference between the average computed by horizontal cells and the local photoreceptors [38]. Hyperpolarizing and depolarizing bipolar cells transmit a differential signal to the retinal ganglion cells.

2.2 Silicon Retina

Several versions of the silicon retina have been previously described [17, 18, 22, 20]. Although each version is different, they have several features in common. For example, because

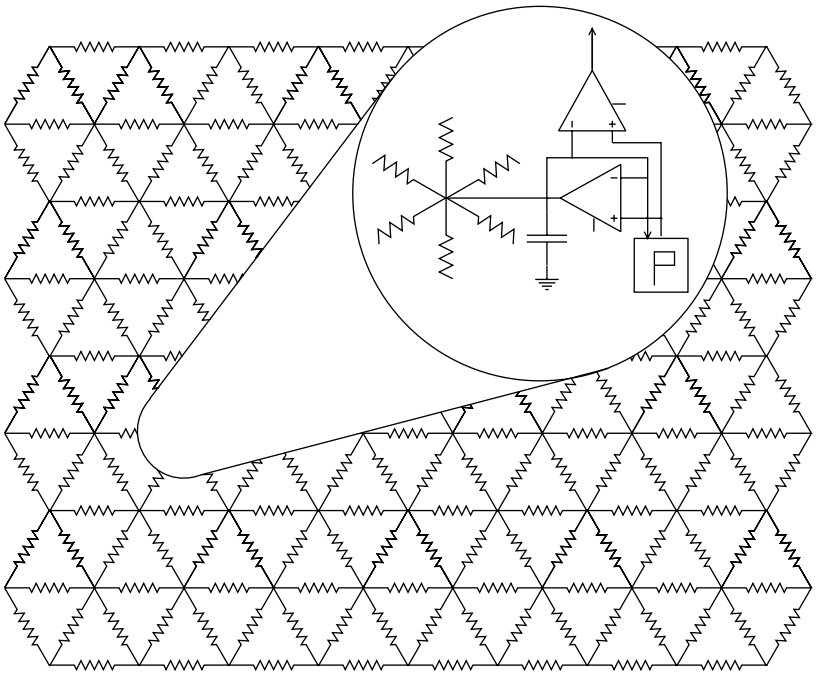

Figure 2.2: Diagram of the silicon retina showing the resistive network; a single pixel element is illustrated in the circular window. The silicon model of the triad synapse consists of a follower-connected transconductance amplifier by which the photoreceptor drives the resistive network, and an amplifier 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.

they are implemented on a physical substrate, they have a straightforward structural relationship to the vertebrate retina. A simplified plan of a typical silicon retina is shown in Figure 2.2. This view emphasizes the lateral spread of the resistive network, corresponding to the horizontal cell layer. The image signal is transduced and processed in parallel by circuitry at each node of the network.

The computations performed by these retinas are based on the interaction between the photoreceptor, the horizontal cells and the bipolar cells in the OPL. These retinas include the following elements:

1. A phototransducing element that generates a current proportional to the light intensity coupled with an MOS transistor whose subthreshold voltage-current relation is logarithmic.

2. A resistive network modeling the horizontal cell layer that spatially and temporally

averages the photoreceptor output.

3. A bipolar cell output that is proportional to the difference between the phototransduced signal and the horizontal cell signal.

The data presented here are taken from two types of silicon retina. These retina are are described in parallel in the text. The first retina is a feedforward retina [17]. The feedforward retina is so called because the signal path is in the forward and lateral directions only. It demonstrates the ability of the resistive network to be used for lateral inhibition. The simple resistive structure gives rise to complex spatiotemporal behavior. The second retina described here [18] is an extension of the first. It includes feedback from the resistive network to the photoreceptors. In addition, the photoreceptor includes a mechanism for light adaptation that also cancels transistor mismatch to improve imaging performance. Analogies between the feedback retina and the vertebrate retina lead to a new interpretation of biophysical phenomena observed in the OPL.

2.2.1 Photoreceptor Circuits

The photoreceptor transduces light into an electrical signal. The logarithmic nature of the response 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 [28]. 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 [8]. The logarithmic nature of the response has an important system-level consequence: 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 pnp bipolar transistor, which occurs as a natural byproduct in the CMOS process [2]. 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 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 holes for every photon absorbed by the structure.

The current from the photodetector is fed into an MOS transistor arrangement. The operation of an MOS trasistor in the subthreshold regime is described by Mead in *Analog VLSI and Neural Systems* [2]. The current-voltage relation of a MOS transistor operating in the subthreshold regime is exponential.

$$I_{DS} = \mathrm{I}_0 e^{\kappa(V_\mathrm{g} - V_\mathrm{s})}(1 - e^{-(V_\mathrm{D} - V_\mathrm{S})})$$

The voltage on the gate of the transistor required to supply a particular current is proportional to the logarithm of the current. The MOS transistor transforms the current from the photodetector that is proportional to the light intensity, to a voltage that is logarithmic in the light intensity.

Logarithmic compression can be used to compress a large input range into a smaller output range. This was the approach adopted in the feedforward silicon retina. However, this compression leads to a lack of sensitivity. The feedforward silicon retina produced a mottled output because the intensity variations within a uniformily illuminated scene are small, roughly the same order as the transistor mismatch. In fact, in biological systems, the range of response of the cones at a particular level of background light adaptation spans only a fraction of the perceptual range [31, 14]. The region of sensitivity of the cones is shifted by feedback from the horizontal cell network. This interaction was incorporated in the feedback silicon retina to increase sensitivity. However, in addition to amplifying the visual signal, feedback of this kind amplifies static transistor mismatch, so an adaptive mechanism

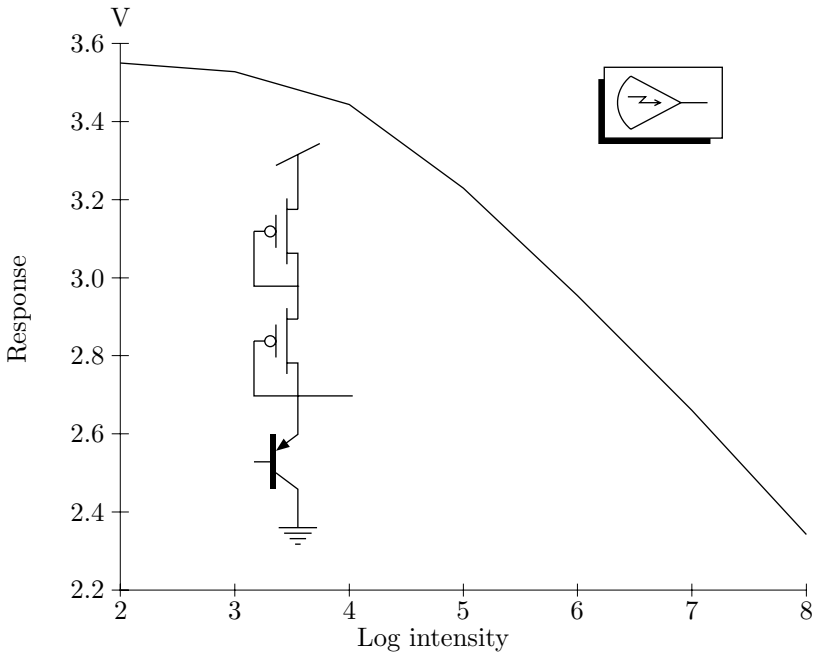

Figure 2.3: Measured response of a logarithmic photoreceptor. Photocurrent is proportional to incident-light intensity. Response is logarithmic over more than four orders of magnitude in intensity. Direct exposure of the chip to room illumination resulted in an output voltage of 2.1 volts. The symbol for the photoreceptor circuit is shown in the inset.

was incorporated into the photoreceptor in the feedback retina. These two photoreceptors are described sequentially in the following text.

The low-gain photoreceptor in the feedforward retina, depicted in Figure 2.3, uses two diode-connected MOS transistors in series to supply the photocurrent. The photocurrent biases these transistors in the subthreshold region. This arrangement produces a voltage proportional to the logarithm of the current, and therefore to the logarithm of the incoming intensity. The constant of proportionality is $V_{\text{out}} \propto \kappa/(\kappa + 1)$, rather than κ as is the case for a single diode, because the change in the output voltage of the second diode must compensate for the change in the gate voltage with current of the first diode. 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 VDD that the resistor bias circuit described by Mead [2] cannot adequately bias the horizontal resistive connections.

The voltage out of the low-gain photoreceptor circuit is logarithmic over four to five orders of magnitude of incoming light intensity, as shown in Figure 2.3. 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. At high light levels, the diode-connected transistors enter the above-threshold operating regime where the output voltage goes as the square root of the current.

The photoreceptor of the feedback retina, shown in Figure 2.4, includes a transducing element embedded in a feedback loop from a high-gain amplifier. The photocurrent is supplied from the power supply through the action of transistor Q1. The current through Q1 clamps the emitter voltage, V_{E}, to be equal to the absolute value of the gate-source voltage on the bias transistor, V_{bias}. Small variations in V_{E} are amplified by the inverting stage comprising the bias transistor and Q2. The output of the receptor, V_{out}, is the voltage output of the inverting amplifier. Because Q1 has an exponential current-voltage relation in subthreshold, the voltage response of the receptor is proportional to the logarithm of the light intensity. To the extent that the gain of the amplifier is effective at clamping the emitter voltage, the gate voltage, W of Q1 is related to the current through the bipolar transducer by the equation

$$W = \left(\frac{1}{\kappa}\right)\left(\ln\frac{I_{\mathrm{photo}}}{\mathrm{I_0}} + (\mathrm{V_{DD}} - V_{\mathrm{bias}})\right).$$

If the horizontal network potential is held fixed, this gate voltage is modulated exclusively by the capacitor, $\mathrm{C_1}$, coupling it to the output of the inverting amplifier. The change in gate voltage, W, is related to a change in the output voltage, V_{out}, by the equation:

$$\delta W = \frac{\mathrm{C_1}}{\mathrm{C_F} + \mathrm{C_2} + \mathrm{C_1}}\delta V_{\mathrm{out}},$$

So the final output of the high-gain receptor is given by:

$$V_{\mathrm{out}} = \left(\frac{\mathrm{C_F} + \mathrm{C_2} + \mathrm{C_1}}{\mathrm{C_1}\kappa}\right)\left(\ln\frac{I_{\mathrm{photo}}}{\mathrm{I_0}} + (\mathrm{V_{DD}} - V_{\mathrm{bias}})\right).$$

The high-gain photoreceptor circuit has higher sensitivity and a commensurately reduced

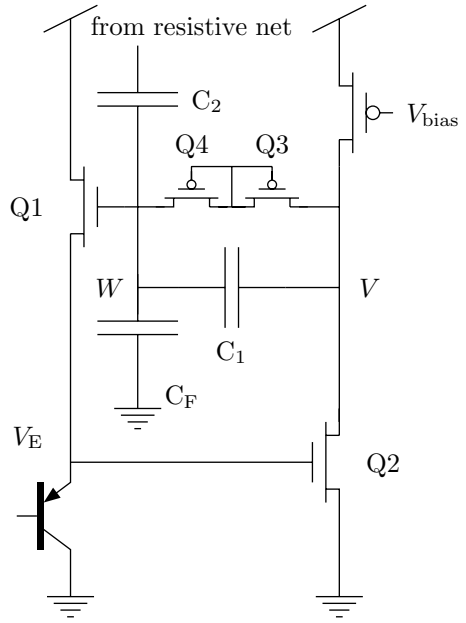

Figure 2.4: Schematic of the high-gain photoreceptor in the feedback retina.

output range relative to the low-gain receptor. In order to function over a wide input range, the operating point of the receptor must shift. The operating point of the receptor is controlled for short times by feedback from the horizontal cell network and in the long term by adaptation within the photoreceptor itself.

2.2.2 Adaptation

A silicon analogue of light-adaptation in cones was first incorporated in a silicon retina by Mead [22] using an ultraviolet-programmable floating gate. As is done in the tiger salamander retina, the operating point of a high-gain receptor was modulated by feedback from the resistive network [31]. Although the transduction processes in cones and in silicon are unrelated, slow adaptation plays an important role in silicon circuits. Slow adaptation was incorporated in the Mead retina in order to keep transistor mismatches from being amplified by the feedback from the resistive network. After adaptation this retina responded quite well to low-contrast images without offsets. However, the adaptation needed to be repeated if the background light level changed significantly. Because ultraviolet could not be used continuously to adapt the chip while it was in operation, this adaptive retina had practical limitations.

Adaptation in the high-gain receptor is mediated by slow negative feedback through the diode connected transistors, Q3 and Q4. Adaptation reduces the gain of the receptor for long times to $kT/q\kappa$ volts per e-fold increase in photocurrent from the transducer. The time scale of this adaptation is set by the leakage current through Q3 or Q4 and the amount of capacitance on node W. These transistors share a common gate that is tied to the well in which the transistors are sitting. No matter which way the light changes, one of the diodes will be reversed biased.

As current flows through the diodes onto node W, current flows back out through the coupling capacitors connected to the node. Adaptation through the diodes allows the photoreceptor and the horizontal network to relax. This slow adaptation insures that offsets between transistors will not be amplified in a sustained way by feedback from the resistive network. Continuous adaptation of this kind provides a "single point correction" at every operating point.

2.2.3 Horizontal Resistive Layer

The retina provides an excellent example of the computation that can be performed using a resistive network. The horizontal cells are connected to one another by gap junctions to form an electrically continuous network in which signals propagate by electrotonic spread [8]. 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.

Inspired by the horizontal cells of the retina, the silicon retina was the first VLSI system to incorporate a resistive network to perform computation. Each photoreceptor in the network is linked to its six neighbors with resistive elements, to form the hexagonal array shown in Figure 2.2. 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. 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 and temporally weighted average of photoreceptor inputs. The spatial scale of the weighting function is affected 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. The time constant of intergration is determined by the capacitance at each node of the network and the magnitude of the conductance. The space constant and time constant of integration can be varied independently.

The spread of activity in a passive resistive network is analyzed extensively in *Analog VLSI and Neural Systems* [2]. This analysis applies to the feedforward retina in which the voltage output of the photoreceptors is unaffected by the voltage of the network itself. A short summary of the analysis of a one-dimensional passive network, shown in Figure 2.5, is provided here. This analysis is extended to the feedback retina, in which the network activity modifies the magnitude of the voltage sources driving it. The equations show that signals propagate with exponential decay in the feedback network just as in the feedforward network, with an appropriate change in variables. Whereas in the feedforward network, signal propagation depends only on the passive components, R and G, in the feedback network, signal propagation depends also on the active gain of the feedback loop.

The analysis of the passive network begins with conservation of charge. By Kirchoff's current law, the network obeys the equation:

$$\mathrm{G}(V_n - U_n) = \frac{U_n - U_{n+1}}{\mathrm{R}} + \frac{U_n - U_{n-1}}{\mathrm{R}}$$

Rearranging to get the driving term on one side, the equation becomes:

$$\mathrm{GR}(-V_n) = U_{n+1} - 2\alpha U_n + U_{n-1},$$

where $2\alpha = 2 + \mathrm{RG}$. Set all the V_n equal to zero and guess that the form of the solution for U_n will be $\mathrm{A}\gamma^n$. Then divide out a factor of $\mathrm{A}\gamma^{n-1}$ to derive the characteristic equation:

$$0 = \gamma^2 - 2\alpha\gamma + 1.$$

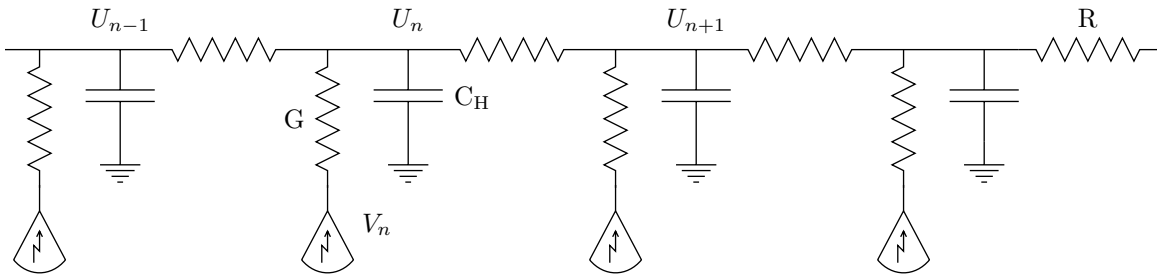

Figure 2.5: One-dimensional passive resistive network driven by photoreceptors acting as voltage sources.

The solution to this equation, derived with the quadratic formula, is:

$$\gamma = \alpha - \sqrt{\alpha^2 - 1}.$$

Substituting for α, the solution becomes:

$$\gamma = 1 + \mathrm{RG}/2 - \sqrt{\mathrm{RG}}\sqrt{1 + \mathrm{RG}/4}.$$

This root is chosen to insure that the solution decays to zero for infinite n. The solution for negative n must be the same by the symmetry of the network.

The solution decays exponentially from the point of drive. As it is difficult to grasp intuitively the behavior of such expressions, Mead [2] has compared this discrete case to the more familiar continuous case. The continuum approximation to the solution for a one-dimensional network is:

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

where $1/\mathrm{L} = \sqrt{\mathrm{RG}}$. The variable x is analogous to n. The factor $e^{-\frac{1}{\mathrm{L}}}$ is analogous to γ.

Substituting 1/L for $\sqrt{\mathrm{RG}}$ in the expression for γ gives:

$$\gamma = 1 - \frac{1}{\mathrm{L}}\sqrt{1 + \frac{1}{4\mathrm{L}^2}} + \frac{1}{2\mathrm{L}^2}$$

This expression is close to the Taylor expansion for $e^{-\frac{1}{\mathrm{L}}}$:

$$e^{-\frac{1}{\mathrm{L}}} \approx 1 - \frac{1}{\mathrm{L}} + \frac{1}{2\mathrm{L}^2} \cdots$$

The continuum approximation is very good, even for values of L as low as 1. The intuitive interpretation of $1/\mathrm{L} = \sqrt{\mathrm{RG}}$ is that a signal spreads farther in the network when R is small and G is small because it meets little resistance to its spread within the network and it has small opportunity to escape.

The amplitude of the response of the network to a single input of magnitude V is calculated by considering the effective impedance of the network. The current through the resistor connecting two adjacent nodes of the network is given by:

$$I_n\mathrm{R} = U_n - U_{n+1}$$

Dividing both sides by U_nR the equation becomes:

$$\mathrm{G_{IN}} = \frac{I_n}{U_n} = \frac{1-\gamma}{\mathrm{R}}.$$

$\mathrm{G_{IN}}$ is the effective conductance seen by node U_n of the network going to ground through the network accessed by that resistor. At each node of the network there are two such conductances being driven by a voltage source V_n through the conductance G, as shown in Figure 2.6. To solve for the voltage on the network U_n set the currents flowing in and out of the node equal to each other.

$$\mathrm{G}(V_n - U_n) = 2\mathrm{G_{IN}}U_n$$

Solve this equation for U_n to calculate the response of the network to a single input, V_n.

$$U_n = \frac{\mathrm{G}}{2\mathrm{G_{IN}} + \mathrm{G}}V_n$$

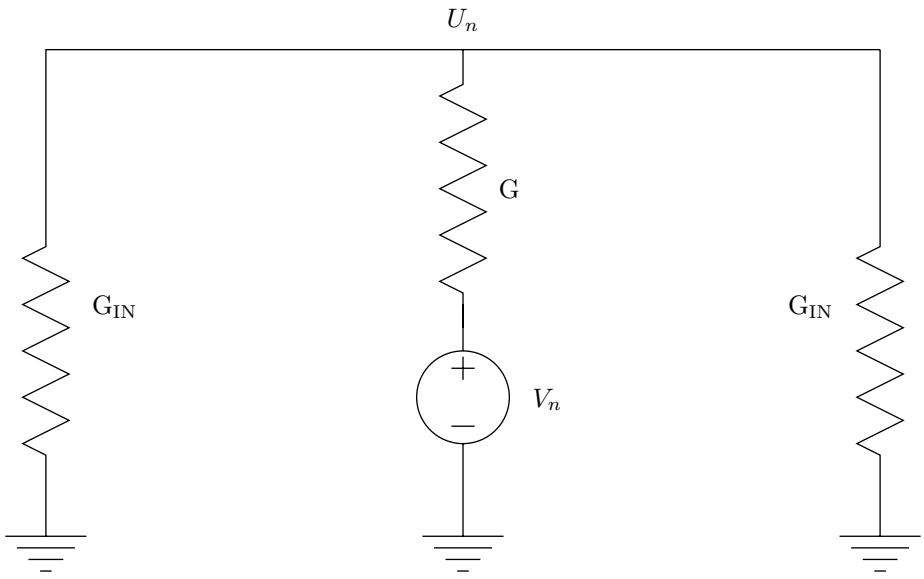

Figure 2.6: Calculation of the amplitude of response of a one-dimensional network driven by a single voltage source. A voltage source, V_n, drives a network node U_n through a conductance, G. The network has been replaced by an equivalent input conductance, G_{IN}.

Substitute for G_{IN} to obtain the solution in terms of R and G.

$$U_n = \left(\frac{1}{\sqrt{1+\frac{4}{RG}}}\right)V_n$$

U_n approaches V_n when RG is large; the more effective an input is at driving the network to its own voltage, the less distance signals will be able to propagate in the network.

Linear superposition can be used to calculate the response at any point in the network to a complex input pattern. The effects of the inputs sum together at each node weighted by the distance between each node and the input.

$$U_n = \left(\frac{1}{\sqrt{1+\frac{4}{RG}}}\right)\sum_{i=-\infty}^{\infty}\gamma^{|i-n|}V_i$$

This analysis is easily extended to the feedback retina, in which the response of the network modulates the output of the receptor. A one-dimensional version of the circuit is shown in Figure 2.7.

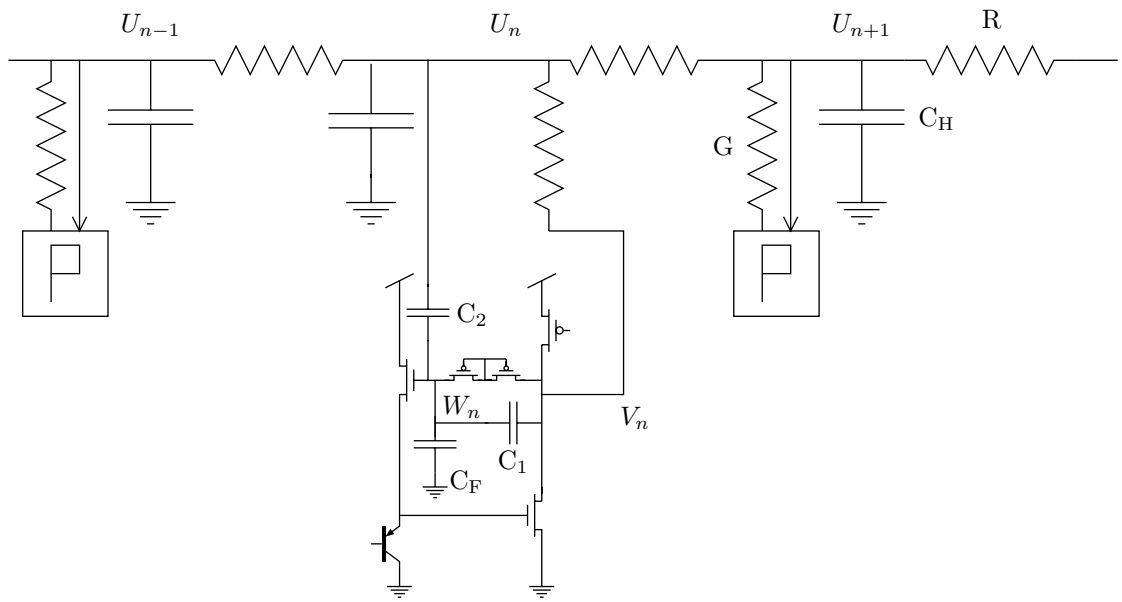

Figure 2.7: A one-dimensional network provides feedback inhibition for the high-gain photoreceptor.

The resistive network is capacitively coupled via C_2 to the control voltage of the photoreceptor, W_n. The whole circuit satisfies the constraint that the voltage W_n is sufficient to supply the current being drawn by the phototransistor. W_n is logarithmic in the incoming light intensity, I_{light}.

$$W_n \propto \ln I_{\text{light}}.$$

The capacitor C_2 acts as a reference point for W_n, much as the capacitor C_F, but in this case, the reference voltage, computed by the resistive network, changes with time. The equation that governs steady-state changes in the output voltage of the amplifier is:

$$\delta V_n = \delta W_n \frac{C_T}{C_1} - \delta U_n \frac{C_2}{C_1},$$

where $C_T = C_1 + C_2 + C_F$. The output of the photoreceptor amplifier, which is the input voltage to the resistive network, is a function of the *difference* between the local light intensity, W_n, and the average, U_n, computed by the resistive network.

This system can be represented in a simplified visual way by replacing the detailed circuit elements with abstract amplifiers. The abstracted system is depicted in Figure 2.8. The values for the gains and the inputs to the system are derived from the equation for δV_n.

$$P = \frac{C_T}{C_1}; H = \frac{C_2}{C_1}.$$

The multiplier, P, multiplies the phototransducer voltage, W_n, while the multiplier, H, multiplies the network voltage, U_n. P is analogous to the gain of the isolated photoreceptor, described in the previous section, when the horizontal network potential is held constant. H is a measure of the extent to which changes in the network voltage affect the photoreceptor output. If the light level is held constant and the network voltage is charged, the photoreceptor must compensate so that the voltage W_n stays constant. This value is simply the ratio of the capacitor coupling the network to W_n and the capacitor that couples the output of the photoreceptor to W_n.

Feedback from the resistive network prevents saturation of the receptor when the background illumination level changes. The network voltage acts as a moving reference point for the photoreceptor by adjusting the reference voltage for the C_2 capacitor. If the input

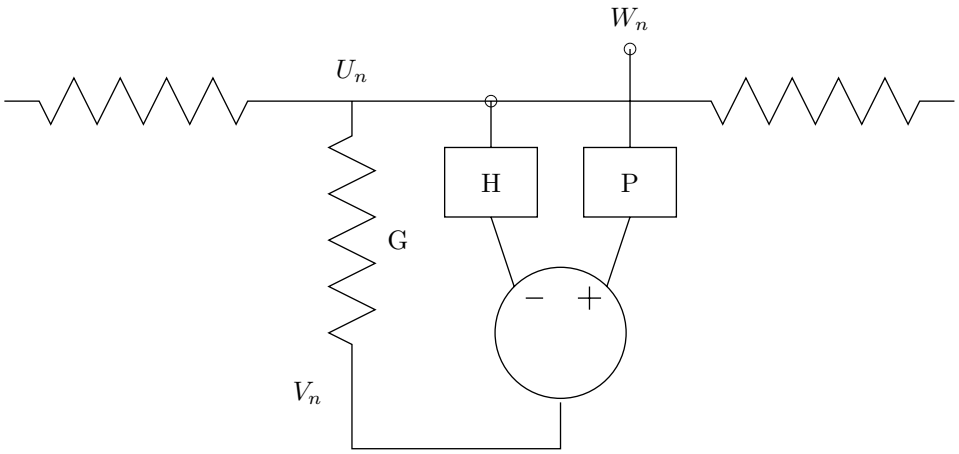

Figure 2.8: Abstract representation of the one-dimensional feedback network.

to the retina is spatially uniform, then the voltages on all of the nodes of the network are identical and so no current flows through the lateral resistors, R. Setting $\delta V_n = \delta U_n$ in the equation for δV_n gives:

$$\delta V_n = \frac{\mathrm{C_T}}{\mathrm{C_1 + C_2}} \delta W_n = \frac{\mathrm{P}}{1 + \mathrm{H}} \delta W_n.$$

Spatially uniform changes in intensity will give rise to changes in the network voltage that are proportional to changes in W_n. As described in the photoreceptor section, W_n is proportional to the logarithm of the light intensity. The amount that the receptor and the network have to shift to sink a change in the photocurrent is determined by the size of the fixed capacitor, $\mathrm{C_F}$. If the fixed capacitance on the W_n node of the photoreceptor, $\mathrm{C_F}$, is much less than $\mathrm{C_1 + C_2}$, then the system responds with approximately unity gain for spatially uniform changes in intensity. Feedback from the network extends the operating range of the photoreceptor by centering its operating point around the response to uniform illumination.

The equation describing Kirchoff's Current Law in the resistive network is:

$$\mathrm{G}[(\mathrm{P}W_n - \mathrm{H}U_N) - U_n] = \frac{U_n - U_{n+1}}{\mathrm{R}} + \frac{U_n - U_{n-1}}{\mathrm{R}}$$

This equation is identical to the equation for that of the passive resistive net except that the driving term is now P W_n and the term multiplying U_n on the left-hand side is (1 + H) instead of 1. The solution of this equation is therefore identical to that of the passive RG network, except that the effective conductance is given by:

$$\mathrm{G_{Eff}} = \mathrm{G}(1 + \mathrm{H}).$$

Just as in the feedforward network, the solution for the spatial spread of signals in the network is an exponential decay. However, the decay of signals is more rapid with distance, with $1/L = \sqrt{\mathrm{RG_{Eff}}}$.

The origin of the increased effective conductance at each node of the network is illustrated in Figure 2.8. The definition of effective conductance is the amount of current that needs to be injected to charge the network by a fixed voltage increment.

$$\frac{\delta I}{\delta U} = \mathrm{G_{Eff}}$$

If a current is injected into node U_n, and the change in voltage measured to determine the conductance of the node, the conductance will appear larger than the physical conductance because no account has been taken of the fact that V_n has changed by an amount δU_nH, so the total voltage change driving current through the physical conductance is

$$\delta I = \mathrm{G}\delta V_{\mathrm{drive}} = \mathrm{G}(\delta U + \delta U\mathrm{H})$$

So the effective conductance, $\frac{\delta I}{\delta U} = \mathrm{G}(1 + \mathrm{H})$.

The photoreceptors in the feedback retina have a center-surround response, shown in Figure 2.9. The decay of signals in the network is due not only to passive decay through a conductance, G, but also to active absorption by the receptors themselves. As a current propagates and charges the network, it affects the voltage output of the receptor and a larger fraction of the current leaks out of the feedback network than would have leaked out of a passive network. The larger the response of the photoreceptor to differences between itself and the network, the harder it forces the resistive network to its own voltage and, therefore, the more quickly the signal in the network decays.

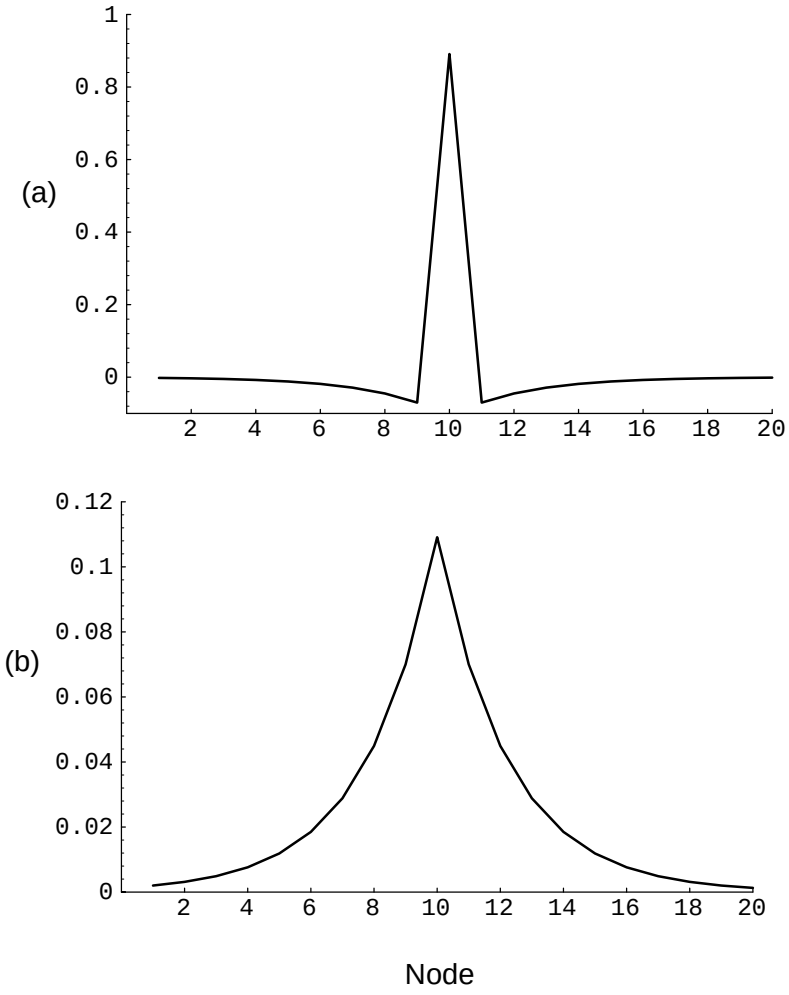

Figure 2.9: Mathematically computed spatial response of a one-dimensional array of receptors (a) and horizontal cells (b) in a feedback arrangement. The ordinate is distance in the array; the coordinate is the response of the cell at that location. Input to the system is a unit delta function at position 10. Notice that the receptor response has a gentle inhibitory surround.

The temporal response of the network is a result of the natural physical properties of the medium. 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. The effect of the capacitance of the horizontal cells is to delay their response. Even changes in the background intensity of the image are passed through the bipolar cells since the horizontal cell surround signal takes time to catch up with the photoreceptor center signal.

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. The temporal integration time of the network is determined by the magnitude of the conductance, G, and the capacitance of the network.

The capacitive coupling of the horizontal cell node to the photoreceptor in the feedback retina must be properly ratioed to the horizontal cell node capacitance to achieve the proper temporal response from the pixel. In order to maintain a large response from the pixel for short times, C_H must be much larger than C_2. In this case, the network acts as an effective reference voltage. The capacitors C_H and C_2 combine like conductances in series. If $C_H >> C_2$ the network acts as a fixed reference, because charge can be drawn from C_H to hold W_n fixed without changing U_n very much.

2.2.4 Bipolar Cell

The receptive field of the bipolar cell shows an antagonistic center-surround response [35]. The center of the bipolar cell receptive field is excited by the photoreceptors, whereas the antagonistic surround is due to the horizontal cells [38]. The gain of the bipolar cell is larger than that of the photoreceptors or the horizontal cells and so the bipolar cell saturates over a smaller range of inputs. The center-surround organization keeps the high-gain of the bipolar cell centered around an appropriate operating point.

The final outputs of both silicon retinas are analogous to the output of a bipolar cell in a vertebrate retina. The bipolar cell analog is a transconductance amplifier that 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 the bipolar cell analog thus represents the difference between a center intensity and a weighted average of the intensities of surrounding points in the image.

Schematic diagrams of all circuits in the feedforward and feedback pixels are shown in Figure 2.10.

2.3 Accessing the Array

The floorplan for the retina is shown in Figure 2.11. 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 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 off-chip. A more complete description of the data scanning methods, including particular circuitry, is provided in [2, 29]. Improvements on these circuits and associated current-sensing amplifiers are described in [23].

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. This method is equivalent to an intracellular electrode recording from a single cell. 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

(a)

(b)

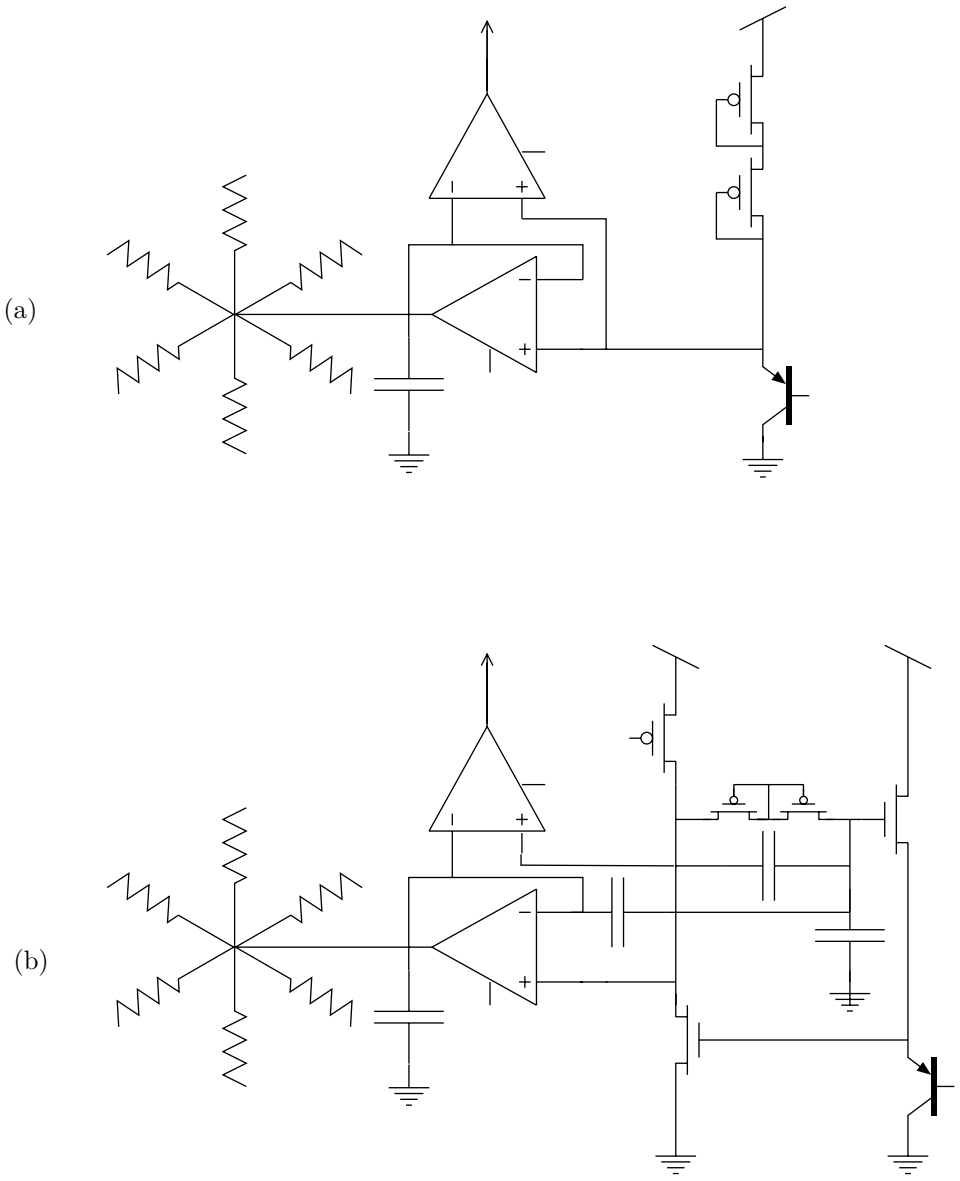

Figure 2.10: Schematics of single pixels of the feedforward (a) and feedback (b) silicon retinas. The pixel is hexagonally tiled to generate the retinal array. Each pixel contains the photoreceptor, the transconductance amplifier coupling the photoreceptor into the resistive network, and the resistors that couple the node of the resistive network to adjacent pixels.

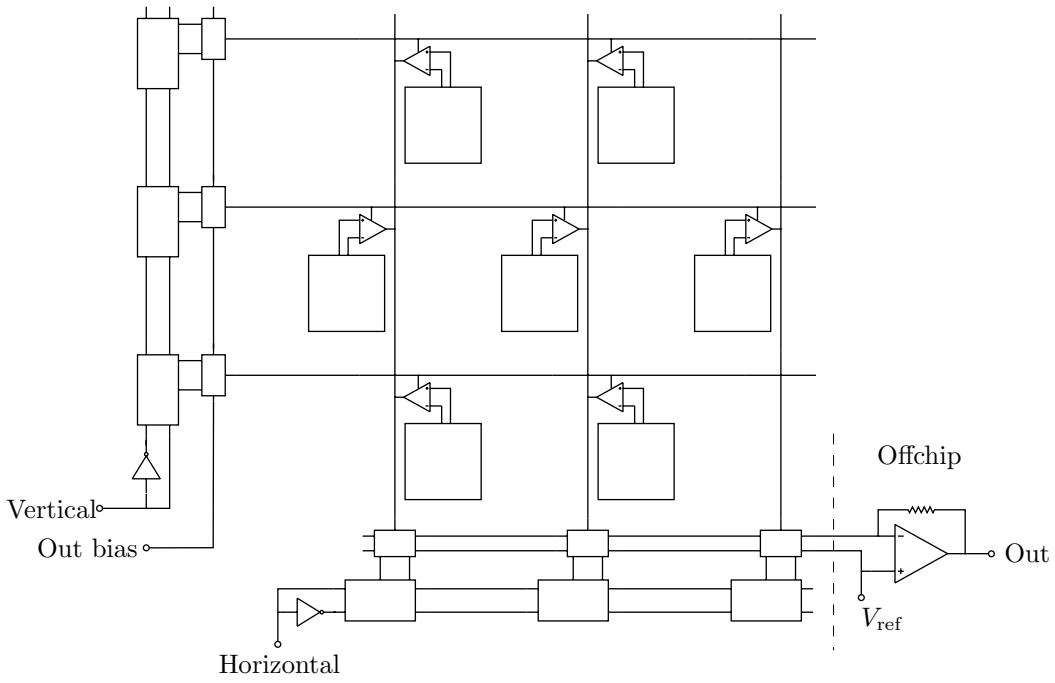

Figure 2.11: Layout of the retina chip. The main pixel array is made up of alternating rows of rectangular tiles, arranged to form a hexagonal array. The scanner along the left side allows any row of pixels to be selected. The scanner along the bottom allows the output current of any selected pixel to be gated onto the output line, where it is sensed by the off-chip current-sensing amplifier.

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. This method is like a brain scan with single cell resolution; the outputs of all the cells are displayed simultaneously from the perspective of a person viewing the monitor.

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 2.10, 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 VDD, 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. Covering the chip with a solid sheet of second-layer metal, with openings directly over the photoreceptors protects the processing circuitry from the effects of stray minority carriers. This second-layer metal covering also distributes the ground of the power supply to the pixels.

2.4 Data–An Electrode's Eye View

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 en-

hancement, and statistical optimization [32, 1]. 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.

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 the 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.

2.4.1 Sensitivity Curves

The computation performed in the distal portion of the retina prevents the output from saturating over an incredible range of illumination levels. Feedback from the horizontal cells to the cones provides a varying amount of current to compensate for the current flowing into the cell through the light-sensitive channels. The cone can avoid saturation over six orders of magnitude change in light level. The shift in photoreceptor output is mediated by feedback from the horizontal cells, which compute a spatially averaged version of the photoreceptor outputs. The cone response is dominated by contrast in the image, rather than absolute light level.

In addition to keeping the cones out of saturation, the horizontal cell response defines the gray-level for the image by feedforward inhibition onto the bipolar cells. The bipolar cell senses the difference between the photoreceptor output and the potential of the horizontal cells, and generates a high-gain output. 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.

The effects of feedback from the horizontal cells to the receptor are illustrated in Fig-

ure 2.12. The response of the photoreceptor was measured in isolation on a small test chip so that all of the nodes in the circuit could be instrumented. Feedback to the pixel was provided by an external pad that emulated the response of the network. The receptor response, measured at node V depicted in Figure 2.4, is similar around all four operating points. The response has slightly lower gain at light levels lower than tenths of a milliwatt/mm^2. (However, the absolute light level is unreliable.) The gain of the adapted response of the receptor, mediated by the diode-connected transistors, Q3 and Q4, coupling the receptor output, V, directly to W, is 30 mV/e-fold change in light level. This value is consistent with a value of 0.85 for κ. The gain of the receptor was measured in two different conditions at each operating point and compared to the gain predicted from estimates of the circuit capacitance values derived from the layout. Estimated capacitance values are:

$$\mathrm{C_1 = 70fF; C_2 = 422fF; and C_F = 315fF}$$

The value for $\mathrm{C_F}$ was larger on the test chip than in the two-dimensional array due to parasitic capacitance introduced by the instrumentation pad. The receptor is operating in its high-gain condition when $\mathrm{C_2}$ is tied to a fixed potential. This condition emulates the response of the retina to a small test flash. A small flash does not significantly affect the average computed by the horizontal cell network, so the network voltage is nearly constant. The photoreceptor gain in this condition is 430 mV/e-fold. This response is 20% larger than predicted by the capacitance values estimated from the circuit layout. The receptor operates in its low-gain condition with $\mathrm{C_2}$ driven directly by the photoreceptor output, V. This condition emulates the response of the system to full-field illumination. The gain of the receptor in the low-gain condition averaged over the three brighter illumination trials 61 mV/e-fold. This value is 20% higher than that predicted by the capacitance values estimated from the layout. The deviation between estimated and measured photoreceptor gain may be due to the finite gain of the clamping amplifier.

In the feedback silicon retina, the photoreceptor in isolation demonstrates invariance to overall changes in illumination. This invariance does not appear in the feedforward retina until the bipolar-cell level. Figure 2.13 shows the shift in operating point of the bipolar-cell output of both a biological and a two-dimensional feedforward silicon retina, as

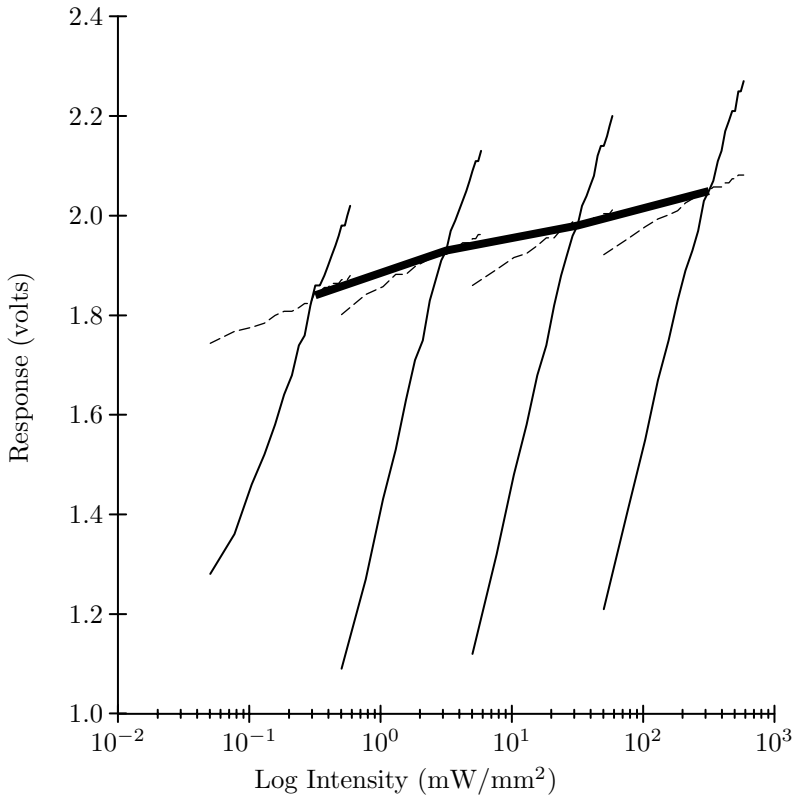

Figure 2.12: Response of the photoreceptor to changes in light intensity of a green light emitting diode (LED). Each set of curves was taken within the same range of currents through the LED. Light energy from the LED was calibrated with a photometer, in milliwatts/mm^2. Due to the light level calibration method, the light intensity values are low. The estimated area of the detector is 336 μm^2. Neutral density filters were used to shift the light intensity over four orders of magnitude. The heavy line indicates the receptor's DC response to illumination measured after fully adapting to that illumination level. The response of the photoreceptor was measured at each adaptation level under a high-gain (solid line) and low-gain (broken line) condition (see text). Light level was briefly displaced and peak voltage response was measured.

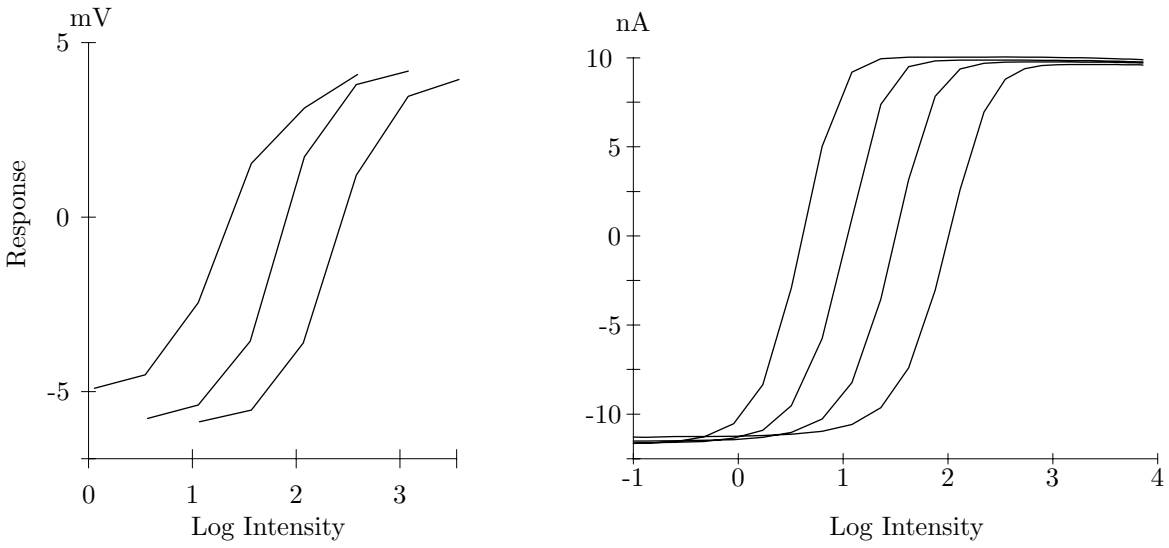

Figure 2.13: Curve shifting. Intensity-response curves shift to higher intensities at higher background illuminations. (a) Intensity-response curves for a depolarizing bipolar cell elicited by full-field flashes. The test flashes were substituted for constant background illuminations. These curves are plotted from the peaks of bipolar response to substituted test flashes. Peak responses are plotted, measured from the membrane potential just prior to response. (Data from Werblin, 1974 [35].) (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.

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 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. Image features are reported with high gain 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 [34]. 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.

2.4.2 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.

Figure 2.14 shows the response of a single bipolar cell output of the feedforward retina 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 [2], 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 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 temporal

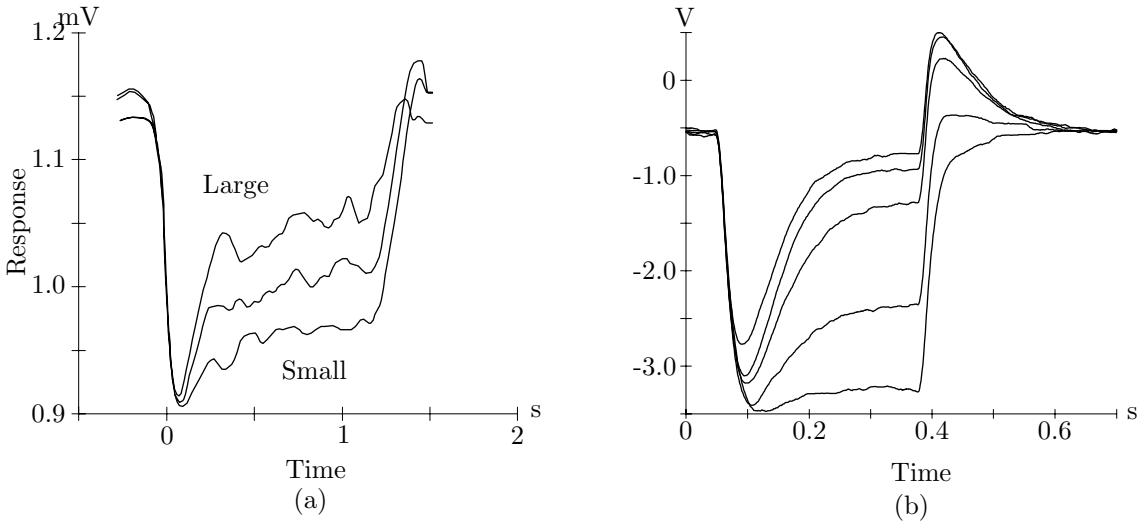

Figure 2.14: Temporal response to different-sized test flashes. (a) Response of a bipolar cell of the mud puppy, *Necturus maculosus*. (Data from Werblin, 1974 [35].) (b) Output of a pixel in the silicon retina. Test flashes of the same intensity but of different diameters were centered on the receptive 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.

differentiator circuit [2], 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.

2.4.3 Edge Response

The suppression of spatially and temporally smooth image information acts 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 [35]. Although the formation of the receptive field of the X-type ganglion cells of the cat is somewhat more complex [8], the end result is qualitatively similar. The receptive fields of these cells are described as 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 2.15. 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 2.15. The way the second spatial derivative is computed is illustrated in Figure 2.16. 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.

Figure 2.17 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

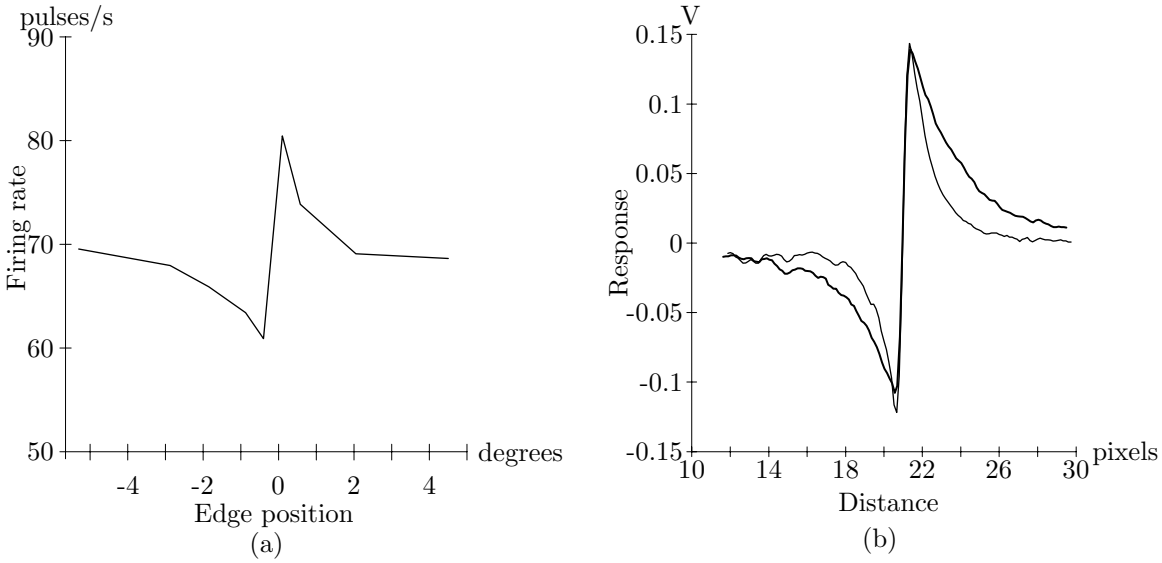

Figure 2.15: Spatial-derivative response of a retinal ganglion cell and of a pixel to a contrast edge. The vertical edge was held stationary at different distances from the receptive-field center. Contrast of the edge was 0.2 in both experiments. (a) On-center X-type ganglion cell of the cat. The contrast 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. (Data from Enroth-Cugell et al. 1966 [9].) (b) Pixel output measured at steady state as the edge was moved in increments of 0.01 centimeter at the image plane. Interpixel spacing corresponded to 0.11 centimeter 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.

