Pergamon

Neural Networks, Vol. 8, No. 7/8, pp. 1323–1347, 1995

0893-6080/95 $9.50+.00

0893-6080(95)00098-4

1995 SPECIAL ISSUE

Analog VLSI Neuromorphic Image Acquisition and Pre-processing Systems

A. G. ANDREOU,[1] R. C. MEITZLER,[1] K. STROHBEHN[1] AND K. A. BOAHEN

[1]Johns Hopkins University and [2]Caltech

(*Received and accepted* 17 *July* 1995)

Abstract—*We consider the problem of automatic object recognition by small, light-weight, low-power, hardware systems. We abstract from biological function and organization and propose hardware architectures and a design methodology to engineer such hardware. Robust, miniature, and energetically efficient VLSI systems for AOR can ultimately be achieved by following a path which optimizes the design at and between all levels of system integration, i.e., from devices and circuit techniques all the way to algorithms and architectural level considerations. By way of example, we discuss two experimental systems for image acquisition and pre-processing fabricated in standard CMOS processes. The first one is a large scale analog system, a contrast sensitive silicon retina, with over 590,000 transistors operating in subthreshold CMOS. The second system is a mixed analog–digital system for image acquisition and tracking compensation that incorporates a contrast sensitive silicon retina in the image sensing area.*

Keywords—Analog VLSI, Subthreshold CMOS, Vision chips, Image processing, Silicon retinas, Neuromorphic systems, Regularization, Spatiotemporal filtering, Neural computation, Automatic target recognition.

1. INTRODUCTION

Automatic target recognition (ATR) is the task of automatic object recognition (AOR) within the domain of Department of Defense (DOD) applications. AOR is defined as the process of detection, classification, identification and recognition of specific objects by *autonomous* human engineered systems. More often than not, the specific objects are not in isolation: rather, they exist in real world environments that include both natural and human made clutter. The recognition task must be performed with a high probability of success while maintaining a low probability of false alarms, and must be completed in real time, often within highly constrained time limits.

Application considerations necessitate operation from mobile vehicles and portable systems and thus the hardware must be miniature, lightweight, low power, and reliable. Aside from numerous DOD-related applications, AOR has applications in machines for face recognition in security systems, automatic fingerprint classification and identification, and autonomous robot operation in "real" world environments as well as collision avoidance systems for vehicles and aircraft.

AOR often involves different sensing modalities and data sources such as natural electromagnetic radiation emitted by the objects (infra-red), reflected electromagnetic radiation (synthetic aperture radar), reflected pressure waves (sonar), etc. In this paper we focus our attention on systems that employ image forming devices such as visible spectrum (VS), polarization vector (PV), and forward looking infra-red (FLIR) imagers.

Despite many years of research and significant contributions in this field, it is widely accepted that ATR as defined above, i.e., performed robustly by truly autonomous, highly mobile and portable hardware systems, is still an open problem (U.S. Army ATR Report, 1994). The inherent challenges

Acknowledgements: The research was supported by NSF grant ECS-9313934; Paul Werbos is the program monitor, and by a contract from Army Night Vision Laboratory at Fort Belvoir and by the CLSP at Johns Hopkins University. R.M. was supported by an Applied Physics Laboratory Fellowship. The authors would like to thank Professor Carver Mead of Caltech for encouraging this work. The authors are thankful to Mark Martin for helping with image acquisition and processing. Chip fabrication was provided by MOSIS.

Requests for reprints should be sent to Andreas G. Andreou, Johns Hopkins University, 34th & Charles Sts., 105 Barton Hall, Baltimore, MD 21218, USA.

within AOR can be appreciated if it is placed within the canonical model of a communication channel and contrasted with traditional communication devices and systems for which we see a plethora of small size, light weight, and low power hardware solutions.