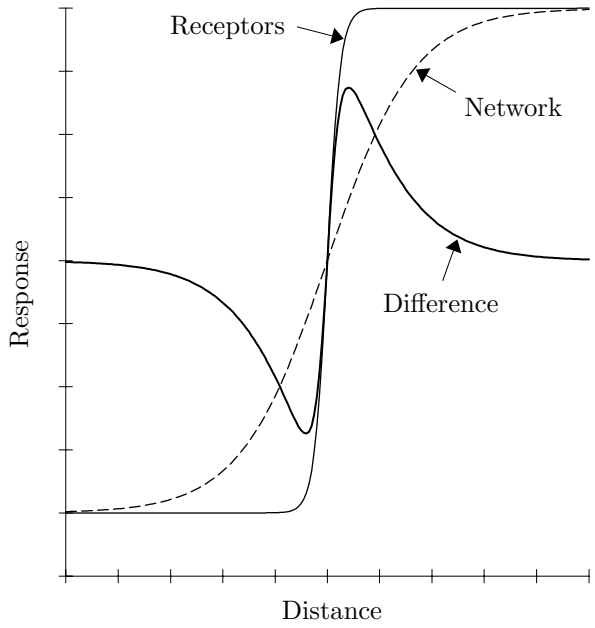

Figure 2.16: Model illustrating the mechanism of the generation of pixel response to spatial edge in intensity. The solid line, labeled *receptors*, represents the voltage outputs of the photoreceptors along a cross-section perpendicular to the edge. The resistive network computes a weighted local average 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.

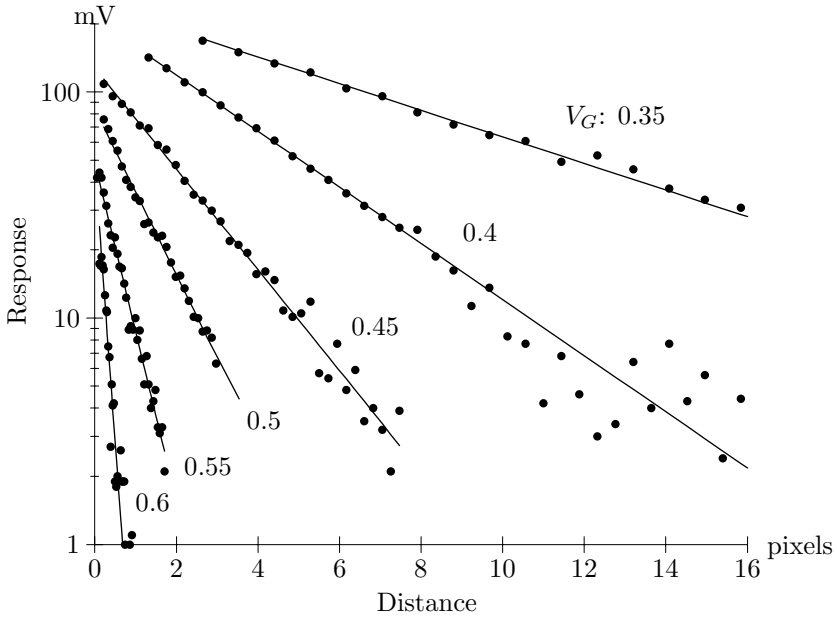

Figure 2.17: Exponential decay of one side of the response to an edge, as shown in Figure 2.15. Each curve was taken with the setting of the V_G control shown. For all curves, V_R was 0.55 volt. The slope of the decay corresponds to the space constant of the network.

network therefore is a good approximation to the response of the two-dimensional network to an edge.

In the feedforward silicon retina, 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:

$$G \propto e^{V_G}$$

and

$$R \propto e^{-V_R}$$

Substitute these expressions into the equation for the space constant to get the space constant in terms of the bias control voltages:

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

The space constant thus should be a function of $V_G - V_R$, and should not be dependent

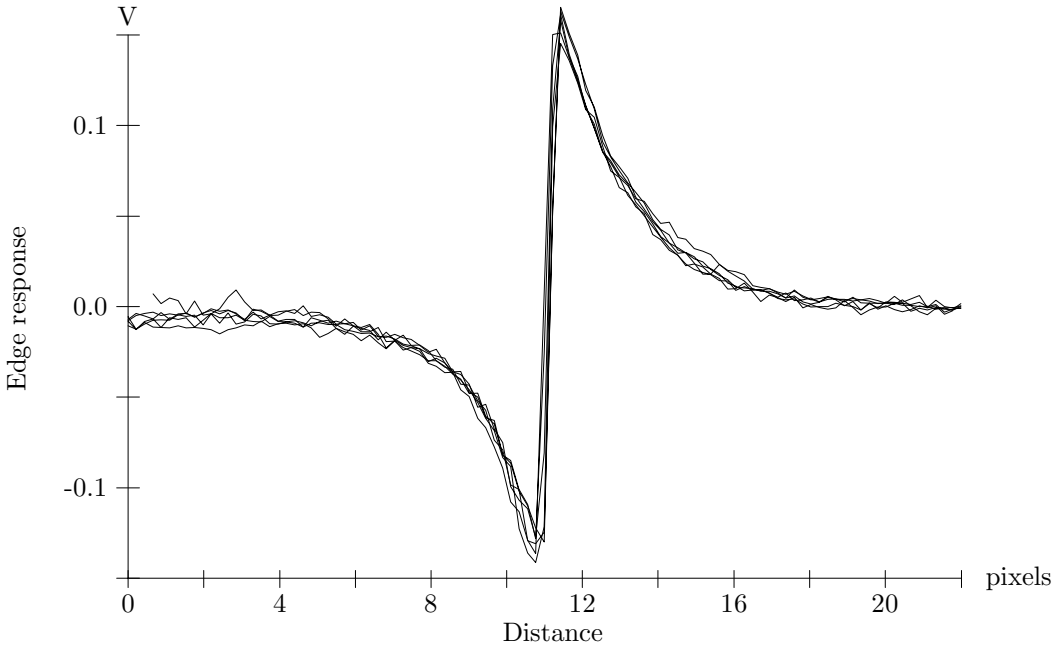

Figure 2.18: The response of a pixel to a 0.2 contrast edge measured for a fixed difference between the conductance bias voltage and the resistor bias voltage. (DC offsets in the response were subtracted out.) The space constant of the network depends on only the ratio of conductance bias current to 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.

on the absolute voltage level. The constant of proportionality 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 2.18 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.

The continuum form of the resistive decay is a good approximation to the horizontal network when the space constant is greater than one and the slopes of the decay curves in Figure 2.17 can be compared to the theoretical expression, where all voltages are expressed in terms of $\mathrm{kT/q}\kappa$. The comparison is shown in Figure 2.19; the voltage dependence of the decay constant is in excellent agreement with the theoretical prediction. The absolute value of the curve in Figure 2.19 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 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 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 2.17. 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 [10].

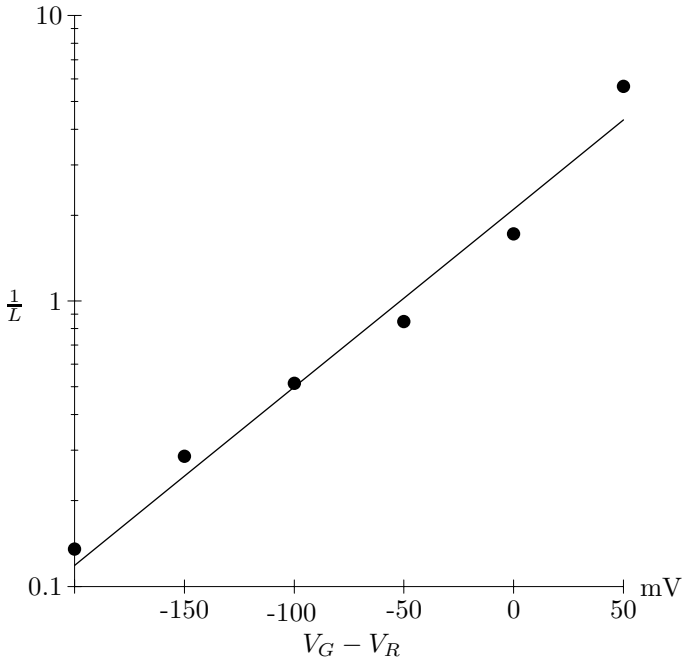

Figure 2.19: Space constant of the response data of Figure 2.17, plotted as a function of $V_G - V_R$. The straight line is the theoretical expression, using the measured value of $\kappa = 0.73$. The magnitude of the curve was adjusted for best fit to the data, and is about a factor of two higher than expected from the width-to-length ratios of transistors in the transconductance amplifier and in the resistor bias circuit.

2.4.4 Adaptation

Adaptation is a much slower process than the spatio-temporal edge enhancement previously discussed. Adaptation takes place in the photoreceptor circuit itself. It is mediated by the two diode-connected transistors, Q3 and Q4, shown in Figure 2.4. Although it is not depicted in the diagram, the transistor gate node is also tied to the well containing Q3 and Q4, which eliminates the back-gate effect. The adaptation of a single photoreceptor to large steps in illumination is illustrated in Figure 2.20 and Figure 2.21. The time-course of adaptation is set by the leakage current through these transistors. In principle, the leakage current should be the same for both transistors and so the temporal characteristics of adaptation to either direction of step should be the same. However, Delbrück (personal communication) has shown that both photo- and thermally generated carriers flowing from the well to the substrate are significant. In a dark-going transition, the output of the photoreceptor, V, goes low. Thus V acts as the drain for Q3 and the source, gate, and bulk of Q3 are tied together, and Q3 acts as a diode. In contrast, the gate, bulk, and drain of Q4 are tied together and the source is node W. Thus transistor Q4 is in a conducting state. Any carriers that are flowing from the well to the substrate are pulled off of W through Q4. Therefore, the time-course of adaptation to dark-going steps is not set by the leakage current through Q3, but by the photo- and thermally generated carriers in the well. In spite of this asymmetry in temporal response, the circuit still adapts to the proper level with a time course longer than necessary for proper spatio-temporal edge enhancement.

The performance of the adaptive feedback retina to static edge detection is compared to the performance of the nonadapting feedforward retina in Figure 2.22. Adaptation adjusts the operating point of the receptor to the appropriate level. Adaptation is driven by the voltage V at the output of the receptor. The photoreceptor amplifies the difference between its own phototransducer and the average computed by the horizontal cells. If the phototransducing elements could be perfectly calibrated with respect to each other by some omniscient external agent, this computation would be straightforward. Without external calibration there is no guarantee that two receptors will respond identically when stimulated with the same amount of light. Differences in their responses are amplified by the output circuitry as if there were real differences in the incoming light intensity. The system is faced with the problem of having to calibrate itself provided only with its own response. It does

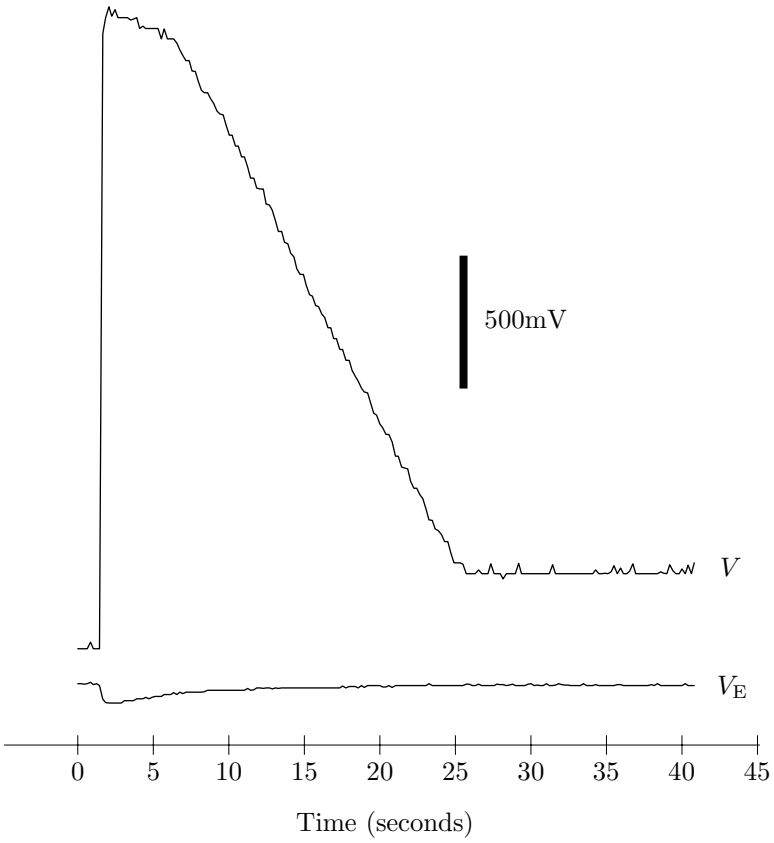

Figure 2.20: Response of a single photoreceptor to a step increase in light from a green LED. The LED intensity was varied from darkness to 315 mW/mm^2. Top trace shows the output of the photoreceptor, V, and the bottom trace is the response of the emitter voltage of the phototransistor, V_E (see Figure 2.4). The adapting current flowing onto W is limited by the diode-connected transistor, Q4. The data were collected at room temperature.

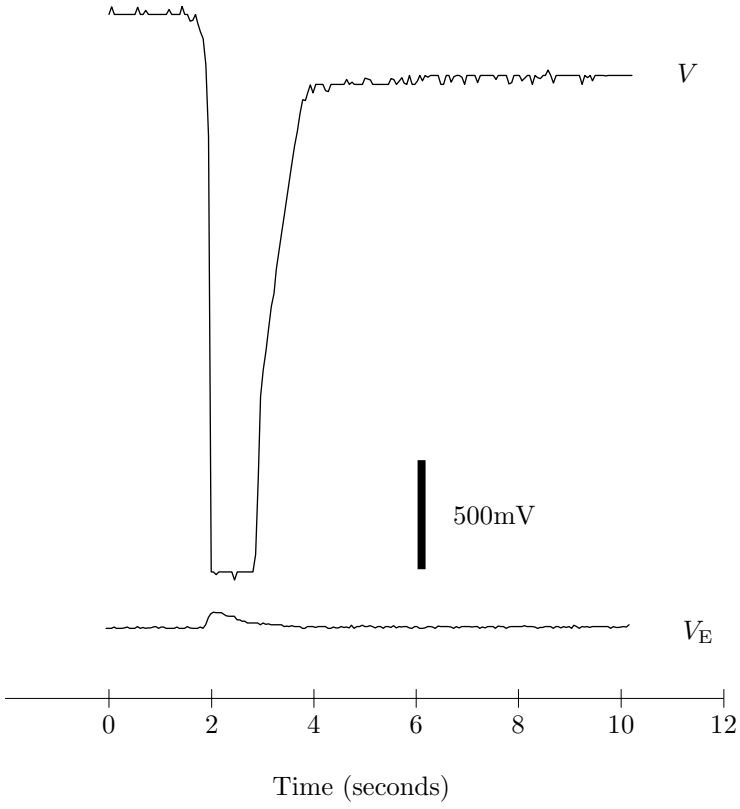

Figure 2.21: Response of a single photoreceptor to a step decrease in light intensity from a green LED. The LED intensity decreased from 315 mW/mm^2 to darkness. Top trace shows the output of the photoreceptor, V, and the bottom trace is the response of the emitter voltage of the phototransistor, V_{E} (see Figure 2.4). The adapting current flowing onto W is limited by the diode-connected transistor, Q3. The data were collected at room temperature.

this calibration via slow adaptation, which is a fundamental component of neural systems. The outcome of this process is that any static image, which cannot be differentiated from offsets in the detectors, is canceled out. If the image is removed, a negative afterimage appears that reveals the pattern of adaptation.

2.5 Form and Function: Encoding Information with a Physical System

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 center-surround receptive field encodes visual information in an optimal way when the amount of correlation in the image is large. Using measures derived from information theory, several investigators have provided a definition of visual information and examined the efficacy with which the retina transmits this information to the brain [1, 32]. These analyses show that the retina makes highly efficient use of the bandwidth of the optic nerve and adapts its encoding to be appropriate at different light levels. The retina devotes its limited output dynamic range to transmitting visual information; it excludes redundant aspects of the image and minimizes the effects of noise.

At low illumination levels, the major source of noise is in the phototransducers, which are trying to measure a small number of photons in the presence of spontaneous photoisomerization. Noise is a form of redundancy since, by definition, it contains no information. Under these conditions, the receptors themselves average over a larger area by coupling to each other and the effects of the inhibitory surround disappear. The retina reports the actual light level.

The silicon retina operates in the photopic region. In photopic lighting conditions, redundancy in the image comes from correlations. When the number of photons falling on the retina is large, the spatial variation caused by objects of different reflectances is relatively small. If the retina simply tried to report the number of photons received as a function of position, noise in the output would be confused with the properties of objects and the most

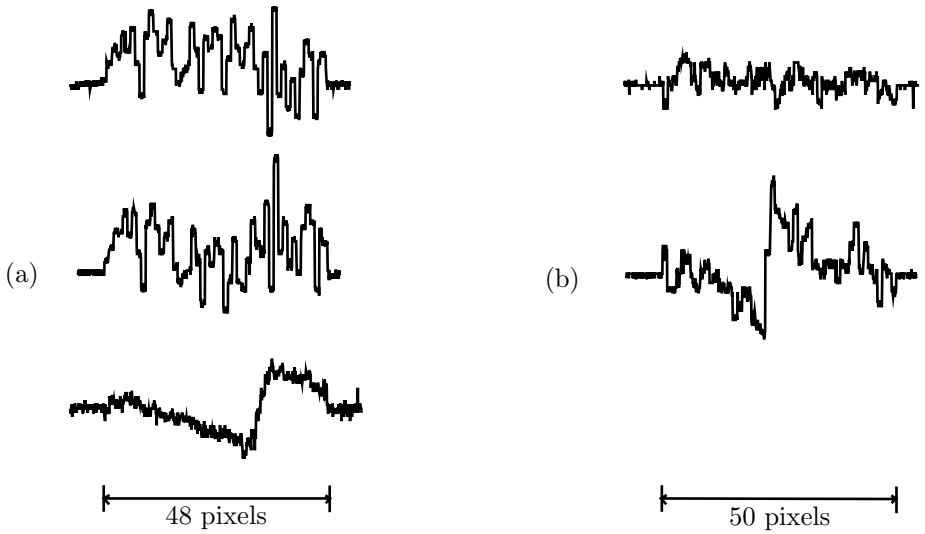

Figure 2.22: Comparison of the edge responses of the Mead and Mahowald retina and the adaptive retina. Stimulus was a 0.2 log unit, one-dimensional step in intensity. Data were taken by multiplexing the analog responses of a row of pixels perpendicular to the intensity edge to a digital storage scope. The current output of the chip was converted to a voltage by an off-chip sense amplifier. The gain of this amplifier, hence the voltage scale of the response, is arbitrary. The DC offsets of the photoreceptors and the output amplifiers appear as small differences in the responses of different pixels. The spatial averaging areas of the resistive grids in both retinas were large. (a) Response of the Mead and Mahowald retina. Top trace shows the response to a uniform field. Middle trace shows raw edge response. The bottom trace shows the edge response minus the uniform field response. The position of the edge is visible only after performing this differencing operation off-chip. (b) Response of the adaptive retina. Top trace shows the response to a uniform field to which the retina was adapted. Middle trace shows raw edge response. The edge was centered roughly around the intensity of the uniform field. The position of the edge is immediately apparent.

apparent feature of neural activity would be the overall illumination level. Instead of taking this direct approach, the retina removes much of the redundant information about uniform light level and encodes the image as a pattern of changes occurring over particular spatial and temporal scales. The overall illumination level is represented with small variations in low spontaneous firing rate, which occur over large regions in the image. One could imagine that neurons with different spontaneous rates might divide the ambient illumination level into different populations, those with sensitive spontaneous rates encoding lower illumination levels and those with higher thresholds being recruited as the illumination level increased. Contrast information would be superimposed on this background activity with much higher spike rates. This encoding leads naturally to perceptual constancy since the pattern of neural activity in response to a particular image is not greatly affected by the constant illumination level.

The constraints imposed by the physical medium determine the way that information is represented. The graceful transition between one type of information encoding (scotopic-mesopic) and another (photopic) is implicit in the biophysics of the retina. It is possible to abstract these functions into simple circuits in such a way that major aspects of information processing are retained. The description of such a circuit is a compact parameterization of retinal function. For example, the receptive field size of the X-type ganglion cell analyzed by Atick can be characterized by the strength of lateral electrotonic coupling and the strength of the feedback from the horizontal cells to the photoreceptors. Parameterization in circuit terms leads to an understanding of the relationships between functions that might otherwise remain disjoint, such as lateral inhibition and receptor calibration. Finally, the characteristics of the representation that arise from these physical constraints affect further processing, as is evidenced by the existence of several visual illusions.

2.5.1 Wiring

The center-surround computation is the basic feature of information encoding in the retina. In computer vision, a common visual primitive is the Laplacian filter, which can be approximated by a difference of Gaussians [4]. 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 that occurs as a natural result of an efficient physical implementation of local level normalization. The information processing abilities of the retina are a direct result of its physical structure.

It is possible to generate a center-surround function in a variety of ways. As previously shown, the center-surround may be computed by a feedforward or feedback mechanism. The surround may be computed using point-to-point wiring [6] or with a resistive net. If feedback is coupled with a structure in which multiple nodes are coupled to each other, as in point-to-point coupling [6] or two coupled resistive networks, then spatial oscillation may result [3]. Although point-to-point connections and resistive networks can compute the same functions, the resistive net is more economical in terms of wiring required to create the same sized receptive field. A comparison of the wiring density needed to compute a receptive field with discrete connections and the wiring requirements of the resistive net is plotted as a function of receptive field size in Figure 2.23. In the figure, the wiring requirements of the resistive net are constant irrespective of receptive field size. In addition to wiring efficiency, the receptive field size can be easily manipulated by changing the space constant of the network.

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 formed by horizontal cells 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. The topographic nature of the cortex insures that most wires can be short; it is perhaps the single most important architectural principle in the brain.

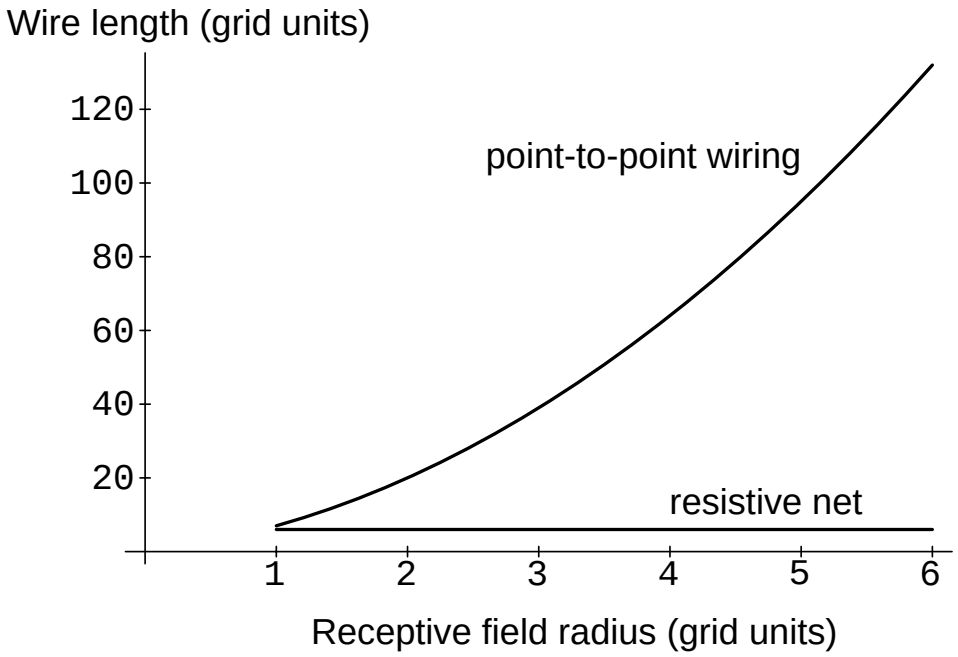

Figure 2.23: Relationship between receptive field size and wire density. The wire density is the wire used per pixel and is equal to the amount of wire necessary per receptive field, assuming one receptive field per pixel. The pixels are assumed to be in a hexagonal array and the radius of the receptive field is in units of the internode spacing of the hexagonal array. The formula for calculating the wire required for point-to-point connectivity is: $W = 6[(\sum_{i=1}^{R} i) + R]$. This function is quadratic in R. The wire density required for a resistive net (nearest neighbor connectivity) is constant with receptive field radius with six units of wire per pixel.

2.5.2 Interpretation of Biological Data

In the previous sections, an argument was made for adopting the principle of center-surround organization in sensory systems, both biological and man-made. An engineering approach to interpretation of biological function leads to new appreciation of the biophysical details of retinal processing.

The retina executes a center-surround computation at the level of the cones by means of feedback from the horizontal cells [2] and again by feedforward synapses onto the bipolar cells [38]. As in the silicon retina, the feedback from horizontal cells to cones balances the current from light-sensitive channels with the average photocurrent computed by the horizontal cells.

Biological systems do not have direct access to either current sources or voltage sources. Instead, they perform most of their computations by modulating the amplitudes of conductances. One major difference between biological neuronal elements and CMOS transistors is the ohmic behavior of individual membrane channels. The current through a transistor biased in subthreshold is exponential in the gate voltage and, for voltages greater than a few $\frac{kT}{q}$, it behaves as an almost perfect current source with respect to the voltage across its terminals. The *number* of channels open in a biological membrane is some function of voltage across the membrane or of neurotransmitter, either exponential or sigmoidal. In addition, the current through an open channel is linear in the driving potential across it.

The modulation of a conductance is an essentially nonlinear operation. Because the biological system modulates conductances in the membrane, it is a challenge to keep the space constant of the horizontal cell resistive network fixed while input from the cones is presumably changing with light level. It is important that the spread of activity in the horizontal cell network remain unchanged as a function of the background illumination level if the center-surround characteristics of the bipolar cells is to remain fixed. The biological retina has come up with clever mechanisms to compensate for its own nonlinearities, thus giving the appearance of a linear system.

It is known that the cones hyperpolarize and decrease their release of transmitter in response to light. The cone transmitter, probably glutamate, holds open a depolarizing conductance in the horizontal cell membrane [8]. In the light, this conductance decreases [36], thus hyperpolarizing the horizontal cell. One expects that the increased membrane

resistance of the horizontal cell would result in a larger electrotonic spreading distance, L, resulting in a larger receptive field. However, in turtle [5] and cat [24] the spreading distance of the horizontal cells appears to be, if not constant, then decreasing slightly with increased ambient illumination.

One straightforward mechanism for keeping the space constant constant is to mediate changes in potential with a resistive voltage-divider mechanism that operates on a push-pull basis, one conductance increasing while another is decreasing. Lasater and Dowling [8] have found evidence for a potassium channel in isolated carp horizontal cells that closes in response to L-glutamate. This channel would act in concert with the sodium sensitive channel to keep the membrane resistance of the horizontal cell constant. Thus the spread of voltage in the network should follow an essentially linear passive electrotonic decay irrespective of light level.

However, many experiments demonstrate that a linear model fails to account for the response properties of horizontal cells. Measurements of the spread and summation of signals in cat horizontal cells [16] have shown that the best-fit estimate of the passive space constant depends on whether a slit or a spot stimulus is used. In addition, the space constant appears to be a function of time [16, 5]. Experiments in which the membrane of the horizontal cell has been artificially polarized [36, 4] show that impedance is a nonlinear function of membrane voltage.

Several mechanisms for how these phenomenon might occur have been proposed, such as nonlinear gap junction resistance [33] and nonlinear voltage-dependent conductances in the non-synaptic horizontal cell membrane [4]. There is no evidence for gap junctions with intrinsically nonlinear conductance properties. Gap junction conductance is modulated by the release of dopamine, probably from interplexiform cells [7]. However, this process, which relates to retinal light adaptation, is too slow to account for the observed temporal changes in the spread of signals in the horizontal cells. Evidence for a voltage-dependent conductance in the non-synaptic membrane of the horizontal cells has been obtained for the most part by experiments in intact retinas, in which the network properties could not be well controlled [36, 4]. Evidence against a voltage-dependent conductance has been obtained in the L-type horizontal cells of the carp retina [13]. These cells receive input from two types of cones, red and green. The spread of activity in the horizontal network is dependent on

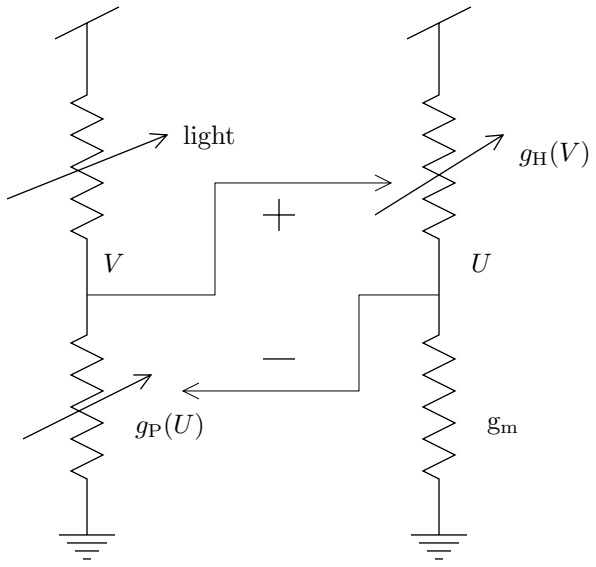

Figure 2.24: A simple model of a single photoreceptor in a feedback loop with a single horizontal cell.

which type of cone was activated, irrespective of absolute voltage level. Kamermans et al. modeled this effect using a network with different strength feedback from the horizontal cell to each type of cone. The feedback to the cones influences the spread of activity in the horizontal cell network [13]. Kamermans's elegant simulations suggest a role for future experimentation on the silicon retina.

In the silicon retina, the feedback contributed an additional effective conductance to the horizontal cell network. This effective conductance was mediated by changing the voltage source that was driving the network. A simplified model of the analogous situation in the biological retina is shown in Figure 2.24.

In order to maintain a uniform space constant, the effective impedance of each node must remain invariant. The crux of this model is to divide the effective impedance into two components, one due to membrane conductance and the other due to the relation between the photoreceptor and the horizontal cell.

This analysis leads to a set of non-linear partial differential equations for the synaptic efficacies of the excitatory and inhibitory synapses between the horizontal cells and the photoreceptors as a function of presynaptic voltages. A simulation technique using discrete

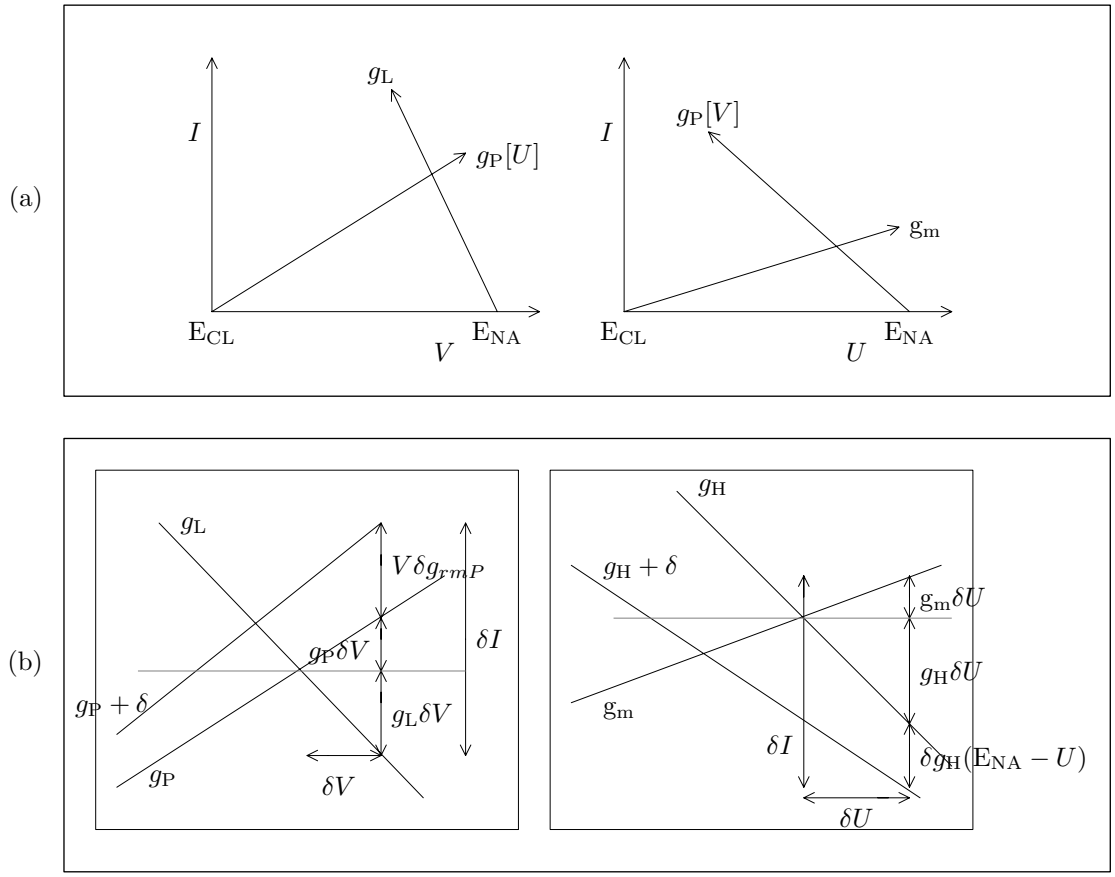

Figure 2.25: Modulation of input impedance by feedback. (a) The voltage, V, of the photoreceptor (left) is set by a balance between the conductance due to light, g_L, and the hyperpolarizing conductance modulated by the horizontal cell, $g_P[U]$. The voltage, U, of the horizontal cell (right) is set by a balance between the passive membrane conductance, g_m. Conductance is the slope of the line on the current-voltage plot. (b) When a test current δI is injected into the photoreceptor (left) or horizontal cell (right), until a fixed δV is produced, some of the current is absorbed by the *change in conductance* produced by the action of the feedback loop from horizontal cells to photoreceptors.

steps was developed by Lloyd Watts was used to examine the simplified case in which current is injected into the the photoreceptor. The horizontal cell interactes were lumped into a black box, so the conductance g_p changes as a function of V. From an initial starting point with g_{L}=g_{P}=0.01, a fixed amount of current I_{inj} was injected into the photoreceptor. The conductance g_{P} was adjusted so that the change in V was constant. In order to take the next step in the iteration, the injected current was removed and the light-sensitive conductance, g_{L} was adjusted to maintain the voltage V that was developed in the previous step. The results are plotted in Figure 2.26. The actual conductance changes by a factor of 100 but the effective conductance evaluated at a particular δV and I_{inj} remains constant.

This simulation demonstrates the qualitative effect of feedback on modulating the effective input conductance of a node. When the physical conductance values are small, the feedback must be large in order to maintain a constant input conductance. When the physical conductance values are large, the feedback must be diminished. This result is consistent with the observed shift in relative strength of the center and surround component of the receptive fields in the retina as a function of light-level. At high light levels (low conductance values) the surround component is prominent, implying large amounts of feedback. At low light levels (high conductance values) the surround component drops out, indicating that there is no feedback. This change in the balance of center surround is predicted by information theoretical arguments about optimal encoding of visual information at different light levels. The modulation of synaptic strength is a simple parameter that is able to generate the required changes in a resistive network architecture. Further work is necessary to determine if the input impedance of both the photoreceptor and the horizontal cell can be held constant by the same mechanism simultaneously. The implications of this model extend to spatial summation in two dimensions [16, 13], and dynamical changes in space constant [5, 16]. The interaction of spatial and temporal factors in the feedback retina is complex and has not yet been explored. curious dynamical behavior has been observed in vertebrate OPL. Changes in the spread of activation in a gap junction coupled cell syncytium have been observed in turtle [6, 5]. This response has been attributed to voltage-sensitive channels [6] and modeled as an inductive element in the cell membrane. However, a delayed feedback mechanism is another possibility [5]. This system illustrates the importance of models in the interpretation of data. The behavior of a neuron embedded

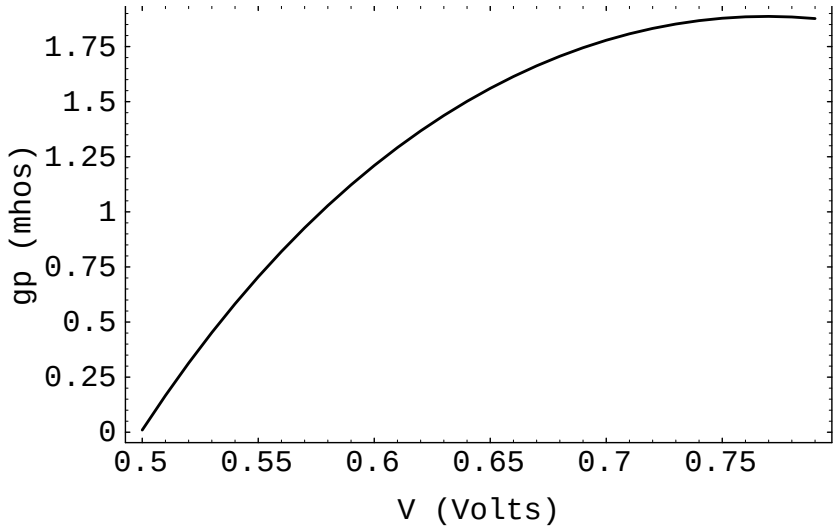

Figure 2.26: Conductance, g_p, as a function of photoreceptor voltage. Although the actual conductance is changing, the effective conductance of the photoreceptor defined as a specific change in potential for a fixed magnitude of injected current is constant. The large change in conductance when the voltage output is small compensates for the small actual conductance.

in a network cannot be taken as evidence for intrinsic membrane channels without careful consideration of the alternatives. I believe that silicon retinas are effective modeling tools in this area and that future silicon modeling efforts will lead to a deeper understanding of these phenomena.

Few experiments have been done on light adaptation as relates to the function of the retinal network. However, the silicon retina suggests that cone adaptation serves to cancel out inter-cone variability. Cone adaptation is mediated by the intracellular calcium concentration, which is a function of the number of light-sensitive channels that are open [26]. This number is in turn a function of the potential of the cone, due to the divalent cation block of the channels. This voltage-dependence allows the electrical feedback from the horizontal cells to affect the chemical concentration of calcium that determines the state of light adaptation in the cone. The voltage-dependence of the current through the light-sensitive channels is unique to cones, which do receive feedback from the horizontal cells, and is absent in rods, which do not receive horizontal cell feedback. The functional significance of the voltage-dependence of the cone channels is unclear. Yau and Baylor [39] state, "It is possible that the peculiar current-voltage relation of the cone conductance has a deeper significance that has not yet been appreciated." The silicon retina has suggested one functional property of this conductance may be a method of inter-receptor calibration. As in the silicon retina, cone light adaptation serves not only to get the cone into the proper operating range, but it cancels out differences between adjacent cone responses through the feedback action of the horizontal network.

2.5.3 Visual Illusions

The brain interprets retinally encoded information to create a model of the objective world. This process remains largely mysterious. However, visual illusions provide some hints about the interaction between retinal output and cortical processing. When the brain perceives an illusion, it is in some sense confusing the real stimulus pattern and another possible stimulus pattern; what one sees *looks like* something else. In fact, the output of the silicon retina correlates well with several well-known visual illusions, such as the simultaneous contrast illusion (Figure 2.27 and Figure 2.28), Mach Bands (Figure 2.29), and the Hermann–Hering grid illusion (Figure 2.30-Figure 2.35). The center-surround encoding process maps the

illusory input to an output that looks like the illusory percept.

There are some cases in which the retinal responses to the illusory stimulus and the actual stimulus are identical. For example, in retinal afterimages (Figure 2.36-Figure 2.39) the response to the removal of the image to which the retina has adapted is identical to what the retinal response would be if it had been adapted to a uniform field and then presented with the negative image. The extent to which the identity between the illusory stimulus and an actual stimulus exists at the output of the retina indicates the relative role of retinal and cortical processing in producing an illusion.

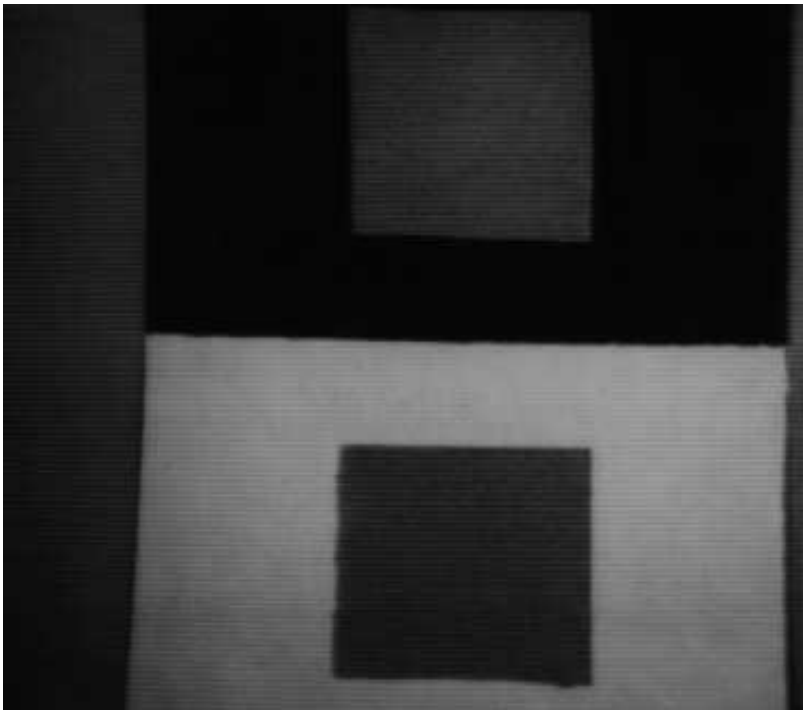

Figure 2.27: The stimulus to produce a simultaneous contrast illusion. The stimulus consists of two identical grey rectangles placed over backgrounds of opposite contrast. The grey square surrounded by black looks brighter than the grey square surrounded by white.

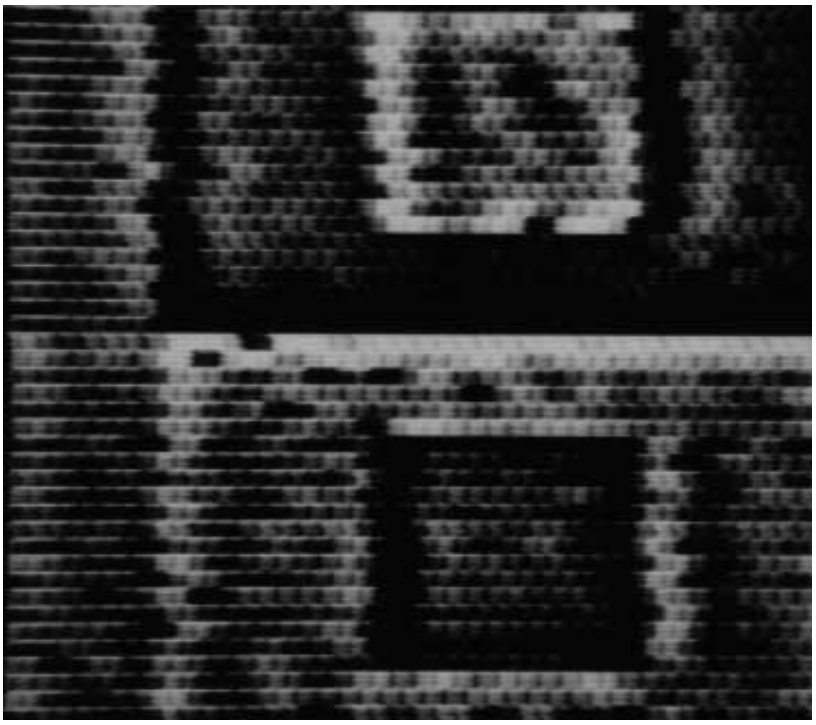

Figure 2.28: The response of the silicon retina to this stimulus. The retina encodes the sign of the contrast, which is positive (bright) for the grey square on black and negative (dark) for the grey square on white.

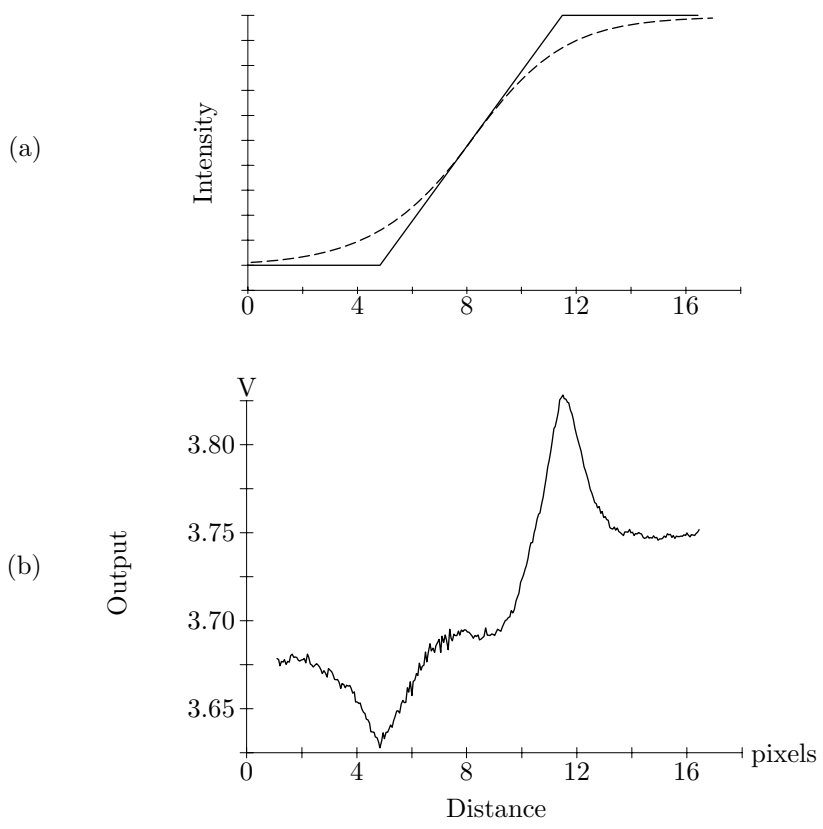

Figure 2.29: **Mach bands are illusory bright and dark bands that appear at the edges of an intensity ramp. The positions of the illusory bands correspond to the positions where the the first derivative of the intensity is changing. Because the retina performs a second-order filtering of the image, changes in the first derivative of intensity are enhanced. (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.

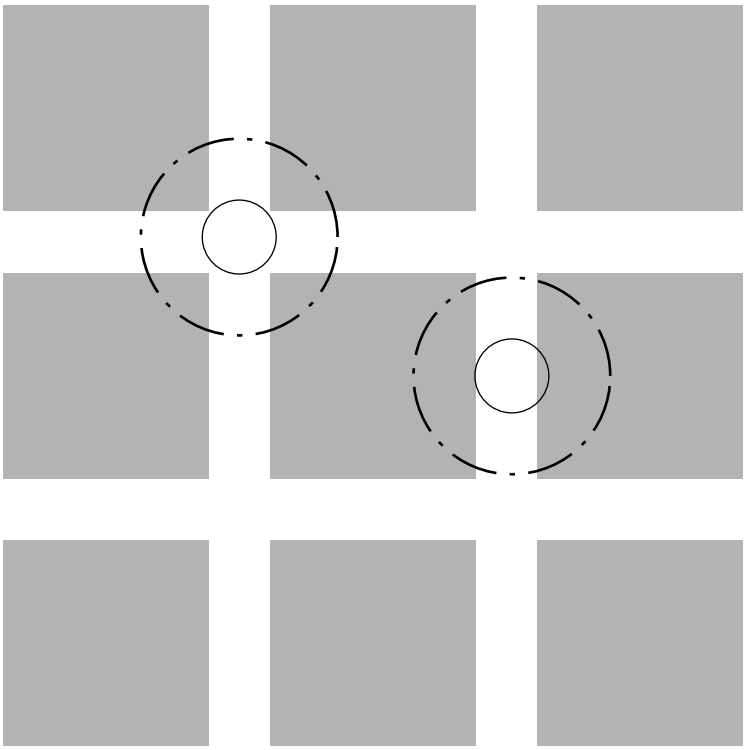

Figure 2.30: The Herring grid illusion appears as grey spots at the intersection points of a square grid viewed at the correct distance. The center-surround receptive fields of the retina compare the average intensity in the surround (dotted outline) to the intensity in the center (solid outline). The neighborhood of the intersections contains more white space and so reduces the apparent brightness of the intersection itself.

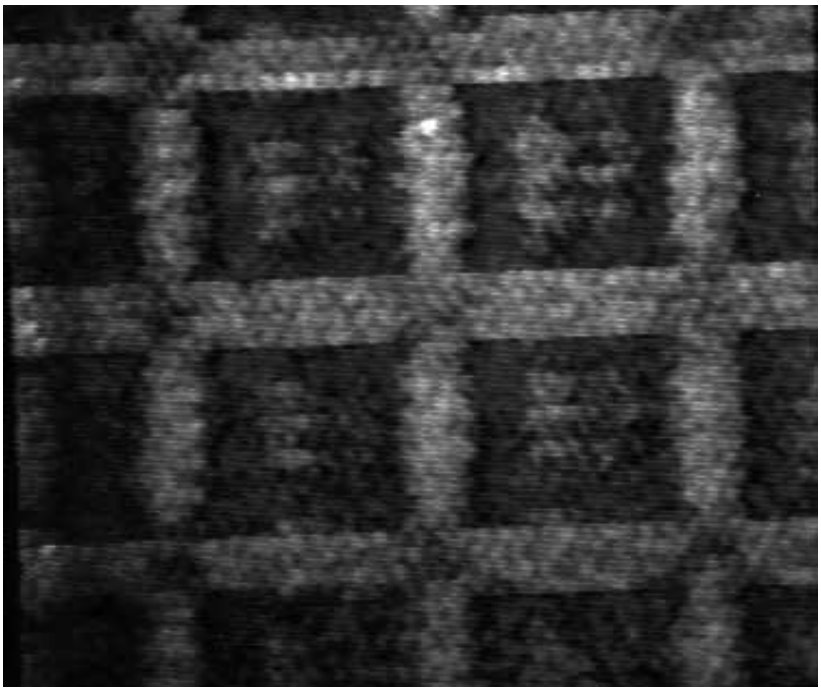

Figure 2.31: Output of the silicon retina. When the center is roughly the size of the white space and the surround is moderately sized, the positive output of the pixels centered at the intersections of the grid are smaller than the outputs of pixels in the boarders. At this viewing distance, this diminished output is interpreted by the brain as dark spots in the intersections.

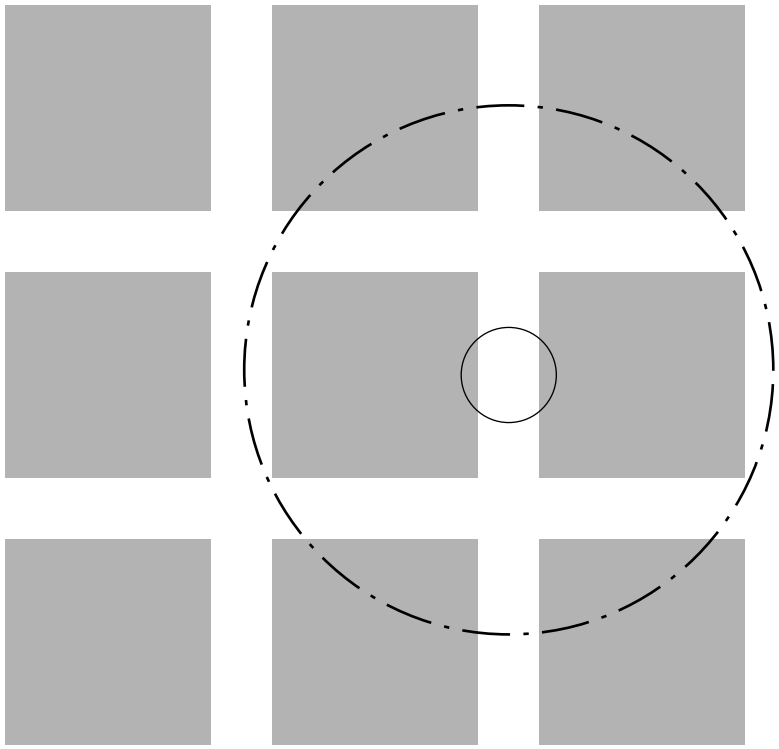

Figure 2.32: **When the surround is large relative to the grid then the retina reports the image intensity reference to the average grey level. The spread of activity in the resistive network is large enough that the average is the same everywhere in the image.**

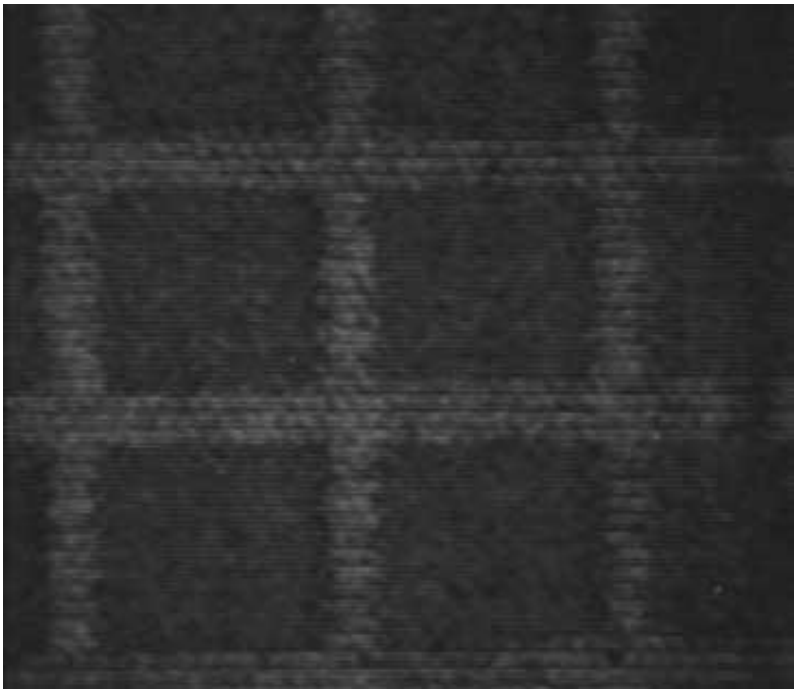

Figure 2.33: Response of the retina to grid stimulus when the averaging distance in the resistive net is very large. In this configuration, the resistive net is essentially reporting the global lighting conditions. The output is not edge-enhanced; intensity is reported relative to the global average.