In most human engineered communication systems where the goal is the *precise restitution* of information, more often than not the engineers have at their disposal the freedom to design a system which is heavily asymmetric in the computational requirements of the encoder and decoder. For example, consider the one way distribution of video over a wireless channel from a client to a multimedia portable server, such as the Infopad (Sheng et al., 1992). This particular task is one with numerous DOD and civilian applications. In this case, the system engineers have the freedom of tailoring the communication channel (radio link) as well as an appropriate source coding to achieve the channel bit-rate requirements. Furthermore, since encoding of the stored information can be done in an environment of almost unlimited computational resources, algorithms that are asymmetric in the complexity of the encoding–decoding process can be selected. Clearly algorithms with high computational demands in the encoder but with *simple decoders* are preferable. Thus, the overall power dissipation in the system can be heavily skewed towards the fixed station where there are no power constraints. It is therefore not surprising at all to see reports in the literature of low-power, custom integrated circuits/systems developed to address this problem (Chandrakasan et al., 1994).

Within the same canonical formulation of a communication channel, an AOR system can be viewed as the decoder aimed at decoding a message (identifying the target) from signals that were received by the sensor(s). The signature/image is generated through a process that involves the interaction between the physical structure of the target with some form of electromagnetic energy. Signals are propagated through the atmosphere where they are distorted and noise is added. Distortion and noise are also added to the scene during the "encoding" so as to hinder the function of the decoder (through camouflage, target-like clutter, and other measures). At the sensor subsystem, energy signals are transduced, amplified, and processed to decode the original message (recognize the target). It should be noted that at the point where the signals impinge upon the transducer arrays, the signals are further degraded by two kinds of noise. The first is noise in a thermodynamic sense, which is always present in electronic systems and therefore must be taken into account. The second source of noise is related to the structural variability in the transducers and signal processing electronics.

We can see why, unlike the former case, the latter scenario describes a much more difficult class of problems of which automatic object/target recognition is a subset. The designers of the system do not have at their disposal the freedom to optimize the encoding or the channel and therefore they are "stuck" with the hard task of decoding a message whose encoding process may be largely unknown and which has been communicated through a channel possessing a great deal of variability. An optimal decoder for automatic object recognition can be designed using statistical methods (Duda & Hart, 1973) and information theory, much as is done for other sensory modalities such as speaker-independent, robust speech recognition (Roe & Wilpon, 1994). In voice communication between human and machine, language modeling is employed to exploit prior knowledge in the design of the decoder (speech recognizer). In the case of AOR, models of the target and the environment are much harder to construct. Furthermore, the space over which such knowledge spans is enormous, thereby making a full search for specific patterns an impossible task. Ironically, the AOR problem in mobile systems is computationally harder because the extra mobility introduces variability in the environment that may have not been accounted for in its internal models. Thus, the system has to deal with this variability through *real-time* adaptation.

Despite the aforementioned obstacles, biological systems excel at object recognition and other sensory communication tasks by sustaining high computational throughput with minimal energy dissipation. These are "real" physical systems which are highly mobile and thus constrained by size, weight, and the availability of energetic resources. The effectiveness and energetic efficiency of natural systems suggest that it may be a worthwhile endeavor to try to understand the principles of *function* and *organization* in biological information processing systems, using these characteristics as a guide for the development of VLSI systems (Mead & Ismail, 1989; Ramacher & Ruckert, 1991; Sánchez-Sinencio & Lau, 1992).

This approach has been pioneered by Carver Mead and his colleagues at Caltech (Mead, 1989, 1990) and pursued independently by other groups (Andreou, 1990; Andreou & Boahen, 1994b; Vittoz, 1994). The neuromorphic/anthropomorphic approach to the engineering of high performance, low power, light weight AOR hardware is the subject of this paper, with the focus being on analog VLSI image acquisition and pre-processing.

The paper is organized into seven sections. The neuromorphic/anthropomorphic approach to the problem is discussed in Section 2. Motivated by the organization of neural systems, we provide a blueprint for system "micro-architecture" in Section 3, where we see the emergence of a *hierarchical* system

organization that evolves around *computational sensors/actuators* and *computational memories*. The technology and design methodology for these systems is outlined in the following section. In Section 5, we discuss analog VLSI systems for adaptive sensing, amplification and feature extraction. Section 6 offers a discussion and Section 7 concludes the paper.

2. THE NEUROMORPHIC APPROACH

Unlike most known forms of computing/calculating activity known to humans, information processing in biological organisms has a rather well defined goal: the detection, decomposition, and transformation of sensory inputs into suitable representations (memory maps). These maps are useful for the ultimate goal of the organism, which is interaction with the environment in a "closed loop" configuration where the environment is an integral part of a loop. This interaction may involve only lower level functions such as early perceptual processing and sensorimotor coordination or it may include a higher level task such as communication through speech which is a unique characteristic to the human species. Recognition and identification of specific objects, tasks within the framework of AOR, are clearly activities that are intimately related to the interaction between an organism and the environment.

The effectiveness of biological systems stems partly from exploiting *prior* knowledge about the problems that they encounter (Barlow, 1989). Such information, in the form of *internal models*, reflects the structural properties of the natural environments in which the systems function. Classical information and communication theory formalisms can be employed to develop theories of how such internal models help to "optimaly" encode signals in the limited capacity neural structures [see for example Srinivasan et al. (1982), Buchsbaum and Gottschalk (1983), Linsker (1986), Li and Atick (1994)].

Further complications arise because both the environment and the physical structures (computational substrate) that process information in the biological organism are not fixed. This necessitates *adaptation* through *self-organization*, to compensate for the *variability*. Thus, an essential aspect of neural computation is adaptation (Kohonen, 1988; Mead, 1990; Gorin et al., 1991; Haykin, 1994).

Adaptation (see Figure 1) is indeed pervasive in neural systems and is found at many different levels of a hierarchical organization. For example, adaptation can be found in the electromechanical properties of sensory transducers, in the network properties of neurons, and even in the abstraction of high level cognitive processes. The importance of adaptation in biological systems has been long recognized by Stephen Grossberg and colleagues; a sample of their early work can be found in the edited volume (Grossberg, 1988).

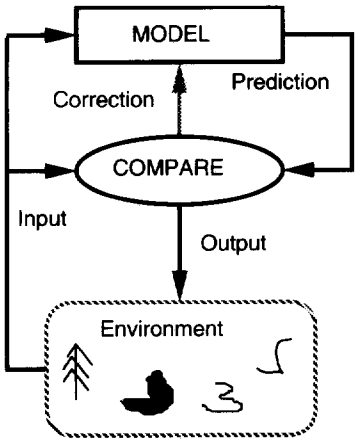


FIGURE 1. An adaptive system and its interaction with the environment. External sensory inputs to the system are compared with its internal state (model), to produce an output. Information from the input and from the comparison process may be used to adjust the parameters of the internal model. [Adapted from Mead (1989).]

With the aforementioned characteristics in mind, neural computation is, from an engineering perspective, fundamentally concerned with the processing of signals in the presence of *noise*. Noise can be *exogenous*, due to the variability in the natural environment (problem related) or *endogenous* to the system due to internal sources of structural variability and fluctuations (in a thermodynamic sense) in the actual physical hardware. Thus it could be argued that the effectiveness and energetic efficiency of biological systems can be associated with their ability to:

1. exploit *prior* knowledge, in the form of internal models,
2. deal effectively with both endogenous and exogenous sources of variability,
3. process information in a hierarchical and parallel manner.

However, the question of *how* such sophisticated processing is actually carried out by the "real" system, and how information is extracted at the different layers of a hierarchical organization still remains open. "Real" computing structures must satisfy strict constraints of size, weight, utilization of energetic resources, and the ability to operate at temperatures where favorable conditions exist for the development of life as "we" know it, $\sim 300\,K$. Algorithms based on statistical methods and self-organizing techniques are notorious for their enormous computational requirements when implemented on digital computers. How is such sophisticated processing done in neural structures?

Carver Mead (Mead, 1989, 1990) has eloquently argued that an answer to this question can perhaps be found if one concentrates on the algorithms and information processing structures that emerge from the physical properties of the computational substrate. Furthermore, Mead and coworkers propose an *analysis by synthesis* approach, where analog methods and VLSI technology can be used to prototype such "not-so-conventional" information processing *systems*.

From the perspective of *science*, analog VLSI technology can be viewed as a modeling tool (Mead, 1989; Mahowald & Douglas, 1991) aimed at capturing the behavior of neurons, networks of neurons, or the complex mechanical–electrical–chemical information processing in biological systems. Computationally, analog VLSI models can be more effective than software simulations. More important, analog VLSI systems are "real" models, constrained by fundamental physical limitations and scaling laws. Constraints such as power dissipation, physical extent of computing hardware, density of interconnects, gain-bandwidth product limitations in the gain elements, precision and noise characteristics of the basic elements, signal dynamic range, and robust behavior and stability may force the development of more realistic models (Andreou, 1991). This work (Andreou & Boahen, 1994b) follows a similar line of thought.