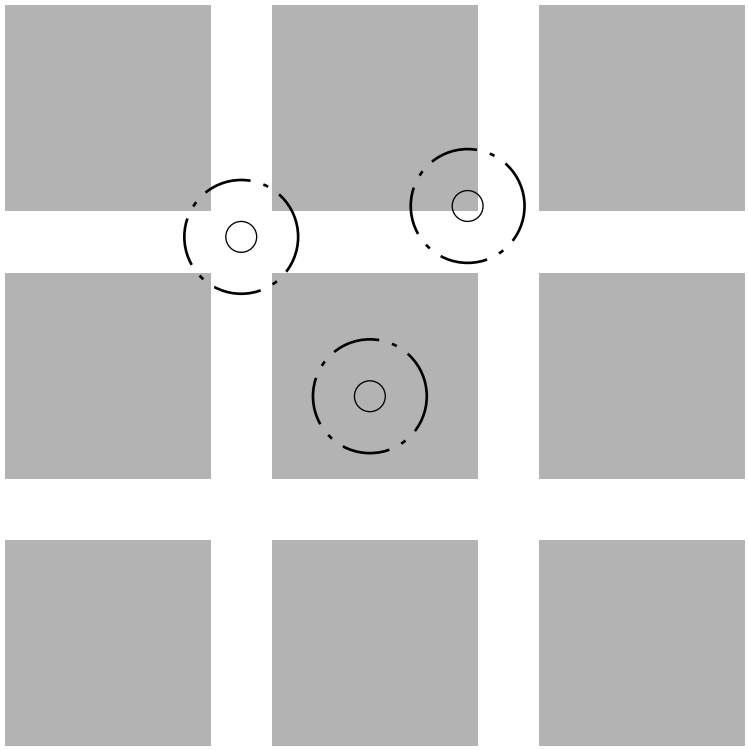

Figure 2.34: When the center and surround are small relative to the grid, then the edges are enhanced. In this viewing condition, no illusion is perceived, although the mechanism used by the brain to interpret the retinal output is unknown.

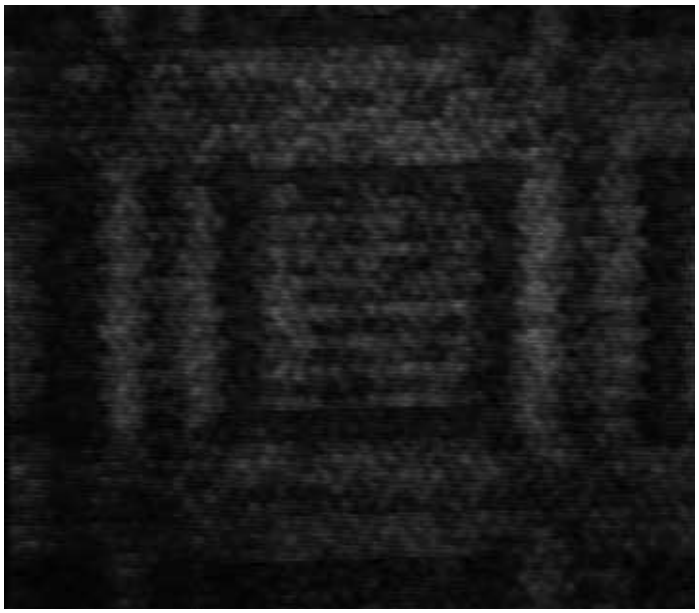

Figure 2.35: Response of the retina to grid stimulus when the center of the receptive field is small compared to the width of the grid. The center lines of the grid are neutral gray except at the outermost edges.

Figure 2.36: High-contrast line drawing of Abraham Lincoln used to illustrate the formation of after-images by the feedback retina.

Figure 2.37: Response of the retina to the stimulus shown in Figure 2.36. The image is positive and edge-enhanced.

Figure 2.38: After 2 minutes, the response has faded due to photoreceptor adaptation.

Figure 2.39: When the image of Lincoln to which the retina has adapted is replaced by a uniform intensity pattern, a netative contrast afterimage appears.

2.6 Summary

The primary task of the retina is to produce meaningful output in a wide range of lighting conditions. It must do so within the constraints of its own physical medium. The power supply limits the output range and the resolution within that range is limited by noise and device imperfections. Furthermore, the nature of optical projection requires that the transducing surface form a two-dimensional sheet with detector-packing density limiting resolution. The solution to this problem under these constraints leads naturally to lateral inhibition via a resistive network.

The silicon retina is a simple physical structure similar to the vertebrate retina. The resistive network computes with minimum wire density a spatiotemporal average that is used as a reference point for the system. By feedback to the photoreceptors, the network signal balances the photocurrent over several orders of magnitude. As the surround mechanism for the bipolar cell receptive field, the horizontal cell also computes the gray-level (zero) for retinal output. The silicon retina's response to spatial and temporal changing images captures much of the complex behavior observed in the OPL.

Analysis of the silicon retina highlights the role of active processes in controlling signal spread in the horizontal network. A comparison between the silicon retina and the biological retina suggests a dual role for the voltage sensitivity of light-gated channels in the cone; the voltage-sensitivity maintains the effective conductance of the active feedback and provides a means of inter-cone calibration through calcium adaptation.

Physical constraints on the operation of the retina determine the way in which information is represented. This point is of further biological significance because the encoding affects later stages of visual perception. The real-time two-dimensional output of the silicon retina illustrates the relationship between lateral inhibition and several visual illusions. From an engineering viewpoint, the efficiency with which the retina encodes visual information encourages a re-evaluation of the way in which image data are transmitted. The silicon retina has inspired the development of a novel protocol for efficient transmission of neural-like signals between chips, which is described in Chapter 3.

References

[1] Atick, J., and Redlich, A. (1990) Quantitative test of a theory of retinal processing: contrast sensitivity curves.

[2] Baylor, D., Fuortes, G., and O'Bryan P. (1971) Receptive fields of single cones in the retina of the turtle. *J. Physiol.* **234**, pp. 256–294.

[3] Boahen (1991)

[4] Byzov, A., and Trifonov, Y. (1981) Ionic mechanisms underlying the nonlinearity of horizontal cell membrane. *Vision Res.* **21**, pp. 1573–1578.

[5] Byzov, A., and Shura-Bura, T. (1983) Spread of potentials along the network of the horizontal cells in the retina of the turtle. *Vision Res.* **4**, pp. 389–397.

[6] Detwiler, P., Hodgkin, A., and McNaughton, P. (1978) A surprising property of electrical spread in the network of rods in the turtle's retina. *Nature*, **274**, pp. 562–565.

[7] Dong, C., and McReynolds, J. (1991) The relationship between light, dopamine release and horizontal cell coupling in the mudpuppy retina. *J. Physiol.* **440**, pp. 291–309.

[8] Dowling, J. (1987) *The Retina: An Approachable Part of the Brain.* Cambridge, MA: Harvard University Press.

[9] Enroth-Cugell, C. and Robson, J.G. (1966) The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology*, **187**, p.517.

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

[11] Hille, B. (1984) *Ionic Channels of Excitable Membranes*. Sunderland, MA: Sinauer Associates Inc.

[12] Julesz, B. (1971) *Foundations of Cyclopean Perception*. Chicago, Illinois: The University of Chicago Press.

[13] Kamermans, M., Van Dijk, G., Spekreijse, H., and Zweyphenning, R. (1989) Lateral feedback from monophasic horizontal cells to cones in carp retina. *J. Gen. Physiol.* **93**, pp. 681-694.

[14] Kamermans, M. (1989) *The Functional Organization of the Horizontal Cell Layers in Carp Retina: An Electrophysiological and Model Study* Doctoral Disseratation, University of Amsterdam, Amsterdam, the Netherlands.

[15] Kamermans, M., Van Dijk, B., Spekreijse, H., and Werblin, F. (1990) A model for the changes in coupling and kinetics of cone driven retinal horizontal cells during light/dark adaptation. *Proceedings of Analysis and Modeling of Neuronal Systems,* July 25-27. Berkeley, CA (in press).

[16] Lankheet, M., Frens, M. R., and Van de Grind, W. (1990) Spatial properties of horizontal cell responses in the cat retina. *Vision Res.* **30**, pp. 1257-1275.

[17] Mahowald, M. A. and Mead, C. A. (1989) "Silicon Retina," In Mead, C. A., *Analog VLSI and Neural Systems*, pp. 257–278, Addison-Wesley, Reading, MA.

[18] Mahowald, M. A. "Silicon Retina with Adaptive Photoreceptors," SPIE's International Symposium on Optical Engineering and Photonics in Aerospace Sensing, Orlando, FL, April 1-5, 1991.

[19] Marr, D. (1982) *Vision*. New York: W. H. Freeman.

[20] Mead, C. A. and Mahowald, M. A. (1988) "A silicon model of early visual processing," *Neural Networks*. **1**, pp. 91–97.

[21] Mead, C.A. (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

[22] Mead, C. (1989) Adaptive retina, In C. Mead and M. Ismail (Eds.), *Analog VLSI Implementation of Neural Systems* (pp. 239–246) Kluwer Academic Publishers, Boston.

[23] Mead, C., and Delbrück, T. (1991) "Scanners for visualizing activity of analog VLSI circuitry," *Analog Integrated Circuits and Signal Processing* (in press).

[24] Nelson, R. (1977) Cat cones have rod input: A comparison of the response properties of cones and horizontal cell bodies in the retina of the cat. *J. of Comparative Neurology* **172**, pp. 109-136.

[25] Normann, R. A., and Werblin, F. S. (1974) Control of retinal sensitivity. I. Light and Dark Adaptation of Vertebrate Rods and Cones, *Journal of General Physiology,* **63**, pp. 37–61.

[26] Pugh, E., and Altman, J. (1988) A role for calcium in adaptation. *Nature*, **334**, pp. 16-17.

[27] Ratliff, F. (1965) *Mach Bands: Quantitative Studies on Neural Networks in the Retina*. San Francisco, California: Holden-Day, Inc.

[28] Shapley, R. and Enroth-Cugell, C. (1984) 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.

[29] Sivilotti, M., Mahowald, M., and Mead, C. (1987) Real-time visual computations using analog CMOS processing arrays. In Losleben, P. (ed.), *Proceedings of the Stanford Conference on Very Large Scale Integration* (pp. 295–311), Cambridge, MA: MIT Press.

[30] Sivilotti, M. (1991) Wiring considerations in Analog VLSI Systems, with Application to Field-Programmable Networks. Doctoral dissertation, Department of Computer Science, California Institute of Technology, Pasadena, California.

[31] Skrzypek, J. (1991) Light Sensitivity in cones is affected by the feedback from horizontal cells, Technical report, Department of Computer Science, UCLA.

[32] Srinivisan, M., Laughlin, S., and Dubs, A. (1982) Predictive coding: A fresh view of inhibition in the retina. *Proc. R. Soc London Ser. B*, **216**, pp.427–459.

[33] Usui, S., Mitarai, G., and Sakakibara, M. (1983) Discrete nonlinear reduction model for horizontal cell response in the carp retina. *Vision Res.* **23**, pp. 413–420.

[34] von Békésy, G. (1967) *Sensory Inhibition.* Princeton, NJ: Princeton University Press.

[35] Werblin, F. S. (1974) Control of retinal sensitivity. II. Lateral interactions at the outer plexiform layer, *Journal of General Physiology,* **63**, pp. 62–87.

[36] Werblin, F.S. (1975) Anomalous rectification in horizontal cells. *J. Physiol.*, **244**, pp. 639-657.

[37] Winslow, R., and Knapp, A. (1991) Dynamic models of the retinal horizontal cell network. *Prog. in Biophys. Molec. Biol.* **56**, pp. 107–133.

[38] Yang, X., and Wu, S. (1991) Feedforward lateral inhibition in retinal bipolar cells: Input-output relation of the horizontal cell-depolarizing bipolar cell synapse. *Proc. Natl. Acad. Sci. USA* **88**, pp. 3310–3313.

[39] Yau, K., and Baylor, D. (1989) Cyclic GMP-activated conductance of retinal photoreceptor cells. *Annual Reviews of Neuroscience* **12**, pp. 289–327.

Chapter 3

The Silicon Optic Nerve

3.1 Introduction

Communication between neuronal elements is a principal limiting factor in the design of VLSI neuromorphic systems. This fact is not surprising considering that a large fraction of the volume of the nervous system is composed of myelinated axons. The degree of convergence and divergence of single neurons is staggering in comparison with man-made computers. It might appear impossible, even in principle, to build such structures in VLSI circuits, which are limited to an almost two-dimensional plane of silicon. Surprisingly, the cortices of the brain are nearly two dimensional as well. In fact, it has been shown that the degree of connectivity in a system whose wires occupy space cannot be increased by employing a structure in which nodes are arrayed in three dimensions [27]. There is nothing fundamental about the structure of neural tissue that cannot be embedded in silicon. The thickness of cortical structures can be represented with a correspondingly larger silicon surface area. However, silicon surface area is available on small die, which are several millimeters on a side and so the number of neurons that can be fabricated on a single die is limited. Consequently, connections between silicon neurons located on different chips are essential for building even moderately sized artificial neural systems.

3.2 Summary of Existing Techniques

The degree of connectivity and the real-time nature of neural processing demand different approaches to the problem of interchip communication than those used in traditional digital computers. VLSI designers have adopted several strategies for interchip communication in silicon neural networks. Each strategy has advantages and the choice of method depends on which factors are most crucial to the system.

One of the most literal approaches to interconnecting processing nodes has been adopted by Paul Mueller's group [19]. Mueller uses a direct physical connection between nodes on different chips through a cross-bar switching array. A major advantage of this approach is that it allows continuous time communication between nodes. In addition, the switching arrays provide flexible connectivity and can be programmed digitally by a host computer. The system is able to handle large connectivities because the dendrites of a single artificial neuron can extend over multiple chips. However, this approach requires many chips to model even a small number of neurons. The number of artificial neurons on each output chip is limited to roughly half the number of pins that are available. Current technology supports 84-pin grid arrays, and in the near future will be extended to 128, meaning at most 64 neurons per chip. A further disadvantage of this design is that, in order to achieve a reasonable degree of matching between the analog performance of the different chips in the system, the transistors are used in their above threshold regime, where power dissipation is great.

Some applications, such as sensory transduction [16] in which the silicon surface acts as a sensory epithelium, require many neurons to be placed on the same chip. The total number of neurons in such a structure greatly exceeds the number of pins available for transmitting their outputs to off-chip targets. The standard approach to resolving this difficulty is to sample and transmit the states of the neurons in sequence. In this case continuous time communication must be sacrificed in order to time-multiplex the outputs of many neurons onto a small number of wires. The output of each neuron is sampled and transmitted for a brief time. The speed at which data can be transmitted determines the frequency above which information will be lost due to temporal aliasing.

Traditional multiplexing schemes are serial access. Each node is polled in fixed sequence

and its output sent off-chip. Each time slot is allocated to a particular node and the receiving device must be synchronized with the sending device in order to preserve the identity of the transmitting node. Most multiplexing schemes rely on a global clock to perform this synchronization. Global clock signals may be skewed to the point of dysfunction if the chips comprising the system are too far from each other.

The choice of multiplexing technique depends on how the neural elements in the system encode information. Some systems use analog-valued outputs, which encode several bits of information on a single wire. In analog multiplexed systems, the receiver chip samples the data stream and holds the data in a buffer until the next frame [10, 18]. This approach is particularly useful for interacting with video equipment as such equipment is designed to work with analog-valued image frames [26]. However, analog data transfer is difficult between chips, in part because the analog data are easily perturbed by noise due to multiplexing. More importantly, the variations in the parameters of fabrication on different wafers means that different chips will have disparate interpretations of analog voltages. These difficulties are avoided by transmitting digital amplitude signals.

Both synchronous and asynchronous techniques have been used to time-multiplex digital amplitude data [20]. Digital signal transmission can be very fast because the settling time for an analog amplifier is avoided. Furthermore, digital signals are noise resistant and independent of variations in fabrication parameters. Synchronous transmission of multiple bits of information has the drawback that synchronous switching of many elements causes noise on the power supply. Asynchronous serial digital communication methods in which the duration of the digital pulse encodes several bits of information have been used [3, 20]. In the voltage-controlled-oscillator encoder used by Murray and collaborators [20, 3], the duration of the pulse is inversely proportional to the analog value of the output. Rather than using a global clocking mechanism to allocate specific time-slots to particular nodes, the identity of the sending neuron is determined by its position in the pulse stream. The node position is computed from the number of transitions in the stream itself. The pulse stream provides its own clock. The pulse stream technique uses time to encode analog state, rather than to communicate explicitly temporal information.

3.3 The Address-Event Representation

The interchip communication protocol that we have developed is an asynchronous digital multiplexing technique which uses an *address-event representation*. The address-event representation has much in common with the action-potential representation used by real neurons. Like neuronal action potentials, events in this system are stereotyped digital amplitude events and the interval between events is analog. Information is encoded in the time between events. The principle of this encoding scheme is that N axonal fibers, with one active at a time, can be replaced by $(1 + \log N)$ wires, which are simultaneously active. Several fibers in a real nerve bundle may be simultaneously active and so violate the encoding condition. This situation can be dealt with in the address-event representation by making the event duration very short (approximately 1 μsecond) compared with the width of neural action potentials (approximately 0.5 millisecond). Short-duration events have small opportunity to overlap. Since, as in a real neuron, the maximum firing rate of a node is limited, even if events from several nodes did occur synchronously, they could be arbitrarily arranged so that they occurred in close succession with little loss of information.

The address-event representation is illustrated in Figure 3.1. The neurons in the sender array generate a temporal sequence of digital amplitude events to encode their output, a representation conceptually equivalent to a train of action potentials. Each neuron is associated with a digital address which uniquely identifies it. Whenever a neuron signals an event, the multiplexing circuitry broadcasts that neuron's address on the inter-chip data bus. The nodes have a refractory period that limits the frequency at which they can issue events. The inter-event interval at a neuron is much longer than the time required to broadcast the neuron's address. Therefore, many addresses can be multiplexed on the same bus. The receiver interprets the broadcast of the address as an event that corresponds to the occurrence of an action potential from the neuron identified by that address. For this reason, we have named our communication code an address-event representation.

Although I have chosen to transmit only the neuron address, which corresponds to a digital amplitude event, it is possible in principle to use the address-event representation to transmit explicitly analog signals. In such a system, the address would be transmitted along with one or more analog values associated with the identified pixel. The pixel would

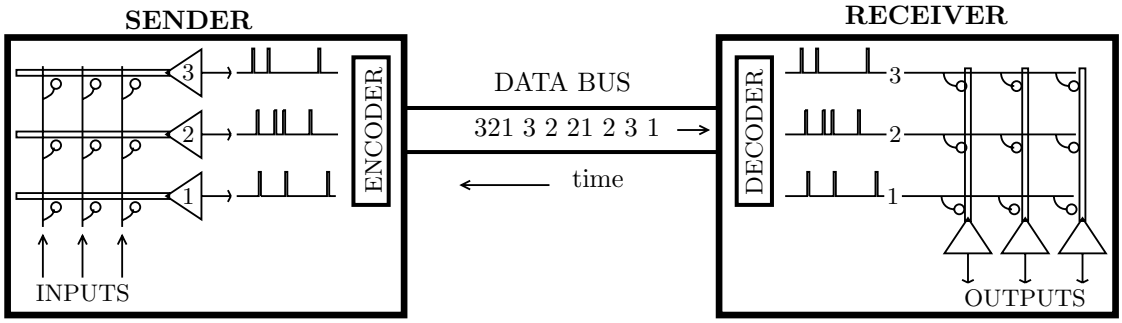

Figure 3.1: The address-event representation. Self-timed neurons on the sending chip generate trains of action potentials. The neurons request control of the bus when they generate action potentials and are selected to transmit their addresses by the multiplexing circuitry. A temporal stream of addresses passes between the sender chip and the receiver chip. This temporal stream is decoded by the receiver into trains of action potentials that reach their proper postsynaptic targets. The detailed timing of the events is preserved.

make a request to transmit the analog data when some control signal indicated a threshold had been passed.

The address-event representation is designed to provide high-bandwidth communication between large arrays of neuron elements. Time-multiplexing is the only way to transfer data from several thousand output nodes within the pin limitations of existing packaging technology. The premise underlying the address-event representation is that the channel bandwidth should be devoted to the transmission of significant signals. For example, the silicon retina [16] has roughly 4000 output nodes. Conventional scanning techniques require that each node be sampled once every frame. Since the retina generates output only at areas in the image where there is spatial or temporal change in the image, most of the nodes will have almost no output, but are sampled anyway. The address-event protocol, in contrast, is data driven. Only pixels that have something to report are transmitting their output over the data bus. Therefore, areas of uniform illumination do not contribute to the communication load. A further major advantage of the address-event communications framework is that it minimizes temporal aliasing by transmitting events as they occur. It need not introduce the degree of sampling inherent in a sequential scanning technique. At low data rates, the bandwidth of the bus is completely devoted to accurate transmission of relative timing of events.

3.3.1 Model of Data-Transfer Timing Efficiency

The temporal efficiency of a traditional, sequentially scanned data-multiplexing system is easy to evaluate because it is exclusively a property of the machine and not a property of the data. The data will occur at random within the frame, and so the average error introduced by waiting to scan the data out of the array is half a frame time. The frame time increases linearly with the number of elements in the array.

Since the address-event communications protocol specifically synchronizes data transfer with the timing of the data, the details of timing efficiency cannot be analyzed without a model of the data to be transmitted. However, an analysis of the average behavior of the system can be performed by assuming that the elements in the array are initiating data transfer requests independently of each other, each at some rate.

A simple model is shown in Figure 3.2. In this model, all of the elements in the array

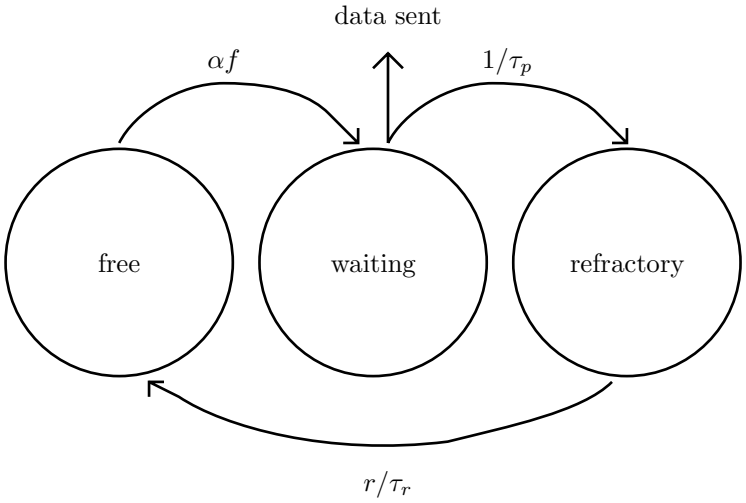

Figure 3.2: Model of the address-event data transfer process. Each sending neuron can be in one of three state, free to generate an event, waiting to transmit an event that it has already generated, or refractory, having just transmitted an event.

are initiating requests at the same rate, α, which has units of events/unit time/element. An element can be in one of three states: it can be free to initiate a request, it can be waiting to have its data transmitted, or it can be in a refractory state. The refractory period is a time in which the element is prohibited from making another request for data transfer after it has successfully transmitted an event. This send time plus the refractory period sets the absolute maximum event rate that an element can attain. Since all of the elements must be in one of these three states, the sum of the elements in all of the states is equal to the total number of elements.

$$N = f + w + r$$

where N is the total number of elements, f is the number free, w is the number waiting and r is the number refractory. The equations governing the movements of elements from one pool to another are:

$$\frac{df}{dt} = -\alpha f + \frac{r}{\tau_r}$$

$$\frac{dw}{dt} = \begin{cases} \alpha f - \frac{1}{\tau_p} & \text{if } w \geq 1 \\ \alpha f - \frac{w}{\tau_p}) & \text{if } 0 \leq w \leq 1 \end{cases}$$

$$\frac{dr}{dt} = \begin{cases} -\frac{r}{\tau_r} + \frac{1}{\tau_p} & \text{if } w \geq 1 \\ -\frac{r}{\tau_r} + \frac{w}{\tau_p} & \text{if } 0 \leq w \leq 1 \end{cases}$$

None of the pools is allowed to contain a negative number of elements. Elements move from the free pool to the waiting pool at an average rate of αf. The waiting elements are serviced one every τ_p, the data transfer time. If there is less than one element waiting on average, the event will be transmitted as soon as it occurs. Of course, the time between events is longer, which is reflected in the equations when $0 \leq w \leq 1$. Elements enter the refractory pool as they are serviced. They leave the refractory pool and re-enter the free pool at a rate $\frac{r}{\tau_r}$, the number of elements that are refractory divided by the refractory time. This term depends on the system having reached steady state, so that the elements are hopping in and out of the refractory pool at the same rate. If events stopped entering the refractory pool, all of the elements that were in the pool would be gone after one refractory time, τ_r, had elapsed. Therefore, in time dt, $\frac{dt}{\tau_r}$ fraction of them will leave the refractory pool.

In steady state, all of the derivatives are zero, and the solutions for w and r become:

$$w/N = \begin{cases} 1 - \frac{1}{N\alpha}\frac{1}{\tau_p} - \frac{\tau_r}{N\tau_p} & \text{if } w \geq 1 \\ \frac{\alpha\tau_p}{1+\alpha(\tau_r+\tau_p)} & \text{if } 0 \leq w \leq 1 \end{cases}$$

$$r/N = \begin{cases} 1 - \frac{1}{N\alpha}\frac{1}{\tau_p} - \frac{\tau_r}{N\tau_p} & \text{if } w \geq 1 \\ \frac{\alpha\tau_r}{1+\alpha(\tau_r+\tau_p)} & \text{if } 0 \leq w \leq 1 \end{cases}$$

When the data rates are low, α is small, and $w \leq 1$. When the system is operating within its design limits, the total number of events generated per second, $N\alpha$, is smaller than the data transfer rate, $\frac{1}{\tau_p}$. If the refractory period is fairly short, the equations for w and r have denominators approximately equal to 1. The number of neurons waiting is just equal to the number of events in one data transfer time, and the number of neurons refractory is

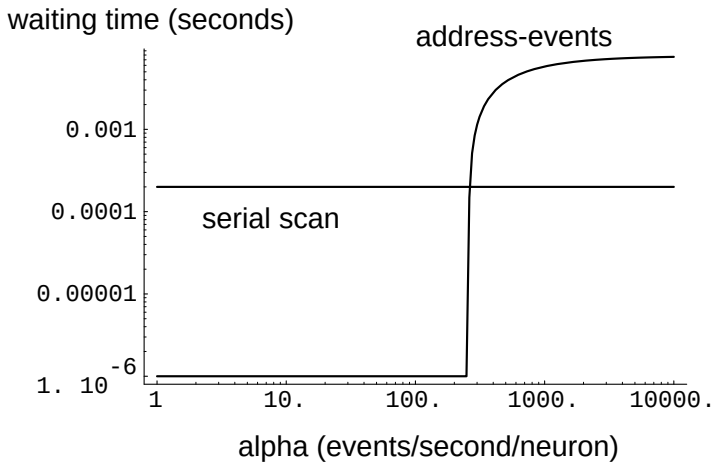

Figure 3.3: Comparison of address-event data transfer timing error with that of a serial scanning system. Ordinate is the average event-rate of the neurons in the array. Coordinate is the average waiting time for the transmission of the event.

just equal to the number of events in one refractory time. At these data rates, the system functions ideally and barring correlations in the data stream, each event is transmitted as soon as it is generated.

When the data rate is high, there is more than one element waiting to transmit its data. In this case, the refractory period is holding as many elements as it can, which is $\frac{\tau_r}{\tau_p}$. If the refractory period is so large that N events can be transmitted in one refractory time, the number of neurons waiting must be none, because they can be serviced as soon as they fall out of the refractory state. At higher rates, the system fails gracefully, since the neurons that cannot be serviced are taken out of the pool that is free to generate new events and placed in the waiting pool and the refractory pool.

If the neurons are selected at random from the waiting pool, w, then the mean waiting

time will be $2w\tau_p$. In the implementation of the address-event communications framework described in this chapter, the data transfer time, τ_p, scales logarithmically with the number of neurons. A comparison of a sequential scanning system with the address-event system is shown in Figure 3.3. The mean event delay is plotted against the events per second generated by each neuron; time is given in units of seconds. The data transfer time, τ_p, is 1 microsecond. This plot shows the data rates for which the address-event protocol gives shorter time delays than sequential scans. The sequential scan data transfer rate per pixel is estimated to be $0.05\tau_p$. The refractory time is set to $(N/4)\tau_p$. This plot shows that the data transfer delay in the address-event system is much better than that in the sequential scan system when the data rate is less than the critical value, after which the address-event system becomes rapidly worse than sequential scan. As mentioned previously, this model does not take into consideration correlations between events either due to random chance or correlated input. It simply illustrates the maximum allowable data rate for the system. The critical point is reached when the number of events generated per second is larger than the number of data transfer times per second. The refractory period can absorb some neurons so that, at the point of failure, the number of neurons that are free to generate events at rate α is $N - \tau_r/\tau_p$. When the system fails, it simply transmits data as quickly as it can, but all of the neurons may be waiting to send more data. If the system briefly exceeds the maximum spike rate, the neurons enter the waiting queue and are removed when their data are transferred. Transient periods of high spike rate cause a loss of temporal resolution, but do not cause irrevocable failure.

This model is over-simplified in several respects, one aspect being that the data are unlikely to be evenly distributed over all of the elements in the system. The model can be extended to include several sub-populations of elements generating events with different rates. Each sub-population follows its own conservation law:

$$N_i = f_i + w_i + r_i$$

where N_i is the number of neurons in the sub-population that generates events at a rate, α_i events/second/element. The total number of elements is $N = \sum i = 1\mathrm{M}N_i$, where M is the number of different sub-populations. The populations are coupled together by the fact that

they are serviced by a common data transfer mechanism. If this mechanism is unbiased then the equations that govern the distributions of elements within each population are:

$$\frac{df_i}{dt} = -\alpha_i f_i + \frac{r_i}{\tau_r}$$

$$\frac{dw_i}{dt} = \begin{cases} \alpha_i f_i - \frac{1}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k} & \text{if } w_i \geq 1 \\ \alpha_i f_i - \frac{w_i}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k}) & \text{if } 0 \leq w_i \leq 1 \end{cases}$$

$$\frac{dr_i}{dt} = \begin{cases} -\frac{r_i}{\tau_r} + \frac{1}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k} & \text{if } w_i \geq 1 \\ -\frac{r}{\tau_r} + \frac{w_i}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k} & \text{if } 0 \leq w_i \leq 1 \end{cases}$$

Once again, none of the f_i, w_i, or r_i is allowed to be negative. This system is non-linear in the case of $0 \leq w \leq 1$, and is difficult to solve. Some progress, however, can be made in the case where $w \geq 1$. Setting all of the derivatives equal to zero gives:

$$f_i = \frac{1}{\alpha_i \tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k}$$

$$r_i = \frac{\tau_r}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k}$$

$$N_i = \frac{1}{\alpha_i \tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k} + w_i + \frac{\tau_r}{\tau_p}\frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k}$$

Let $W = \frac{w_i}{\sum_{k=1}^{\mathrm{M}} w_k}$.

$$WN_i = (\frac{1}{\alpha_i \tau_p} + W + \frac{\tau_r}{\tau_p})w_i$$

Rearranging to solve for w_i yields:

$$w_i = \frac{WN_i}{\frac{1}{\alpha_i \tau_p} + W + \frac{\tau_r}{\tau_p}}$$

The fraction of N_i that is waiting is larger for the populations that are generating events at a higher rate, α_i.

To calculate the delay time, it is necessary to solve for W. Unfortunately, even with such a simplified model, the mathematics once again becomes intractable. The equation for

W is polynomial of order M, and can only be solved when $M = 2$. Summing both sides of the previous equation over all the populations gives:

$$\sum_{i=1}^{\mathrm{M}} w_i = W = \sum_{i=1}^{\mathrm{M}} \frac{W N_i}{\frac{1}{\alpha_i \tau_p} + W + \frac{\tau_r}{\tau_p}}$$

Dividing through by W gives:

$$1 = \sum_{i=1}^{\mathrm{M}} \frac{N_i}{\frac{1}{\alpha_i \tau_p} + W + \frac{\tau_r}{\tau_p}}$$

I have used these equations to calculate the waiting time for event transfer in one system in which I intend to use the address-event communications framework, namely, in the silicon retina. The silicon retina has roughly 4000 pixels. In a typical image about a quarter of them are activated above the spontaneous level. The spontaneous rate for our silicon neurons is 15 Hz and a fast event rate is 300 Hz. The refractory time of a neuron is about 1 millisecond. The data transfer time measured for the system described here is $\tau_p = 2 \times 10^{-6}$. With these parameters, there is no queue for data transfer. The average number of spikes/second/neuron is 86, which is within the working range for the event-address system depicted in Figure 3.3. This estimate does not account for correlations in the image that give rise to correlated firing. However, it does indicate that at these data rates, the system is performing as well as it could; it has not reached the domain where neurons are not able to generate new events whenever they wish. Used with a system that has a sparse activation profile, the address-event communication framework is able to preserve timing information orders of magnitude better than a sequential scan.

3.3.2 Advantages of Address-Events

The address-event representation provides a unifying framework for the construction of multi-chip systems. Digital-amplitude analog-time events have been used successfully in many silicon neuromorphic systems: auditory localization and pitch perception [11], electrolocation models [12], central pattern generators [21], sensory-motor systems [5], and prototype real-time learning systems [14]. These existing chips could be easily integrated to form more complex systems by placing them in an address-event design frame.

The use of a digital address to specify the identity of the sending neuron makes the mapping of pre-synaptic signals onto post-synaptic targets extremely flexible because the address-event carries its place of origin within itself. Unlike serial-scanning multiplexors, in which temporal order is easily confused with spatial position, the address-event can be easily decoded into any physical ordering on the receiving chip. The ordering can be specified when the chip is designed, particularly if the technique of silicon compilation is used to specify the design. Alternatively, the connectivity pattern can be specified dynamically when the chip is being tested by using static digital latches. In the latter case, specification of the mapping between input and output can be controlled by a host digital computer. The mapping of input to output is itself a complex computation in the nervous system [22] and is a task more easily performed by computer than by hand wiring.

The address-event multiplexing method bears a close resemblance to the action potential representation that is the common coinage of communication in the nervous system. It is likely that the underlying reasons are similar. In an event-based communication scheme, the amplitude of a signal is represented by the *number* and times of events. Time is the same everywhere in the system, and number is an abstract quantity that is also the same everywhere in the system. For example, a signal may be at maximum value when there are 200 events generated per second. The actual voltage value that this maximum corresponds to may be specified independently for each unit in the system. This normalized encoding is useful because an actual analog value is difficult to transmit when the ground potential is not the same everywhere. The lack of a common ground is like the problem of transistor mismatch, which can be modeled to first order as an offset voltage on the transistor gate. It is certainly the case in the nervous system that the ground potential is not uniform in different areas of the brain. In addition to reducing the impact of such static noise, the problem of dynamic noise on the axonal "wires" is ameliorated by using a strongly restored signal.

The richness of this biological representation is not fully understood. Sensory processing has been shown in some cases to take full advantage of the event-like nature of the action potential. For example, the timing of action potentials in the auditory system is crucial in auditory localization [17]. Psychophysical studies indicate that event timing is significant in visual stereo and motion processing [4]. Recently, interest in the spatio-temporal processing

capabilities of cortical neurons has given rise to several hypotheses of information processing in the brain. For example, it has been proposed that information is encoded as synchronized neuronal activities over populations of action-potential generating neurons [8]. This type of synchronization cannot be emulated with multiplexing systems whose frame rate is on the same time scale as the neural oscillation.

The choice of representation of information for inter-chip communication is critical because it determines the way that the system can easily evolve. I believe that this choice of representation can lead to the development of silicon systems whose fundamental information processing strategies are similar to those of neuronal systems. For example, learning based on spatio-temporal processes within the dendritic tree is under investigation [2] and may turn out to be a key issue in neuronal information processing. The flexibility of the digital address allows individual synapses in an artificial dendritic tree to be mapped to their presynaptic elements after fabrication. The address-event representation preserves the temporal order of events. The role of placement of inputs along the dendritic tree in learning spatio-temporal patterns can therefore easily be investigated. If the computational primitives are correctly chosen, the processes of understanding biological systems and of building silicon systems are complementary.

3.4 Data Transfer in One Dimension

A more complete review of self-timed systems can be found in [23]. A few definitions and basic principles are described here to provide background for the remaining discussion. A *self-timed system* generates its own idea of time, independently of an external clock, by keeping track of a sequence of events. The nature of sequence is exemplified by a *handshake*. Like an ordinary handshake, involving two people, a simple handshake involves two chips. One chip, the *sender*, initiates the process, by the equivalent of putting forth its "hand" initiating a *request*. The second chip, the *receiver*, must *acknowledge* the request by "shaking hands" with the sender. To complete the handshake, the sender drops his "hand," removing the request, and the receiver drops his "hand" by removing the acknowledge. The system is returned to its initial state. Both parties are quiescent until some process within the sender initiates another request. The address-event protocol is said to be *data driven* because the

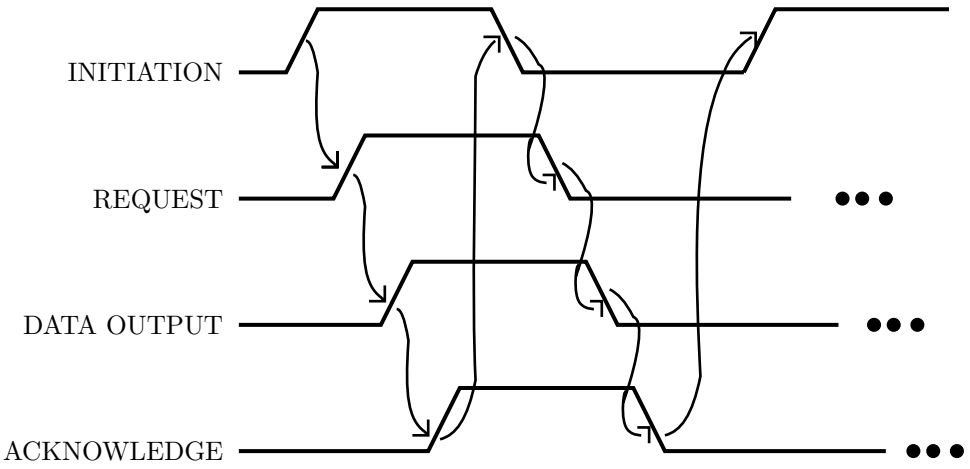

Figure 3.4: The handshake. The INITIATION signal is strictly within the sending chip, while the REQUEST, DATA OUTPUT, and ACKNOWLEDGE signals travel between the sender and the receiver. A pixel with data to transmit initiates data transfer by prompting the sender to make a request. After the sender makes the request, it places the output data on the bus without waiting for the receiver to do anything. The receiver acknowledges receipt of the data. The initiation signal is reset so the sender drops the request. The data are removed from the data bus and the acknowledge is withdrawn.

initiation of the handshake depends on the neural nodes in the sender trying to transmit an event.

We have fabricated a sender retina with 64×64 pixels and a receiver chip with 64×64 nodes in a $2u$ p-well CMOS process. Data transfer between the retina and the receiver is an asynchronous procedure, which is driven by the data generated by the pixels on the sender chip. The request initiated by a pixel begins a cycle of events that results in the transfer of that pixel's address to the receiving chip. When the data transfer cycle terminates, the state of the system is reinitialized so that the cycle is free to occur again when there is another data event. This data transfer protocol is illustrated in simplified form in Figure 3.4. The protocol used by the implementation of address-event data transfer described in this chapter is based on the absolute voltage levels of the signals, rather than their transitions.

In this section, the transfer of address-events between one-dimensional neuronal arrays

is developed. The data transfer process resembles the sequence of events that takes place in the generation of an action potential in real neurons. The communications framework is described in terms of a simple circuit described by Mead [16] as the "axon hillock" circuit which has been used extensively in VLSI neuromorphic systems that use action potential-like communication within and between chips [11, 12, 21, 5, 14].

3.4.1 The Action Potential

The action-potential of a neuron is generated by two main currents, the sodium current and the potassium current. The sodium current is activated when the membrane voltage crosses a threshold level. It depolarizes the membrane and generates the rising phase of the voltage spike. The membrane is repolarized by the delayed potassium current. This function is captured abstractly by the basic circuit shown in Figure 3.5. It is like a three-inverter oscillator except that instead of being fully connected head to tail, the closing link is split into a pull up transistor PP, and a pull down, NA. Since the tail activates the pull-down transistor, the oscillator goes through a single cycle and stops. The oscillation, which is similar to the generation of an action potential, is equivalent to the data transfer process. A single cycle of oscillation (i.e. a single datum transfer) is initiated by the pull up transistor PP.

As a starting point for analysis of the circuit, assume that PP is off and the capacitor on the Initiation node is discharged to ground. In this resting state, the Request node is high and the Acknowledge node is low. The data event initiating a cycle activates PP, which pulls up the Initiation node. In this analysis, we are assuming that PP supplies enough current to pull the Initiation node well above the inverter threshold before the signal can propagate through the oscillator. Because the real system has many stages of delay which have been lumped together in the inverting amplifiers in this diagram, the circuit cannot hang in a state in which PP is just balanced by NA. (The "axon hillock" circuit was prevented from hanging by positive feedback through a coupling capacitor between the Initiation node and the Acknowledge node.) When the Initiation node goes high, the signal propagates through the oscillator. The Request node goes low and the Acknowledge node goes high, activating NA. Assume that NA is stronger than PP, so that the Initiation node is pulled down, independent of the gate voltage controlling PP. When the Initiation node is pulled below

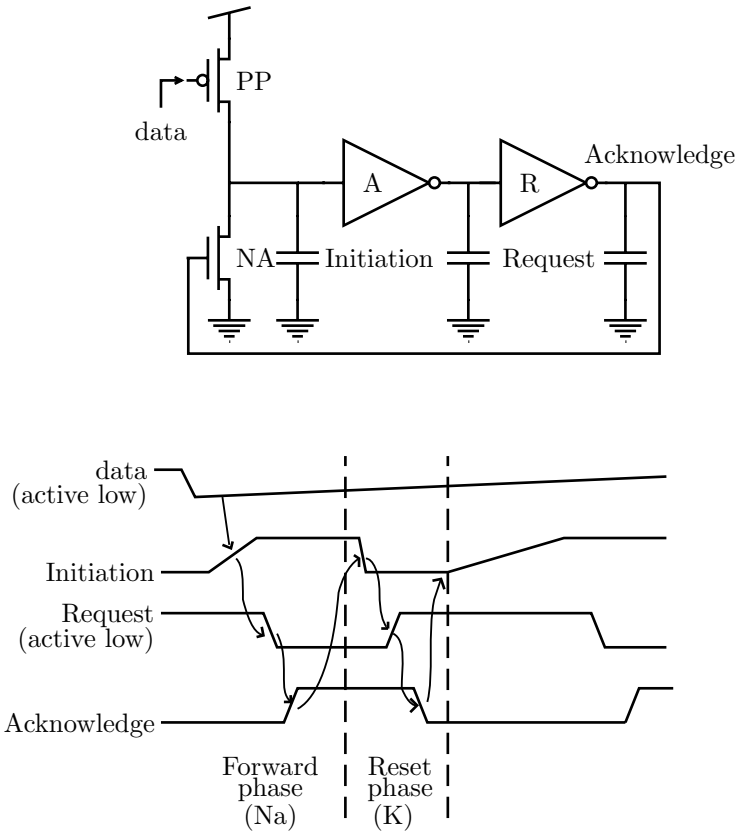

Figure 3.5: A modified three-inverter oscillator and timing diagram illustrating the sequence of events in a single cycle of the oscillator. Inverter A is part of the sending chip. Inverter B is part of the receiving chip.

the inverter threshold, that transition cycles through the oscillator, the receiver withdraws the Acknowledge and turns off NA. The circuit is now ready to begin another cycle, as soon as the Initiation node is charged up again past the inverter threshold.

The voltages on the nodes of the oscillator are shown as a function of time in Figure 3.5. The Initiation node is analogous to the membrane voltage of a neuron before an action potential is generated. Current is integrated on the Initiation capacitor until it passes the inverter threshold. The request, which is amplified by the inverter, is analogous to the sodium conductance in an active membrane. The delayed rectifier potassium current that repolarized the membrane is analogous to the Acknowledge signal. When the Acknowledge node is pulled up, it begins the second phase of the cycle by discharging the Initiation capacitor through NA and resets the system to its initial state. In previous applications of

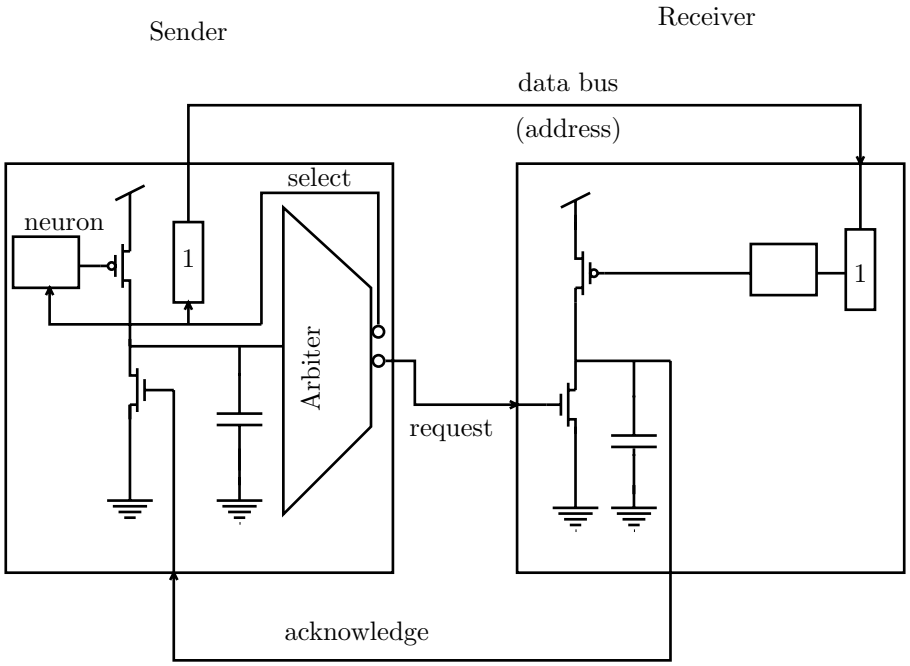

Figure 3.6: Address-event communication system with a single neuron and single-bit address (1). The output of the A inverter has been split into two halves, one of which is directly transmitted to the receiver as the request, the other of which goes back into the neuron to reset the neuron's state and places the neuron's address on the data bus. These two outputs from the A inverter are recombined on the Receiver to generate the Acknowledge signal.

the axon hillock circuit, the action potential waveform was compared to the fully restored digital signal that has been called the Acknowledge in this discussion. The temporal course of the Acknowledge signal is a digital amplitude pulse whose onset is triggered by the Initiation node going above the inverter threshold.

Our data transfer procedure must transfer an address, rather than a single digital amplitude pulse. The axon hillock circuit has been adapted to this end. The adapted circuit is shown in Figure 3.6. The output of the inverter, A, is broken into two parts to be passed on to the Receiver: the select, which places the address on the data bus, and Request signal, which indicates that the data transfer process is activated. The Request signal is low (active) while data transfer is in progress.

The Receiver, which was originally a simple inverter, has been extended to accept the address-event passed on by the sender. The address-event is decoded on the receiver chip into a current that stimulates the post-synaptic neuron. In addition to stimulating the post-

synaptic target, the decoded address pulls up on the Acknowledge node. The pull down transistor driving the Acknowledge node is turned off because the active Request from the A inverter indicates that data transfer is in progress. The successful transfer of an address is the culmination of the forward phase of the data transfer cycle. The reset phase is initiated by the Acknowledge being pulled high by the decoded address. The Acknowledge indicates that data have been transferred. It returns to the Sender and discharges the Initiation node below the inverter threshold.

3.4.2 One-Dimensional Arrays

This circuit is generalized to perform event multiplexing and transmission for many neurons. A one-dimensional sender array is illustrated in Figure 3.7. It is possible to transmit events simply as they happen; however, when events overlap temporally, spurious addresses might be generated. In order to preserve the fidelity of the data without simply discarding the colliding data, arbitration necessary to resolve contention for the bus. The Arbiter, an extension of the A inverter, is responsible for multiplexing events that occur nearly simultaneously onto a single data bus by forcing the neurons to take turns sending their data. In order to perform its multiplexing function, the Arbiter, described in the next section, is extended to have as many inputs as there are neurons in the array. Each neuron controls its own Initiation node. Several neurons may drive their respective Initiation nodes above threshold nearly simultaneously. As in the single pixel case, the output of the Arbiter is split into two types, a single Request signal that is transmitted to the Receiving chip, and the select signals, one for each neuron in the array. The Request signal is activated whenever any event has been supplied to the Arbiter, even if the Arbiter has not selected which event to process. In each data transfer cycle, the Arbiter activates a single Select signal. The Select signal transfers the address of the chosen neuron onto the data bus.

The address encoder is illustrated in Figure 3.8 for a simple two-bit address with the particular value 01. The bits are added to the address encoder as needed and the layout is arranged in such a way that 1's and 0's are interchangeable. This encoder was developed by John Wawrzynk. The address bits are driven onto the address lines by activating the select signal. I have incorporated pull-down transistors at the ends of the address lines so that the address goes to all zeros, which is a null address, when none of the neurons is

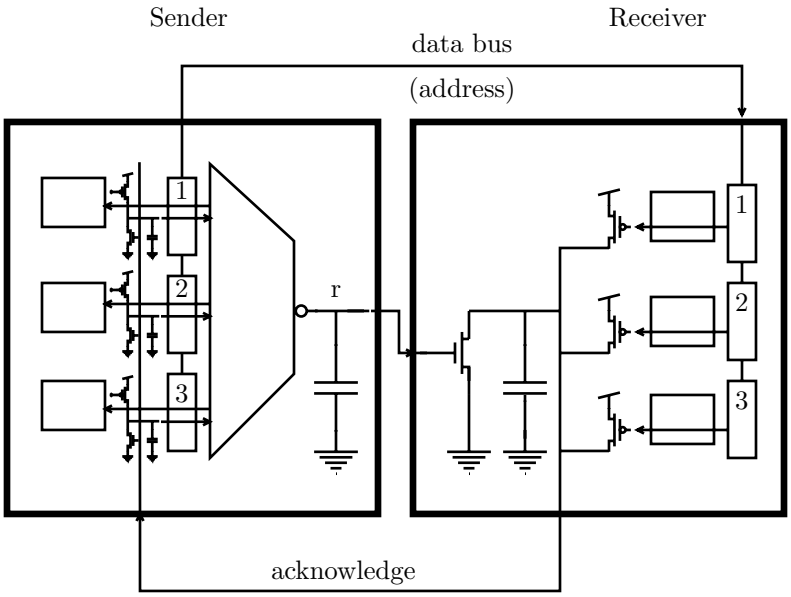

Figure 3.7: A one-dimensional data transfer system. The select signals coming from the Arbiter are depicted as coming from the left side of the Arbiter amplifier. The Request that the arbiter transmits to the Receiver is labeled r.

selected. Since the address lines may take different amounts of time to stabilize and, in the process, take on spurious valid addresses, a DATA VALID signal has been incorporated into the address encoder. The DATA VALID signal is another address bit whose settling time is manipulated by making its pull-down current, through PD1, stronger than that of the other bits, PD2. The stronger pull down makes the DATA VALID bit to go high more slowly than the address bits and also makes it go low more quickly than the address bits. When the DATA VALID line is high, the address should have stabilized. The encoder is redundant for this application since the bits that are low in the address are already pulled down.

The Receiver is generalized so that the Acknowledge signal can be pulled up by the receipt of *any* valid address. The address is decoded by a circuit shown in Figure 3.9. This example shows the decoding of the address 01, corresponding to the encoder in Figure 3.8. The decoded address pulls up directly on the Acknowledge node. The Acknowledge node is a wired-OR structure. The Acknowledge signal returns to the Sender and marks the reset phase transfer process.

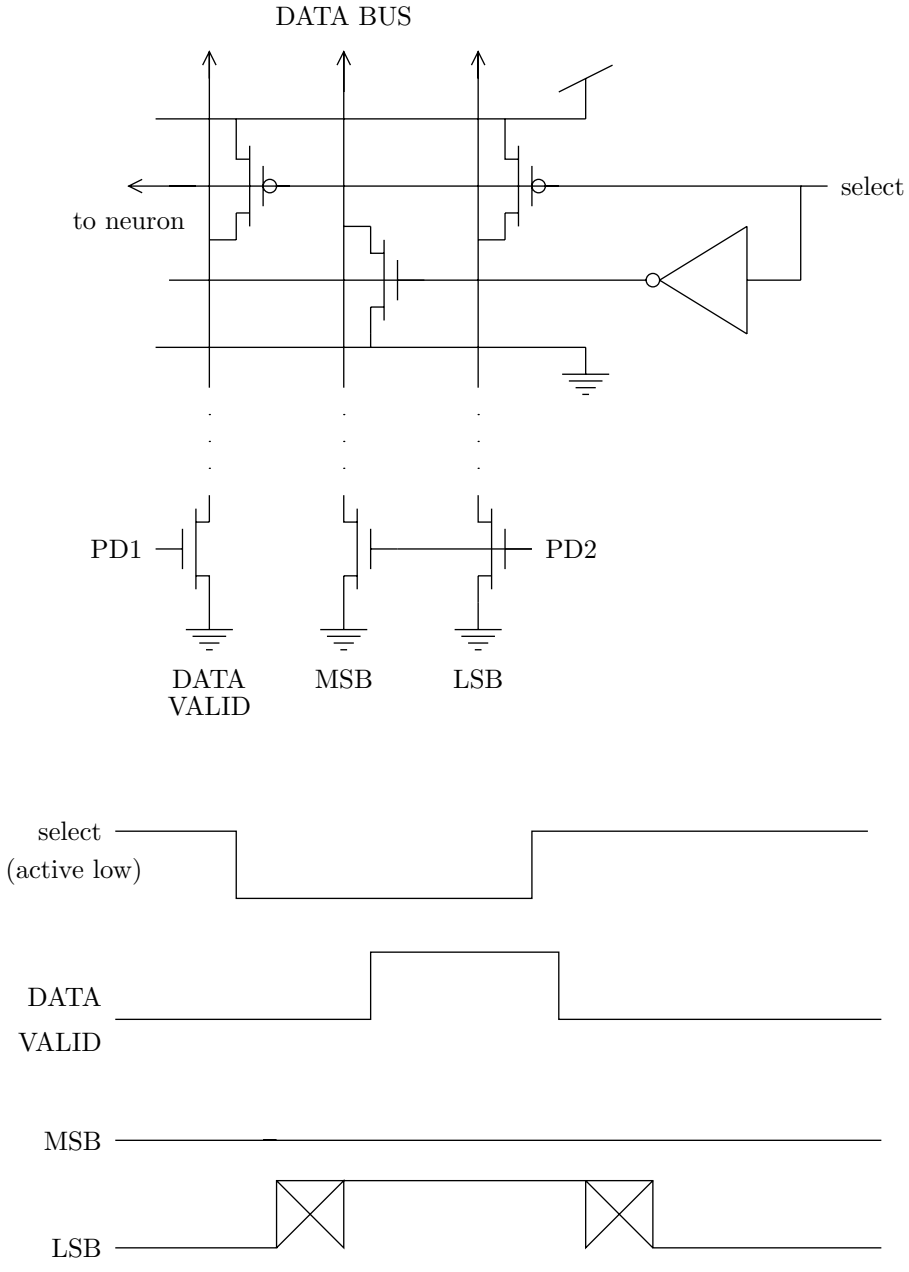

Figure 3.8: An address encoder for the address 01 with a DATA VALID bit. The DATA VALID pull-down is biased by a DC voltage PD1 and the pull-downs on the address bits are biased by a DC voltage PD2. The expected time course of the DATA VALID signal relative to that of the address bits is shown below.

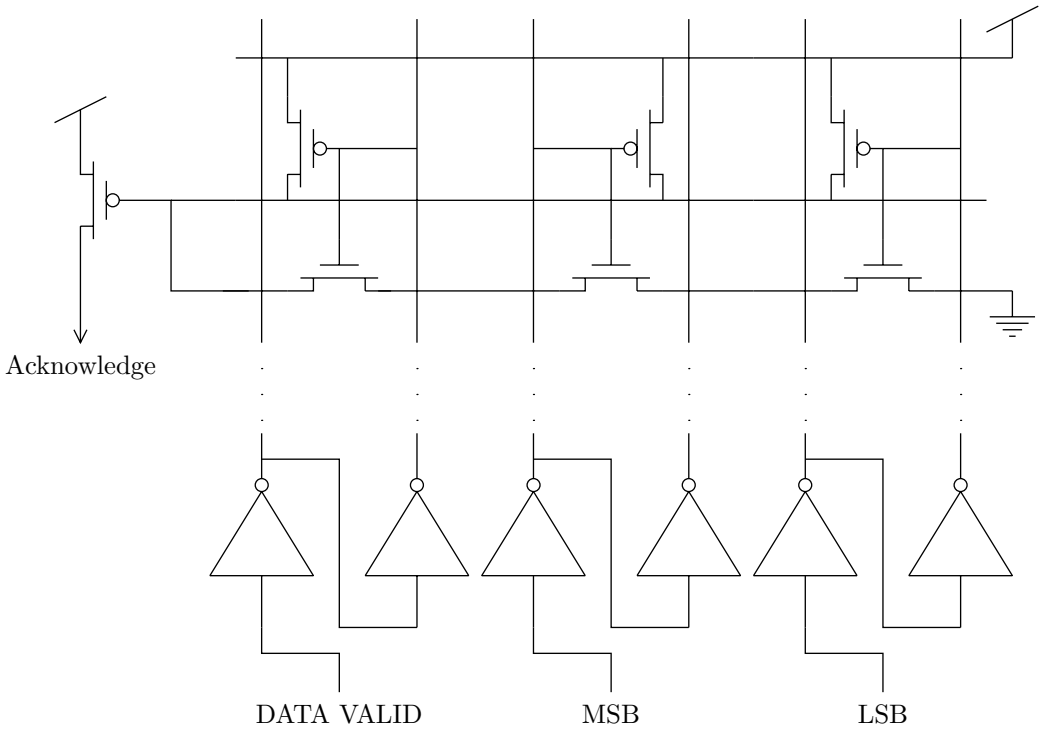

Figure 3.9: An address decoder for the address 01 with an additional DATA VALID bit. The decoded address pulls up directly on the Acknowledge node.

In a one-dimensional system, the Acknowledge need only reset the initiation node of the neuron whose data were transferred. The reset transistor, NA, shown in Figure 3.6, may be put in series with a gating transistor that allows current to flow only when the neuron is selected. This implementation is particularly space efficient because part of the data processing of the neuron can be incorporated in the data transfer machinery. The initiation node is analogous to the membrane capacitance of a biological neuron. The membrane is hyperpolarized by the delayed-rectifier potassium current even though the synaptic input (the current through the PP transistor in Figure 3.6) is still flowing. The discharge of the initiation node terminates the data event from that neuron. This implementation is also temporally efficient because events from other neurons that occurred during this data transfer period and have propagated some distance into the Arbiter can be selected with minimum delay because they are not reset. However, some consideration must be given to the method of distinguishing individual events, since there will be no reset signal from the Arbiter to remove the Acknowledge between events. I have not implemented such a system.

The implementation described in this chapter is conservative. The entire state of the system is reset at the end of each data transfer cycle. In this implementation, the Acknowledge signal returns to the Sender and resets all of the Initiation nodes. This protocol is necessary for generalization to arbitration in two dimensions. When all of the Initiation nodes have been reset, the Arbiter reinitializes itself. Upon reinitialization, the select signal is terminated and the Request signal from the Arbiter goes high, indicating that there is no data transfer in progress. When the select signal is terminated, the data are removed from the bus. Although the Acknowledge is no longer pulled up by PR, it will remain high until it is pulled down by the withdrawal of the Request. The Arbiter ensures that the Request signal will not be withdrawn before the select is terminated. When the Request signal is withdrawn, the state of the entire Arbiter has been initialized. At this point, the data transfer cycle is completed, the Acknowledge goes low, and the Initiation nodes can once again be activated by the neurons.

Because the Acknowledge must reset all of the initiation nodes, a problem arises that was not evident when only a single neuron generated events. The problem is that the system must keep track of which neuron has succeeded in broadcasting its address in such a way that it does not send the same data more than once and that it does not erase any data that

are still waiting their turn to be transmitted. This problem has several possible solutions. The one described here is suited for extension to two-dimensional arbitration.

The problem of deciding who has transmitted data is solved in this system by creating an additional state variable inside the neuron. This state variable is reset only when the neuron is selected and has presumably transmitted its data. Although all of the Initiation nodes must be discharged by the Acknowledge signal from the Receiver in order to complete one cycle of data transfer, the neurons that have not been selected remember that they would still like to transmit their addresses. Their data are not erased by the data transfer cycle.

The internal state variable of the neuron must be regulated in such a way that one and only one event is transmitted during the data transfer process. The event which initiated the data transfer process must be terminated by the time the process is complete, and no new events may be generated before the process is completed. The mechanisms by which these conditions are enforced are depicted in Figure 3.10. The select signal going back to the neuron from the Arbiter activates these mechanisms.

To ensure that the state of the pixel has been reset before another data transfer cycle is initiated, the Acknowledge signal resets the Initiation nodes in one of two ways, depending on whether the pixel is selected or not. The select signal is active low. If the pixel has not been selected, the Acknowledge signal is able to pull down on the Initiation node through transistor NA and forcibly reset the Initiation node. If the pixel is selected, the pull down transistor limits the current that the Acknowledge signal can apply through transistor NA2. The Initiation node associated with the selected pixel will not be discharged until the pixel itself has removed its data from PP. This mechanism is similar to one that is seen in real neurons and has to do with the strength of the potassium current. A neuron cannot fire a second action potential unless it has been hyperpolarized sufficiently to reactivate its sodium channels [9]. This feature has been used to advantage by the amacrine cells of the retina, which have a potassium current that turns off before the sodium inactivation is released [1]. These cells generate a single spike in response to a persistent bipolar cell input. By making the potassium current sufficiently weak, the cell is prevented from generating another event until the depolarizing current into the cell is sufficiently reduced that it hyperpolarizes enough to reactive its sodium channels. Unfortunately, in this multiplexing system, waiting for a single neuron to hyperpolarize means that a single recalcitrant neuron can hold up

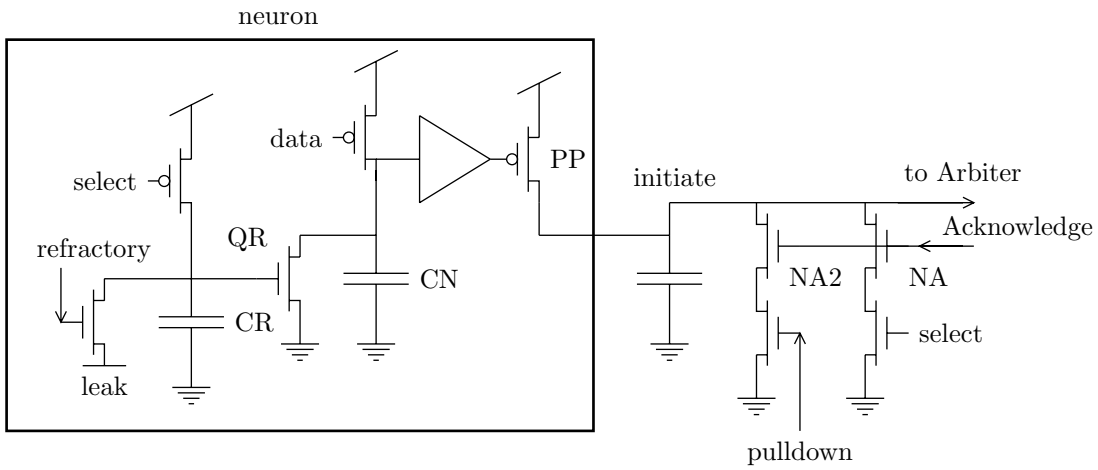

Figure 3.10: Mechanisms for resetting data transfer. The internal state variable of the neuron is the voltage on capacitor CN. This state variable integrates the input data with a time constant set by CN and the leak voltage. The state variable CN is put through a non-linear threshold and the output of that threshold function drives the initiation node of the Arbiter. The neuron contains a circuit to reset CN and make it refractory if it is selected by the Arbiter. The duration of the refractory period is set by the size of capacitor CR and the magnitude of the refractory control voltage. The reset of the initiation node proceeds independently of that of CN. Once activated, the initiation node remains activated until the Acknowledge is returned. The initiation node is reset through transistor NA if this neuron is not selected. If the neuron is selected, reset is accomplished by transistor NA2.

data transfer from the whole array. Although this waiting mechanisms has been included in the design, it is probably not really necessary, since the mechanism for resetting the neuron's internal state variable is rapid and can be made arbitrarily strong.

The resetting of the neuron's internal state variable, CN, is mediated by transistor QR whose gate is connected to capacitor CR. Capacitor CR is charged when the neuron is selected. Since the select is a rapid digital-amplitude signal, CR is quickly charged up so that the current through QR is larger than the current through the data transistor and the voltage on CN drops below the inverter threshold. When CN has been reset, the neuron turns off transistor PP. However, the select signal is not removed until the Acknowledge signal resets the Initiation node. The selected pixel cannot initiate another event until the current through QR has become smaller than the data current so that CN can be charged up past the inverter threshold. The reset variable, CR, provides an opportunity to create a refractory period for the neuron. If the charge leaks off of CR slowly, the neuron will be unable to charge CN above threshold for some time after the select has been withdrawn. The current through QR is similar to the delayed rectifier potassium current of the biological neuron. It limits the maximum spike rate of the cell. The refractory period allows arbitration between coincident events to proceed more effectively than if neurons were allowed to fire at arbitrarily high firing rates.

When the data transfer cycle is completed, competition for the bus begins again, as if all nodes were requesting for the first time. Because the pixels that were not selected have not had their states reset, their PP transistors have remained on. When the Acknowledge goes low, their Initiation nodes will go high. Arbitration in this system is not *fair* since no attempt is made to keep a list of who has initiated data transfer previously and in what order. Making the refractory period of the neuron long prevents it from reengaging in competition with the neurons whose data events have not been transmitted. The maximum desirable refractory period considering multiplexing constraints alone is one that will allow all of the events that could possibly occur simultaneously to be transferred in rapid succession, before a new event is generated. All of the neurons are able to send all of their data if the refractory period of a neuron is longer than the number of neurons sharing the bus multiplied by the data transfer period. Addresses must be transferred faster than the the maximum event frequency of the neuron multiplied by the number of neurons in order to

guarantee that all the events will be transferred. However, if the system is operating in the intended regime, in which the number of events to be transmitted is sufficiently small, the length of the refractory period should be set equal to the number of anticipated coincident events. A refractory period of 2 milliseconds, which is a biologically plausible time, would be sufficient to transmit about 1000 effectively synchronous events, before a neuron that had already had a turn could get back into the queue.

3.4.3 Arbiter

The Arbiter itself is central to the success of the address-event protocol. It selects one of many requests for transmission by using a high gain positive feedback element to resolve contention. The Arbiter was designed and the basic circuit element analyzed by Mass Sivilotti [27]. The arbiter described here was slightly modified for more robust behavior.

Binary Tree

The Arbiter was designed to scale well, in terms of both area and speed, as the size of the

pixel array is increased. The basic one-dimensional Arbiter is a binary tree of simple arbiter elements as shown in Figure 3.11. For a linear array of size N, the total number of Arbiter elements required is N-1. The entire Arbiter thus occupies only a thin strip along the edge of the array. I have implemented a silicon compiler written in WOLCOMP [25], which is described in Appendix A. The compiler to automatically and reliably construct Arbiters for any sized array from a library of cell types included in the WOLCOMP module. Each Arbiter element receives two input request lines from the lower level and sends a single request line to the next level of the tree. Each element receives a single select line from above and sends two select lines to the lower level. The job of each element is to choose one of the two incoming request signals, and to pass along the select from above to the chosen request. If the select is not received from above, then neither of the incoming requests is selected. Starting from a completely initialized state, the time required to complete the arbitration is determined by the amount of time required for a request to propagate to the top level of the tree and for the select to propagate back down. Arbitration occurs in

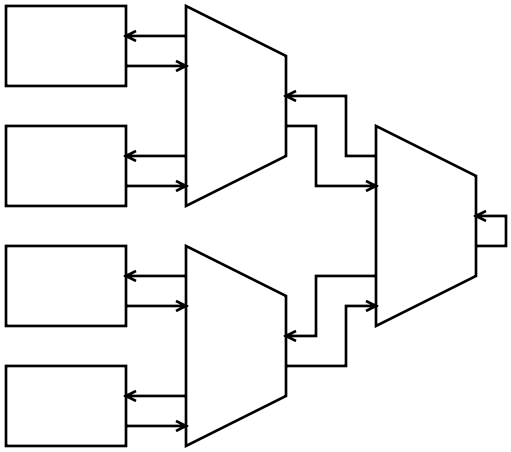

Figure 3.11: The Arbiter is a binary tree of two-input arbitration cells. Each cell receives two requests from below and an Acknowledge from above. It transmits a request to the higher level of the tree and passes down two Acknowledge signals.

parallel at each level of the tree, so the delay through the Arbiter increases only slowly with the size of the array. The total delay through the Arbiter is proportional to log(N).

Circuit

A circuit schematic of the Arbiter element is provided in Figure 3.12. The circuit is composed of three parts. The first part is an OR gate that transmits a request signal to the next level of the tree if either incoming request is activated. The second circuit chooses one of the two incoming requests. This circuit is composed of two cross-coupled NAND gates. The cross-coupled element ensures that only one request will be chosen even if both requests are active. The incoming requests are labeled R_1 and R_2. The lines indicating which request has been chosen are labelled R_1' and R_2'. Unlike the request lines, the choose variables are active when they are at a low voltage. If R_1' is low, it indicates that R_1 has been chosen by this Arbiter element. There are eight possible incoming signal states, listed in Table 3.1. The third circuit directs the select signal coming from the next level of the tree to the descending select output corresponding to the chosen request. This circuit acts as a differential amplifier whose power is turned on by the incoming select. The chosen variable that is in the more active state will drive the corresponding select signal high.

The interaction between the choosing circuit and the select steering circuit is the crux of safe arbitration. The problem is to prevent a select from propagating down the tree before

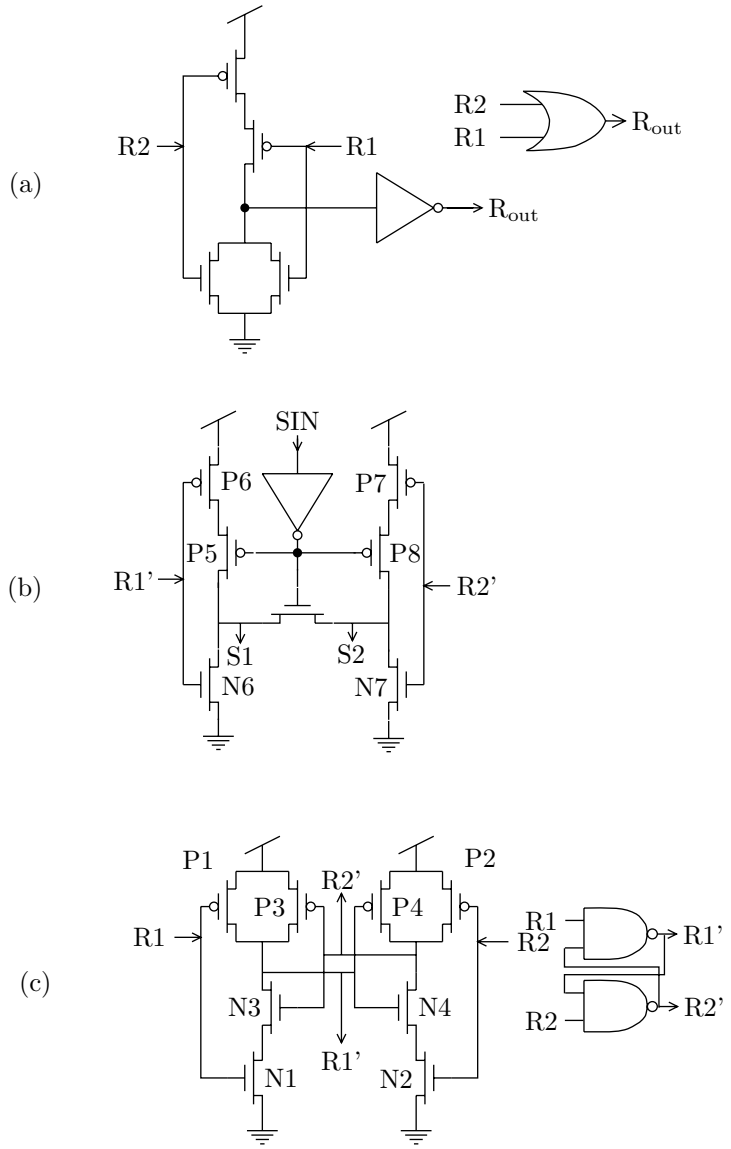

Figure 3.12: Schematic of two-input Arbiter element. (a): the request-generating circuit. (b): the choosing circuit. (c): the steering circuit.

R_1	R_2	S_{IN}	R'_1	R'_2	S_1	S_2	R_{OUT}
0	0	0	1	1	0	0	0
0	1	0	1	0	0	0	1
1	0	0	0	1	0	0	1
1	1	0	0	1	0	0	1
1	1	0	1	0	0	0	1
0	0	1	1	1	0	0	0
0	1	1	1	0	0	1	1
1	0	1	0	1	1	0	1
1	1	1	0	1	1	0	1
1	1	1	1	0	0	1	1

Table 3.1: Truth table for a single arbitration element in the Arbiter. Input parameters are: incoming request from below R_1, incoming request from below R_2, incoming select from above S_{IN}. Output parameters are: outgoing request to above R_{OUT}, outgoing select to below S_1, and outgoing select to below S_2. The intermediate results indicating which of the incoming requests have been chosen are: R'_1 and R'_2. These two signals are active low. The table is divided in half for convenience; all of the states in which this Arbiter cell has not been selected from the higher level of the tree, and which therefore have no active outgoing select signals, are shown in the top half of the table.

a clear choice has been made. Since the OR gate that issues a request to the higher level of the tree can do so while the choosing circuit is hung in a metastable (and undecided) state, it is possible that the select could be issued before the choice has been made. The select can be kept from propagating down the tree if the choice lines do not cross threshold while the choice circuit is in its metastable state. Even when SIN is active, the outputs S1 and S2 must be low when R'1 is equal to R'2. (R'1 equal to R'2 is the metastable state of the cross-coupled nand gates if they have identical geometries.) This condition can be met by making N6/N7 strong relative to P6/P7 and/or P3/P4 wide relative to N3/N4. Using conservative estimates, Sivilotti [27] calculated that safe arbitration could be achieved if the P3/P4 transistors were six times stronger than N3/N4. This ratio is satisfied by the current Arbiter.

In the forward phase of the data transfer cycle, the requests propagate from the lowest level of the tree to the top. At the top level of the tree, the outgoing request is tied to the incoming select. This signal is the request that goes to the receiver chip. When the select propagates back to the bottom level of the tree, the selected neuron address is placed on the data bus. In the reset phase of the data transfer cycle, the neuron Initiation nodes are reset at the lowest level of the tree by the Acknowledge from the receiver. When both of the requests coming into an Arbiter leaf cell are off, the select signal does not pass through that leaf cell. Therefore, the select to the pixel is inactivated before the state of the whole Arbiter has been reset. In the communications protocol that I have implemented, the request to the receiver is terminated only when the reset of the requests has propagated to the top of the tree. If the Acknowledge from the receiver remains active until the request has been terminated, the state of the system is fully reset at the end of a data transfer cycle.

3.5 Data Transfer in Two Dimensions

The example system is a 64x64 pixel retina that uses the address-event representation to copy its image onto a receiving chip. The data transmission protocol for this system is complicated by the fact that the retina is a two-dimensional structure. The complexity arises because of geometrical constraints in implementation of the circuit. The multiplexing machinery is best kept to a small area at the edge of the data processing array. Not only

does this arrangement save area, but the delicate analog machinery responsible for light transduction within each pixel is best kept as well isolated as possible from the fast digital signals involved in multiplexing. The consequence of restricting the multiplexing machinery to the periphery of the chip is that each pixel is specified by an x- y-coordinate address.

This encoding system has relatively little impact on the receiver chip, depicted in Figure 3.13. The core of the receiver chip is a 64x64 square array of nodes. The circuitry at each node is shown in Figure 3.17 and will be described in the next section. Each node on the receiver is driven by the pixel in the corresponding position in the sender array. The address-events are decoded into a position by a set of digital decoders located on two edges of the array. Input to the node requires that the decode line in the x-dimension and the decode line in the y-dimension be activated by the proper address. The ANDing of the address coordinates in the two dimensions is a straightforward extension of the decoding process described in the one-dimensional system. An additional modification for a two-dimensional receiver is that the pull up of the Acknowledge of this system must be aggregated in two dimensions, as shown in Figure 3.13. The coincidence of activation on the x- and y-decode lines pulls down a line that runs along that column. The column lines correspond to the individual node pull-ups in the one-dimensional system. In this way, if any of the nodes in the array is activated, the Acknowledge is pulled-up to indicate that the address-event has been received.

The generalization of the data transfer protocol is more difficult for the two-dimensional sender. The selection of the pixel which will transmit its address must be coordinated in the two dimensions. If there were two contending pixels, (x_1, y_1) and (x_2, y_2), and the arbitration in the two dimensions were allowed to proceed independently, two ghost events at (x_1, y_2) and (x_2, y_1) might be transmitted. In order to avoid this problem, arbitration in the two dimensions proceeds sequentially.

The sender is illustrated in Figure 3.14. The core of the chip is a 64 x 64 array of pixel elements. One pixel is depicted in Figure 3.20. The circuitry of the pixel will be described in detail in the next section. The portion of the circuit involved in data transfer is identical to that illustrated in Figure 3.10. Two sides of the array are occupied by the sequential analog multiplexors for video display, which have been described previously [26]. The remaining two sides of the array are occupied by the data transfer mechanism.

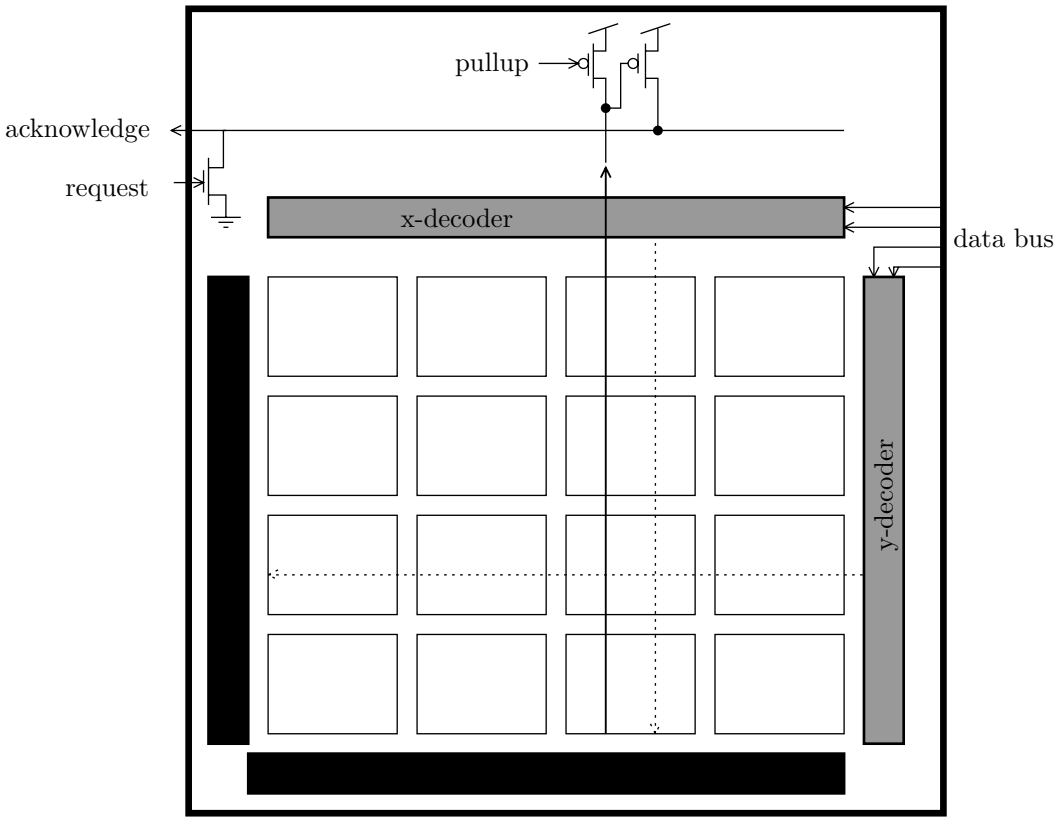

Figure 3.13: Schematic of receiver. The address is decoded independently in the x- and y-dimensions. When the address has been successfully decoded, the Acknowledge signals from all the pixels are aggregated by a wire OR structure, first along columns and then along rows. Because only one address can appear on the data bus, only one node will be pulling on the wire OR at any time (see Figure 3.17).

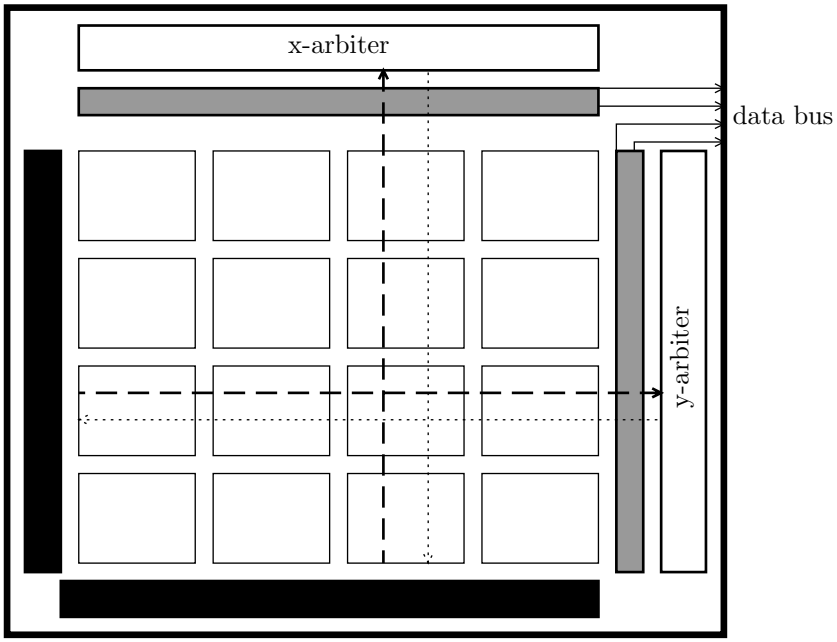

Figure 3.14: The sending chip contains an array of pixels surrounded by multiplexing circuitry; the Arbiter, two white boxes, which decides which pixel has control of the data bus at each instant; two gray boxes adjacent to the Arbiter, which include the address encoders and circuitry involved in coordinating the data transfer process between the two chips; and analog scanning circuitry, depicted as two black boxes along the remaining two sides of the chip.

The data transfer process is initiated by a pixel. The initiation process is sequential, occurring first in the vertical, then in the horizontal dimension. When the data processing circuitry inside a pixel decides that it would like to transmit an event, it pulls up on the initiation line which runs the length of the row. If that row is selected by the vertical Arbiter, the select signal on that row is activated and the y-dimension address bits of that row are placed on the bus. The row select allows all of the pixels along that row to pull up on initiation lines running the length of the columns. In the second stage of the initiation cycle, the horizontal Arbiter selects an initiating pixel on the row that was just selected by the vertical Arbiter and activates the appropriate column select line. This places the x-dimension address bits on the bus. The completed address can then be decoded by the receiver.

The two-dimensional data transfer protocol requires some modification from the one-dimensional case. The pixel must have an internal state variable and threshold, as described in Figure 3.10. There are several reasons for this additional state variable. For example, the initiation process is asymmetrical in the two dimensions. The initiation lines in both dimensions have one pull-up transistor for each pixel. Because the effects of the pull-up transistors sum, it is possible for several pixels on a *row* in combination to bring the row initiation line above threshold. However, since only one pixel per column is enabled by the row select, only one pixel may pull up on the column initiation line of the horizontal Arbiter. If the pixel outputs are small analog values, they may sum to initiate an event on the row but none of them individually may be able to bring the column line above threshold. Therefore, the pixel must have an internal threshold amplifier with enough gain to ensure that it is either fully on or off. This internal state variable provides a mechanism for generating a refractory period for the pixel once it has been selected. As in the one-dimensional case, the state of the selected pixel is reset, this time by the AND of a row and column select signal.

There are more possible reset protocols for the Arbiter in the two-dimensional system than there were in the one-dimensional system. I have chosen to implement an extremely conservative protocol, which resets the state of the entire system, including all of the intermediate nodes in both the horizontal and vertical Arbiter trees, after each data transfer cycle. However, more temporally efficient mechanisms are possible. I will describe two such

hypothetical protocols before describing what I actually implemented. One hypothetical protocol would not reset the selected row of the vertical (row-selecting) arbiter until all of the neurons making column requests had transmitted their data. This sequence is necessary so that the proper x- and y-addresses remain associated. Only the selected column and row initiation nodes would be reset, and they would be reset with weak NA2 transistors so that the neurons would have to have been transmitted before their initiation nodes could be reset. This protocol has the disadvantage that one row might control the bus indefinitely if it had a persistently active pixel on it.

An alternative Arbiter reset protocol, suggested by Alain Martin (personal communication) entails resetting the entire horizontal (column-selecting) Arbiter and resetting only the selected row. The vertical Arbiter would be forced to choose a new row and the initiation nodes of the horizontal Arbiter would be reset so that the new row could enter into fresh competition. The address stream would be punctuated by the reset of the horizontal Arbiter, which would toggle the request to the Receiver chip. The selected vertical Arbiter initiation node could be reset by the Acknowledge signal, which would also reset all of the horizontal Arbiter initiation nodes. If necessary, the method of resetting the initiation nodes used in the one-dimensional case could be applied to the reset of the horizontal Arbiter because the selected row is essentially a one-dimensional system. This reset mechanism would be faster than the one that I implemented because the partial state of the vertical Arbiter tree would be conserved. In light of my present experience, this protocol appears to be preferable to the one that I have implemented, which is described next.

In the implemented system, all of the initiation nodes of the the vertical Arbiter are forcefully reset by the AND of the horizontal Arbiter top-level request, indicating that all of the column-initiation lines have been reset, and the Acknowledge. It is not necessary, nor is it possible, to determine at this point whether or not the internal state of the selected pixel has been reset. This determination is made previously in the reset protocol by the horizontal Arbiter reset, as described in the one-dimensional case. When the vertical Arbiter has been reset, the withdrawal of the request pulls down the Acknowledge and completes the data transfer cycle. The reset of the initiation lines is terminated and the pixels are free to reinitiate requests at the base of the vertical Arbiter tree. A single complete data transfer cycle performed by the sender and the receiver chips is shown in Figure 3.15.

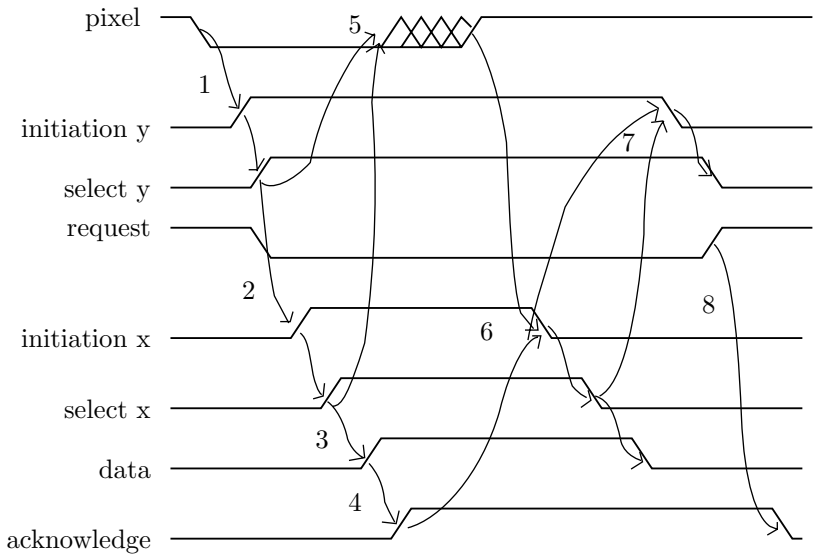

Figure 3.15: Timing diagram for data transfer between sender and receiver.

3.6 Image Transfer

The system used to demonstrate the address-event protocol transfers a time-derivative image from a transmitting retina to a receiver. The circuits on the sender that encode the image and the circuits on the receiver that reconstruct the image are described here. A node in the receiver array is illustrated in Figure 3.17. The node that is tiled to make the receiver array contains both multiplexing circuits and the actual integrator that reconstructs the address-event stream into an analog potential. The address is decoded independently in the x- and y-dimensions and ANDed inside the pixel. When both the x- and y-decode lines are active, the Acknowledge aggregation line is pulled low through transistors Ax and Ay. In addition, some current flows onto the state capacitor C through transistors Sx and Sy. The generation of an Acknowledge in response to the decoding of an address-event and the placing of an increment of charge on the state capacitor is shown in Figure 3.18. The magnitude of the current is controlled by *delta* and the total amount of charge is a function of the length of time the address data are valid. Current leaks from the capacitor at a rate set by *tau*. The data from the capacitors are scanned out serially for display on a video monitor.

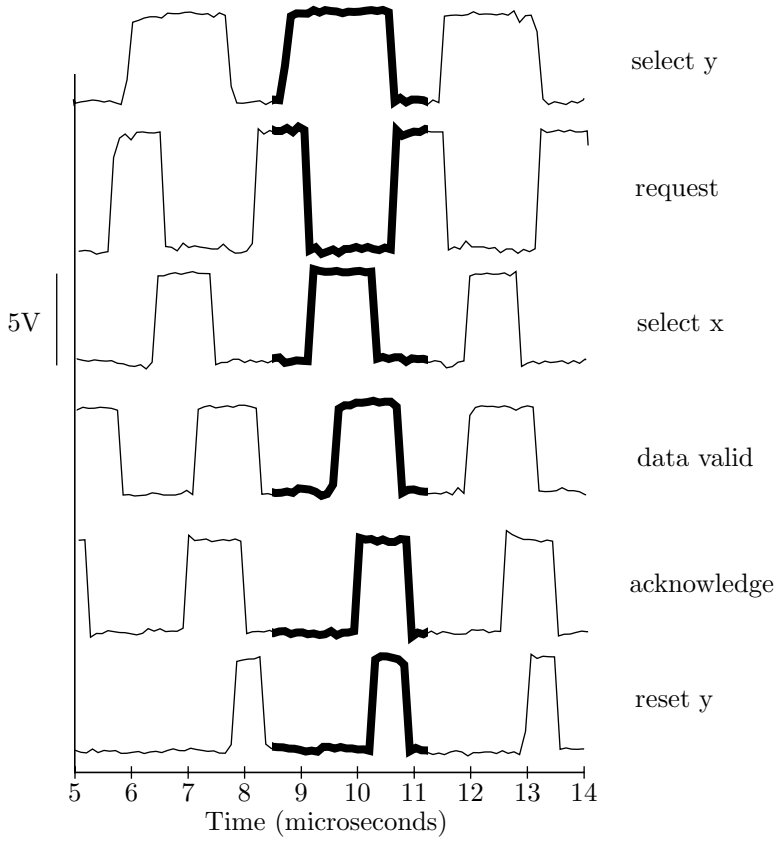

Figure 3.16: Data transfer between sender and receiver. Data taken from real system. [Not all the signals shown in the timing diagram (Figure 3.15) are instrumented on the chips]. Data were collected with a digital scope triggering off of the falling edge of the x-select signal and are synchronized to that. The chip was configured in such a way that all the neurons were firing at a high rate, in order to measure the minimum data transfer period. The minimum period was approximately $2 \cdot 10^{-6}$ seconds.

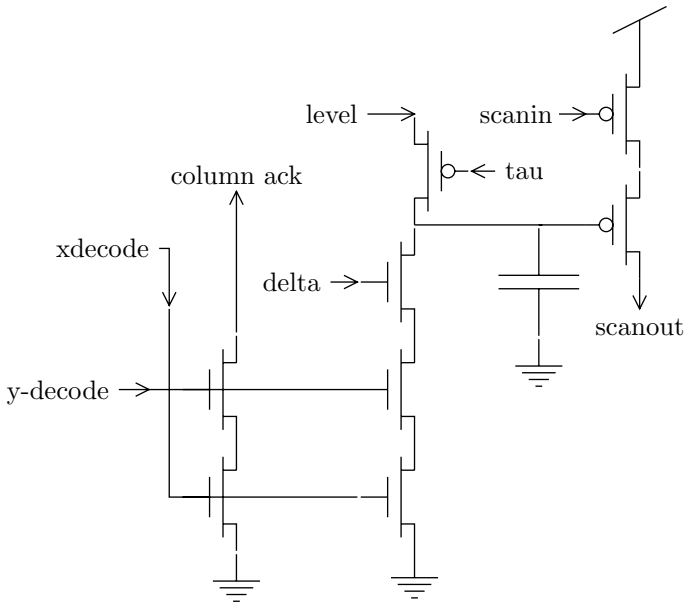

Figure 3.17: Schematic of single node in the receiver array.

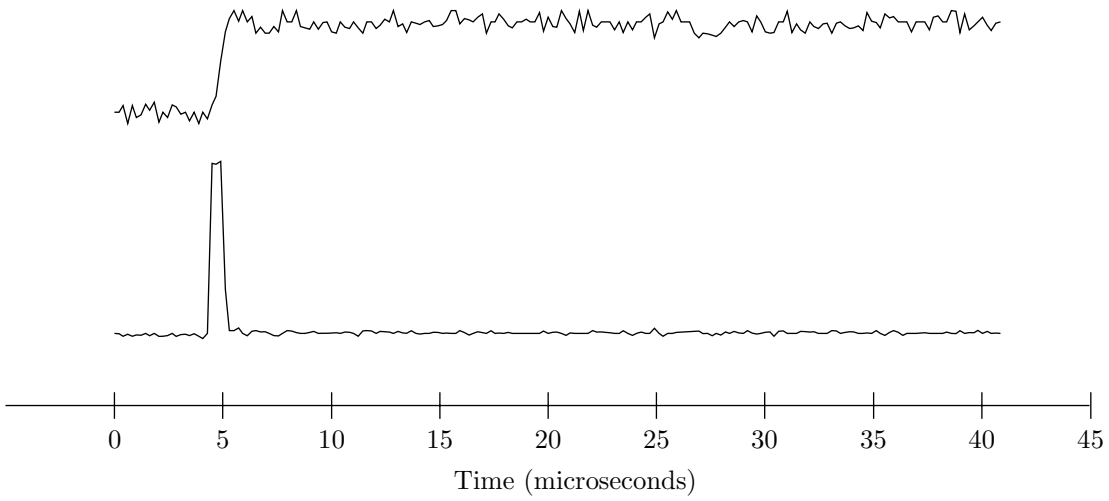

Figure 3.18: The receiving element generates a step in potential in response the arriving address-event. The voltage step size was arbitrarily scaled by the off-chip current-sensing amplifier. The bottom trace is the Acknowledge signal that terminates the data transfer. The Acknowledge signal is 5 volts in amplitude and 1 microsecond in duration.

The behavior of the receiving pixel is illustrated in Figure 3.19. Synthetic data in the form of a temporal stream of digital addresses were generated on an HP 9836C workstation and transmitted to the chip with a custom hardware interface board. The time constant of integration on the receiver determines the time over which spikes can be averaged. This time constant of integration, controlled by *tau*, can be varied over several orders of magnitude. A long integration time is advantageous for integrating a small signal that is contaminated with sporadic random noise. A short integration time increases the temporal resolution of the system and a simple threshold is able to detect spike coincidence to within the integration time of the integrator.

This receiving pixel should be modified to include a leak whose magnitude is a function of the voltage level of the integrator. This feature would allow a stable translation of event frequency into analog voltage level. Additional circuitry may be necessary to control better the quantity of input charge for each event. In the existing design, the event duration is linearly related to the amount of charge that is deposited on the integration capacitor for a given event. Events of different duration will result in different amounts of current being integrated on the capacitor. One solution to this problem would be to put a timing element in each pixel that regenerated a long-duration spike, triggered by the event. If such a long-duration spike mechanism were incorporated the fractional variation in event width would be caused by transistor mismatch on the receiving chip rather than transmission variability. A longer spike would presumably have less fractional duration variation. However, this solution requires more area. Event durations do not appear to vary by more than 50 percent. Until it can be demonstrated that there is a significant impact on the computation, there is no reason to include such a mechanism.

The first step in image transfer is the creation of the image on the retina. The retinal pixel incorporated into the self-timed data transfer system generates events when the light level increases. In this way, it is similar to the on-transient retinal ganglion cell [7]. A schematic diagram of the pixel circuitry is shown in Figure 3.20. The drive to the spike-generating pixel generated by a circuit similar to that of the feedforward retina described in the previous chapter, but the resistors have been omitted to reduce the size of the pixel. The drive circuit averages the output of the logarithmic photoreceptor with a follower-integrator whose time constant is controlled by *tau*. The output of the drive circuit is a

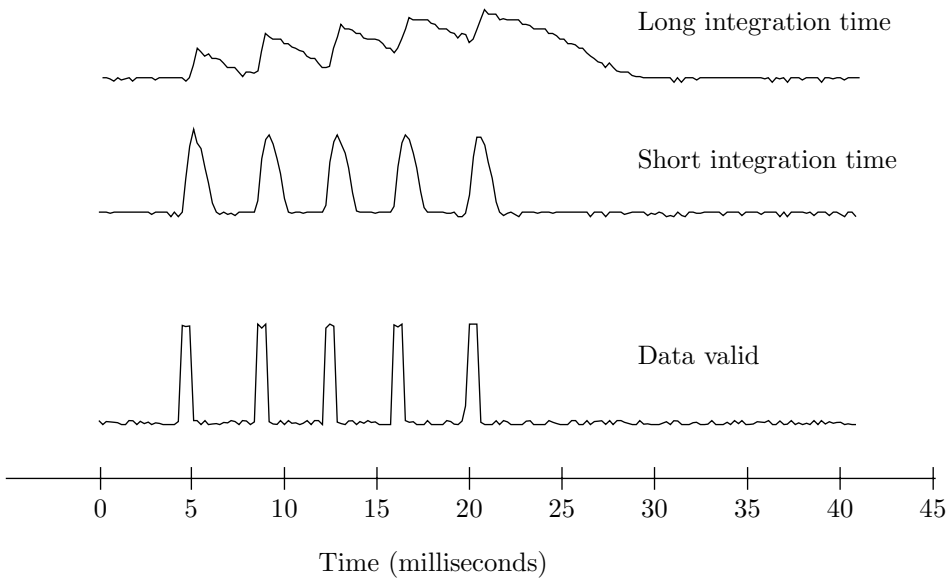

Figure 3.19: The output of a single pixel on the receiver chip shown for two different integration times. Address-events were generated with a custom hardware interface board and an HP9836C computer for the receiver pixel accessed by the serial scanner. The integration time of the pixel was modified by changing the bias voltages on the tau and delta controls. Level was set to 4.046 volts. For the fast integration time trace, delta was set to 0.67 volts and tau was set to 2.756 volts. For the slow integration time trace, delta was set to 0.62 volts, and tau was set to 2.85 volts. See Figure 3.17.

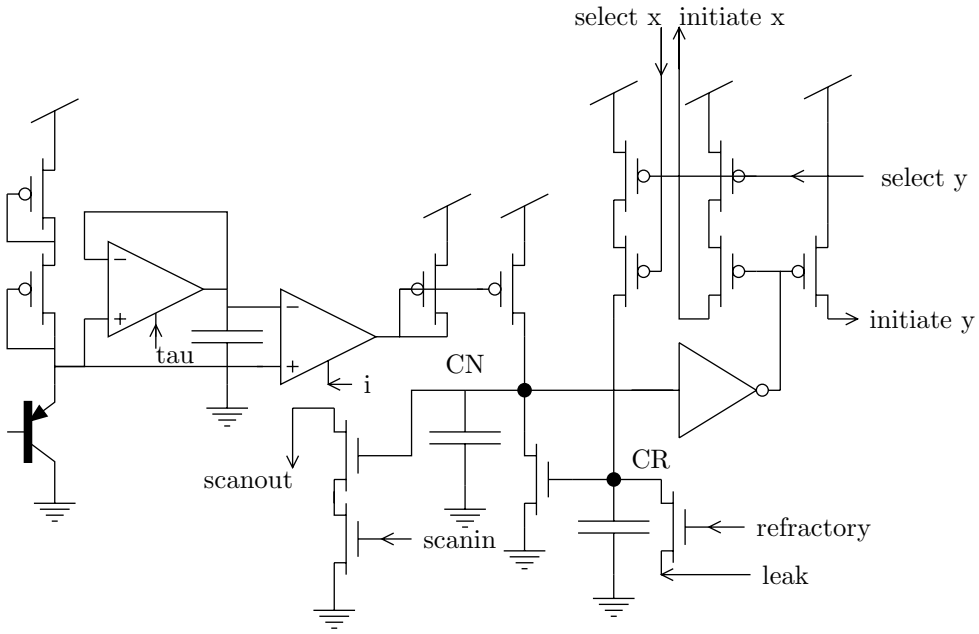

Figure 3.20: Schematic of a single pixel in the sender array.

current proportional difference between the average intensity and the instantaneous values of intensity, which is scaled by the control voltage i. This system is analogous to the bipolar cell of the outerplexiform layer of the retina. The spike-generation circuitry of the pixel is like that shown in Figure 3.10, except that data transfer sequence takes place in two dimensions. The primary state variable, CN, is like the membrane capacitance of a retinal ganglion cell. This capacitor is a leaky integrator with a time-constant set by the leak parameter, which determines the quiescent voltage on CR. Capacitor CN integrates the charge supplied by the drive circuitry until its voltage reaches the inverter threshold. The inverter initiates the data transfer process. Once the pixel is selected in both the x- and y-dimensions, capacitor CN is discharged.

The parameter settings of the pixel affect the number of spikes that it generates in response to a particular stimulus. The responses of a pixel to a flashing LED for several different settings of the time-constant of the differentiator are shown in Figure 3.21. If the follower integrator is able to follow the stimulus intensity more quickly, less current is produced by the differencing amplifier and so fewer spikes are produced.

Figure 3.22 shows the difference in response caused by the refractory period, which is analogous to the duration of the delayed-rectifier current, I_{K_D}, in biological neurons. Like

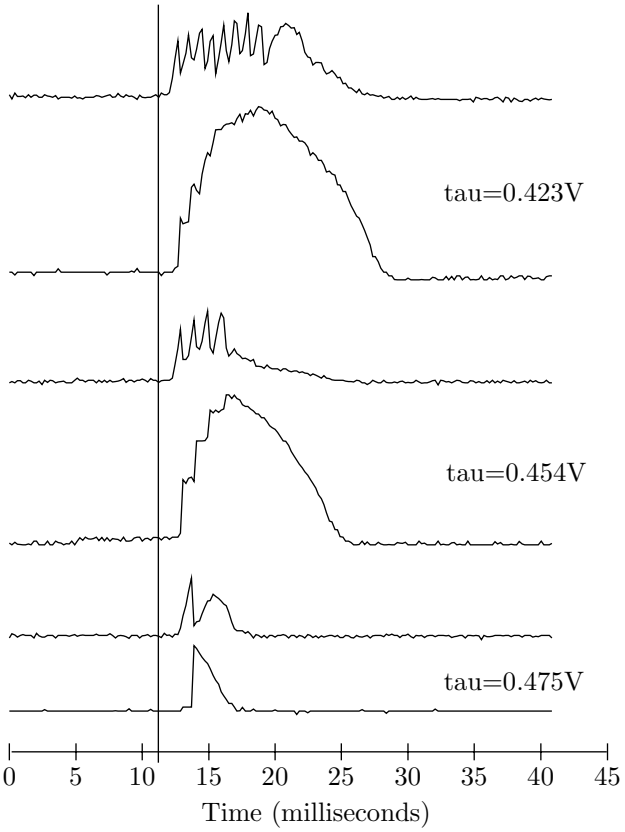

Figure 3.21: The response of the complete sender-receiver system to a flashing light-emitting diode (LED) of intensity 63.2 $\mathrm{mW/mm^2}$ **with three different time-constants for the differentiator. Stimulus onset is indicated by a vertical line. The output of the sending pixel and the corresponding node on the receiver are shown as a pair, the sender pixel waveform above the receiver response. Responses were averaged by the digital oscilloscope over eight stimulus presentations. The voltage,** *tau***, controlling the time constant of the differentiator in the sending pixel is shown next to each pair of responses. All other parameters were held constant.**

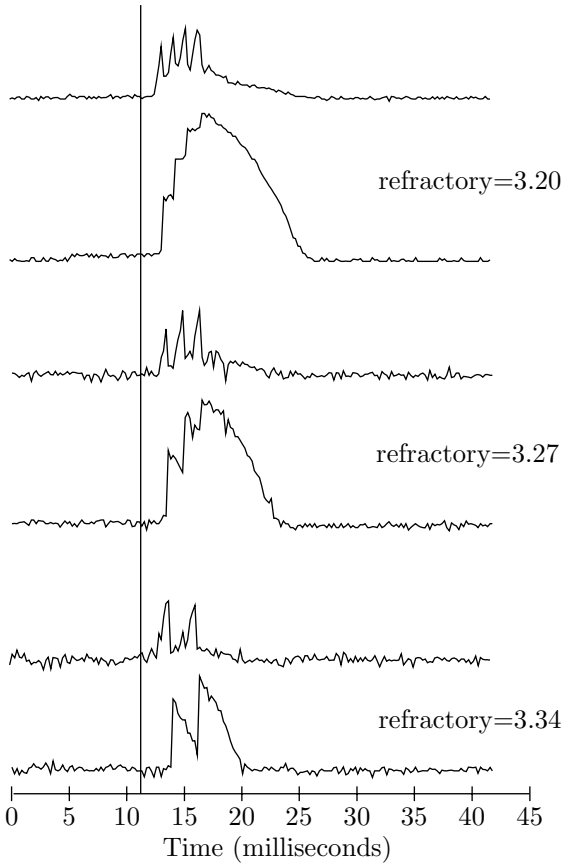

Figure 3.22: The output of a single pixel on the sender chip and the corresponding node on the receiving chip. Stimulus was a flashing LED of intensity 63.2 $\mathrm{mW/mm^2}$**. Stimulus onset is indicated by a vertical line. Values of the refractory transistor gate voltage are shown next to each pair of responses. As the refractory period decreases, the maximum event rate increases so the number of events per stimulus presentation increases.**

the $\mathrm{I_{K_D}}$ current, the reset current is sensitive to the voltage of the pixel. The reset current increases in amplitude until the voltage on capacitor CN is discharged below the inverter threshold. When the Acknowledge has reset either the x- or y-initiation node, the select signals that are contributing to the reset current are withdrawn. The magnitude of the reset current set by the voltage on capacitor CR decays at a rate set by the refractory control. When the reset current is smaller than the current from the differentiator, the pixel voltage begins to increase. When the reset current is of longer duration, the voltage on the pixel capacitor remains low longer. Fewer spikes are produced in response to the same stimulus and thus there is less activity in the receiving node.

The gain of the action-potential is generated in biological neurons by the positive feed-

back from the sodium spike channels. The sodium phase of the action potential is generated by the digital circuitry on the chip. Even with the gain of the inverter, the parameter i that scales the difference current driving the state capacitor must not be too small, or else the inverter will not cross threshold quickly enough for the data transfer process to proceed quickly. Several of the inverters in a row may be approaching their transitions and their effects sum to initiate a horizontal request. Once the row is selected, there is a delay until one of the inverters crosses threshold far enough to initiate data transfer in the column. This delay is apparent only when the current flowing into the state capacitor is very small. The gain problem may be ameliorated by incorporating positive feedback from the row select to the pixel. However, all of the pixels along the row would receive this positive feedback. Any such feedback mechanism should be capacitive, so that the feedback cannot be integrated by the initiation mechanism into an entirely new event. The magnitude of the feedback should be small enough not to bring all the pixels in the row past the inverter threshold.

The major drawback to this particular pixel is that it is not sufficiently sensitive with low-offset to make a practical imager using this communications protocol. The gain of this photoreceptor is low and the DC offsets are integrated by the pulse generation mechanism so that much of the bandwidth is occupied transmitting offset data. The data that were taken in this chapter were taken with the chip configured to have a large quiescent leak, which reduced DC offset problems. However, the stimulus needed to be high contrast enough to elicit an above-threshold response.

Figure 3.23 shows the response of the system to increasing intensity steps. The magnitude of the step in light intensity is encoded by the number of spikes generated. Event timing as well as total number of events carry information about the image since the latency of response is increased when the stimulus has lower contrast. A similar phenomenon is observed in biological visual systems. It forms the basis of the Pulfrich effect, a stereoscopic depth illusion. Placing a neutral density filter in front of one eye causes a delayed response to the stimulus from that eye. This delay is interpreted by the motion-interpolation processing in the cortex as a shift in the position of the target between the two eyes. This artificially induced disparity is indistinguishable from real depth. A pendulum bob swinging back and forth in a plane in front of the viewer is seen to move in a circle in depth.

The representation of temporal change is natural for the address-event representation;

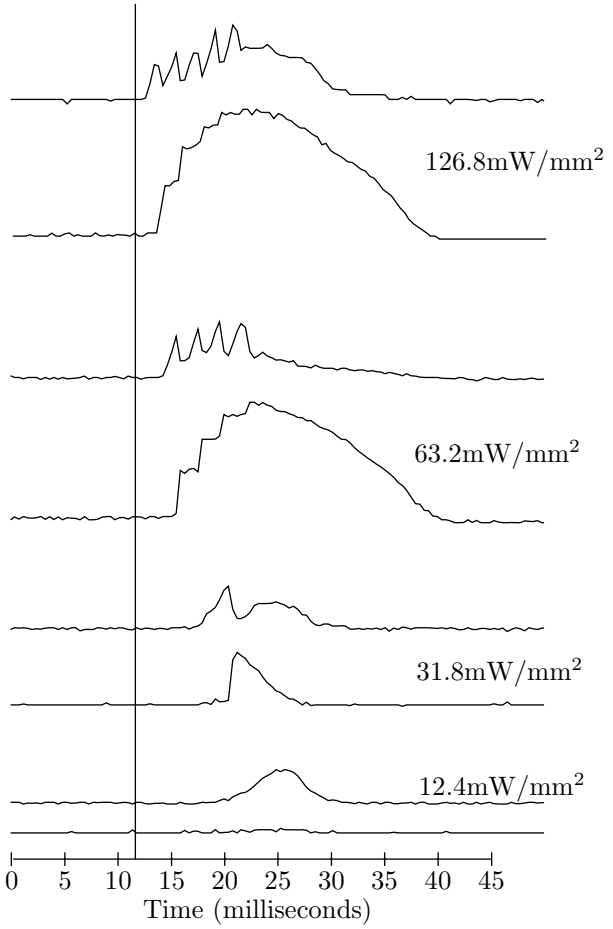

Figure 3.23: Analog data from a single sender pixel and the corresponding receiver node to flashing LED of different intensities. Light onset is indicated by a vertical line. The intensity of the flash is shown next to each pair of traces. The number of spikes and the response latency are a function of the step size. The bottom pair of traces shows the response of the sending pixel to a small intensity flash. Current is integrated on the state capacitor, but the pixel fails to reach threshold. The current decays away at a rate set by the leak voltage. In this case, the leak voltage was 0.65 volt.

temporal accuracy is important and events are sparse in the retinal array. In general, a delta-modulated encoding of data is best for this communication protocol. The full signal must be represented by changes in the signal, and the effects of these changes integrated by the receiver, if the full DC value is to be reconstructed. For such a reconstruction, the time constant of integration on the receiver should be long. In contrast, the time constant of integration on the receiver should be short for the detection of temporal coincidence of events. Both of these regimes of operation are easily achieved within the range of current levels in subthreshold CMOS transistors. Both can be done in parallel on the same receiving chip, or on different receivers. Different time constants of integration or frequency characteristics are observed in parallel streams of the visual system. The magnocellular system is responsible for transmitting high temporal frequency information and has a lower integration time, while the parvocellular system is responsible for higher spatial frequencies but with longer integration times.

Of course, it is desirable to instrument the entire imaging array. The retinotopic nature of image transfer is best illustrated by comparing images scanned from the sender and the receiver chip using traditional analog video scanning techniques. The image of a flashing LED as it appears on the sending retina is depicted in Figure 3.24 and the corresponding image on the receiver is shown in Figure 3.25.

3.7 Future System Development

I have used the particular example of the retina to illustrate the use of the address-event representation. However, the address-event representation can be used to advantage by any system whose event generation rate is sufficiently low. The significance of the representation lies in its generality, which makes possible the modular design of multi-chip systems. The general technical issues faced in the development of multi-chip systems are discussed in the following section. Some biologically motivated example systems are then discussed.

3.7.1 Extensions of the Address-Event Representation

The arbitration procedure that has been described for a single sender and a single receiver can be extended to systems with multiple senders and receivers. In the one-dimensional

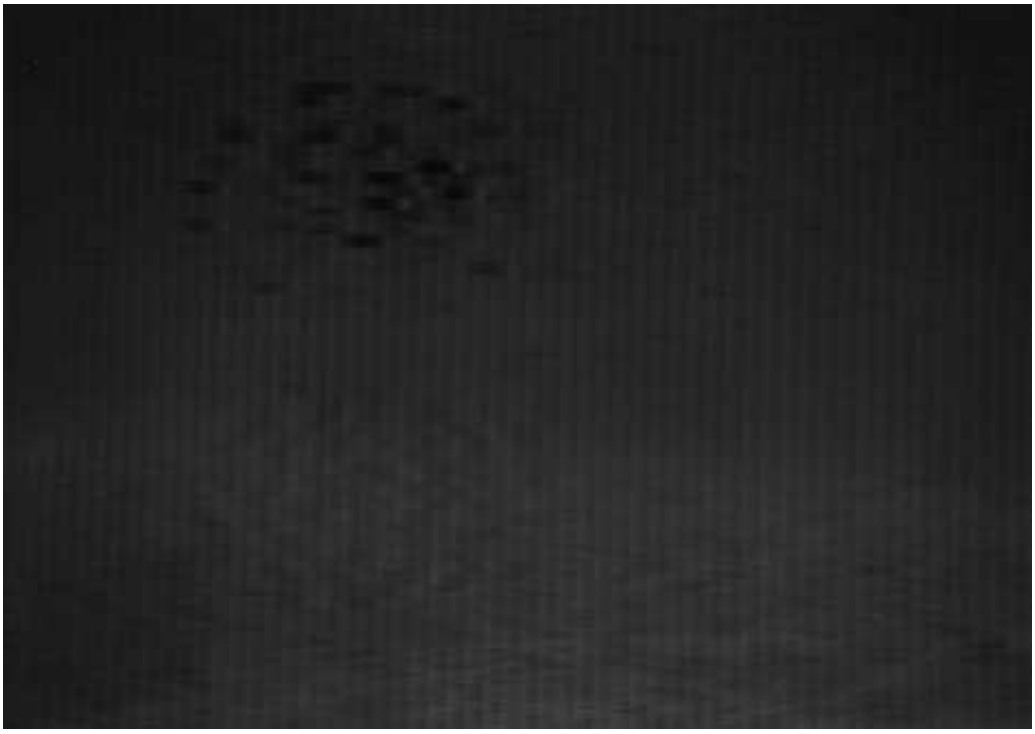

Figure 3.24: Response of the retina to a flashing LED. The voltage on the pixel state variables is sequentially scanned to the video monitor. The origin of the addresses is in the upper-left-hand corner of the image. The response is indistinct because the voltage of the sending pixel only rises to the threshold level before it is reset by data transmission.

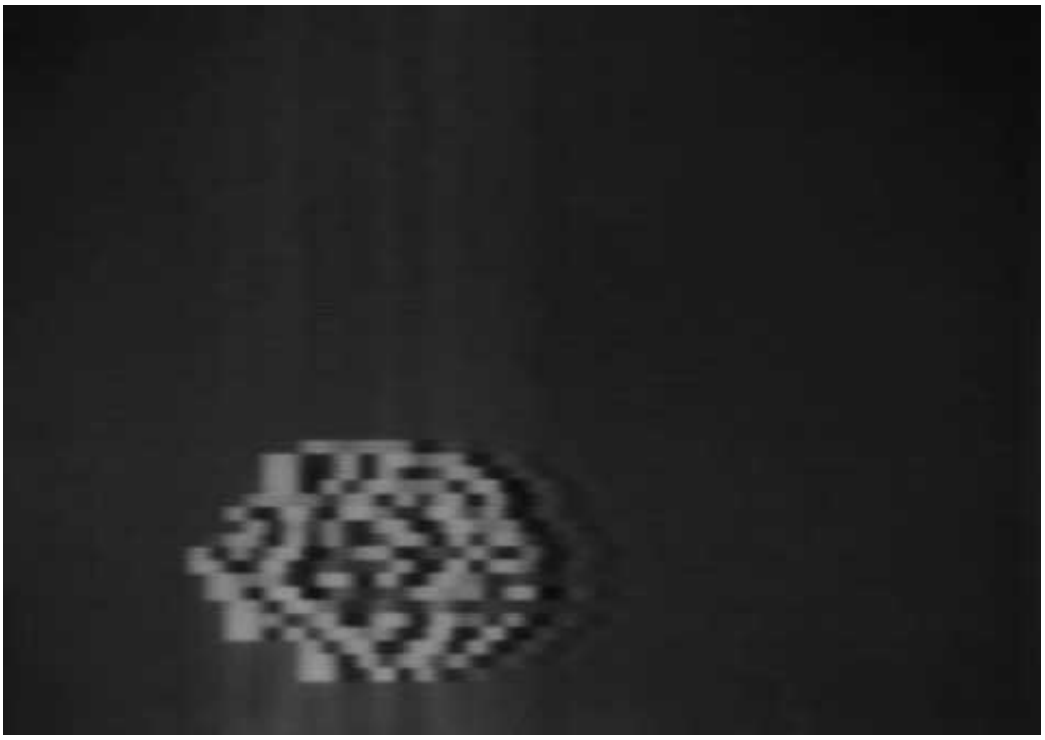

Figure 3.25: The receiver integrates the address-events coming from the sender. The voltages on the nodes of the receiver are sequentially scanned for display to the video monitor. The origin of the addresses is in the lower-left-hand corner of the image. Consequently, the pattern of activity on the receiver is mirrored around the horizontal axis from that of the sender in Figure 3.24.

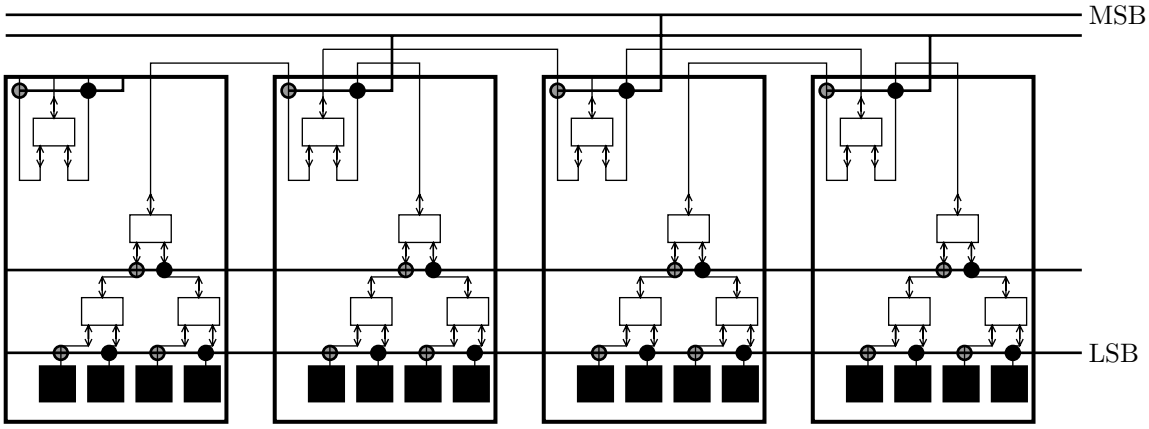

Figure 3.26: System of 16 neurons (black squares) distributed over four identical chips. The binary arbitration tree is distributed over all the chips and constructed by the wiring pattern between chips. Each bit of the address bus is driven by the Arbiter elements at the appropriate level of the tree. When the element to the right is selected, a zero is placed on the bus (filled circle) and when the element to the left is selected, a one is placed on the bus (open circle).

case, the binary arbitration tree can be distributed over multiple transmitters, as shown in Figure 3.26. All of the transmitters compete for control of a common bus. A major constraint on the size of the system is, as in the single chip case, the number of events generated by the combination of all the neurons in the system per address broadcast time. The layout of the Arbiter elements on each chip is identical; each chip contains a complete binary tree of Arbiter elements for its own neurons and a single additional Arbiter element whose inputs and outputs are brought off-chip. Binary trees of any size can be concatenated from the appropriate number of chips. The address is generated by logic distributed over all of the chips. The bits of the address are set by the Arbiter element at the appropriate level of the tree.

The Acknowledge generation in a multi-receiver system may be made conditional on all of the chips having received the data; however, this approach is useful only when all of the receiving chips would like to listen to all of the data. The decoding structure on each chip

would have to span the entire address space. Alternatively, each address might go to only one receiver, in which case, the Acknowledge would simply be the OR of the Acknowledges from all of the receivers. The simplest and most general method for terminating data transfer is to have the transmitter simply generate its own Acknowledge after some predefined waiting period. This procedure does not guarantee that the data have actually been received, but it may be perfectly adequate for the neuromorphic systems for which the address-event representation was designed. Unlike digital-logic-based systems, the occasional failure to transmit an event should not change the outcome of the computation. Even if the handshaking protocol is dropped, procedure for decoding the address performed by the receiving chips depends on the desired connectivity pattern of the system. One possible connectivity pattern is illustrated in Figure 3.27. In this case, the connectivity pattern is semi-local, with each receiver configured to accept data from a local address space. The higher-order bits of the desired address are externally established for each chip. This address is subtracted from the address on the bus to transform the address into local coordinates.

The address-event representation is particularly attractive for use in multi-chip systems because it transforms a concrete signal associated with a particular time and place into the abstract domain of digital logic. Space in the abstract domain can be manipulated to extend across the physical chip boundaries and time can be divided so finely that many digital events can go by in an instant on the timescale of the macroscopic biological world. The address-event representation attempts to overcome the volumetric wiring deficiencies of VLSI relative to neural tissue by using the strengths of digital VLSI medium, its speed, and its abstract symbolic manipulation efficiency.

In terms of circuit architecture, time and space are often interchangeable. This principle is the foundation of multiplexing, which reduces the number of wires needed to transmit the activity of an array of elements. The number of wires in the address-event representation can be further reduced by multiplexing the bits of the address. Since arbitration proceeds sequentially, the same data bus could transmit the y- and x-address bits for a two-dimensional array as they are selected without delaying transmission of the event. This address encoding is used in commercial dynamic RAM circuits and requires only half as many address pins. This procedure could be extended to all of the bits of the address, since the address is really determined one bit at a time as the select signal proceeds down the

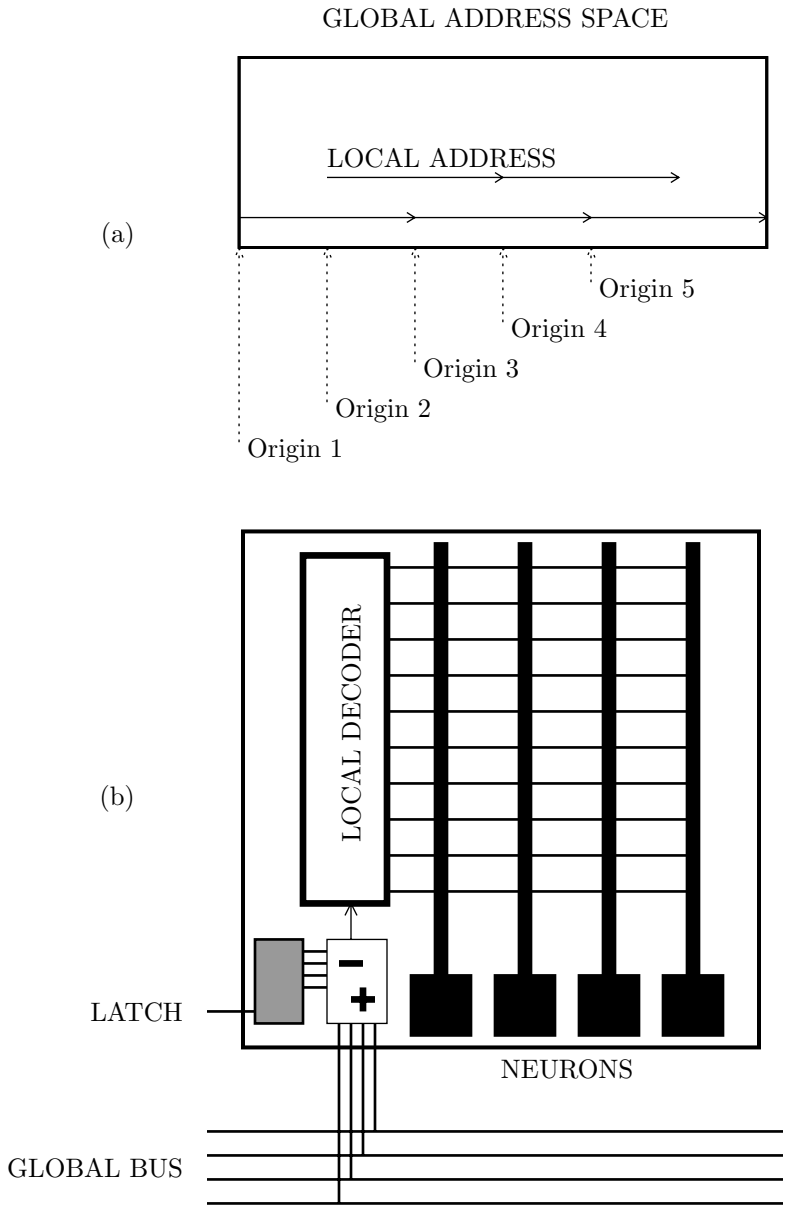

Figure 3.27: Semi-locally connected multi-receiver system. (a) Five receivers accept events from a contiguous, local region of the global address space. The origin of the contiguous region is programmed onto each chip. The regions spanned by the various chips overlap. (b) The decoding of the address by the receiver is accomplished by subtracting the origin of the local coordinate system, stored in the static latch (gray rectangle), from the incoming address.

binary arbitration tree.

The trade-off of time and space can be taken advantage of in a multi-chip address-event-based system by using digital logic and multiplexing to construct artificial dendritic and axonal arborizations, which are extended in time rather than in space. One idea would be to construct a look-up table that would transform a single event from pre-synaptic address to a stream of post-synaptic target locations, as illustrated in Figure 3.28 The post-synaptic target locations would be transmitted as a stream of events before the next pre-synaptic event could be issued. Temporal resolution is compromised for the space saved by reducing the number of synapses necessary for the post-synaptic neurons. Since only one pre-synaptic event is transmitted at a time, the same post-synaptic synapse can be used to receive events from a number of pre-synaptic neurons. The synaptic weight connecting each pre- and post-synaptic pair could be stored on the digital transforming chip and transmitted along with the event to the receiver, either as an analog voltage or as the duration of the post-synaptic address.

3.7.2 Systems Examples

The computational significance of temporal relationships between action-potentials in neural systems has not been extensively explored in a systems context. A VLSI neuromorphic system based on the address-event representation would allow experimentation in this area.

In any sizable neural system the axonal conduction delays must be taken into account if timing relationships are to be preserved. Conduction delay is critical in auditory localization [17] and has been incorporated into silicon auditory models [11]. Delays can be incorporated in the address-event representation. Events might propagate through several chips that are only locally interconnected, as shown in Figure 3.29. In this very simple system, each chip simply loads events into synchronous delay lines as they are received. This method reduces the temporal resolution of the address-event encoding to the synchronous clock period. One delay line runs in each direction. Each chip inserts its locally generated addresses in the center of both delay lines. The inserted addresses have their chip coordinate set to zero and the chip coordinate is incremented as the data are passed from chip to chip. The data bus therefore encodes the original address bits, plus bits that indicate how many chips the data had passed through. The address changes in time since it is expressed in local coordinates.

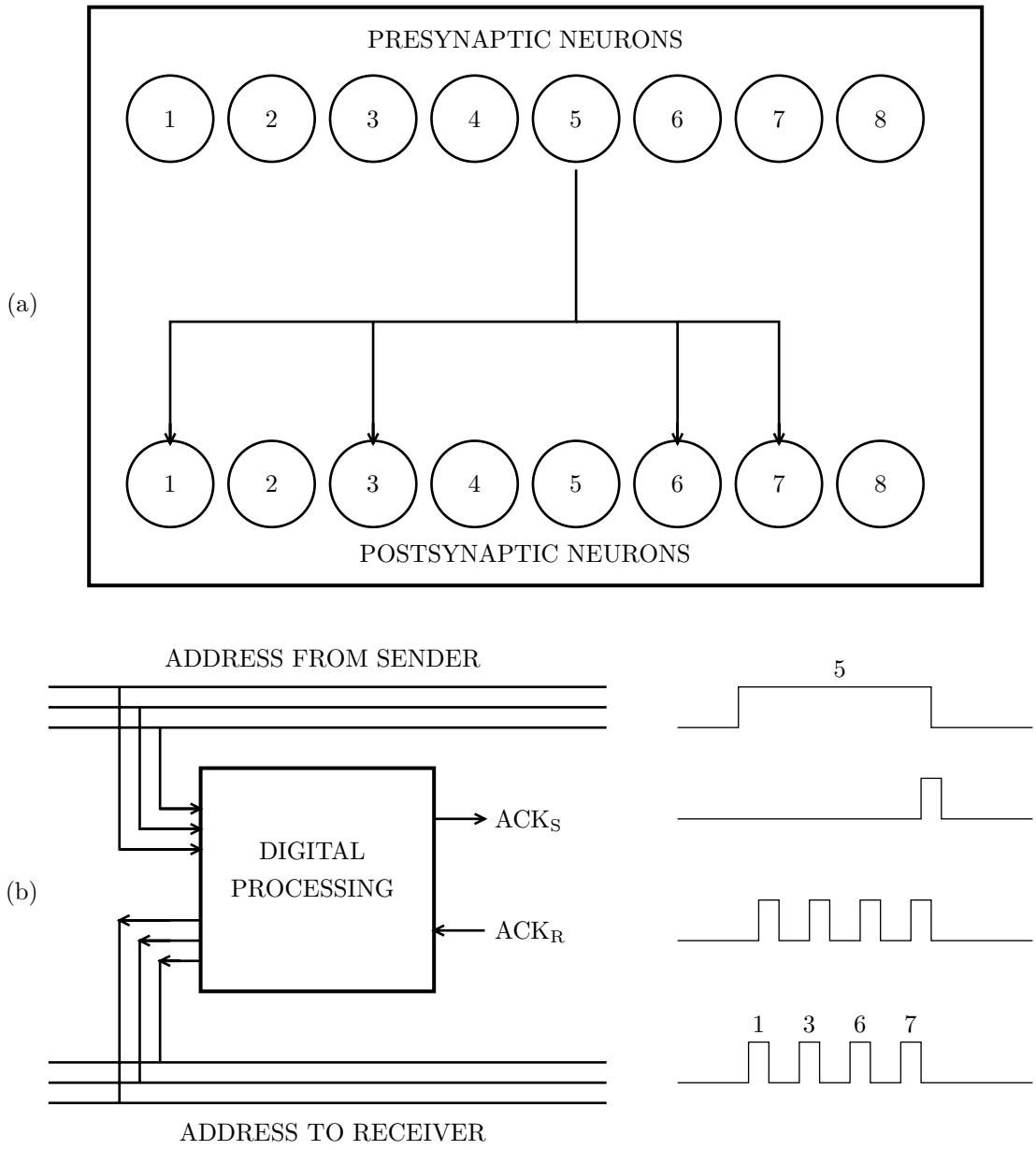

Figure 3.28: Constructing artificial receptive fields using address-events and digital processing to store the receptive field structure. The projection pattern illustrated in part (a) is recalled by the digital processor (b) when neuron 5 generates an event. The digital processor transmits events to the post-synaptic neurons on the receiver to which neuron 5 projects. When all of the postsynaptic neurons have acknowledged receipt of an event, the digital processor acknowledges the sender, which then is free to transmit another event.

One advantage to this encoding scheme is that receptive fields are easily made translationally invariant. As the event propagates away from the source, its coordinate bits get bigger and bigger. The events must be rejected after they have propagated some distance in order to make room for new events to be inserted into the delay lines. Alternatively, locally generated addresses could simply overwrite non-local events in the delay line. The probability of being overwritten would increase with distance. The number of events that can be stored and their temporal resolution are determined by the number of stages in the delay line. The system can be extended to two-dimensions. The number of delay lines necessary for events to propagate to each node in the array via a unique path is eight for a square grid and twelve for a hexagonal grid.

Investigation of the interaction of the morphology and conduction properties of a neuronal dendritic tree requires spatio-temporally patterned inputs. The address-event representation is flexible enough to allow reconfigurable spatial decoding of the event into a position along the post-synaptic dendrite by means of digitally-programmable static-latch decoders. Although these latches are larger than hard-wired decoders, they are critical to this application. In fact, a number of such configurable synapses may be desirable in general systems if the space used by including them is less than the space taken up by the unused synapses in a general system hard-wired to allow full connectivity.

The next chapter describes stereopsis chips that perform one-dimensional matching using an address-event representation. Stereomatching of real images requires that the imaging foci be separated by a distance much larger than that available on the surface of a single chip. Although in principle this problem can be solved using optics, it is much more convenient to separate the image planes and communicate the information electronically. Furthermore, stereopsis is believed to rely heavily on the output of transient (change-detecting) magnocells [13] and be influenced strongly by the temporal order of events [24, 4]. The task of stereomatching provides a system-level test of the ability of the address-event representation to preserve salient temporal and spatial sensory information.

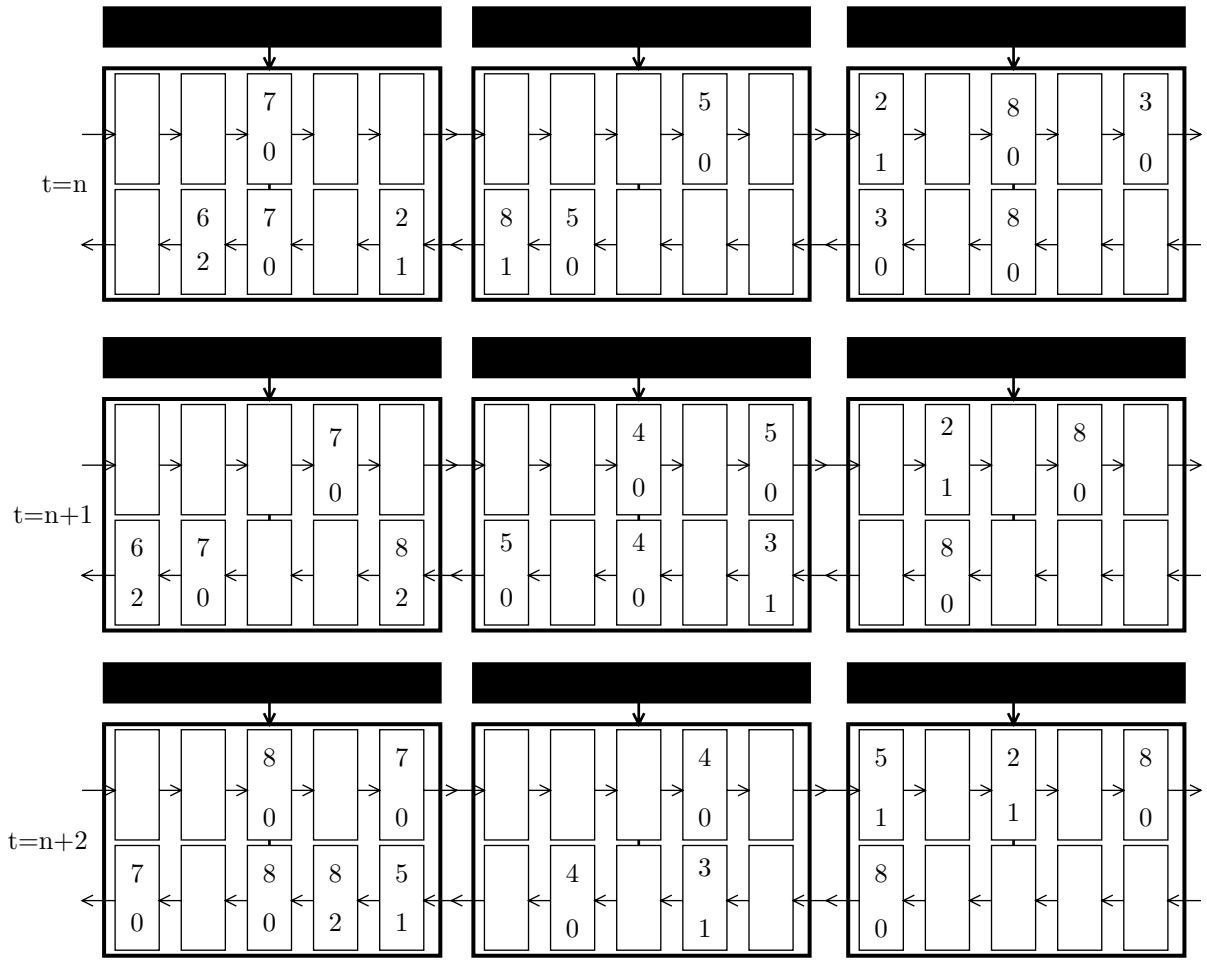

Figure 3.29: Address-event encoding scheme that generates axonal conduction delay. A linear array of three chips (black rectangles), each containing neurons and an associated address-event generator, and associated delay-lines, is depicted for three timesteps. When a chip generates an address-event, the event is loaded into the delay lines. The event propagates with delay throughout the system. In this architecture, the delay lines are synchronous.

References

[1] Barnes, S., and Werblin, F. (1986) Gated currents generate single spike activity in amacrine cells of the tiger salamander retina. *Proc. Natl. Acad. Sci. U.S.A.* **83**, pp. 1509-1512.

[2] Brown, T., Zador, A., and Claiborne, B. (in press) Hebbian computations in hippocampal neurons. In T. McKenna, J. Davis, and S. Zornetzer, (Eds.), *Single Neuron Computation*, Orlando, FL: Academic Press; Harcourt, Brace and Jovanovich.

[3] Brownlow, M., Tarassenko, L., Murray, A., Hamilton, A., Han, I.L., and Reekie, H.M. (1990). Pulse-firing neural chips for hundreds of neurons. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 785-792). San Mateo, CA: Morgan Kaufmann.

[4] Burr, C., and Ross, J. (1979) How does binocular delay give information about depth? *Vision Res.*, **19**, pp. 523–533.

[5] DeWeerth, S., Nielsen, L., Mead, C., and Astrom, K., (1991) A simple neuron servo. *IEEE Transactions on Neural Networks*, **2**, pp. 248–251.

[6] Douglas, R., personal communication.

[7] Dowling, J. (1987) *The Retina: An Approachable Part of the Brain.* Cambridge, MA: Harvard University Press.

[8] Gray, C., and Singer, W. (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **86**, pp. 1698–1702.

[9] Hille, B. (1984) *Ionic Channels of Excitable Membranes.* Sunderland, MA: Sinauer Associates Inc.

[10] Hutchinson, J., Koch, C., Luo, J., and Mead, C. (1988) Computing motion using analog and binary resistive networks. *IEEE Computer*, **21**, pp.52–63.

[11] Lazzaro, J., and Mead, C. (1989) Silicon modeling of pitch perception. *Proc. Natl. Acad. Sci. USA*, **86**, pp. 9597–9601.

[12] LeMoncheck, J., personal communication.

[13] Livingstone, M., and Hubel, D. (1988) Segregation of form, color, movement and depth: Anatomy, physiology and perception. *Science*, **240**, pp. 740–749.

[14] Mahowald, M. (in press), Evolving Analog VLSI Neurons. In T. McKenna, J. Davis, and S. Zornetzer (Eds.), *Single Neuron Computation*, Orlando, FL: Academic Press; Harcourt, Brace and Jovanovich.

[15] McCormick, D.A. (1990) Membrane properties and neurotransmitter actions. In G. Shepard (Ed.), *The Synaptic Organization of the Brain (3rd edition)* (pp. 220-243). New York: Oxford University Press.

[16] Mead, C.A. (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

[17] Moiseff, A., and Konishi, M. (1981) Neuronal and behavioral sensitivity to binaural time differences in the owl. *Journal of Neuroscience*, **1**, pp. 40–48.

[18] Moore, A., Allman, J., and Goodman, R. (1991) A real-time neural system for color constancy. *IEEE Transactions on Neural Networks* **2**, pp. 237–247.

[19] Mueller, P., Van der Spiegel, J., Blackman, D., Chiu, T., Clare, T., Donham, C., Hsieh, T., Loinaz, M. (1989). Design and fabrication of VLSI components for a general purpose analog neural computer. In C. Mead and M. Ismail (Eds.), *Analog VLSI Implementation of Neural Systems* (pp. 135–169). Boston: Kluwer Academic Publishers.

[20] Murray, A., Del Corso, D., and Tarassenko L. (1991) Pulse-stream VLSI neural networks mixing analog and digital techniques. *IEEE Transactions on Neural Networks*, **2**, pp. 193–204.

[21] Ryckebusch, S., Bower, J., and Mead, C. (1989) Modeling Small Oscillating Biological Networks in Analog VLSI. In D. Touretzky (Ed.), *Advances in Neural Information Processing Systems 1* (pp. 384-393), San Mateo, CA: Morgan Kaufmann.

[22] Schwartz, E. (1977) Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, **25**, pp. 181–194.

[23] Seitz, C.L. (1980) System timing. In *Introduction to VLSI Systems* by C.A. Mead and L. Conway (pp. 218–262). Reading, MA: Addison-Wesley.

[24] Shimojo, S., Silverman, G., and Nakayama, K. (1988) An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature* **333**, pp. 265–268.

[25] Sivilotti, M. (1986) WOLCOMP CS technical report, California Institute of Technology, Pasadena, California.

[26] Sivilotti, M., Mahowald, M., and Mead, C. (1987) Real-time visual computations using analog CMOS processing arrays. In Losleben, P. (ed.), *Proceedings of the Stanford Conference on Very Large Scale Integration.* (pp. 295–311), Cambridge, MA: MIT Press.

[27] Sivilotti, M. (1991) Wiring considerations in Analog VLSI Systems, with Application to Field-Programmable Networks. Doctoral dissertation, Department of Computer Science, California Institute of Technology, Pasadena, California.

Chapter 4

Stereopsis

4.1 Introduction

Stereopsis is the combination of visual information from two eyes for the determination of depth. Stereopsis has been studied using psychophysics, neurophysiology and computational vision. The analog stereo-matching chip presented in this chapter represents a new experimental approach to the study of stereocorrespondence, a primary subtask of stereopsis. It organizes much of what has been learned about stereocorrespondence using more traditional approaches, in a physical framework supplied by the basic circuits that underlie the computation.

Stereoscopic depth is a derived quantity, not immediately present in the two-dimensional images formed by the retinae. Neuroanatomical studies place the most peripheral locus at which stereopsis may occur at the primary visual cortex, the first site at which information from the two eyes is combined in higher animals. The step into cortex opens a Pandora's box of possibilities. It could be that stereopsis relies on object recognition, semantic knowledge and consciousness. Fortunately, psychophysical studies show that the problem of stereopsis may be approached without addressing the full complexity of the brain. Indeed, Julesz [14] has described stereopsis as a process mediated by a centrally located "cyclopean retina," not so different from the monocular retina. By using random dot stereograms (see Figure 4.1), Julesz has shown that stereofusion can occur without cognitive cues. Stereopsis is a fascinating problem that lies in the alluring region somewhere between passive sensation and active imagination.

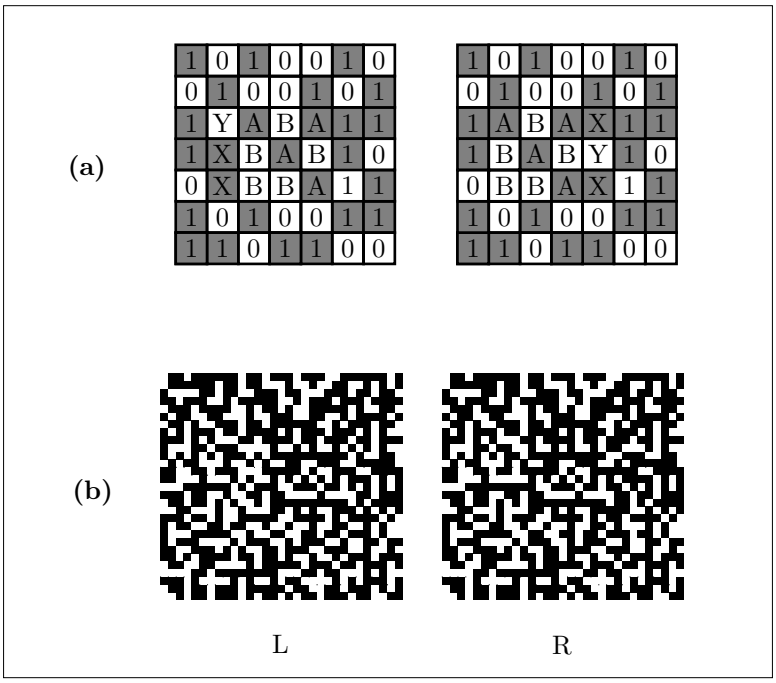

Figure 4.1:

Random-dot stereograms. (a) Making a random-dot stereogram. A random pattern of 1s and 0s is generated to be presented to the left eye. An identical copy of the pattern is made for the right eye, except that a central square region within the image (labeled with As and Bs) is displaced to the right. When the two images are fused, this square region will appear closer than the background. Occluded areas (areas having no counterpart in the opposite eye's image) are labeled with X's and Y's. (Modified from Julesz, 1971[14].) (b) A random-dot stereogram showing a raised square. You can fuse the stereogram by letting your eyes diverge as though you were looking at infinity. Your left eye should see the pattern on the left, and your right eye should see the pattern on the right. The primary difficulty is focusing on the paper while your eyes are diverged. Myopic readers may find it helpful to remove their glasses.

Because the random dot stereogram reduced the problem of stereopsis to comprehensible primitives, it has become the canonical test case for the computational vision community, which has proposed a variety of algorithms for its solution. Each algorithm is rooted in a different tradition. These traditions include: robotic vision [16, 10], psychophysics [26, 41, 40] and computational theory [25]. By and large, these algorithms have been expressed in and constrained by the language of the digital computer and are thus often difficult to relate to analog neurophysiological function. In spite of the gap between experimental electrophysiology and theoretical model, the existence of working algorithms makes stereopsis an attractive arena for the study of cortical function; there is some hope that computational functions might be associated with individual neuronal response.

The study of the neurophysiological basis of stereopsis has indicated that stereoscopic fusion has correlates in neuronal response as early as primary visual cortex (for a review see [36]). Primary types of response to binocular stimuli (including random dot patterns) have been identified and individual neurons are classified on this basis. As is often the case, it is not possible to definitively assign a computational function to a particular class of neurons. Aside from the teleological difficulties that arise from consideration of the computational purposes of neurons, it has been difficult just to gather enough information using single electrode recording to place the neurons in a network context.

Network interactions are a critical part of stereopsis because disparity tuning is fundamentally unlike the classical problems of orientation tuning [11] or velocity tuning, which can in principle be performed by spatiotemporally oriented receptive fields that are convolved with the retinal input [28]. Computational and psychophysical experiments indicate that stereofusion of a random dot pattern is an inherently nonlocal and nonlinear operation, which probably requires positive feedback [14, 25]. Although anatomy and basic neuronal biophysics reveal that these operations are consistent with the predominant features of cortical circuitry, few models explaining neuronal response characteristics have attempted to incorporate them because they are difficult to analyze or even to simulate numerically.

This chapter explores the interaction between computational algorithm and physical implementation. The system described is an analog CMOS stereo-matching circuit based on a new stereocorrespondence algorithm. The algorithm was devised under the constraints of the analog electronic medium. It is embodied in a compact circuit that is able to solve

one-dimensional random dot patterns. The circuit implementation is efficient since the algorithm requires relatively low wiring density and takes advantage of the device physics. The circuit includes nonlinear positive and negative feedback elements and converges to a solution in less than 25 milliseconds. Unlike previous circuits [22] that were based on Marr and Poggio's cooperative stereocorrespondence algorithm [24], the new algorithm/circuit performs well on surfaces that are tilted in depth. Individual electrical nodes in the circuit can be related to the types of stereo-tuned neurons found in primary visual cortex. The tuning curves of the electrical nodes in the circuit are explained in terms of the function of the whole network.

4.2 The Problem of Stereocorrespondence

The problem of recovering the three-dimensional geometry of space from two-dimensional projections can be broken down into several related subtasks, such as feature extraction, eye vergence control, computation of real distance from image disparity and eye position, etc. The subtask solved by the circuitry described in this chapter is called stereocorrespondence, which allows the determination of image disparity. A more complete description of the problem of stereocorrespondence can be found in [14, 37].

Binocular vision generates two images of a scene, one from each eye. Because the two eyes regard the scene from different points of view, they differ in their impression of the relationships between objects. Figure 4.2 shows two eyes of an observer in cross-section. The lens of the eye focuses an image of the scene composed of discrete targets located in three-dimensional space onto the surface of the retina. Stereocorrespondence is the pairing of features in one retinal image with features on the other retinal image that arose from the same target in three-dimensional space.

The task of finding matching features in each eye would be straightforward if features could be identified uniquely. However, random dot stereograms [13] demonstrate that the human visual system can compute disparity even when there are many identical features in close proximity (see Figure 4.1). Because no pattern is visible monocularly, the determination of correspondence must take place without cognitive assistance. Furthermore, no single pair of targets is sufficient to determine the appropriate correspondences; since all

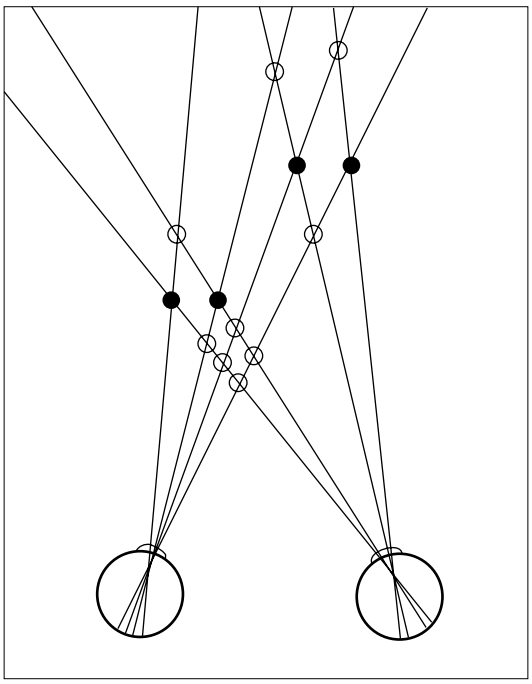

Figure 4.2: Stereopsis. This figure illustrates the projection of images of four identical targets (dark disks) onto the right and left eyes of an observer. The lines going through the lenses connecting each target with the retinas are lines of sight. The intersections of the lines of sight indicate possible target positions in space. False targets (transparent disks) are located at the intersections of lines of sight that originate from different targets in the two eyes.

targets are identical, they could be matched with any of the others. The intersection of lines of sight for features that do not correspond represents false targets. There is no way to differentiate a false target from a real target without making some assumptions about the three-dimensional structure of the targets. The determination of appropriate correspondence is a cooperative process that must consider simultaneously many possible feature pairings.

Calculation of stereocorrespondence is simplified by the fact that the search need not take place over the entire two-dimensional image. The features in the right and left image corresponding to the same target are confined to lie along lines on each of the retinae, as shown in Figure 4.3. These lines, called epipolar lines, are the locus of points that must be searched to establish the stereocorrespondence of the features that line on them. The origin of the epipolar lines can be understood by imagining that the image position of a single feature is known, as are the positions of the two eyes, but that the 3-D position of the target giving rise to the image feature is not. The feature in the image projects through the nodal point of the eye along a line of sight. The target could lie anywhere along this line. That line of sight is imaged through the nodal point of the other eye to the corresponding epipolar line. Corresponding epipolar lines in the two images result from intersection of the plane defined by the nodal points of the two eyes and the target, with the image planes.

When the eyes are verged to infinity so that the optical axes of the eyes are parallel, the epipolar lines are all parallel to the horizontal axis (assuming that the image planes are flat). Features in the right image at a particular elevation must correspond to features in the left image at that same elevation. However, as vergence changes, the epipolar lines tilt. All of the epipolar lines intersect at the point defined by intersection of the line connecting the nodal points of the eyes and the (infinitely extended) image plane [10]. In this case, the search for corresponding features must extend over different *vertical* displacements, depending on the state of vergence of the eyes. Although the region of possible correspondence shifts as the epipolar lines are tilted, for any given state of fixation, the search for possible correspondence is a one-dimensional problem.

Once the stereocorrespondence of the targets has been determined, the disparity can be calculated. In this chapter, disparity is defined geometrically, as if the points in the retina were assigned coordinates $(x_{\mathrm{l}}, y_{\mathrm{l}})$ in the left eye and $(x_{\mathrm{r}}, y_{\mathrm{r}})$ in the right eye. The geometry

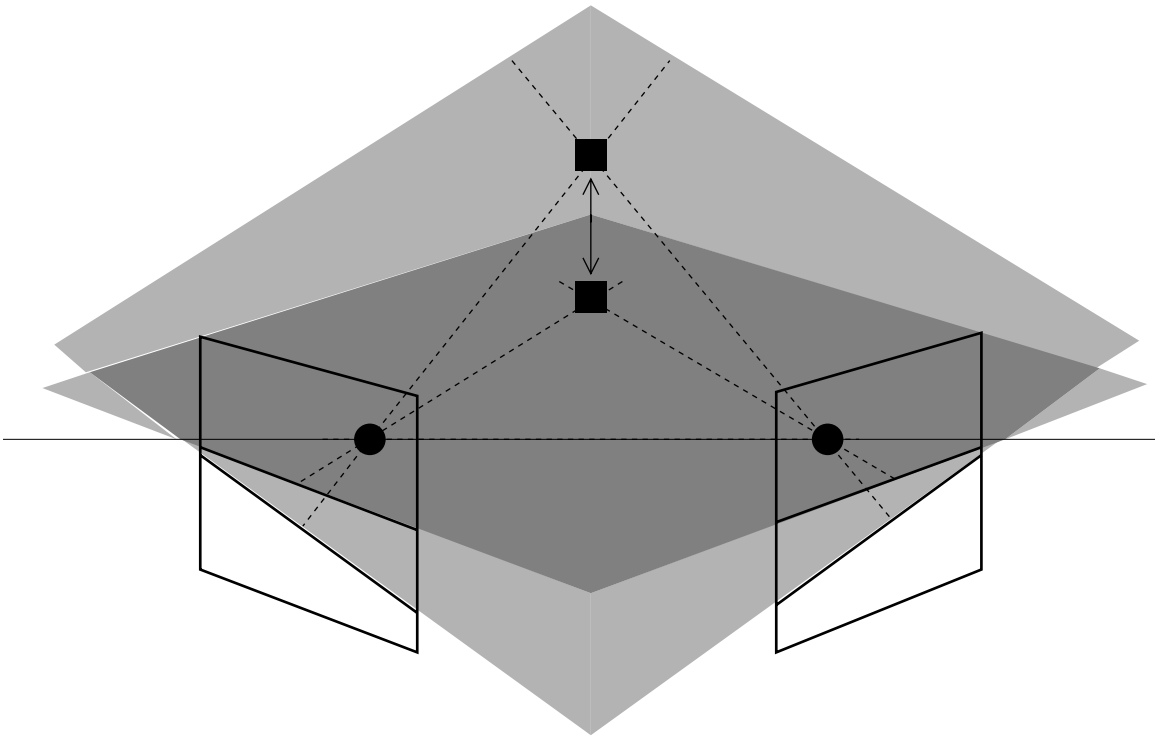

Figure 4.3: The epipolar lines for two targets at different elevations. The retinas (vertically oriented image planes) are shown symmetrically verged about the midline. The nodal points of the lenses are shown as filled circles. Two targets (filled squares), one above the other, are shown with associated lines of sight (dotted lines). Each target, along with the nodal points, defines a plane, which intersects the retinas to form epipolar lines. The epipolar lines intersect at the point of intersection between the line joining the nodal points of the eyes and the image plane.

of the epipolar lines depicted in Figure 4.3 demonstrates that, in general, receptors with the same coordinates on the two retinae cannot be stimulated by the same target. The locus of points in space that stimulate the same coordinates on the two retinae is called the horopter and is the zero-disparity surface of fixation. The horopter exists over an entire two-dimensional image only when the optical axes of the eyes are parallel. Otherwise, the geometrical horopter exists only in the horizontal plane that intersects the nodal points of the eye and is perpendicular to the image plane. This horopter is known as the Vieth-Müller circle. The simplest interpretation of the one-dimensional stereocorrespondence chip is that it is computing correspondence on the epipolar line of this circle. All the image features corresponding to targets in the plane of the Vieth-Müller circle have the same retinal y-coordinate. Targets closer to the viewer than the horopter have crossed (negative) disparity, ($x_\mathrm{l} < x_\mathrm{r}$). (See Figure 4.5.) Targets more distant than the horopter have uncrossed (positive) disparity, ($x_\mathrm{l} > x_\mathrm{r}$).

4.3 Overview

A number of algorithms for the computation of stereocorrespondence have been proposed, several of which are based in part on psychophysical measurements and/or neurophysiological recordings from single cells. The stereocorrespondence algorithms that are executed on digital machines have provided new insight into the function of neural systems by creating a level of abstraction that organizes individual measurements. The stereomatching chip described in this chapter differs from digital algorithms because, in addition to taking account of neurophysiological and psychophysical data, it is constrained by the properties of the analog electronic medium. These constraints create another level of correspondence between form and function. Because it is designed to function in the real world, it must deal with unnormalized, continuous time input. In contrast to a sequential digital simulation, the cost of connectivity in the analog medium is higher than the cost of iteration, so a feedback structure with local connectivity is cheaper than a globally connected feedforward structure. Electronic analogs make new links between single-cell physiology, algorithm and psychophysics because they incorporate electrical behavior, purposive design, and real-time system performance.

4.3.1 Neurophysiology

The mechanisms that biological systems use to compute stereodisparity are unknown. However, some of the neurophysiological characteristics of the neurons believed to participate in the computation have been elucidated. Gian Poggio has summarized the physiologic responses of cell types sensitive to binocular disparity in macaque monkey, which are illustrated in Figure 4.4 [36, 38]. He lists five major categories of cell: the tuned excitatory cell, which is stimulated strongly only by binocular stimuli that are close to zero disparity; the tuned inhibitory cell, which is typically strongly stimulated by monocular targets presented to one of the two eyes and is always inhibited by binocular stimuli at zero disparity; the tuned-near/tuned-far cells and the near/far cells, which are driven by stimuli of larger crossed or uncrossed disparity; and the disparity flat cells that are stimulated by targets presented through either eye alone, or by binocular targets at any disparity. All of these cells are usually also tuned to other stimulus parameters, such as contrast, spatial frequency, orientation, and direction of motion, and may be classified as simple or complex based on the structure of subregions in their receptive fields. Similar proportions of simple and complex neurons are sensitive to the disparity of narrow bars. Only complex cells, however, appear to be sensitive to random dot stereograms.

4.3.2 Computational Algorithms

Many stereocorrespondence algorithms that are more or less consistent with neurophysiological and psychophysical data have been proposed, a number of which are reviewed by Poggio and Poggio [37] and Blake and Wilson [2]. These algorithms fall into two major classes: those that discriminate true targets from false targets based on cooperative interactions, and those that pre-filter the input across multiple spatial scales and restrict the search area in order to reduce the probability of a false match.

Cooperative algorithms typically include arrays of units that are narrowly tuned for disparity, similar to the tuned-zero neurons. These elements participate in two forms of interaction:

1. A nonlinear inhibitory mechanism that suppresses false targets.

2. A nonlocal interaction that gathers evidence to guide this decision.

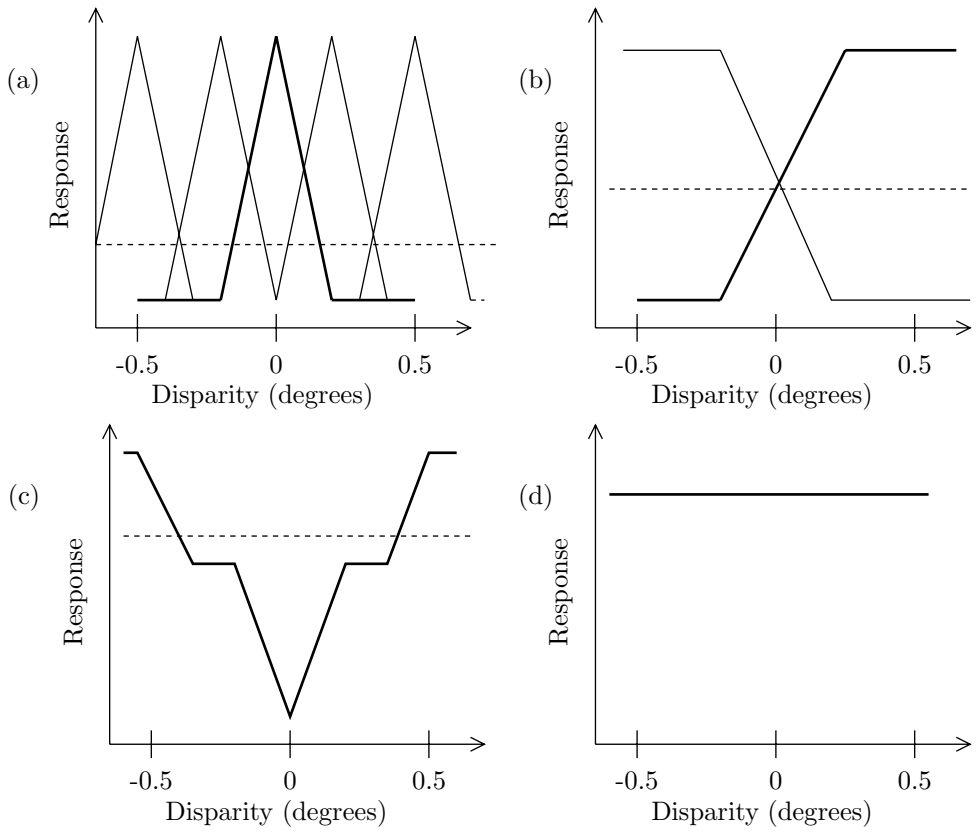

Figure 4.4: Schematization of the disparity tuning curves of several of the major cell types believed to be involved in the computation of disparity in the macaque monkey. (a) TE: tuned excitatory cell; (b) NE/NF: tuned near and tuned far cells. (c) TI: tuned inhibitory cell; (d) FL: disparity flat cell. Dotted lines show the response of the cells to monocular stimulation. Adapted from Poggio et al. [38].

A diagram of a network structure that supports a typical cooperative algorithm is shown in Figure 4.5. The algorithm is implemented with an array of correlators (circles) depicted beneath two retinas. The correlator array represents internally the space outside the observer. Each row of correlators responds to targets at a particular depth. Columns in the array correspond to horizontal image position, or cyclopean angle. The outputs of each retina are projected into the correlator array at 45 degrees. The lines of retinal projection correspond to lines of sight. There is a non-linear competition between correlators to select the true matches. The correlators in competition with each other are either those in a column or those along the same line of sight. This competition is based on the idea that each feature in one retina should only correspond to one feature in the other retina. This constraint arises because one object along a line of sight occludes another object behind it.

The accumulation of evidence favoring true targets over false ones is gathered for some distance across the image. This interaction shown in Figure 4.5 by heavy lines coupling correlators at the same disparity. Candidate matches support other candidates that are consistent with themselves. Which candidates are consistent with each other depends on some assumption about the structure of physical objects. Typically, the assumption is that objects are continuous in depth. (Pollard et al. [40] have employed a constraint on the solution based on psychophysics that has a similar effect–namely, that the correct solution should not include targets whose depth changes too quickly with horizontal distance.) Since the candidates of a consistent solution mutually support each other, this interaction is a form of positive feedback. Positive feedback may be implemented explicitly [7, 24, 40] or via disinhibition [46].

The Marr and Poggio cooperative stereomatching algorithm [24] was translated into an analog VLSI circuit [22]. There were several shortcomings to the algorithm. First, the positive coupling must pass from node to node in the correlator array. The depth solution is filled in and must exist at every point in the image. Targets must be close together if they are to influence each other, or else the excitation must be strong enough to propagate over large distances through the correlator array. Second, the algorithm only works well on fronto-parallel surfaces, since the excitation propagates only within a single disparity.

Another class of algorithm, illustrated in Figure 4.6, relies on spatial frequency filters to eliminate false matches. Marr and Poggio [25] pointed out that false matches occur

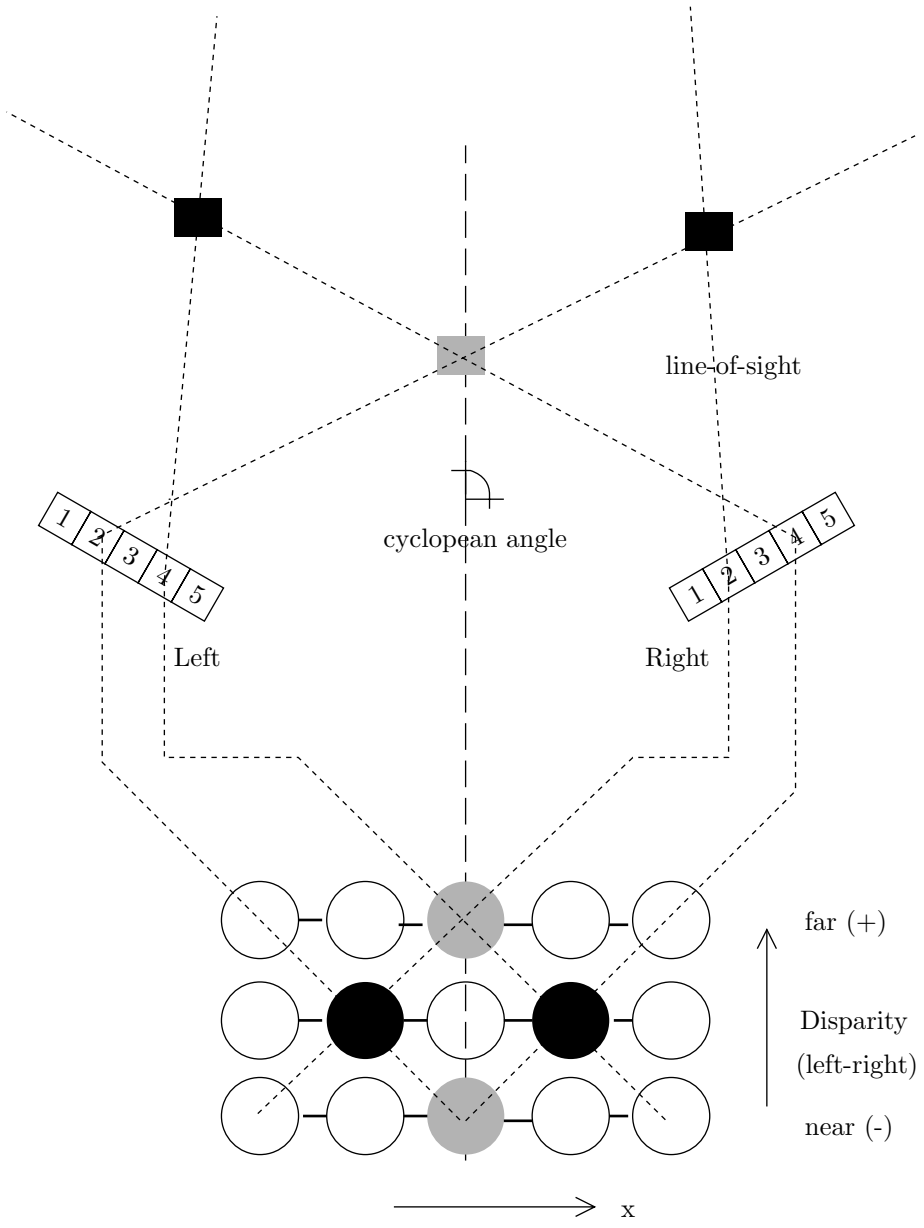

Figure 4.5:

General framework of the cooperative algorithms. The left and right retinas with five pixels are depicted looking out at a scene (above) with two targets at zero disparity and a false target visible at -1 disparity (another false target is implied at +1 disparity but is not shown). The external scene is reconstructed beneath the two retinas in a correlator array (circles). Gray circles are correlators responding to false matches, and black circles are correlators responding to true matches. Inhibitory interaction among correlation elements run along lines of sight (dotted lines) or along lines of equal cyclopean angle (dashed line). Solid lines along disparity planes indicate positive coupling between correlators.

only at disparities on the order of the width of the spatial frequency of the filter. If the peaks of spatial intensity modulation are considered targets, then another target closer than the spatial frequency of the filter would constitute a higher spatial frequency and would have been filtered out. Observing that cortical cells respond at roughly half amplitude for frequencies twice their peak frequency, Blake and Wilson [2] have proposed a "quarter-cycle limit." To eliminate false matches, the maximum disparity for fusion should be less than a quarter cycle of the spatial wavelength of the stimulus.

Combining the outputs of filters tuned to different spatial frequencies gives rise to a range of disparity tuning curves. Filters tuned to high spatial frequency have narrow disparity tuning, like the tuned-zero cells, and filters tuned to low spatial frequency have broad disparity tuning, similar to that of the tuned-near/tuned-far or near/far cells [27]. Low spatial frequency filters with shallow tuning curves respond in an analog fashion over a range of stimulus disparities.

Because only low-spatial-frequency units can be used at large disparities without introducing false matches, the stereoacuity at large disparities is poor. In order to achieve fine disparity resolution over a large range of disparities, these algorithms shift the range of the high-spatial-frequency filters to be centered around the disparity indicated by the coarser resolution channels. This shift can either be accomplished with eye movements [25] or by gating the activity of the high resolution units with the activation of the low-spatial-frequency units [35].

The multi-resolution algorithms are practical and effective. Nishihara [35] has implemented a multi-resolution algorithm with special-purpose digital signal processing hardware, which is able to find the disparity in real scenes illuminated with speckled light, which increases target density. The disparity map is generated at a rate of 30 seconds per frame.

However, comparison with human performance on stereocorrespondence tasks suggests that these multi-resolution algorithms are incomplete. Psychophysical measurements indicate that the quarter-cycle limit reasonably obeyed for spatial frequencies less than about 2 cycles per degree. However, at higher spatial frequencies, the fusional limit is constant at about 15 minutes of arc, large enough for significant numbers of false matches to occur [44]. Humans are able to process images with large disparities even in the absence of low-spatial-frequency information.

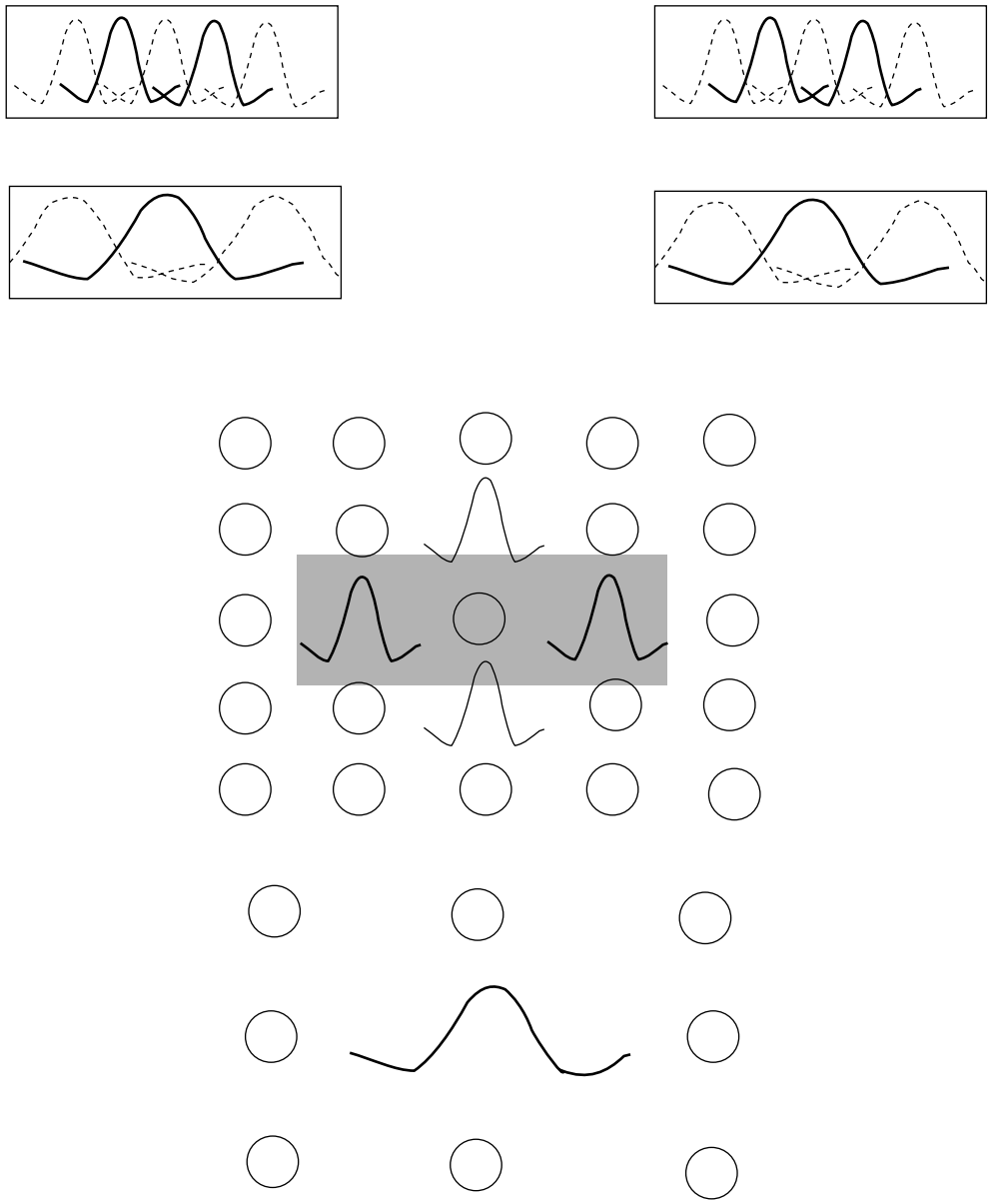

Figure 4.6: Multi-resolution algorithm. Two one-dimensional retinae with two arrays of band-pass filters are shown above. The left retina shows highlighted the receptive fields, corresponding to the position of retinal features. The outputs of the filters of the same type in the two retinae are convolved by binocular cells to produce the disparity tuning curves depicted below. Correlation of the activated filters of the right and left retinae produce candidate matches in the correlator array. The low-spatial-frequency binocular cells activate the portions of the fine disparity array that is correlated with their own response (grey box). This activation selects the true targets, which are unambiguously indicated in the coarse disparity array.

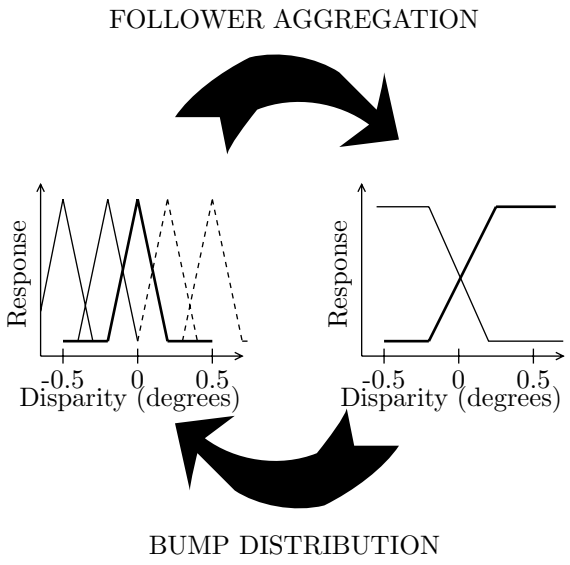

Figure 4.7: Transformation between place-value encoding of disparity in the correlator array and analog-valued encoding of disparity.

4.3.3 Electronic Analog

I have invented a new algorithm, embodied in a VLSI circuit, that generates electrical responses similar to those found in neurobiology. The algorithm links multiple-scale disparity algorithms and cooperative algorithms to remedy some of the shortcomings of each. The principal innovation is the addition of analog-encoded disparity units that interact via positive feedback with the correlators in the disparity array. Unlike the multi-scaled algorithms which are feedforward from low-spatial-frequency to high-spatial-frequency encoding, my algorithm generates a low-spatial-frequency estimate of disparity from high-spatial-frequency units. The system as a whole performs a transformation of the representation of disparity from a place-valued encoding of disparity in the correlator array to an analog-valued encoding of disparity [1], which is analogous to a low-spatial-frequency estimate of disparity. This transformation, illustrated in Figure 4.7, allows interpolation to occur in the analog domain where it is implemented more easily.

A block diagram showing the major components of the algorithm is depicted in Figure 4.8. The largest block is the correlator array. The units in this array are analogous

to the tuned excitatory cells. They receive input from binocular receptive fields that are slightly displaced from each other on the two retinae. The magnitude of the displacement determines the peak disparity tuning of each unit.

Beneath the correlator array is an array of monocularly driven units that have electrical responses that are similar to the tuned inhibitory cells. The response characteristics of these units arise because they are in competition with the correlator array. When the correlator array is stimulated by a binocular input, it suppresses activity in the monocular units.

Competition between the cells in the correlator array and the monocular units is mediated by an array of Winner-Take-All (WTA) units shown beneath the monocular units. The WTA units provide feedback inhibition to all the correlators and monocular units at a particular horizontal position in the image. Because they receive input from all the correlators, they respond to input at all disparities, like the disparity flat neurons. The effect of the WTA inhibition is to suppress the activity driven by a retinal feature at the WTA horizontal position in all but one correlator or monocular unit. This non-linear inhibition suppresses false targets.

The final array of units is the analog-valued disparity units, which are analogous to the near/far neurons in that their response magnitude is monotonic in disparity over a larger range than that of the tuned cells. These analog-valued units are part of a positive feedback loop with the tuned units in order to lend support to feature matches that are smoothly varying in disparity.

The major interactions between the elements of the circuit are illustrated in Figure 4.9. Most of the interactions take place within the correlator array. A correlator receives input from a single position on each retina. Inputs from the right and left retinas are multiplied before summing into the output node, V.

The correlators compete with each other at each horizontal position via negative feedback from a common winner-take-all circuit. The negative feedback pathway is indicated by dashed lines in Figure 4.8. The winner-take-all (WTA) circuit suppresses false matches. The inhibition level is averaged over horizontal position since the WTA circuits are spatially coupled.

The WTA circuit cannot discriminate between correlations of equal strength. Therefore, false matches are distinguished from true matches by providing additional input to the true

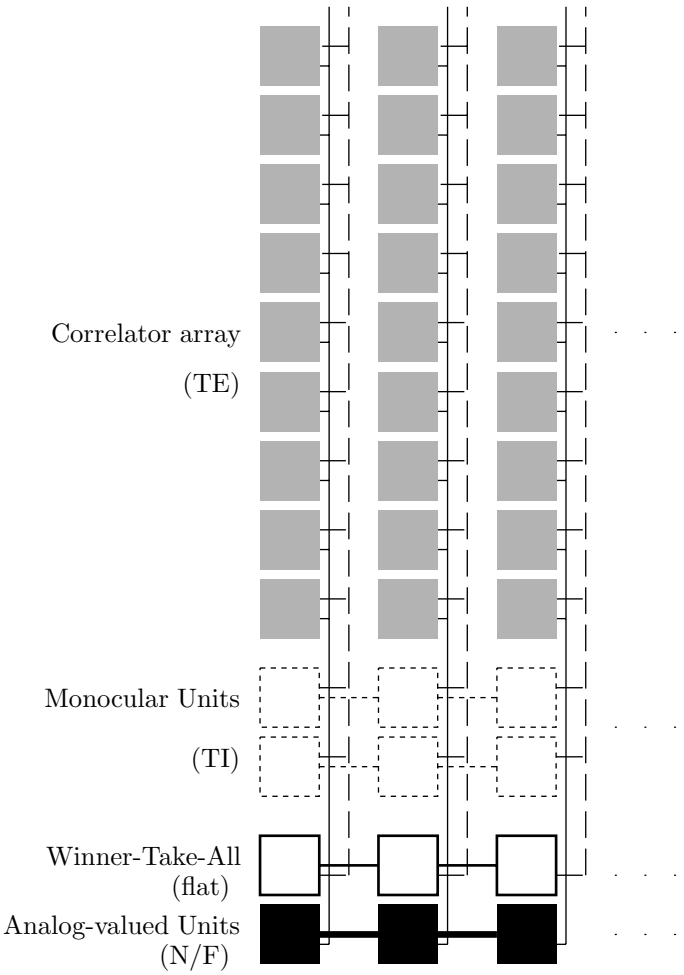

Figure 4.8: Block diagram of the major components of the algorithm. This diagram presents a narrow horizontal slice through the chip. The chip contains 57 retinal positions in the horizontal dimension (3 are shown). The retinas are not shown. There are 9 $\times$ 57 correlators (grey boxes), 2 $\times$ 57 monocular units (one linear array for each eye) shown in dotted outline, and linear arrays of analog-valued units (black squares) and WTA units (squares in heavy outline). Each component corresponds to a cell type observed in biological systems. (See text.) Dashed vertical lines indicate negative feedback from the WTA circuit to the correlator array and the monocular units. Solid vertical lines indicate positive feedback from the analog-disparity units to the correlator array. The WTA and the analog-valued units are coupled to their neighbors at adjacent horizontal image positions.

match via a positive feedback from the analog disparity unit. The analog disparity unit estimates the true disparity in the image by converging to a non-linear, non-local average of the correlations.

Although in principle the analog units could receive low-spatial-frequency retinal input, in this implementation, input to the analog disparity units is provided exclusively by the correlator array. Each correlator drives the analog-valued unit associated with horizontal position through a variable conductance, which is modulated by that correlator's activity. The value to which the correlator drives the analog-valued unit is determined by the correlator's peak disparity tuning. Each disparity peak is assigned a disparity reference value, which is computed by a resistive voltage-divider. Correlators tuned to larger disparities drive the analog-disparity unit to larger voltages. The analog-disparity unit thus computes a weighted average of activity at that point in the image.

The analog units are coupled to each other in a one-dimensional resistive network, which averages the local average of disparity across horizontal image position. The voltage on each node of the resistive net represents the best estimate of the image disparity at that position. The resistive coupling allows the disparity solution to be linearly interpolated so that objects tilted in depth can be properly resolved. The interpolated value in the analog net represents the depth at that point in the image, although there are no retinal targets present at that location. In this way, depth is represented at every point, without the necessity of activating the correlation units directly.

The discrimination between true and false targets performed by the WTA circuit is biased to the smooth solution by feedback from the analog-disparity units into the correlator array. The analog-disparity unit maximally stimulates the correlator whose disparity reference voltage most closely agree with its own voltage. This interaction is represented in Figure 4.9 by the element represented as a circle enclosing a tuned response. The peak response of the element occurs at the disparity reference voltage. The maximally stimulated correlator begins to win the competition and the losing correlators make less and less contribution to the voltage of the analog-disparity unit. The system converges to a solution in which at most one correlator is activated at each horizontal position. The amplitude of the tuned feedback is in proportion to the magnitude of the input at that horizontal position, which is measured by the WTA circuit. This scaling insures that the positive feedback will

be in proportion to the level of retinal input. When the input is large, the feedback must be large enough to elevate the true matches above the false ones. Yet when the input is weak, the feedback must not be so strong that the system locks into a state from which it cannot escape.

In addition to the analog units, a set of units that report the existence of unmatched monocular targets was incorporated in the network. Monocular units receive input from only one retina. The retinal input is summed with positive lateral inputs from neighboring monocular cells so that regions of unmatched features in the same eye support each other and compete with unmatched regions in the other eye. The monocular units of the right and left retinas compete with each other and with the correlator units via the WTA inhibition. The system tries to decide if a retinal feature has a match in the other retinal image; however, it has a place to represent the feature if there is no match. Unmatched targets arise in real images from occlusion events at depth discontinuities. The monocular units are used to break the disparity interpolation by controlling a fuse circuit that disconnects the analog units from each other where there is an occlusion event. The analog estimate of depth therefore is not averaged across occluded targets that signal a depth discontinuity. The use of fuses in stereodisparity computations has been previously proposed in the context of an analog network interpolation algorithm based on a variational principle [5]. In my algorithm, however, the fuse is controlled by the monocular unit driven by retinal input, rather than the voltage representing disparity at the terminals of the fuse. This external control allows the system to remain continously sensitive to changing retinal input.

The convergence properties of this chip are difficult to analyze, as the behavior of all the elements depends nonlinearly on the stimulus. An attempt to characterize convergence empirically is presented after the circuit details are discussed. The performance of the chip is compared to human psychophysical performance on stereo disparity tasks. Stimuli include random dot stereograms with abrupt transitions and occluded features, disparity patterns that are tilted in depth, and periodic patterns whose disparity is unambiguous only at the end points. Also, the disparity gradient limit for fusion of two targets is measured.

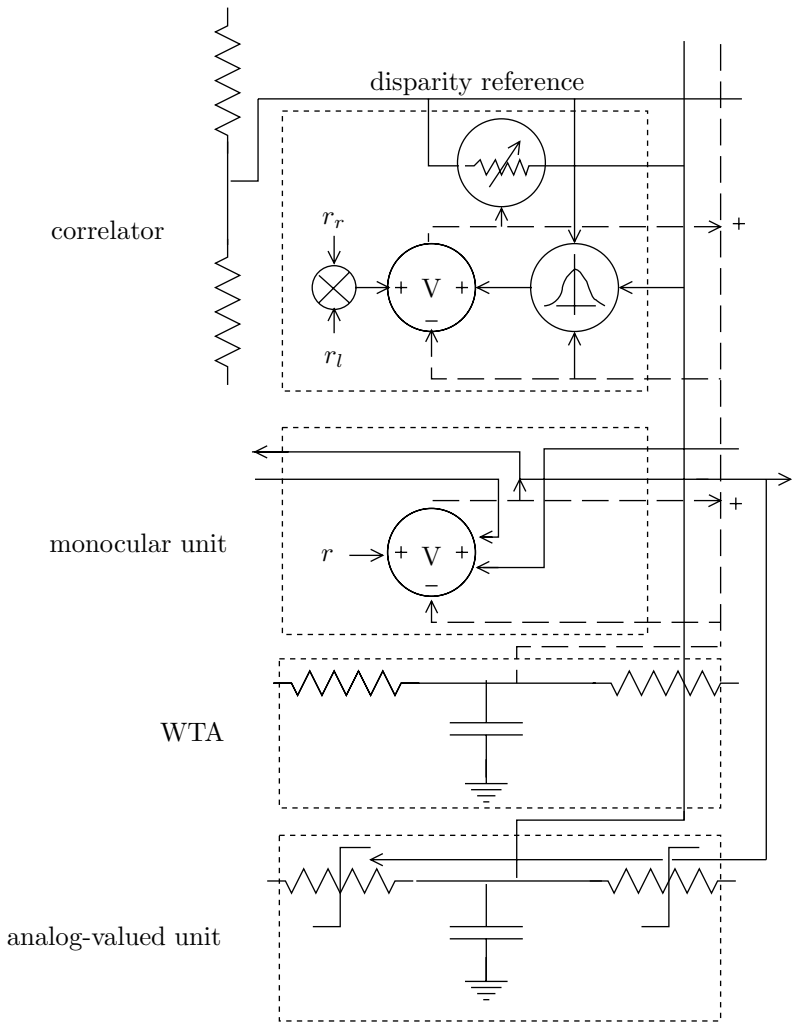

Figure 4.9: Summary of the major interactions between elements of the stereocorrespondence chip. A single correlator element is shown in gray at the top of the figure. The resisitive voltage-divider that generates the disparity reference voltage is shown at the left. Beneath the correlator is a dotted box containing a monocular unit. Monocular units are coupled to their neighbors. Beneath that is the WTA element and beneath the WTA is the analog-valued unit. Both the WTA elements and the analog-valued units are resistively coupled to their neighbors. The resistors coupling the analog-valued units are drawn with fuses that are controlled by the output of the monocular units. Interaction with the WTA takes place along the dashed line. Interaction with the analog-valued unit takes place along the solid line. r_r: right retinal input, r_l: left retinal input.

4.4 The Chip

The architecture of the stereo matching chip is shown in Figure 4.10. The chip correlates the outputs of two one-dimensional retinas, and the representation of the solution expands into the second dimension of the silicon surface.

There are two principle paths along which information propagates, lines of sight and lines of average retinal position, illustrated in Figure 4.5. Retinal input units, of course, project into the correlator array along lines of sight. Most of the interesting computation, however, takes place along lines of average retinal position. If the eyes are symmetrically verged, these lines are lines of equal cyclopean angle. Cyclopean angle is equivalent to horizontal cyclopean image position. If i is the location of a pixel in one retina and j is the location of a pixel in the other retina, then the correlators associated with the nth cyclopean angle line receive input from the retinal coordinates $i = n - d/2; j = n + d/2$ where d is the disparity to which the correlator is tuned. Monocular units are assigned a cyclopean position that is equal to their retinal position. This assignment places them on the same cyclopean angle line as the zero-disparity cell with which they have a common retinal input. There are twice as many cyclopean angles as there are monocular retinal positions, since both pixel positions and disparities are integer valued. The negative and positive feedback interactions in the algorithm take place along lines of cyclopean angle.

4.4.1 Input

The input to the stereo chip is buffered in two one-dimensional retinas which represent the corresponding epipolar lines from each eye along which matching can take place. Each retinal array has 57 elements. These retinas are not themselves light sensitive, rather they are driven by address-events like the receiver chip described in Chapter 3. The circuit diagram of the retinal node is shown in Figure 4.11. Two eight-bit-wide input ports supply addresses and DATA VALID signals. Although the structure was designed to be used with two retinae that generate address-events, the chip was characterized with computer-generated stimuli, as described in Chapter 3.

The address-events represent features that has been detected by some earlier stage of processing. The features could be spatio-temporal image contrast derived by a center-

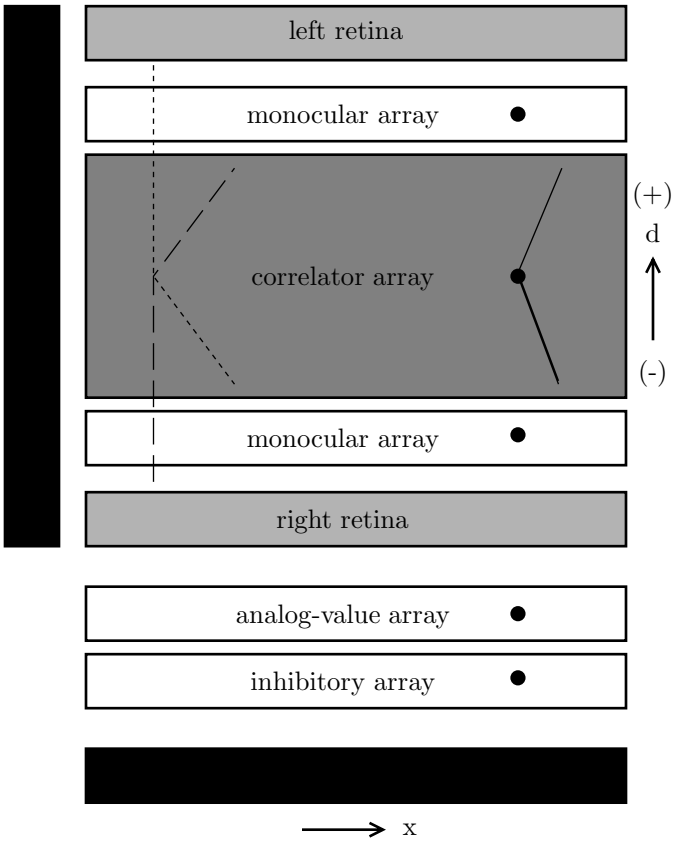

Figure 4.10: The architecture of the stereocorrespondence chip. The topography is deformed from the conceptual correlator arrays shown in Figure 4.6 and Figure 4.5 to facilitate the layout. The retinas project into the correlator array and monocular arrays along lines of sight. Line of sight from the lower retina is dashed and the line of sight from the upper retina is dotted. The analog-value resistive network and the inhibitory resistive network are below the lower retina. An analog-valued unit, indicated with a dot, interacts with the correlator array along the line of equal cyclopean angle shown in thin, unbroken line. Corresponding retina and monocular positions are indicated by dots. The shape of the line of equal cyclopean angle (average retinal position) on the surface of the chip is the average of the shapes of the upper and lower lines of sight. The output of the chip is scanned with a two-dimensional analog scanner shown in black. Images scanned from the chip for video display show activity on the retinae, the monocular arrays, and the correlator array. The responses of the analog-valued array and the inhibitory array are monitored on an oscilloscope.

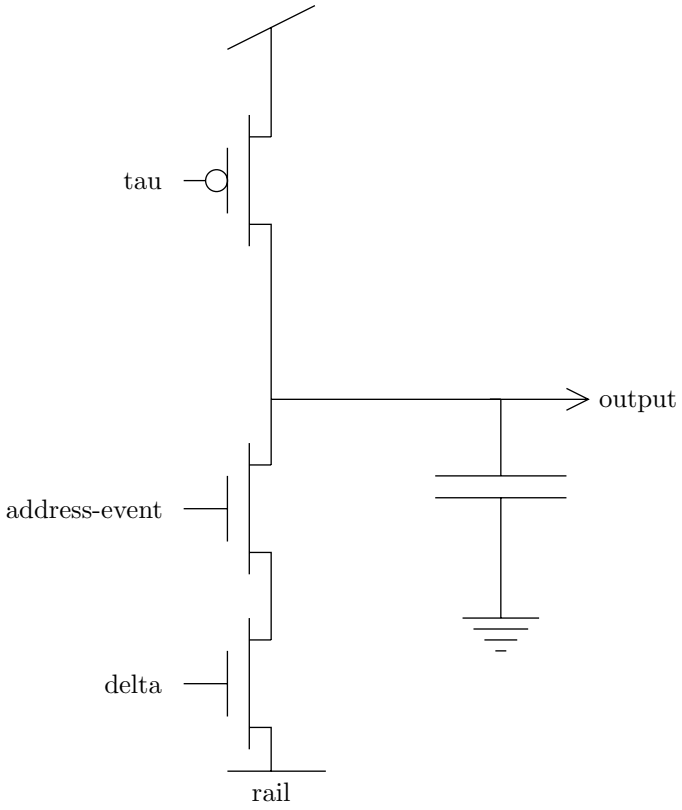

Figure 4.11: Schematic of retinal pixel that receives address-events. When an address-event is decoded, it removes an amount of charge from the output capacitor. The amount of charge removed is regulated by *delta*. Charge gradually leaks onto the capacitor at a rate controlled by *tau*. The output of the pixel is a voltage that is active low.

surround silicon retina, or a more complex signal, like a bandpass oriented edge. The image feature that drives stereocorrespondence is entirely a function of the circuitry generating the addresses.

4.4.2 Correlators

The heart of the chip is the correlator array, analogous to the tuned excitatory cells. The disparity-tuning curve for a single correlator is depicted in Figure 4.12. The correlator response is high at a single value of stimulus disparity and low at all other disparities. The tuning curves of all the correlators are similar except that they are shifted across the disparity axis. The maximum disparity to which correlators are tuned in this chip is ± 4, for a total of 9 rows of correlators. The range of disparity over which these correlators is tuned is the range of disparity over which images can be fused. Psychophysically this area is called Panum's Fusional area. The extent of Panum's area is a function of the spatial frequencies present in the stimulus [44]. These units represent high-spatial-frequency features and the constant disparity range of the chip is similar to the 15 minutes of arc fusional limit at frequencies greater than 2 cycles per degree.

The circuit schematic of a correlator is shown in Figure 4.13. The retinal input to a correlator is a nonlinear combination of the output of two pixels, one pixel from each retina. Each row of the correlator array is an iso-disparity plane that represents a point-by-point cross-correlation between the two retinas, at a spatial offset corresponding to that plane's disparity. The combination of signals from the right and left pixels is performed in the correlator cell by means of two serially connected transistors. In subthreshold, these two transistors compute the function:

$$I_{\text{mult}} = \frac{I_{\text{r}} I_{\text{l}}}{I_{\text{r}} + I_{\text{l}}}.$$

I_{r} and I_{l} are the currents through the rectifiers in the right and left pixels, respectively. This operation is a normalized multiplication of the two retinal inputs. If either retinal input is small, the current into the correlator is small. In principle, the algorithm does not depend on the nonlinearity in the combination of retinal inputs; the nonlinear inhibition should be sufficient to eliminate the units that are stimulated by only one retinal input.

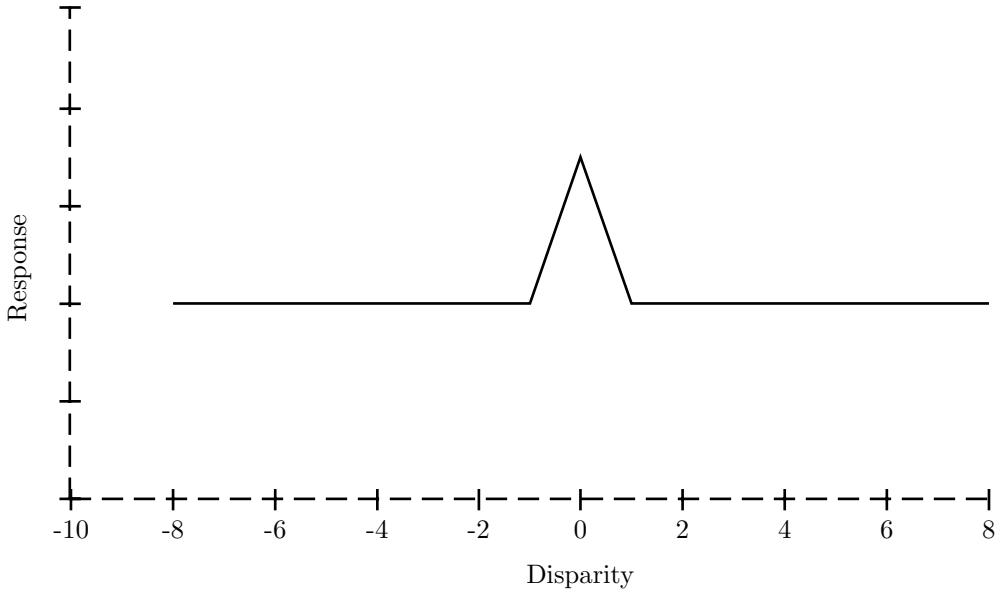

Figure 4.12: Response of a single correlator as a function of disparity plotted for the correlator tuned to zero disparity. The tuning curves of all of the correlators are similar, but shifted on the disparity axis. There is no response to monocular stimulation due to the nonlinear combination of inputs from the two retinae. These tuning curves are similar to those of the tuned excitatory cells (TE).

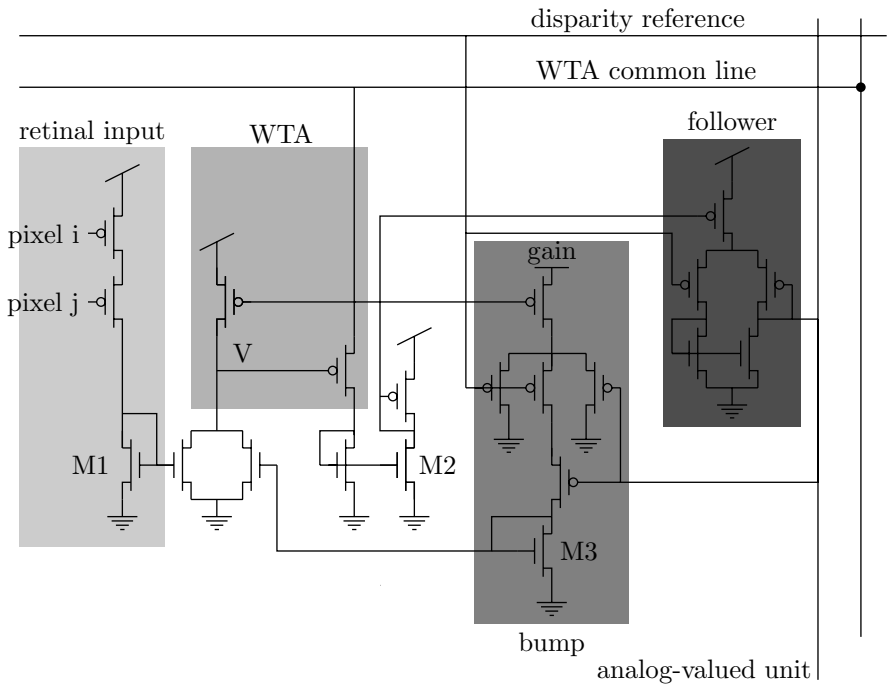

Figure 4.13:
Simplified version of the electrical interactions at a correlation element. The correlator output voltage, V, is determined by the sum of the currents flowing into the node. The retinal input, provided by current-mirror M1, is a current that is a nonlinear AND–type function of the signals from pixel i in the right retina and pixel j in the left retina. Input from analog-valued units through mirror M3 is summed with the retinal input. The WTA circuit provides feedback inhibition that sinks the input current to V_{dd}. Mirror M3 is controlled by a bump circuit whose bias is set by the voltage on the winner-take-all line. The current onto the WTA common line supplied by the correlator is mirrored by M2 to bias the follower driving the analog-valued unit.

However, the signal-to-noise ratio is much higher if this nonlinearity is present. The retinal input is summed into the output node, V, by current mirror M1.

The correlator receives an additional input through current mirror M3, which provides feedback from the analog-valued unit via the bump circuit. The feedback from the analog-valued units to the correlators is the basis for convergence to a solution that is globally optimal. The sum of the currents through mirrors M1 and M3 is the total positive input to the correlator. This current is counterbalanced by inhibition from the WTA circuit.

The feedback pattern from the analog-valued unit into the correlator array is depicted in Figure 4.14. The distribution pattern is along lines of equal cyclopean angle. The bump

circuit measures the difference between the analog-value unit voltage and the disparity reference voltage of the correlator and provides input to the correlator whose disparity is in agreement with the analog disparity estimate. The amount of feedback to a correlator at disparity d is a function of the analog-valued unit voltage that is a bump centered around the disparity reference voltage, which is denoted by V_d. This function is specified by the following equation:

$$I_{\text{feedback}_d} = \frac{I_n}{1 + \alpha \cosh^2 \left[\beta \left(V_d - V_{\text{analog}_n}\right)\right]}$$

where α and β are constants and I_n is exponential in the difference between voltage on the common line of the WTA circuit at cyclopean angle n and the control voltage labeled *gain*. The gain of the positive feedback is adjusted to be sufficiently low to prevent the system from latching up into a fixed state. The fact that the magnitude of the feedback is proportional to the activity at the cyclopean angle of the correlator itself has several consequences. It means that the feedback gets stronger as the solution gains strength. The magnitude of the positive feedback scales with the magnitude of the retinal input so that the input magnitude can vary widely and the feedback strength does not need to be externally adjusted. This scaling could also be accomplished by making the positive feedback proportional to the activity of the correlator itself; however this method would not allow correlators that were losing the competition to receive any positive feedback. Offsets in the magnitude of the retinal input might be able to keep the correct solution from winning the competition by preventing the correct correlator from receiving any positive feedback.

4.4.3 Inhibition

Inhibition in this system is responsible for normalizing the output in the face of inputs of various magnitudes and selecting between competing hypotheses about the true image disparity. It is implemented by a Winner-Take-All (WTA) circuit [17].

A simple, two-channel, WTA circuit is illustrated in Figure 4.15. To understand how the circuit works in subthreshold, imagine that the circuit is in equilibrium and that each channel is receiving an identical input current. In this configuration, $I_1 = I_2 = I_{out_1} = I_{out_2}$. The voltage on the common line, V_c, is therefore constrained to be logarithmic in the input current. The voltages on the output nodes, V_1 and V_2, are constrained to supply the bias

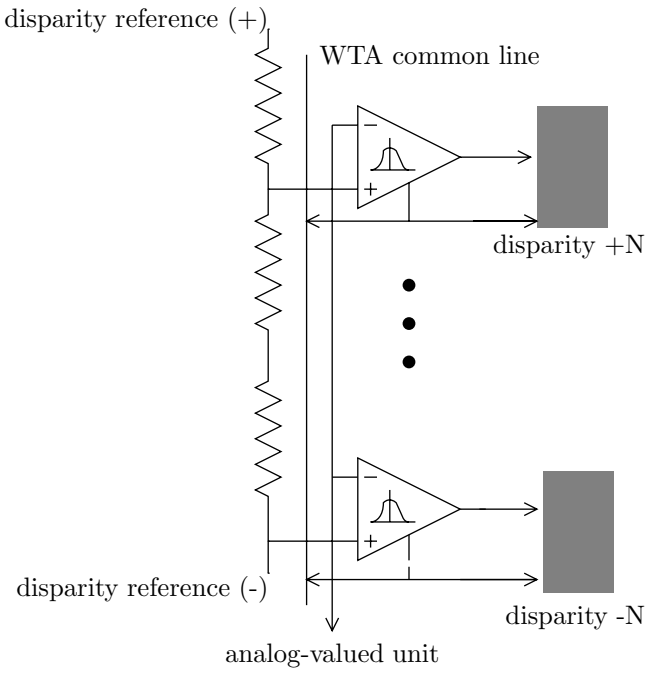

Figure 4.14: Positive feedback from the analog-valued units to the correlator array along a line of cyclopean angle is mediated by bump circuits.

current to the common line through source-follower transistors, T_{2_1} and T_{2_2}. These voltages are above the common line voltage by an amount that is logarithmic in the bias current. To make one channel win over the other, we increase its input current. Increasing the current to one channel charges up that channel's output node. The voltage on the common line follows the output voltage of the winning channel with a voltage difference set by the bias current. The output node stops charging when the current through its T_1 transistor is equal to the new input current. The output voltage of the winning channel increases logarithmically with input current while the loser's voltage decreases.

The loser is suppressed because the inhibition is drawing more current than is being supplied by its input. Since the voltage on the common line, V_c, controls the current out of both channels, the capacitor of the channel with less current is discharged until its T_1 transistor draws only its input current. For current differences between the channels of more than a few percent, the T_1 transistor of the losing channel will come out of saturation; the output voltage is within a few $\frac{kT}{q}$ of ground. When the current difference between channels is small, the output voltage on the losing channel is determined by the Early voltage of the T_1 transistor and by the level of the input current.

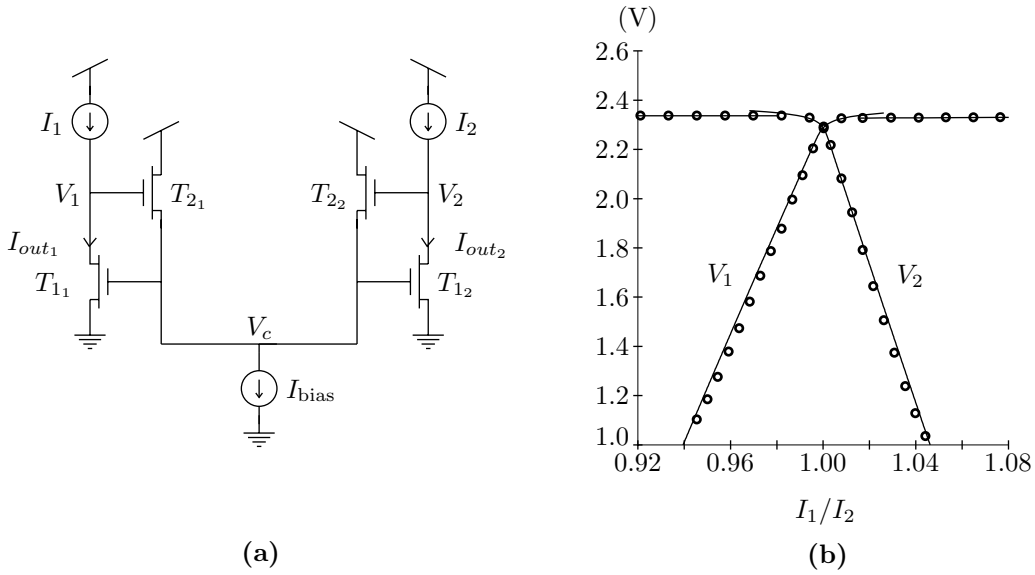

Figure 4.15: The winner-take-all circuit. (a) Schematic of a simple two-channel winner-take-all circuit. (b) Current-voltage characteristic of the two channel WTA circuit. The voltage output of the two channels is plotted against the ratio of their input currents.

The WTA circuit is extended to N input channels by simply connecting each channel to the same common line. The connectivity of such a system is 2N. The common line collects input from all N nodes in the competition and inhibits all of them.

In this stereocorrepondence chip, the WTA circuit establishes a competitive feedback interaction between all the correlators along a line of cyclopean angle [7, 34], rather than along lines of sight [24, 22]. The monocular units also engage in the WTA competition along the same line of cyclopean angle as the zero-disparity correlators that are driven by the monocular units' retinal input. All the correlators along a line of cyclopean angle drive the same common line, thus the disparity-tuning curve of the WTA common line is flat. Since the monocular cells are also in competition with the correlators, the WTA circuit also responds to monocular stimulation. The extent to which it responds to monocular targets compared to binocular targets is a function of the magnitude of the input to the monocular cells relative to the correlator cells. The disparity tuning curve for the voltage on the common line of the WTA circuit is shown in Figure 4.16.

The common lines of the WTA circuits are resistively coupled, thus inhibition is spatially

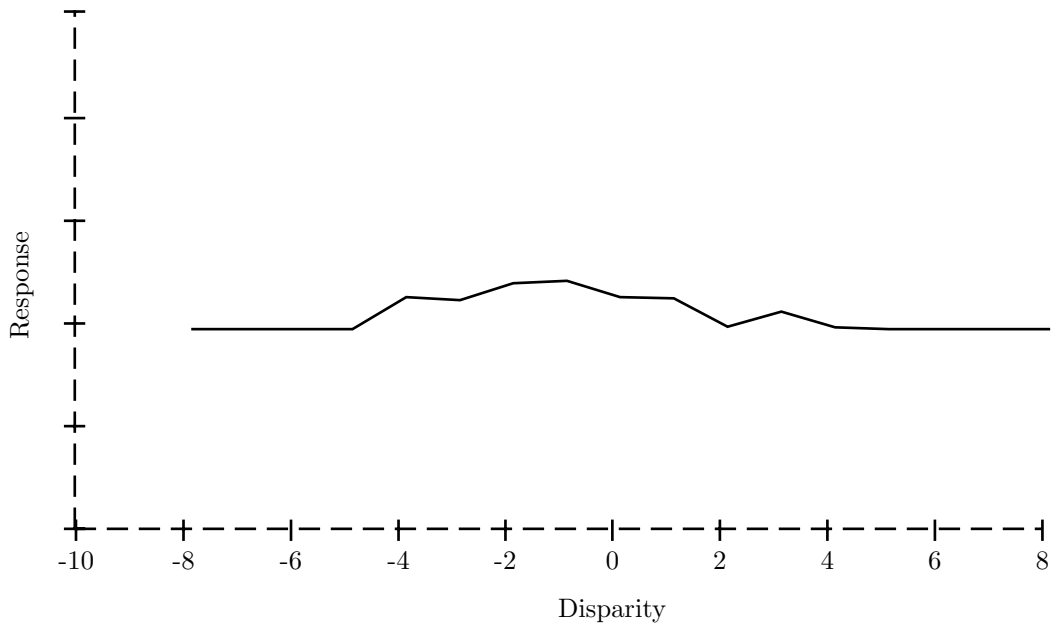

Figure 4.16: Disparity tuning curve for voltage on the common line of the winner-take-all circuit. This tuning curve is analogous to the disparity tuning curve of the disparity flat cell.

averaged across multiple lines of cyclopean angle. The spread of inhibition allows correlators on different cyclopean angles to inhibit one another. The strength of the resistor relative to the bias current on the common line (I_{bias} in Figure 4.15 and WTA bias in Figure 4.17) sets the spread of inhibition across retinal position.

An additional channel with fixed input, depicted in Figure 4.17, participates in the WTA competition at each cyclopean angle and thereby sets a threshold for activation of the correlators. The threshold level is set so that the bump circuit feedback from the analog-valued units is unable to bring a correlator above threshold. The system does not latch into a state and stay there because the correlator input falls below threshold when the retinal input is removed. When the threshold element is winning the WTA competition, it drives the analog-valued unit to a resting potential through a follower that acts as a conductance. The magnitude of the conductance is set by the limit transistor.

4.4.4 Analog-Valued Units

The analog-valued units encode disparity as an analog voltage. Their response to stimuli at different disparities, shown in Figure 4.18, is similar to the near/far cells. In the biological

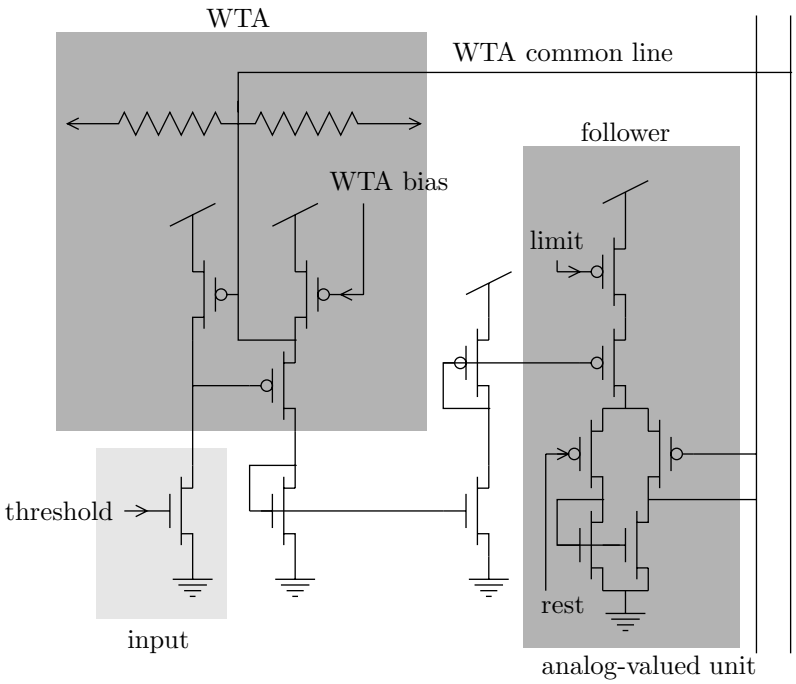

Figure 4.17: Circuit schematic for threshold element. The threshold input level sets the magnitude of input a correlator or monocular unit must attain in order to win the WTA competition. The threshold elements set the resting voltage and passive conductance for the analog-valued units.

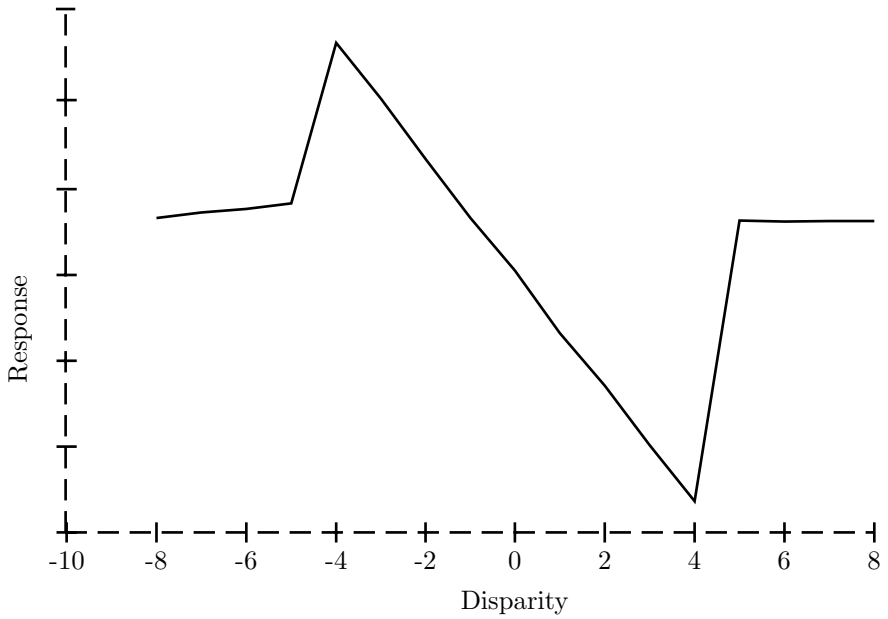

Figure 4.18: Disparity tuning curve of the analog-valued units is most similar to that of the near/far or tuned-near/tuned-far cells.

system, these cells probably receive input from the retina as well as intra-cortical input. However, in this chip, the voltage output of the analog-valued units is derived exclusively from the activity within the correlator array.

The analog-valued units aggregate the activity in the correlator array along equal cyclopean angle lines to form an analog estimate of the image disparity. The activity in the correlator array is transformed into an analog value using a follower aggregation circuit [2] shown in Figure 4.19. Each correlator on the cyclopean angle line controls the conductance of a transconductance amplifier that couples the analog-valued unit to a disparity reference voltage. The voltage of the analog disparity unit is the weighted average of the voltages of the disparities indicated by active correlators. The equation for this average is:

$$V_{\text{analog}} = \frac{\sum_{d=-N}^{N} G_d V_d}{\sum_{d=-N}^{N} G_d}$$

where V_{analog} is the voltage of the analog disparity unit, V_d is the disparity reference voltage associated with disparity d, and G_d is a function of the activation of the correlator tuned to disparity d. G_d is set by the current that the correlator at disparity d is supplying to the

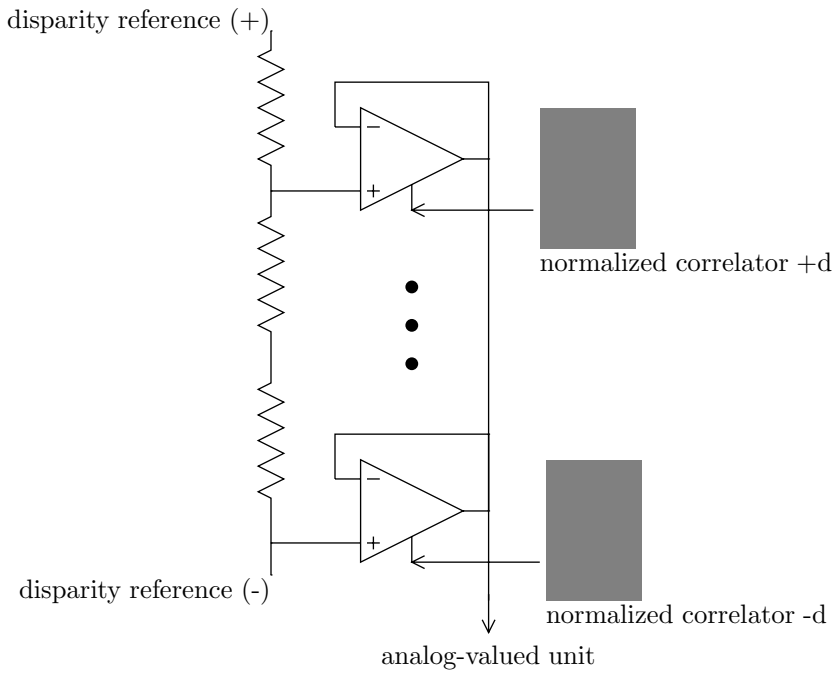

Figure 4.19: Follower aggregation. The correlator at disparity d drives the analog-valued unit voltage to its disparity reference voltage with a normalized driving signal. The normalized signal is the current that the correlator is providing to the WTA. Winning correlators provide more current to the WTA and thus drive the analog-valued unit voltage more strongly.

WTA common line. This current is mirrored into the bias control of the follower by M2 (see Figure 4.13). Since the total current flowing into the common line is set by the WTA bias (Figure 4.17) the total conductance onto the analog-valued unit, $\sum_{d=-N}^{N} G_d$, is constant. In the final state of the system, only one correlator should be active so all the other G_d are zero and the conductance set by the winning correlator is equal to the WTA bias. The final output of the analog disparity unit is simply equal to the voltage that corresponds to the disparity of the active correlator.

In addition to being stimulated by correlators, analog-valued units are resistively coupled to each other across equal cyclopean angle lines. The follower-aggregation circuitry and the resistive coupling between analog-valued units form a one-dimensional resistive network. The coupling between analog units means that their estimate of disparity is no longer purely local, but instead is based on a semi-global average. When the analog-valued units are resistively coupled, their voltage response is influenced by the responses of adjacent analog units. The current from the aggregation circuitry is summed on a capacitor, as

shown in Figure 4.20. The total current into each analog unit is given by this equation:

$$0 = \sum_{k=-N}^{N} G_k \left[\sum_{d=-N}^{N} \frac{G_d(V_d - V_{\text{analog}_n})}{\sum_{k=-N}^{N} G_k} \right] + \frac{V_{\text{analog}_{n+1}} - V_{\text{analog}_n}}{R_{\text{agg}}} + \frac{V_{\text{analog}_{n-1}} - V_{\text{analog}_n}}{R_{\text{agg}}}$$

The space constant of aggregation is set by the magnitude of the resistors and the magnitude of the WTA bias current. However, when there is no retinal input at a particular cyclopean angle, all of the correlators are below threshold. The threshold unit then determines the space constant of the analog-valued network with the follower that drives the analog-valued unit to its resting potential. Typically the limit transistor (Figure 4.17) is set to a lower conductance than the WTA bias, so that the voltage on the analog-valued network decays very slowly where there are no retinal features. In this way, the analog estimate of disparity can propagate for a long way in the network with little attenuation. The distance over which the information travels in the absence of retinal targets is independent of the degree of averaging between targets. Since the conductance at each node in the analog unit array is a function of the activity level of the correlators along its cyclopean angle line, this system is highly stimulus dependent. The spread of activity in the resistive net cannot be calculated without knowing the state of the correlator array.

The qualitative effect of resistive coupling of the analog-valued units is to perform an interpolation in depth. The averaging effect may be undesirable when the change in disparity is too large, indicating a boundary between objects at different depths. In order to investigate this issue, a fuse whose state is controlled by the monocular units described in the next section was included to break the averaging across disparity discontinuities. The principle is that the presence of a binocularly unfused target may represent an area of occlusion that occurs at disparity edges. The monocular units activate a fuse circuit that makes the resistance between two units in the analog-valued network very large. This circuit was invented by Carver Mead. It has the advantage that the activation of the fuse keeps the bias circuit in a well-behaved, well-defined state. Even if the fuse operation of the monocular units is disabled by making the pullup very strong, the current that can be drawn across a disparity discontinuity is limited by the nonlinearity of the resistor. Saturation of the resistor allows a large voltage difference to form across the edge [12].

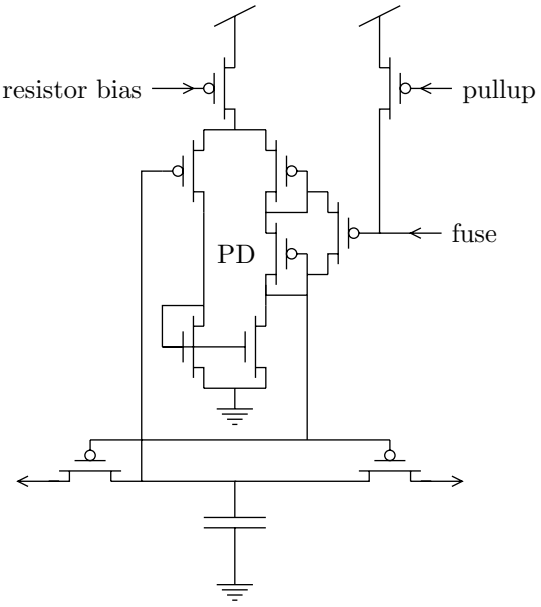

Figure 4.20: The analog-valued units are coupled across lines of cyclopean angle with resistor circuits. The bias circuit sets the magnitude of the resistance by raising the gate voltage of the lateral transistors above the voltage of the central node. The maximum current through the lateral transistors is equal to the current flowing through the diode-connected transistor PD. When the fuse line is pulled low, the current flowing through PD goes to zero so the resistance becomes very large. Although the current through PD is zero, the currents flowing in the bias circuit remain unperturbed so that the circuit does not transit into an undefined state.

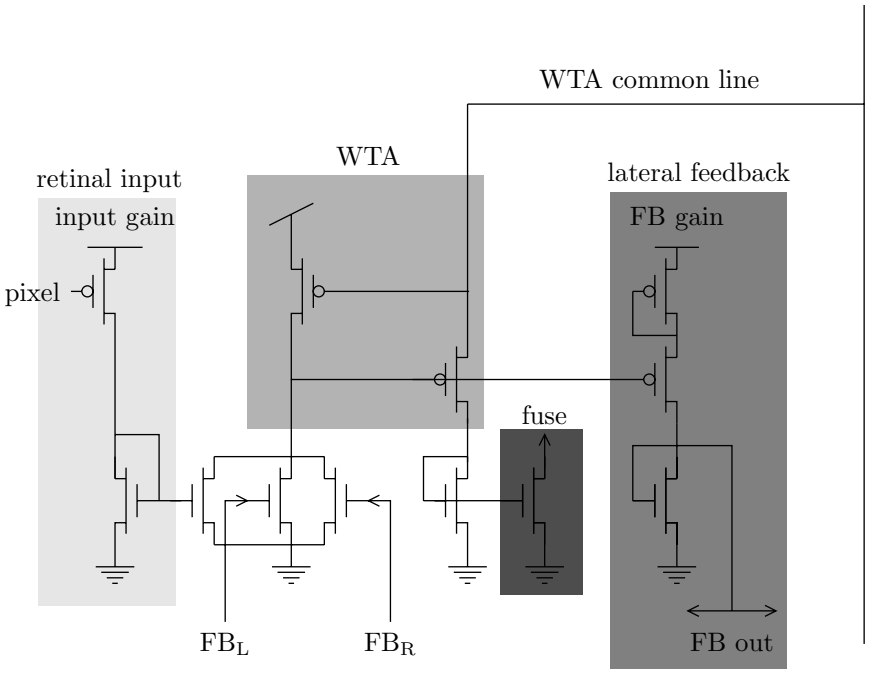

Figure 4.21: Circuit diagram of the monocularly stimulated element.

4.4.5 Monocular Units

In addition to representing binocularly correlated targets, the system also includes state variables that indicate the presence of unpaired, or monocular targets. The circuit diagram for the monocular unit is depicted in Figure 4.21. Monocular units are driven by input from a single pixel. The magnitude of the retinal input to the monocular cell relative to that in the correlator array is set by the input gain control. Monocular units are inhibited by the WTA circuit on the line of cyclopean angle associated with their retinal input. Monocular units excite each other by means of an active lateral excitation circuit. The gain of the positive feedback loop is controlled by the source labelled *FB gain*. As in the correlator array, the magnitude of the positive feedback scales with the magnitude of the retinal input.

The response of these units to stimulation with binocular targets is shown in Figure 4.22. The response is vigorous for disparities that are not fused in the correlator array. The response is inhibited when the disparity range is within the disparity range of the correlator array. This tuning curve resembles that for the tuned inhibitory cells (TI) [36]. Although it is not apparent in this measurement, the inhibition between the monocular units and the correlator array is largest at zero disparity and diminishes at larger disparities. The reason

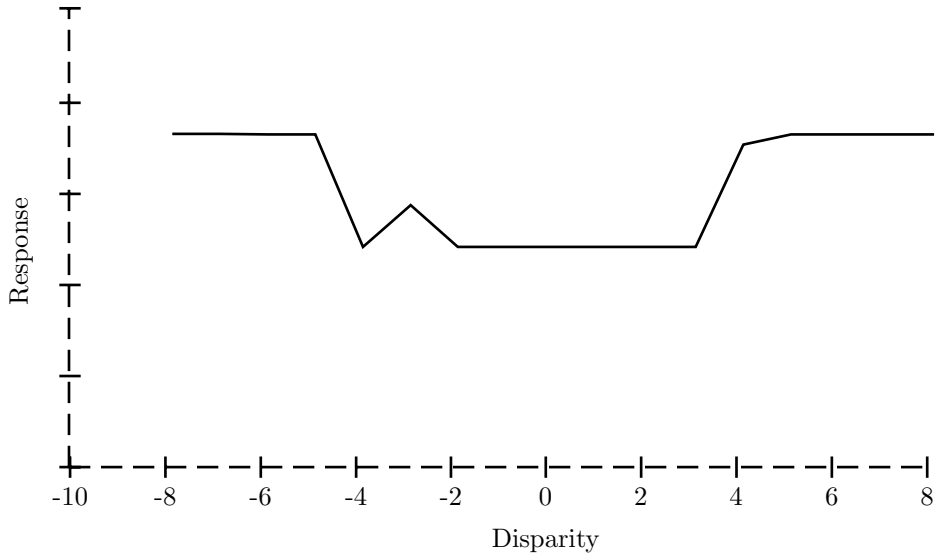

Figure 4.22: Monocular response as a function of disparity.

for this decrement is the geometry of the line of cyclopean angle. The only correlator on the same line of cyclopean angle as the monocular unit that receives input from the same pixel is at zero disparity. Correlators at different disparities that receive input from the monocular unit's pixel must be on different lines of cyclopean angle.

4.5 Analog Psychophysics

In this section, classical stimuli are used to examine the performance of the stereocorrespondence chip and a qualitative understanding of its function is developed. The stimuli were presented as address-events generated by computer. Unless otherwise stated, the parameter settings on the chip were the same for all of the stimulus patterns presented. The input gain of the monocular units was very low except for the measurement of the disparity gradient limit and the random-dot stimulus. The chip is very sensitive to the timing of the presentation of events. The data bus was able to present a single target to the two retinae simultaneously. Care was taken to randomize the pairing of the address-events so that the chip could not use temporal correlation as a method for determining stereocorrespondence. In addition, the order of presentation of targets across the image was randomized as much

as possible, since it was observered that the order of presentation of targets could affect the final state of the system. This occurred because the data presentation rate from the computer was much slower than the convergence time of the chip. To test that stimuli were being fairly evaluated, the positive feedback was disabled until the complete stimulus pattern had established itself on the retina. If the solution obtained in this way was not the same as the one immediately generated upon serial presentation of the stimulus, the order of address generation was changed. The reduction in the gain of the positive feedback has the advantage that it makes most of the false targets visible in the correlator array. The inhibition is unable to determine a winner because the true and false targets receive equal amounts of retinal stimulation.

The performance of the chip on these stimuli suggests interpretations of canonical psychophysical results in terms of the underlying circuitry. The data presented in this section are scanned off the chip using the same method as in previous chapters. The output of the inhibitory units and the analog-values units are captured on an oscilloscope trace and presented separately from the two-dimensional correlator output.

4.5.1 Tilted Surfaces

Many stereo algorithms have been designed to cooperatively solve stereograms. However, they frequently assume assume that the stereogram depicts a surface that is fronto-parallel [24, 7, 34]. The pattern of excitatory interactions between correlators that assist in the determination of true matches are localized to one disparity plane, and so false matches that occur on the same disparity plane will be encouraged instead of true matches that lie along a smoothly varying trajectory.

The stereocorrespondence chip is able to fuse tilted stereograms, as depicted in Figure 4.26 and Figure 4.25. The addition of the analog-valued units allows interpolation to occur in a more natural representation. A comparison of the analog-valued unit output in the false target case (Figure 4.24) and the case in which the positive feedback has been established (Figure 4.26) show that the analog-valued output converges to match the correct solution. The convergence to a correct solution depends to some extent on the depth averaging of the follower aggregation network. Even before the false matches in the correlator array have been suppressed, the analog-valued solution lies in the vicinity of the correct re-

sponse because the false targets on one side of the true solution cancel out the effects of the false targets on the other side of the true solution. Disparity averaging to eliminate false targets has been proposed by Tyler [47]. Analog-valued units have the additional feature that an analog representation of disparity is easier to average across space than the place-valued representation used in the narrowly tuned correlator array. Analog interpolation has been used previously in intensity-gradient disparity algorithms [5]. These algorithms use the spatial derivative of intensity in a single image and the change in intensity at identical spatial locations in the images from the two eyes to compute the disparity. Intensity-gradient algorithms depend on image features with smooth intensity-gradients that are larger than the disparities that you would like to perceive. Like feedforward multi-resolution algorithms, they place a constraint on the perceived disparity and the spatial frequency content of the image. This constraint is not obeyed by one-dimensional dot patterns.

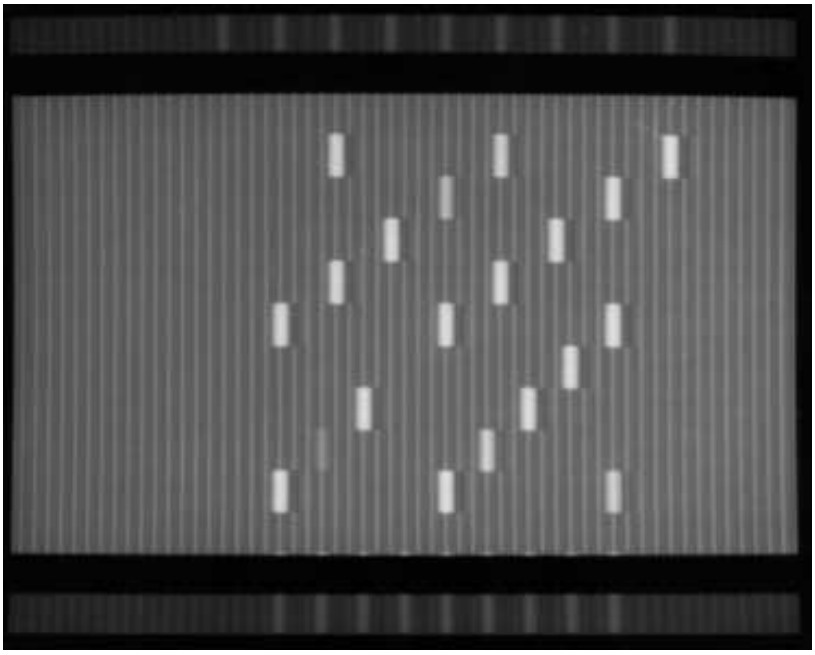

Figure 4.23: The retinal input is depicted at the top and bottom of the figure. Lighter grey indicates activity. The correlator array shows activity at all the possible feature correlations when the positive feedback to the analog-valued disparity units is disabled.

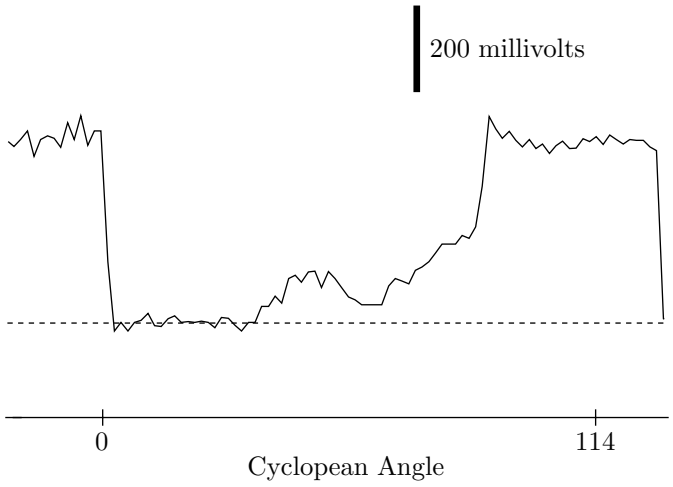

Figure 4.24: Analog-valued disparity output when positive feedback is disabled.

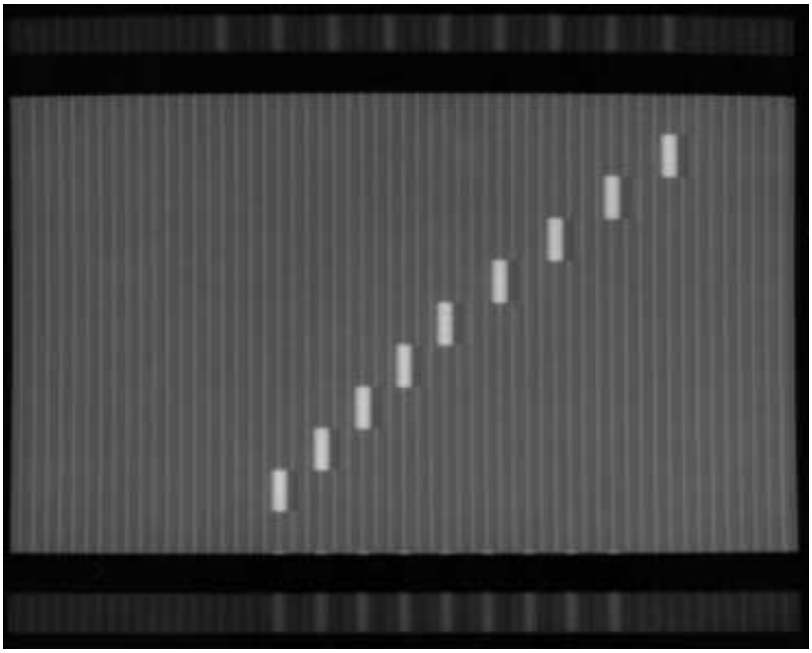

Figure 4.25: Positive feedback from the analog-valued units allows the WTA competition to suppress false targets.

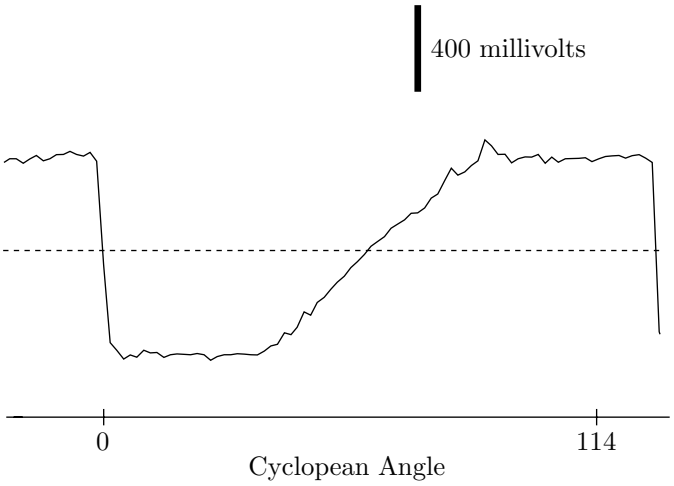

Figure 4.26: Analog-valued disparity output converges to the correct solution when positive feedback to the correlators is enabled.

4.5.2 Interpolation from Unambiguous Endpoints

Mitchison and McKee [30, 31, 32] have shown that the perception of the depth of closely spaced targets can be predicted by interpolation between unambiguous end points. Some of the simpler stimuli used in their experiments, shown in Figure 4.27 through Figure 4.34, were presented to the stereocorrespondence chip. These stimuli consist a periodic array of dots whose stereocorrespondence is ambiguous except at the ends of the array. Like human observers, the chip is able to perceive the stimulus configuration most consistent with the unambiguous endpoints. The analog-valued units perform an interpolation into the target array from the unambiguous endpoints.

The solution to which the chip converges is biased by the average computed by the analog-value units. A comparison of the analog-value unit response shown in Figure 4.28 and Figure 4.32 reveals that the average computed by the analog-value units is closer to the solution at disparity + 2 when the endpoints are set at +1 and that the average is closer to the solution at disparity -2 when the endpoints are at disparity -1. The chip fails to find the correct solution when the array of ambiguous points is too large relative to the averaging distance of the analog-value units. An asymmetry in the design of the bump circuit biases

the solution to larger disparity values.

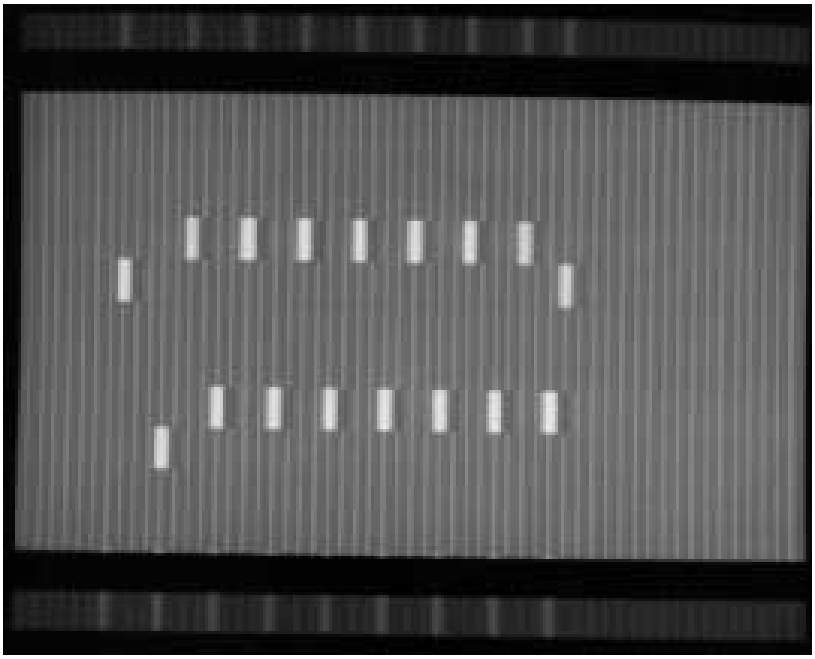

Figure 4.27: Output of the correlator array. Endpoints have disparity +1. Possible targets visible when positive feedback is disabled.

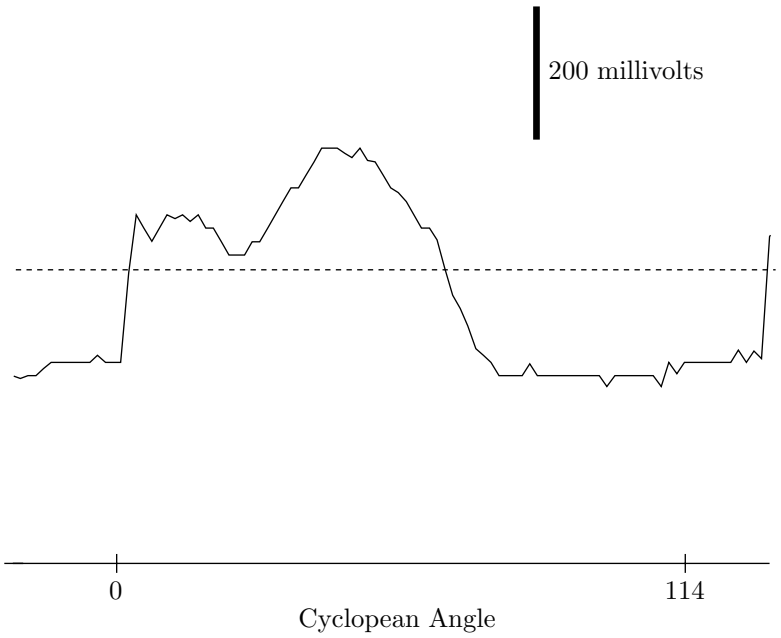

Figure 4.28: Analog-valued disparity output when endpoints have disparity +1 and positive feedback is disabled. The voltage corresponding to zero disparity is shown by the dotted line.

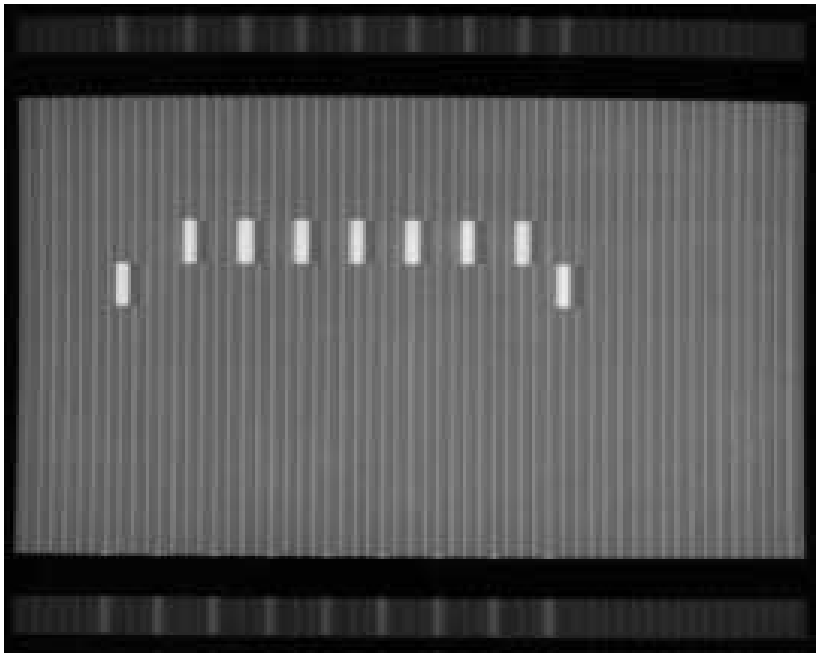

Figure 4.29: Solution at disparity +2 is favored by endpoints at +1 when positive feedback is enabled.

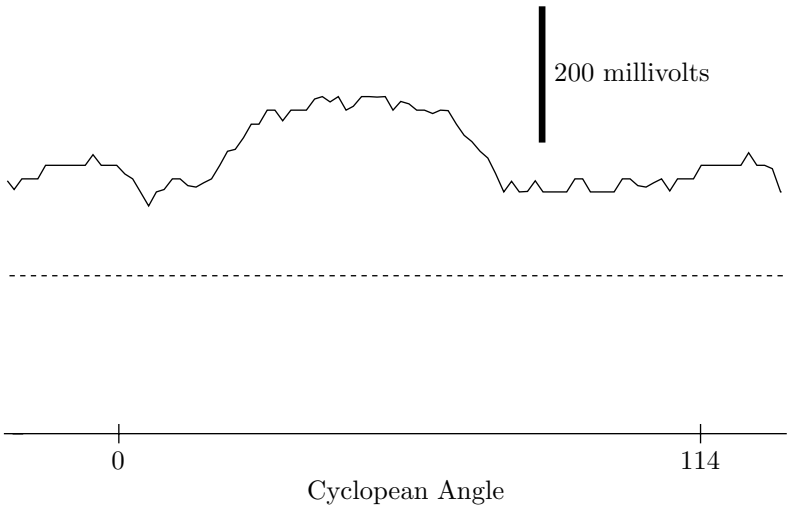

Figure 4.30: Analog-valued disparity output for solution at +2. The voltage corresponding to zero disparity is shown by the dotted line.

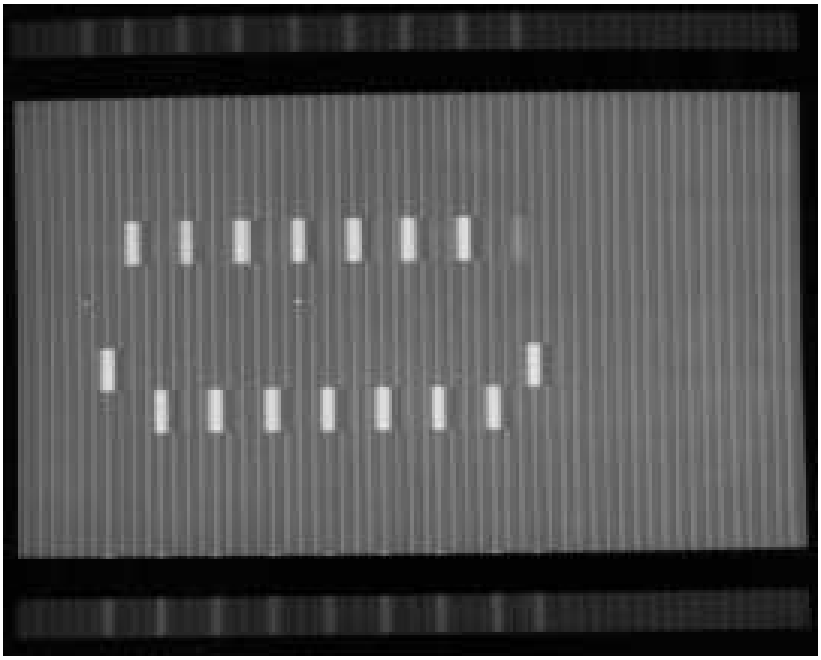

Figure 4.31: Endpoints have disparity -1. False targets visible when positive feedback is disabled.

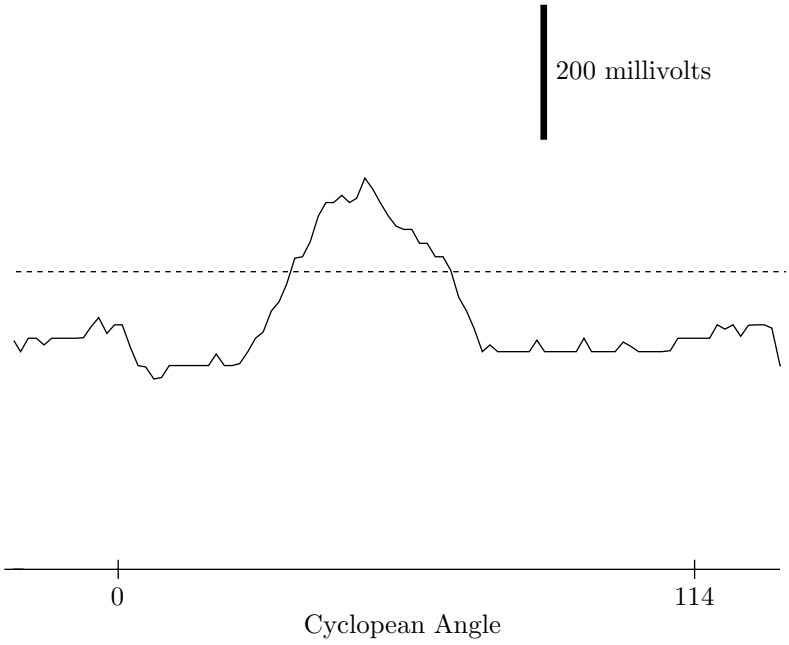

Figure 4.32: Analog-valued disparity output when endpoints have disparity -1 and positive feedback is disabled. The voltage corresponding to zero disparity is shown by the dotted line.

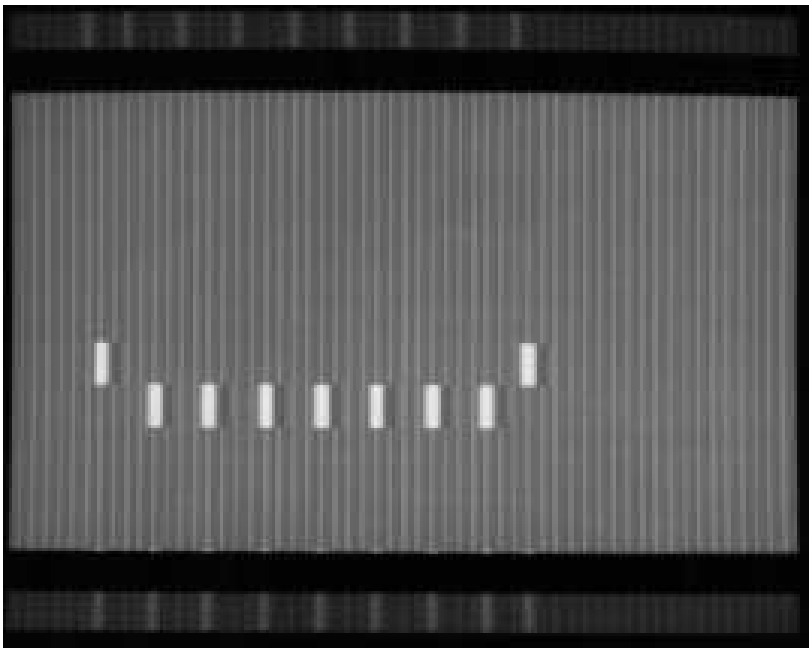

Figure 4.33: Solution at disparity -2 is favored by endpoints at -1 when positive feedback is enabled.

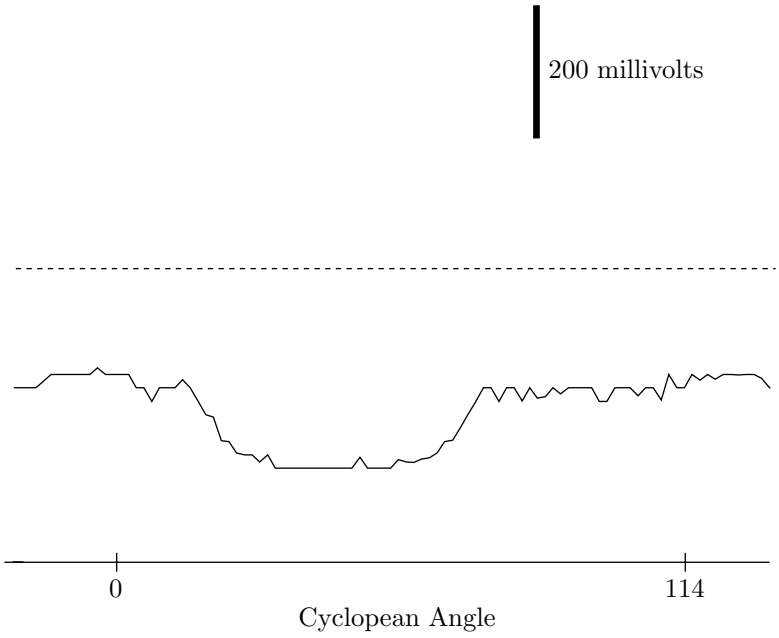

Figure 4.34: Analog-valued disparity output for solution at -2. The voltage corresponding to zero disparity is shown by the dotted line.

4.5.3 Setting Parameters

Identical parameter settings were used for all of these experiments. Convergence was most greatly affected by the spread of activity in the analog-value network and in the inhibitory network. Changes in the final state of the stereocorrespondence chip in response to the same stimulus for different configurations of the analog-value network and the WTA inhibitory network are shown in Figure 4.35 and Figure 4.40 respectively. Due to the non-linearity introduced by the low resting conductance of the analog-network (see Figure 4.17), it is difficult to evaluate whether the inhibition spreads farther than the excitation, although the parameter settings indicate that this is the case. The conductance, and thus the spreading distance of both networks is scaled by the WTA bias voltage (Figure 4.17). A difference in bias voltage of the resistor bias and the WTA bias voltage of 40 millivolts accounts for an approximately $\sqrt{e}$ change in the spreading distance. When the difference between the WTA bias voltage and the resistor bias voltage is large and negative, the averaging distance in the network is large. Activity in both networks must be allowed to spread over several spatial positions in order for the solution to converge properly.

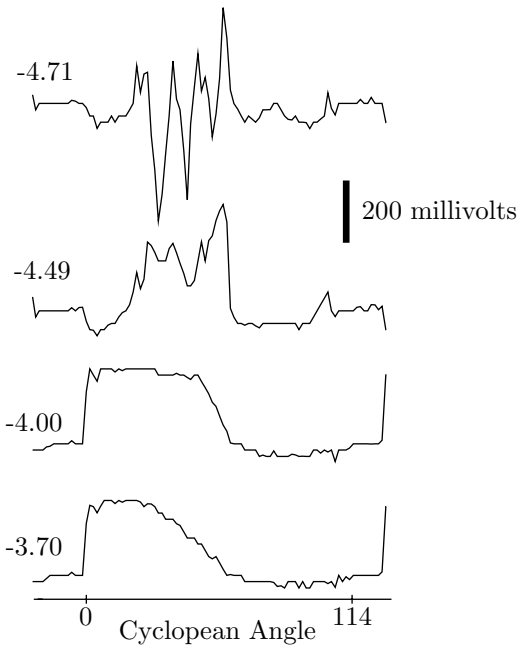

Figure 4.35: Activity in the analog-valued array in response to a random dot stereogram for four different coupling resistor strengths. The voltages on the resistor bias circuits are shown next tot each trace. The WTA bias voltage was -4.26 volts. The WTA-coupling resistor bias was -3.91. The chip was p-well and voltages are reported according to the grounded substrate convention.

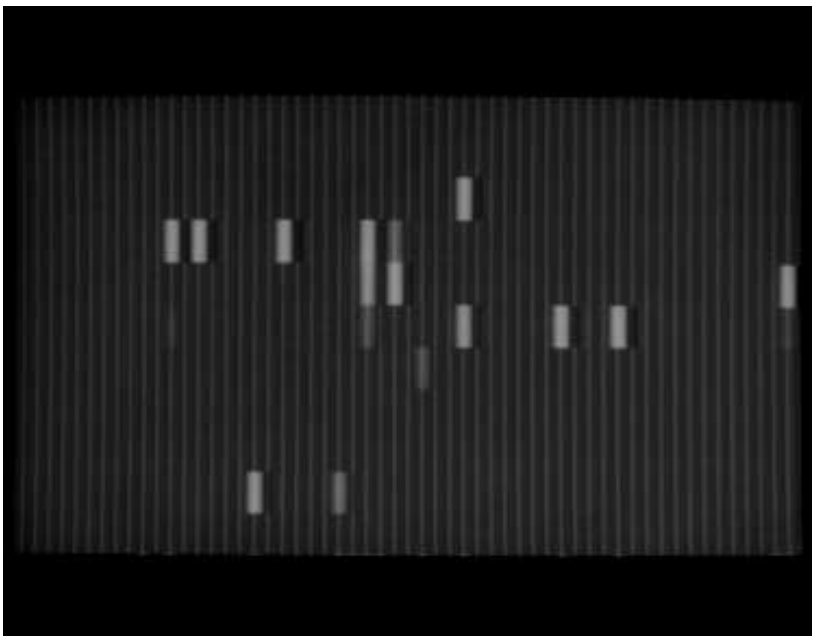

Figure 4.36: Activity in the correlator array in response to a random dot stereogram with analog-valued-unit-coupling resistor bias voltage at -4.71. See Figure 4.35. This solution shows little false target supression.

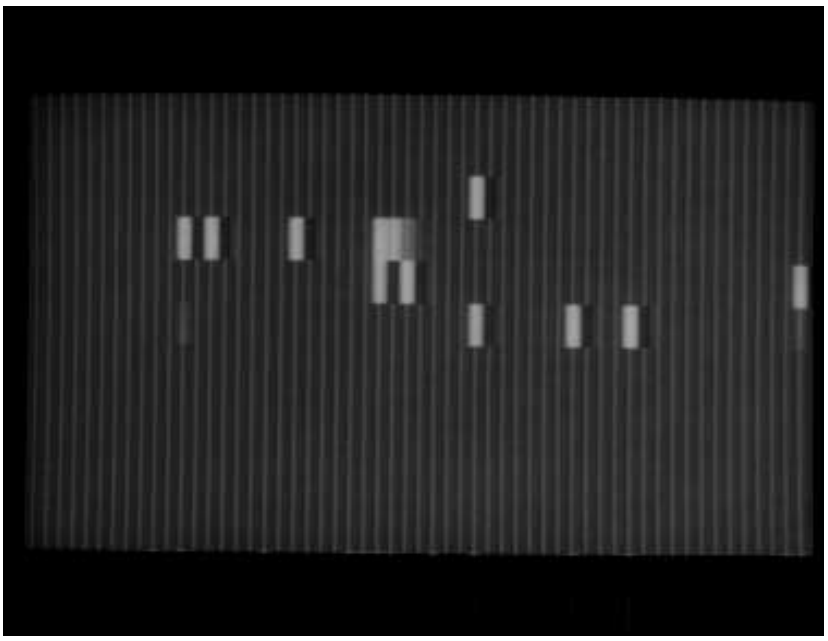

Figure 4.37: Activity in the correlator array in response to a random dot stereogram with analog-valued-unit-coupling resistor bias voltage at -4.49. See Figure 4.35. This solution shows some false target suppression.

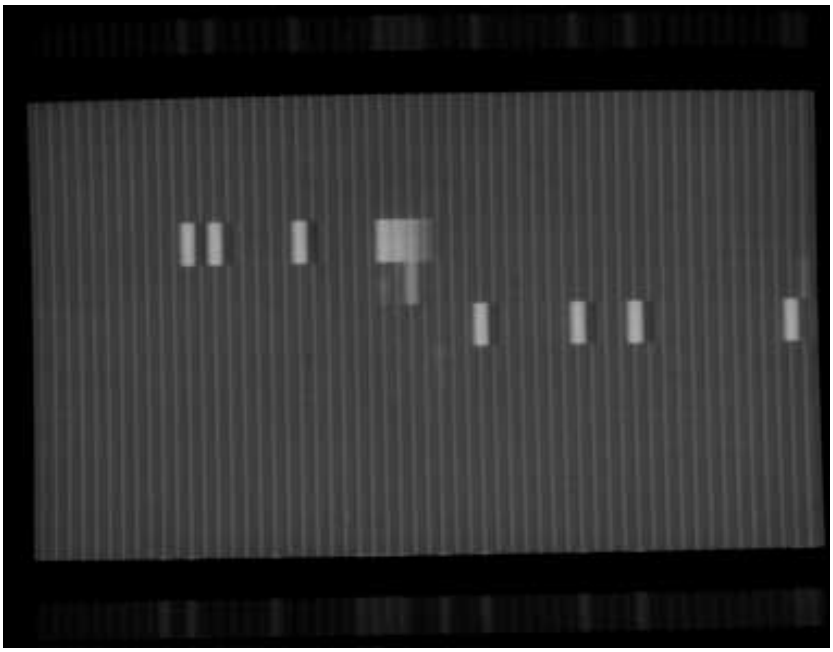

Figure 4.38: Activity in the correlator array in response to a random dot stereogram with analog-valued-unit-coupling resistor bias voltage at -4.00. See Figure 4.35. This solution is optimal.

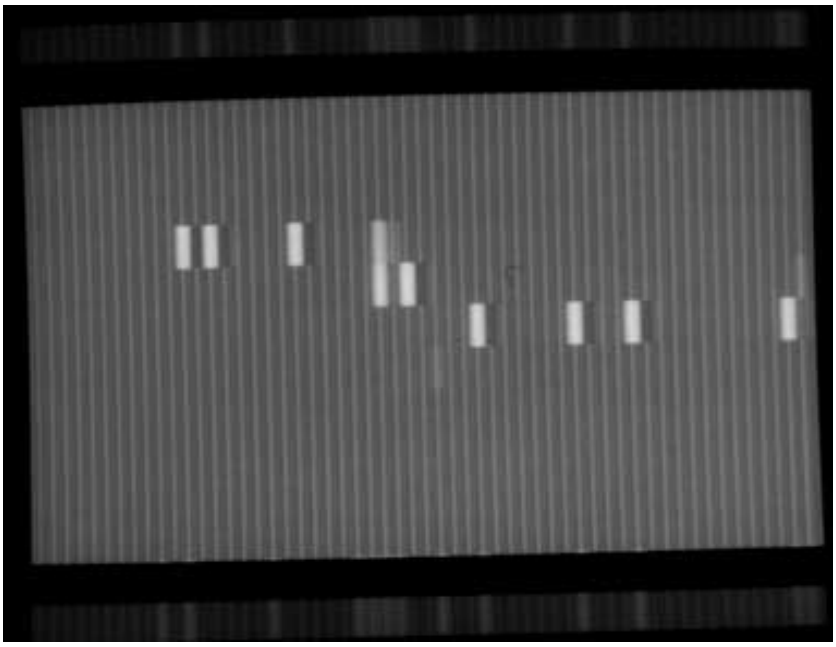

Figure 4.39: Activity in the correlator array in response to a random dot stereogram with analog-valued-unit-coupling resistor bias voltage at -3.70. See Figure 4.35. This solution is smoothed too much across the disparity discontinuity.

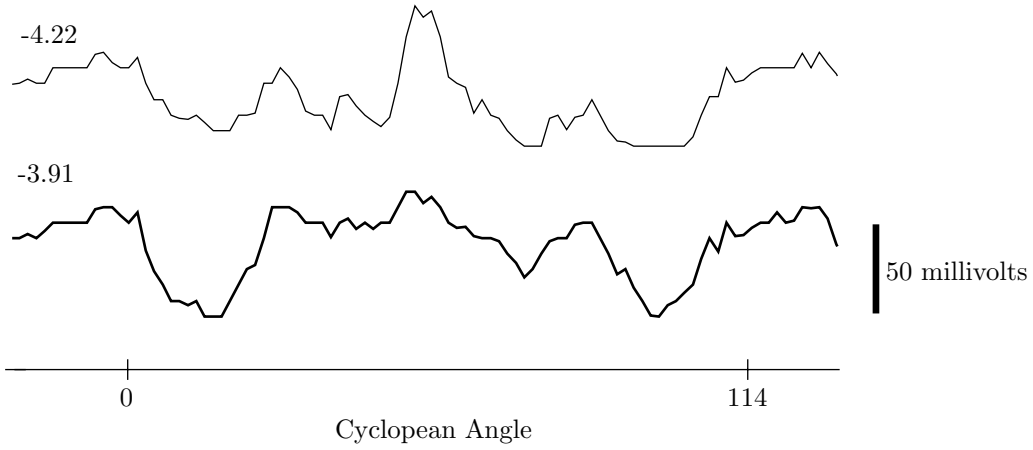

Figure 4.40: Profile of activation of the WTA common-line array for WTA-coupling resistor bias voltages -4.22 and -3.91 volts. The WTA bias voltage was -4.26 volts.

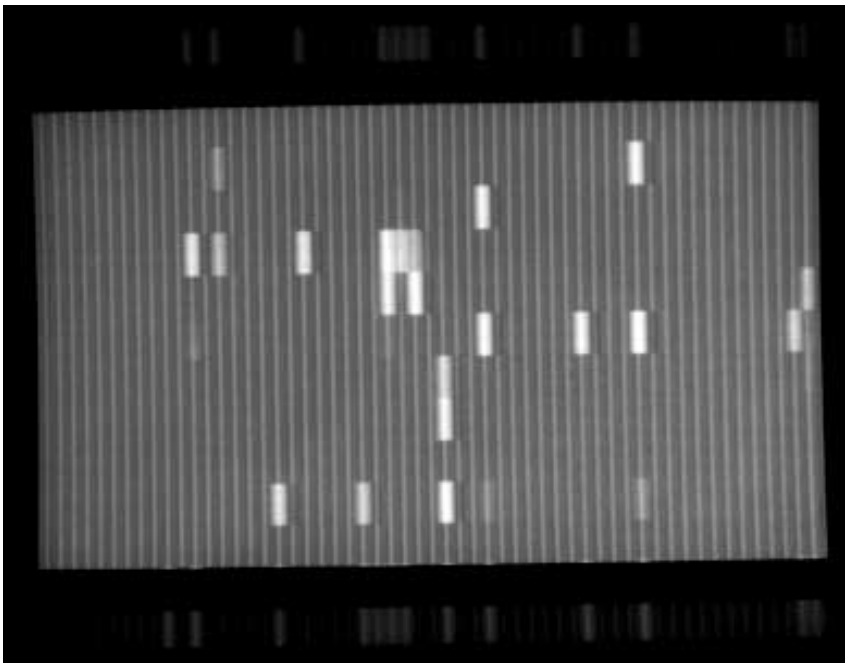

Figure 4.41: Activity in the correlator array in response to a random dot stereogram with WTA-coupling resistor bias voltage at -3.70. Compare with Figure 4.38, in which the WTA-coupling resistor bias voltage was -3.91. This solution is undesirable.

4.5.4 Disparity Gradient Limit

The disparity gradient limit was first described by Burt and Julesz [4]. The disparity gradient is defined as the binocular disparity between two targets divided by their binocular separation. The limiting case in which the two targets are aligned along the same line of sight in one eye, and so appear as a single target, but are visible separately from the other eye's perspective is known as Panum's limiting case and corresponds to a disparity gradient of two. Objects with a higher disparity gradient than this will appear in opposite order on the two retinae. The disparity gradient for fusion in humans is about one. Figure 4.42 depicts the forbidden area resulting from a disparity gradient of two and one in terms of the stereocorrespondence chip architecture. The black square in the center of the correlator array indicates the presence of a target at zero disparity. The shaded areas represent the positions of matches which would be prohibited by the presence of a match at zero disparity, if the chip generated the same disparity gradient limit as the human visual system.

The disparity gradient limit has been the basis for several stereo-matching algorithms [40, 41]. These algorithms assumed that a forbidden region existed around each possible match and chose a set of matches which were consistent with each other. However, the mechanism by which the forbidden region is generated has not received much attention. The disparity gradient limit for fusion arises in this chip through a combination of inhibition that includes the correlators and the monocular cells, and excitatory interactions between the correlators and the analog-valued units.

Several features of these interactions are revealed by data from the stereocorrespondence chip, depicted in Figure 4.43. The limit was measured by placing one target at zero disparity and presenting the additional target. When the disparity gradient was greater than two, a new solution arose that paired one retinal feature of the original zero disparity target with the other retinal feature of the new target. The match at zero disparity disappeared and two new matches of similar near or far disparity appeared. When the disparity gradient was equal to two, the fusion of the new target was often inhibited. (Positions of inhibition are shown as medium grey squares). However, inhibition is ineffective when the positive feedback between the two targets is strong, so that one cannot suppress the other. Thus two adjacent targets on one retina both match with a single target on the other retina a produce a blurred target that activates two correlators, one at zero disparity, the other at

± 1. This output may be interpreted as a target at ± 0.5 and so represents a reasonable interpolation of the disparity of a single blurred target. When the disparity gradient is between 1 and 2, the target is often not fused or incompletely fused. When the target is not fused, it appears in the monocular unit array. In a state of incomplete fusion, the correlator that should represent the match flickers due to oscillation in the circuit. This oscillation may be analogous to the lusterous quality perceived by humans viewing improperly fused targets. Often the monocular cell associated with the new target is partially activated.

The monocular cells extend the disparity gradient limit. If the input gain to the monocular cells is decreased, they cannot compete as effectively with the correlator array, and the disparity gradient limit of the chip diminishes. The monocular units help suppress the formation of a fused match because the correlator that would represent the match is flanked by an active correlator at zero disparity, and a monocular cell on the other side. The combined inhibition from these two units and the lack of positive feedback is sufficient to suppress the fused target. Further experiments must be performed to tease apart the interaction of excitation and inhibition in the stereocorrespondence network. A disparity gradient could be enforced by many connectivity patterns in a network. However, the results in this biologically plausible network suggest that the tuned inhibitory neurons may play a significant role in establishing the psychophysically measured disparity gradient limit.

4.5.5 Occlusion

Occlusion is a significant clue to depth. Nakayama has shown that occlusion alone is able to generate the perception of a raised surface and also generates a sharp edge at the boundary of a surface that has targets on it [33]. The monocularly activated units in this stereocorrespondence algorithm provide a natural opportunity to explore the role of occlusion in stereocorrespondence.

Few network models have considered the role of occlusion in the computation of stereocorrespondence. The question of whether occlusion is computed before, after or during stereocorrespondence is unanswered. Intensity gradient stereo algorithms based on Markov-random field models have used fuses to eliminate disparity interpolation across discontinuities [5]. While this approach cannot generate depth from occluded areas alone, it is easy to implement in a VLSI resistive network.

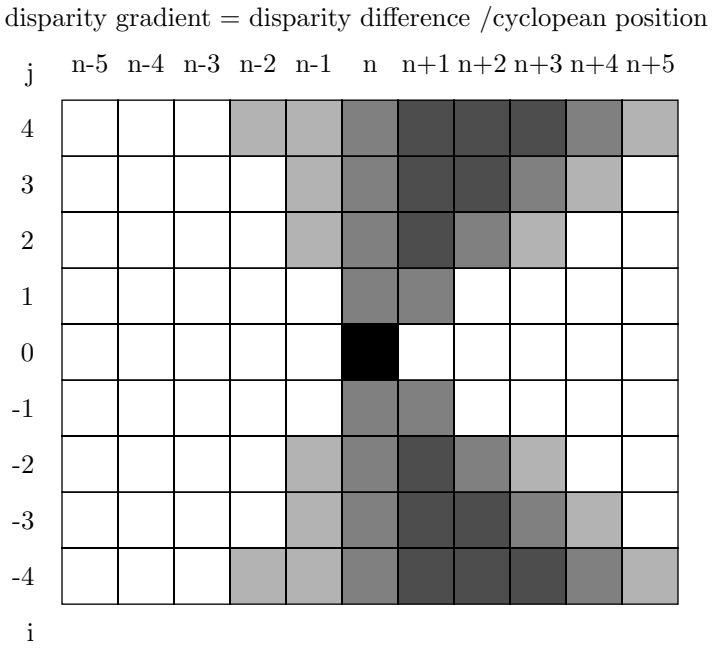

Figure 4.42: Disparity gradient limit. A:Idealized representation showing disparity gradient greater than two (ordering constraint violated) in dark grey, disparity gradient equals two (Panum's limiting case) in medium grey, and disparity gradient greater than or equal to one (human fusional limit) in light grey.

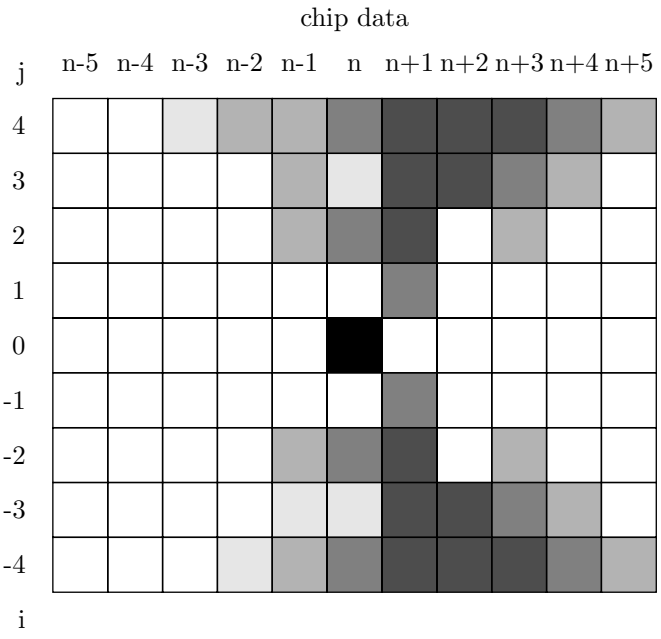

Figure 4.43: Disparity gradient limit. Data from the chip showing areas of target reordering (numbers indicate perceived targets), no fusion (medium grey), and incomplete fusion (light grey). The area of target reordering is equivalent to disparity gradient greater than two. The areas of incomplete fusion nearly correspond to a disparity gradient limit of one.

I have incorporated fuses to interpolation in the analog unit array that are broken by activation of a monocular cell. This strategy improves the disparity estimate computed by the analog-valued units in response to a one-dimensional random dot stereogram, shown in Figure 4.44 and Figure 4.46. The stimulus has a single disparity discontinuity and a single unpaired feature on the lower retina. The disparity estimate computed by the analog-valued units smooths over the discontinuity when the fuse is disabled (Figure 4.47). When the fuse is enabled, however, the disparity solution at +2 is filled in up to the discontinuity (Figure 4.48).

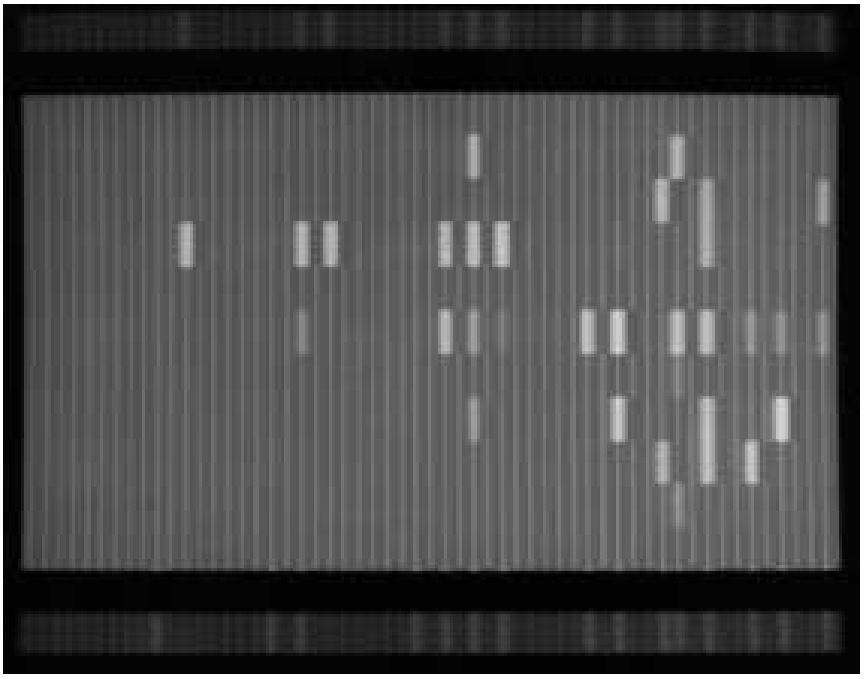

Figure 4.44: Activity in the correlator array in response to a random dot stereogram with single disparity discontinuity and occluded target. Positive feedback is disabled so false targets are visible. The input gain of the monocular units was temporarily reduced so that the monocular units did not suppress the unreinforced targets in the correlator array.

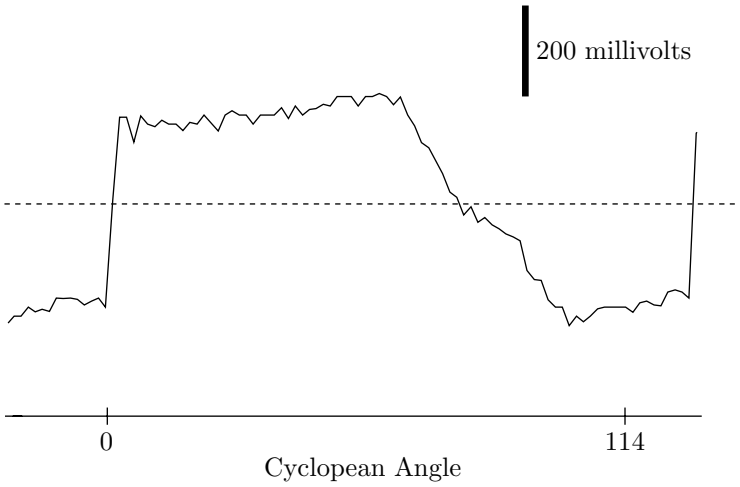

Figure 4.45: Response of analog-valued units with positive feedback disabled. Analog-valued units compute the average disparity of all the active correlators.

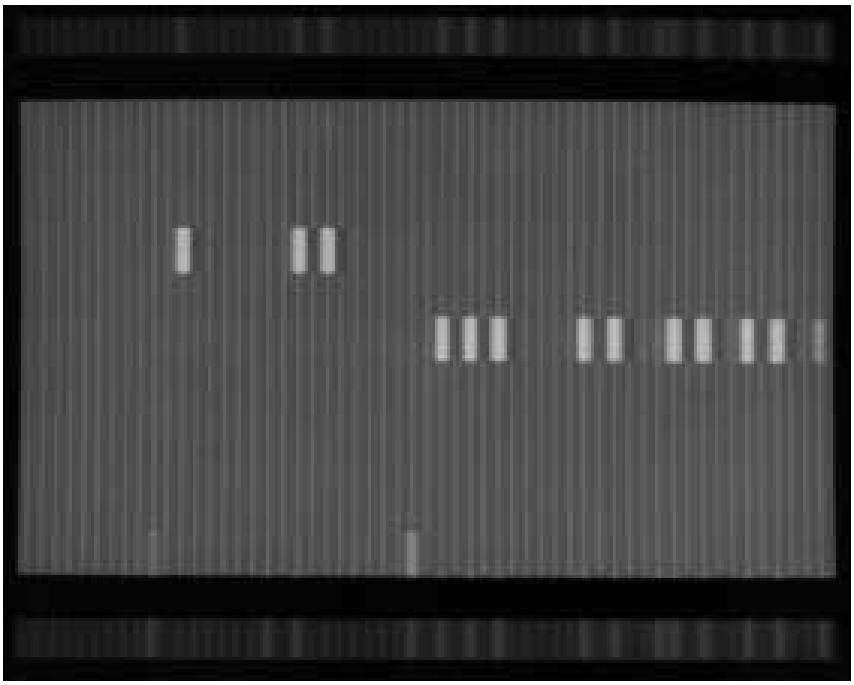

Figure 4.46: Response of correlator array to random dot stimulus when positive feedback is enabled. Compare to Figure 4.44.

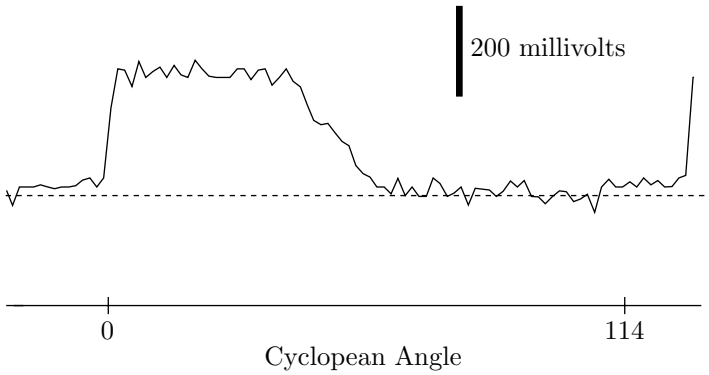

Figure 4.47: Random dot stereogram with single disparity discontinuity and occluded target. Positive feedback is enabled and fuse is disabled. The analog-valued unit output is smoothed across the discontinuity and the occluded target visible as activity in the lower array of monocular units in Figure 4.46. The position of the occlusion had to be determined in an ambiguous string of four contiguous features.

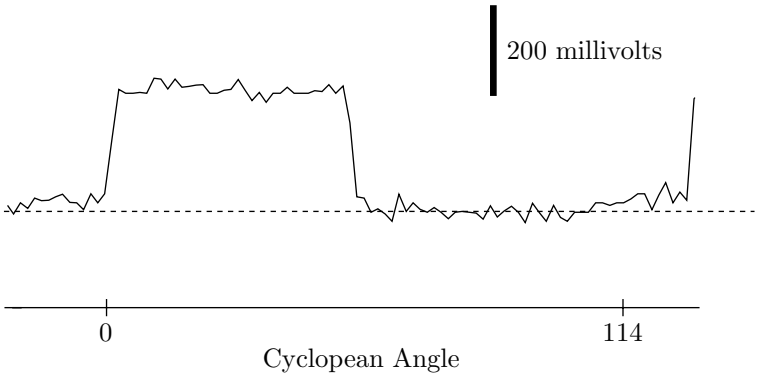

Figure 4.48: Random dot stereogram with single disparity discontinuity and occluded target. Positive feedback and monocular fuse circuits are enabled. An occluded target breaks the interpolation in the analog-valued unit array and allows the solution to be filled in at disparity +2 up to the occlusion event, although there are no additional targets on the surface.

4.6 Discussion

The analog stereo matcher suppresses false matches by a collective interaction that requires a transformation of representation. This transformation allows the generation of stable states of the system in which the units have analog-valued outputs. This feature distinguishes the stereomatcher from the traditional Hopfield network in which the state of the system is pushed to corners of the hyperspace by the positive feedback responsible for reconstruction of the stored memory [9]. However, this stability arises from a neurobiologically unfounded type of "synaptic" interaction; namely, synapses whose magnitude is a non-monotonic function of the value of the presynaptic neuron. This type of interaction has previously been proposed for the construction of radial basis functions [39]. The radial basis function has mainly been used in the context of interpolation in a high dimensional input space, rather than in the context of associative memory. Although the radial basis function may not be the function of a single synapse, it is possible that a network of neurons with proper traditional synaptic connections might compute such a non-monotonic function. The success of this algorithm and circuit at solving the stereocorrespondence problem as well as the successes of radial basis function networks, motivates the search for such a network architecture.

Although this algorithm has used a single "feature" based input, this is not the ideal form. It is likely that there are several interacting populations of neurons that are tuned for different orientations, directions and spatial frequencies. The analog valued units of this algorithm most likely correspond to disparity units tuned to low spatial frequencies. The analog valued units should not be seen as derived or secondary to the finely tuned correlator array. The algorithm was designed to operate specifically on tasks in which the low spatial frequency information had been removed. I believe that the algorithm represents only a small fraction of the interactions normally taking place in cortical computation of stereocorrespondence. The general principle is that the state of cortex must be consistent with itself over small distances in all its dimensions. This algorithm embodies the requirement for consistency between narrowly tuned units and broadly tuned analog units at one physical location and consistency between analog units at different spatial locations. This requirement for consistency was sufficient to perform stereocorrespondence, even when the

low spatial frequency units receive no retinal input.

One of the novel features of the stereocorrespondence chip is the incorporation of monocular units into the cooperative computation. However, this experiment has had an ambiguous result; the role of the monocular units in binocular fusion remains mysterious. Measurement of the disparity gradient limit of the chip indicates that the monocular units play a significant role in the formation of the forbidden zone. However, activation of the monocular units during the normal fusion process in general seemed to produce a negative effect, even if the fuse governing analog-value unit interpolation was disabled. This effect was manifest as an unwillingness of the network that included monocular units to find solutions near zero disparity. It arises because the competition between the zero disparity correlators and the monocular units is direct. The difficulty may well be a result of the inhomogeneity of the positive feedback mechanisms in the array. In order to get the monocular units into the competition, their input gain had to be adjusted to a high level, relative to the retinal input level of the correlator array. This increased feedforward gain places the monocular cells at an unfair advantage early in the convergence process. The development of a more consistent cooperative framework, in which all units participate on an equal footing, will be necessary before the role of monocular units in the computation of stereocorrespondence can be evaluated.

The necessity for a common framework for neural interaction is counterbalanced by the need to limit the possibilities. Investigators in neural networks have adopted a strategy of employing predefined architectures and fixed neuronal representations. For example, Qian and Sejnowski [42] have shown that a back-propagation network initially configured according to the architecture of the cooperative stereocorrespondence algorithms is capable of learning to fuse random dot stereograms and can even handle transparency. Lehky and Sejnowski [20] have demonstrated depth interpolation and representation of transparency using a population encoding of disparity that had both narrowly tuned and broadly tuned features. However, their analysis assumed that the units tuning curves were given and did not address issues of finding stereocorrespondence in random dot patterns. Since the analog circuit relates interpolation and stereocorrespondence, one logical extension of the circuit would be to add a number of analog-valued units as are present in Lehky and Sejnowski's representation and explore the perception of transparency. Transparent surfaces

would gather consistent patterns of activation. As long as both surfaces had equal levels of consistency, they would both rise above the adaptive threshold level set by the common line inhibition of the WTA.

The stereocorrespondence chip can be expanded in many directions. One generalization of the circuit is the incorporation of adaptation in the neurons. I believe that adaptation will allow the monocular or tuned-inhibitory units to exhibit binocular rivalry [18]. The psychophysics of long-term adaptation (i.e. to oriented gratings) and binocular rivalry has been studied by Lehky and Blake [21]. They conclude that binocular rivalry must occur before binocular fusion. However, these experiments do not rule out the possibility of rivalry occurring in the monocular cells that are participating in the fusional process, as is the case for the monocular cells in this algorithm.

Another interesting avenue of investigation is the use of temporal correlation in stereopsis. I have observed that the stereocorrespondence problem that the chip needs to solve is greatly simplified by the addition of temporal correlation of the address-events supplied to the two retinae. The addition of analog delay structures, perhaps based on dendritic morphology, will allow the exploration of motion interpolation and stereopsis. Little research has been done in time-based algorithms for stereopsis because it is difficult to simulate temporal functions using traditional methods. Previous stereomatching chips [22] have used time-derivatives as the input for stereo matching. The use of time derivatives improved the performance of the matcher by amplifying the input signal relative to the offsets. It is known that time is an intrinsic part of the disparity computation in natural systems. Perceptual psychologists have shown that binocular time delay and disparity can be substituted for each other in moving stimuli [3]. Binocular time delay has been used to characterize disparity sensitive neurons in visual cortex [8]. Signals that are time delayed between the two eyes result from motion in a complex environment in which surfaces occlude one another [45]. The address-event communication protocol facilitates investigation of these issues since it does not introduce the kind of temporal aliasing as does a sequential-scanning multiplexing method.

4.7 Summary

The stereocorrespondence chip embodies a new algorithm, intermediate between multi-resolution algorithms and cooperative algorithms. It can find the correct stereocorrespondence in an one-dimensional random dot stereogram depicting front-parallel or oblique surfaces. Its performance on a number of stimuli that have been used in psychophysical research resembles the performance of the human subjects. Furthermore, the disparity-tuning curves of several of the electrical units of the circuit are similar to the disparity-tuning curves of stereo-tuned neurons in primate cortex. Thus, the stereocorrespondence chip links electrophysiology with psychophysical behavior and computational function.

The stereocorrespondence chip has opened a number of avenues for future research in the fields of neurophysiology, computational neurobiology and engineering. The performance of the chip suggests that the disparity tuning characteristics of the disparity flat cell and the tuned inhibitory cell may be a result of network interactions. The algorithm used by the chip motivates the search for neurally plausible architectures that perform a transformation of representation between place-valued and analog-valued encoding. The rapid and robust function of the stereocorrespondence chip raises the possibility of building an analog multi-chip system to compute stereocorrespondence in real time based on the address-event communication protocol. Although the directions and possibilities for future research are many, they all lead towards the development of a vocabulary of realizable circuit elements that form a rich and self-consistent framework for the synthesis of architecturally differentiated neural structures.

References

[1] Ballard, D.H. (1986) Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences* **9**, pp. 67–120.

[2] Blake, R., and Wilson, H. (1991) Neural models of stereoscopic vision. *TINS.* **14**, pp. 445-452.

[3] Burr, D.C., and Ross, J. (1979) How does binocular delay give information about depth? *Vision Research*, **19** pp. 523–532.

[4] Burt, P., and Julesz, B. (1980) A disparity gradient limit for binocular fusion. *Science*, **208**, pp. 615-617.

[5] Chhabra, A., and Grogan, T. (1989) Depth from stereo: variational theory and a hybrid analog-digital network. *SPIE Vol.1076 Image Understanding and the Man-Machine Interface II* conference proceedings January 1989, Los Angeles. pp. 131–138.

[6] Cooper, M.L., and Pettigrew, J.D. (1979) A neurophysiological determination of the vertical horopter in the cat and owl. *Journal of comparative Neurology*, **184**, pp. 1–26.

bibitemdelbruck:89 Delbrück, T., and Mead, C.A. (1989). "An electronic photoreceptor sensitive to small changes in intensity," In Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems* 1., pp. 712–727, San Mateo, CA: Morgan Kaufman.

[7] Dev, B. (1975) Perception of depth surfaces in random-dot stereograms: A neural model. *Int. J. Man-Machine Studies*, **7**, pp. 511-528.

[8] Gardner, J.C., Douglas, R.M., and Cyander, M.S. (1985) A time-based stereoscopic depth mechanism in the visual cortex. *Brain Research*, **328**, pp. 154–157.

[9] Hopfield, J. (1982) *Proceedings of the National Academy of Science, U.S.A.*, **79**, p. 2554.

[10] Horn, B. (1986) *Robot Vision.* Cambridge, MA: MIT Press.

[11] Hubel, D., and Wiesel, T. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**, pp. 106–154.

[12] Hutchinson, J., Koch, C., Luo, J., and Mead, C. (1988) Computing motion using analog and binary resistive networks. *IEEE Computer*, March pp. 52–63.

[13] Julesz, B. (1960). Binocular depth perception of computer-generated patterns. *Bell Syst. Tech. J.*, **39**, pp. 1125–1162.

[14] Julesz, B. (1971) *Foundations of Cyclopean Perception.* Chicago, Illinois: The University of Chicago Press.

[15] Julesz, B. (1986) Stereoscopic vision. *Vision Research*, **26**, pp. 1601-1612.

[16] Kanade, T. editor (1987) *Three-dimensional Machine Vision.* Boston, MA: Kluwer Academic Publishers.

[17] Lazzaro, J., Ryckebusch S., Mahowald, M.A., and Mead, C.A. (1989) Winner-Take-All circuits of $O(n)$ complexity. In Touretzky, D.S. (ed), *Advances in Neural Information Processing Systems* 1. pp. 703–711, San Mateo, CA: Morgan Kaufman.

[18] Lehky, S. (1988) An astable multivibrator model of binocular rivalry. *Perception*, **17** pp. 215-228.

[19] Lehky, S., Pouget, A., and Sejnowski, T. (1990) Neural models of binocular depth perception. *Cold Spring Harbor Symposia on Quantitative Biology* volume LV, Cold Spring Harbor Laboratory Press. pp.765–777.

[20] Lehky, S., and Sejnowski, T. (1990) Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience*, **10**, pp. 2281-2299.

[21] Lehky, S., and Blake, R. (1991) Organization of binocular pathways: Modeling and data related to rivalry. *Neural Computation*, **3**, pp. 44-53.

[22] Mahowald, M., and Delbrück, T. (1989) Cooperative stereo matching using static and dynamic image features. In C. Mead and M. Ismail (Eds.), *Analog VLSI Implementation of Neural Systems*, (pp. 213–238) Kluwer Academic Publishers, Boston.

[23] Marr, D., Palm, G., and Poggio, T. (1978). Analysis of a cooperative stereo algorithm. *Biological Cybernetics* **28:** 223–239.

[24] Marr, D., and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science* **194:** 283–287.

[25] Marr, D. (1982) *Vision* New York: W. H. Freeman.

[26] Mayhew, J. and Frisby, J.P. (1981). Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence* **17:** 349–385.

[27] McKee, S., Levi, D., and Bowne, S. (1990) The imprecision of stereopsis. *Vision Research.* **30:** 1763–1779.

[28] McLean, J., and Palmer, L. (1989) Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat. *Vision Research***29:**675-679.

[29] Mead, C.A. (1989). *Analog VLSI and Neural Systems.* Reading, MA: Addison-Wesley.

[30] Mitchison, G., and McKee, S. (1985) Interpolation in stereoscopic matching. *Nature***315:**402–404.

[31] Mitchison, G., and McKee, S. (1987) The resolution of ambiguous stereoscopic matches by interpolation. *Vision Research* **27:**285-294.

[32] Mitchison, G., and McKee, S. (1987) Interpolation and the detection of fine structure in stereoscopic matching. *Vision Research* **27:** 295-302.

[33] Nakayama, K., and Shimojo, S. (1990) Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research.***30:** 1811–1825.

[34] Nelson, J. (1975) Globality and stereoscopic fusion in binocular vision. *Journal of Theoretical Biology***49:** 1-88.

[35] Nishihara, H. (1984). Practical real-time imaging stereo matcher. *Optical Engineering* **23:** 536–545.

[36] Poggio, G. (1984). Processing of stereoscopic information in primate visual cortex. In Edelman, G. M., Gall W. E., and Cowan, W. M. (eds), *Dynamic aspects of neocortical function*, pp. 613–635, New York: John Wiley & Sons.

[37] Poggio, G. and Poggio, T. (1984). The analysis of stereopsis. *Annual Review of Neuroscience* **7:** 379–412.

[38] Poggio, G., Gonzalez, F., and Krause, F. (1988) Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *Journal of Neuroscience* **8:** 4531–4550.

[39] Poggio, T., and Girosi, F. (1989) A theory of networks for approximation and learning. *A.I. Memo No.1140 M.I.T. Artificial Intelligence Laboratory.*

[40] Pollard, S., Mayhew, J., and Frisby J. (1985) PMF: A stereo correspondence algorithm using a disparity gradient limit. *Perception***14:**449-470.

[41] Prazdny K. (1985) Detection of binocular disparities. *Biological Cybernetics***52:**93-99.

[42] Qian, N., and Sejnowski, T. (198?) Learning to solve random-dot stereograms of dense and transparent surfaces with recurrent backpropagation. *Connectionist Models Summer School* San Mateo, CA: Morgan Kaufmann Publishers. pp.435–443.

[43] Richards, W. (1975). Visual space perception. In Carterette, E. and Friedman, M. (eds.), *Handbook of Perception V: Seeing* pp. 351–386, New York: Academic Press.

[44] Schor, C., Wood, I., and Ogawa, J. (1984) *Vision Research* **24:** 661-665.

[45] Shimojo, S., Silverman, G.H., and Nakayama, K. (1985). An occlusion-related mechanism of depth perception based on motion and interocular sequence. *Nature* **333:** 265–268.

[46] Sperling, G. (1970) Binocular vision: a physical and a neural theory *American Journal of Psychology* **83:** 461-534.

[47] Tyler, C. (1977) Stereomovement from interocular delay in dynamic visual noise: A random spatial disparity hypothesis. *American Journal of Optometry and Physiological Optics* **54:** 374-386.

Chapter 5

Conclusion

This thesis has described the development and testing of a simple artificial visual system fabricated using analog CMOS VLSI. This visual system is composed of three novel subsystems. A silicon retina that transduces light and performs signal processing of kind similar to that observed in simple vertebrate retinae. A stereocorrespondence chip uses bilateral retinal input to estimate the location of objects in depth. A silicon optic nerve provides a communication system between chips by a method that preserves the idiom of action potential transmission in the nervous system. Each of these subsystems illuminates various aspects of the relationship between VLSI analogs and their neurobiological counterparts.

The silicon retina described in chapter 2 is a classical example of the unity of form and function in evolved systems [6, 1]. The purposive function of the retina, to provide relevant visual information to the organism, is performed in the context of physical limitation, such as finite communication bandwidth. The center-surround receptive field structure that is optimal for information transmission is computed with lateral inhibition via a resistive network. The resistance of the network and the gain of the feedback to the photoreceptors are parameters controlling the size of the center-surround operator. Modulation of these intrinsic parameters adapts the retina to different viewing conditions. Photoreceptor adaptation is naturally integrated into the feedback retina. In this case, the feedback serves to calibrate the individual receptors with respect to each other. Relative calibration is the most that autonomous systems, which act without external reference, can achieve. The constraints of form in the retina lead to a representation of visual information that is largely invariant with respect to changes in illumination. Thus, the process of scene abstraction, usually

considered a purely cognitive phenomenon, in fact commences in the most peripheral stages of vision. As evidence for this I have shown that several subjective visual illusions are observed in the output of the chip.

The address-event communications protocol described in Chapter 3 capitalizes on the representation generated by the retina to efficiently transmit information between chips. Both VLSI chips and neuronal systems suffer from bandwidth limitations. The communications bandwidth is set by the product of the number of wires in the channel and the speed of each. Therefore, the strength of silicon technology relative to neurons, its speed, can be used to compensate for its weakness, the small number of pins available to communicate between chips. This trade-off is accomplished while preserving the event-like quality of nerve impulse transmission. The viability of this protocol depends explicitly on the efficiency of information encoding that is used in the retina, and which may be a general characteristic of neural systems. More work needs to be done in this area in order to define a precise inter-chip communication protocol that can be used commonly among VLSI neural network designers. The speed of the arbitration protocol and the interface of arbitration to internal analog circuitry can both be improved substantially. In addition, provisions for interfacing multiple senders and receivers and systems for determining the optimal number of data buses, the width of the data buses, and the partitioning of neurons onto these buses must be devised. The design of buses for interchip communication in analogs of specific neural structures will require a through understanding of their anatomy.

The stereocorrespondence chip described in chapter 4 is based on a novel stereocorrespondence algorithm that unites cooperative and multi-resolution approaches. The electrical elements of the chip have disparity-tuning characteristics similar to those found in biological systems. These characteristics arise from network interactions. The form and function of these electrical elements suggest plausible hypothesis for the mechanism of formation of biological receptive field properties and hints at their role in the computation of stereodisparity. Future research should explore new architectures based on stereotyped elements. The success of the present algorithm, based on a transformation of representation, suggests that we should search for architectures that can perform similar transformations, but have more neurally-plausible subunits. The computation of transparency and the rectification of the imbalance between the correlator array and the monocular units both require the

development of stereotyped subcircuits that can be combined into complex architectures. Just as the silicon retina is able to emulate the spatial and temporal response of the biological retina to arbitrary stimuli within a single physical structure, the evolution of a physical structure for the computation of stereo disparity should result in a system whose behavior is consistent with psychophysical and neurophysiological data over a large range of stimuli.

The silicon retina, the silicon optic nerve and the analog stereocorrespondence chip demonstrate that analog VLSI can capture a significant fraction of the function of neural structures at a systems level, and, concomitantly, that neural architectures can lead to new engineering approaches to computation in VLSI. The relationship between neural systems and VLSI is rooted in the shared limitations imposed by performing computation in similar physical media. The systems discussed in this text support the belief that the physical limitations imposed by the computational medium have as significant an effect on the algorithm. Since circuits are essentially physical structures, I advocate the use of analog VLSI as powerful medium of abstraction, suitable for understanding and expressing the function of real neural systems. The working chip elevates the circuit description to a kind of synthetic formalism. Thus, the physical circuit provides a formal test of theories of function that can be expressed in a circuit language.

Circuit language exists only in embryonic form. Carver Mead [2] has begun development of such a language, but at this time, the definition of the semantics of the primitive circuit elements and the syntax of their combination is still unclear. However, dramatic progress has been made towards standardizing design techniques in the related field of digital VLSI design [3]. There is every reason to believe that similar techniques will emerge in the field of neuromorphic analog design. The address-event communications protocol described in Chapter 2 is a major step towards such a standardization.

Neurobiological systems are sufficiently complex that the transition from description to traditional formal analysis is difficult. These systems consist of large numbers of non-linear elements and are analytically intractable and computation intensive for numerical simulation. Circuit design will play an increasingly significant role in computational neuroscience. One major advantage to building analog VLSI circuits is that, unlike digital simulation, VLSI analogs can be cascaded without affecting their performance. The real-time sensor-driven analog system exists at the same level as its biological counterpart. It can be driven

with real stimuli and generate electrical and even motor behaviors that can be observed with the same tools used to evaluate the performance of biological organisms. The ability to harness such extraordinary computational power will inevitably lead to qualitatively new understanding that will benefit both neuroscience and technology.

References

[1] Dowling, J. (1987) *The Retina: An Approachable Part of the Brain*. Cambridge, MA: Harvard University Press.

[2] Mead, C.A. (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

[3] Mead, C. A., and Conway, L. (1980) *Introduction to VLSI Systems*. Reading, MA: Addison-Wesley.

[4] Marr, D. (1982) *Vision*. New York: W. H. Freeman.

[5] Julesz, B. (1971) *Foundations of Cyclopean Perception*. Chicago, Illinois: The University of Chicago Press.

[6] Ratliff, F. (1965) *Mach Bands: Quantitative Studies on Neural Networks in the Retina*. San Francisco, California: Holden-Day, Inc.

Appendix A

Compiling the Arbiter

Silicon compilation is an essential tool for constructing VLSI chips that use the binary tree arbiter described in Chapter 3. Constructing a large binary tree structure is difficult in VLSI because the structure is regular between scales but not at any particular scale. Simple tiling of small elements cannot capture the large scale structure. In order for the Arbiter to be used in a practical sense as a design frame, it must be automatically scalable to any size array. I have written a WOLCOMP program in Pascal that automatically places a fixed set of small geometry cells to build an arbiter tree of whatever size is specified by the user. The compilation of the arbiter means that the design frame is essentially transparent to the user, although currently the library of geometry cells must be modified by each user to have the same pitch as user's base neuron element.

A well-commented program is listed at the end of the text. It will build arbiters of any size that have an even number of input neurons. However, Arbiters that are not a power of 2 will have timing asymmetries and thus favor some pixels over others. The tree generated by this routine is folded so that it occupies minimum space at the edge of the neuron array. A simple folded tree that arbitrates between four neurons is depicted in Figure A.1. The program is designed to fill the tree in from the bottom up. A tree with six neurons is shown in Figure A.2. The wiring for the tree is composed of small routing cells, illustrated in Figure A.3. The cell configuration for the four neuron tree is shown in Figure A.4.

```
import
$search '/LIB/WOLLIB/WOLCOMP/WOLCOMP'$ wolcomplib,
```

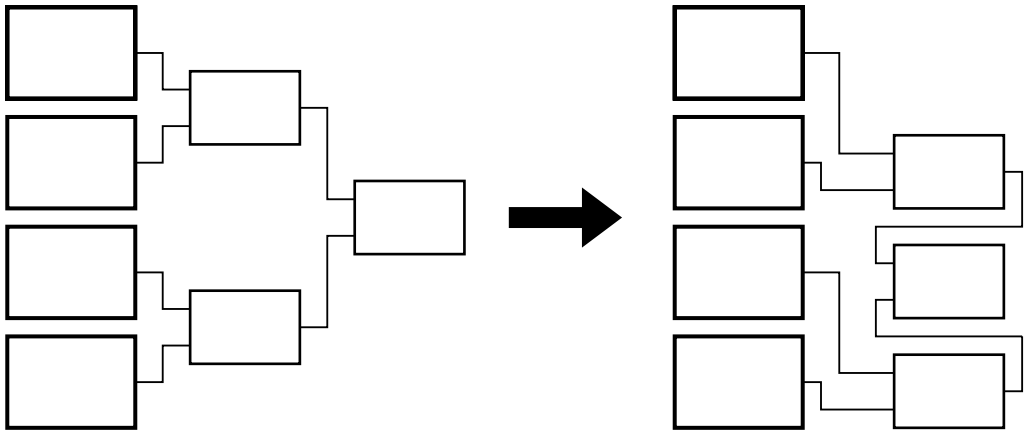

Figure A.1: Folded four-neuron tree arbiter.

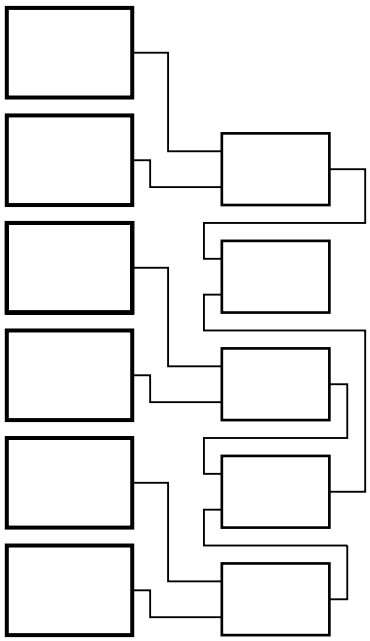

Figure A.2: Folded six-neuron tree arbiter.

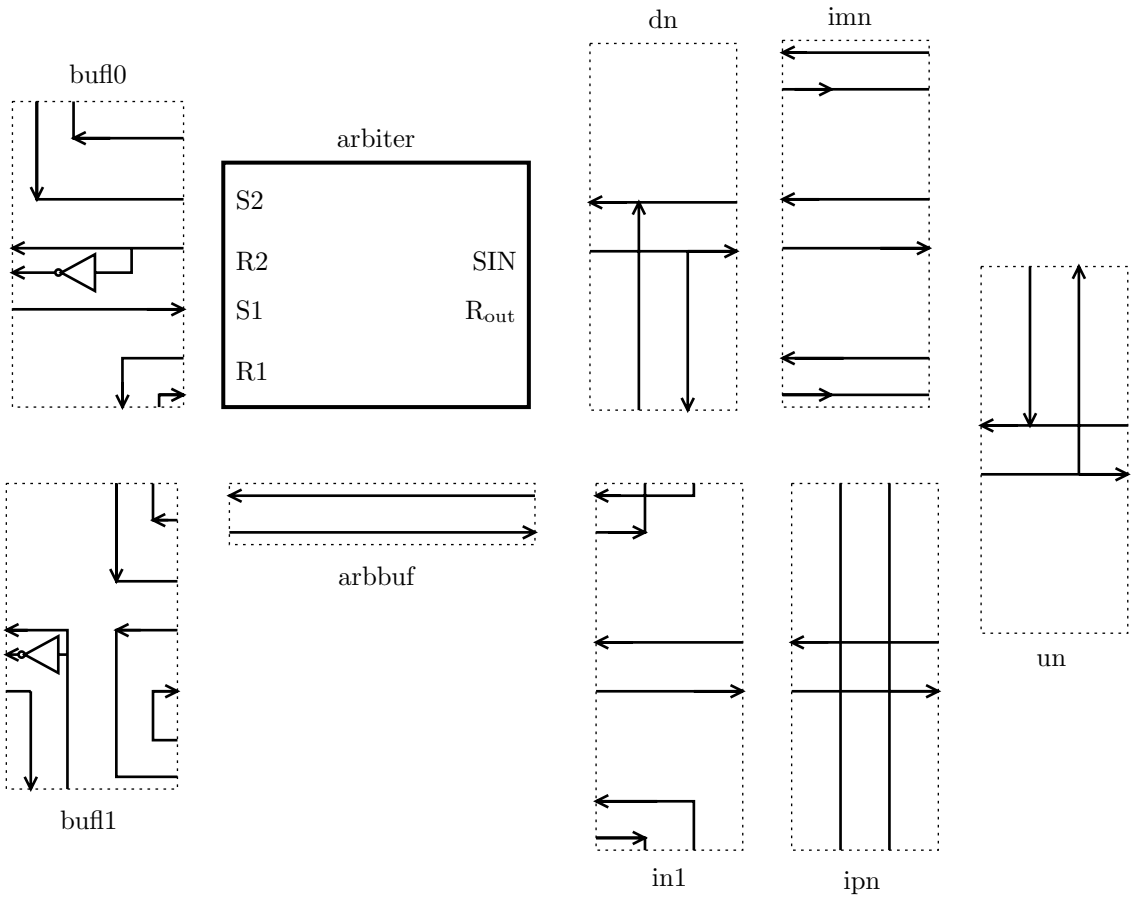

Figure A.3: Geometry cells for arbiter tree.

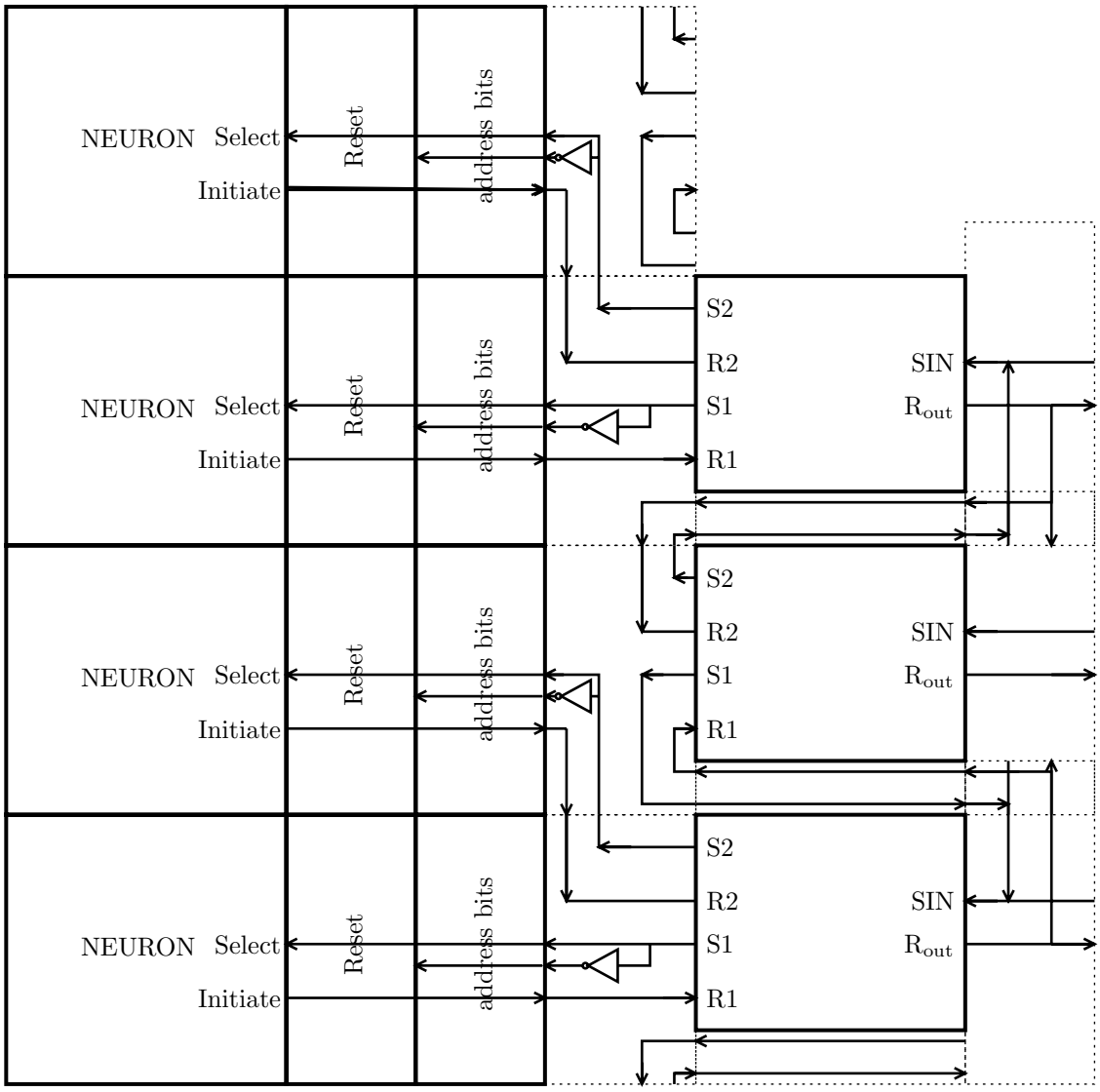

Figure A.4: Cell composition for four-neuron tree.

```
$search '/LIB/WOLLIB/WOLCOMP/WOLCOMPII'$ wolcompII,
const
  maxdepth=9; allows for 512 arrays
type
  intarray= array[1 .. maxdepth] of integer;
var
  numlayer : intarray; number of cells in each layer of the arbiter tree
  digit : intarray; binary encoding of the number of the current pixel
  dig : intarray; binary encoding of the number of pixels
  depth: integer; tells the depth of the array
  depth2: integer; tells the number of address bits
  numcells_x: integer;
  numcells_y: integer; function expon(base,power:integer):integer;
var
  i,temp:integer;
begin
  temp:=1;
  for i:=1 to power do
    begin
      temp:=temp*base;
  end;
  expon:=temp;
end; procedure binary(num:integer);
deals with the structure of the binary tree of arbiter elements
it calculates the number of two-input arbiters at each layer of the tree
and stores the result in numlayer
it also calculates the structure of the tree and stores it in dig array
when dig[n]=1, then the tree is complete at level n
complete means that all of the inputs coming up from n-1 have been allocated
to a two-input arbiter element at the nth level of the tree
the tree is built to be complete from the bottom
```

```
odd men out are left to the top of the tree structure
var
  temp,i:integer;
begin
  temp:=num-1;
initialize dig variable that stores the binary encoding of num
  for i:=1 to maxdepth do
    begin
      dig[i]:=0;
    end;
  i:=0;
  while temp¿=1 do
    begin
start with LSB level of the tree–call it level 1
      i:=i+1;
      numlayer[i]:=temp div 2;
      if (temp mod 2)=1 then
        begin
          dig[i]:=1;
          numlayer[i]:=numlayer[i]+1;
        end
      else
        dig[i]:=0;
      temp:=temp div 2;
    end;
  depth:=i;
depth is the number of levels in the binary arbiter tree
  if expon(2,depth)=num then
figures out how many address bits (depth) are needed for num pixels
since counting begins at one, 2^N pixels need N + 1 address bits
    depth2:=depth+1
```

```
  else
    depth2:=depth;
end; procedure binary2(num:integer);
calculates the addresss bits for the numth pixel and stores it in digit
var
  temp,i:integer;
begin
  temp:=num;
  for i:=1 to maxdepth do
    begin
      digit[i]:=0;
    end;
  i:=0;
  while temp¿=1 do
  begin
      i:=i+1;
      if (temp mod 2)=1 then
        begin
        digit[i]:=1;
        end
      else
        digit[i]:=0;
      temp:=temp div 2;
    end;
end; procedure arbiter_make(num:integer; dnum:integer; horizontal:boolean);
num is the actual number of pixels in the array
dnum controls the desired width of the address bus, typically equal to num
the width of the address bus is as wide as if there were dnum pixels
num and dnum must be even
var
  tmpvar1,tmpvar2,tmpvar3,pitch,level,toplevel,i,j,k:integer;
```

```
  noconnect,try:boolean;
begin
decide to build arbiter for horizontal side of pixel array, or vertical
since the neurons may not be square, a different set of buffers may be necessary
  if horizontal then
    pitch:=cell_width('pixel')
  else
    pitch:=cell_height('pixel');
  binary(dnum);
calculate the structure of the tree
  for i:=1 to num-1 do
place the arbiter element that abbuts pixel i at the correct level in the tree
for all but the last pixel in the array–begin counting pixels at 1
    begin
      j:=depth;
guess that the arbiter cell associated with pixel i
is at the top level of the tree
the guess is decremented at the end of the routine
      try:=true;
      while try do
keep trying until you get it right
        begin
          if (i mod expon(2,j-1))=0 then
if this is the position for an arbiter at depth j
            begin
arbiter is built from bottom to top (1 st pixel to num th pixel
and from left to right
order of cells is pixel-reset circuits (hreset, vreset)
- address bits (1 addr, 0 addr)- interface (*arbufl1, *arbufl0)-
two input arbiter element(scan)-
arbiter tree wiring cells (in1, ip, imn, dn, un)
```

```
add initiation node reset circuitry to aribter
            x0:=0;
            y0:=(pitch*(i-1));
calculate vertical coordinate of the i th pixel
            if horizontal then
              place('hreset')
            else
              place('vreset');
            x0:=x1; make address bits of the ith pixel
since the pixel numbering starts at 1, there are "depth2" bits
            binary2(i);
            for k:=1 to depth2 do
              begin
                if digit[k]=1 then
                  place('1addr')
                else
                  place('0addr');
                x0:=x1;
              end;
place buffers to interface cells to arbiter
            if (i mod 2)=0 then
              begin
                if horizontal then
                  place('harbufl1')
                else
                  place('varbufl1')
              end
            else
              begin
                if horizontal then
                  place('harbufl0')
```

```
            else
              place('varbufl0')
          end;
        x0:=x1;
        if horizontal then
            place('harbbuf')
        else
            place('varbbuf');
        y0:=y1;
        place('scan');
        x0:=x1;
        place('inv');
        x0:=x1;
        y0:=(pitch*(i-1));
re-calculate vertical coordinate of the i th pixel
find the level in the tree that you need to send connection to
provision for non-2^N trees
        if i¿=(num-expon(2,(j-1))) then
test if i is close enough to the top of the tree to be irregular
          begin
            k:=j+1;
            noconnect:=true;
            while noconnect do
find the level of the tree that you will connect to
              begin
                if k¡=depth then
                  begin
                    if dig[k]=0 then
there is no one for you to connect with at this level
                      begin
                      noconnect:=true
```

```
                    end
                    else
dig[k]=1 and you are connecting at level k-1
                    begin
                    level:=k-1;
                    noconnect:=false;
                    end;
                    end
                  else
if k ¿ depth you are at the top and should connect
                   begin
                    noconnect:=false;
                    level:=k;
                   end;
                  k:=k+1;
                end;
              end
            else
you are regular and the level in the tree you are connecting to is your depth
              level:=j;
do routing
there are (depth-1) routing channels that need to be placed
            for k:=1 to (depth-1) do
              begin
                if k¡(j-1) then
if you will connect to an arbiter element farther out than
the current wiring track level k
                  place('imn')
make sure your inputs will get to you and break wiring tracks that shouldn't pass
                else
                 if k=(j-1) then
```

```
make your connection to one level below you on the tree
                        place('in1')
                      else
                       if k=level then
send your output to the arbiter element you should connect to
                         begin
test to see if lower or upper branch in the arbiter tree
                          tmpvar1:=trunc(i/expon(2,j));
                          tmpvar2:=i+expon(2,j-1);
                          if ((tmpvar1 mod 2)=0) and (num¿tmpvar2) then
you are reaching up
                          place('un')
                          else
you are reaching down
                          place('dn');
                         end
                       else
k ¿ your connection level and you should provide wiring tracks
and send your output up in case you are the top level arbiter
                          place('ipn');
                     x0:=x1;
                   end;
                 try:=false;
you have successfully built pixel i's segment of the arbiter
so you can stop trying
              end
            else
else this is not a position for arbiter level j
              j:=j-1;
decrease your guess and try again
          end;
```

```
    end;
top off the arbiter
place reset circuitry
  x0:=0;
  y0:=pitch*(num-1);
  if horizontal then
    place('hreset')
  else
    place('vreset');
  x0:=x1;
  binary2(num);
find the address of the top bit
place address bits
  for i:=1 to depth2 do
    begin
      if digit[i]=1 then
        place('1addr')
      else
        place('0addr');
    x0:=x1;
    end;
place interface circuitry
  if (num mod 2)=0 then
    begin
      if horizontal then
        place('harbufl1')
      else
        place('varbufl1')
    end
  else
    begin
```

```
        if horizontal then
          place('harbufl0')
        else
          place('varbufl0')
      end;
place dummy cells at the top
  x0:=x1;
  place('arbtop');
  x0:=x1;
  place('arbtinv');
  x0:=x1;
  y0:=(pitch*(num-1));
  for i:=1 to (depth-1) do
    begin
      place('arbtch');
      x0:=x1;
    end;
end;
end of making arbiter
```