From an *engineering* viewpoint, such ideas have shown promise towards the development of VLSI systems that are more effective in solving sensory communication problems. In the next section, we will introduce an *architectural framework* for autonomous object and target recognition systems.

3. AOR SYSTEM MICRO-ARCHITECTURE

Developing an information processing system's architecture for sensory communication (of which AOR is one modality) requires careful consideration. Calculating/computing activity as we know it today is aimed at solving particular problems—for instance, doing a spreadsheet calculation or drawing a picture using one's favorite drawing program. The correct answer is the ultimate goal and computation is continued until the problem is solved. The availability of results in a timely fashion is necessary but the perfect answer is rarely traded off for quick response.

Information processing for AOR has rather different objectives. The value of the computation resides in the availability of results in a timely fashion. Severe time constraints impose a mode of operation where one does not seek the perfect answer but an answer that will be available when it is needed. Advances in information technologies have resulted in computing systems with enormous capabilities and certainly one should aim at getting the perfect answer. However, this goal may not be achievable as fundamental limitations due to power dissipation, communication bandwidth constraints, as well as system complexity and reliability are important issues that must be dealt with when assuming the availability of unlimited computational resources.

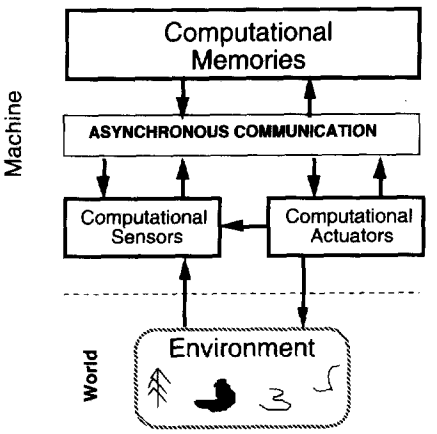


FIGURE 2. "Micro-architecture" for an AOR system. This is a simplified diagram depicting single instances of the various subcomponents within the larger system. A more realistic diagram would involve multiple instances of each subcomponent organized *hierarchies*.

Abstracting on biological *organization*, the adopted computer architecture takes the form shown in Figure 2. The proposed organization evolves around three different types of subsystems and enforces a *parallel distributed processing* paradigm by incorporating only the essential local computation at each stage in the system hierarchy. *Locality of reference* and, associated with it, a minimal amount of long distance communication follows directly from this organization. We will now briefly discuss each subcomponent of this architecture.

3.1. Computational Memories

Sensory communication problems (including AOR) are inherently memory intensive as a priori information (internal models) must be stored in the system. In a typical recognition problem, *searching* large spaces for pattern matching is one of the most common operations.

Given that vision problems can be mapped nicely on parallel hardware, it is natural to consider parallel processing architectures and in particular the ultimate SIMD example, a *memory based* associative processor. Experimental systems with digital storage but analog processing, designed and optimized for low power dissipation, have been reported in the literature (Boahen & Andreou, 1993; He et al., 1993; Pouliquen et al., 1993). Computational memories with analog storage capabilities have also been reported in the

literature (Holler et al., 1989; Boser et al., 1991; Cauwenberghs et al., 1992; Horio & Nakamura, 1992; Yang et al., 1992). Unfortunately, the technologies for prototyping such large scale memory based systems are not readily available for experimentation and therefore the capabilities of reported systems are limited and have not yet been able to address the complexity of real world problems.

3.2. Asynchronous Communication

Since events in the real world are by their nature asynchronous, computation for AOR can also be naturally asynchronous. As new data are received and processed, evidence for the presence or absence of a target must be re-evaluated. Thus, the communication between system subcomponents can be asynchronous, thereby offering savings in terms of power and available bandwidth. One could take the view that the architecture for AOR will resemble a modern ATM (asynchronous transfer mode) packet switching network.

Other forms of asynchronous communication in vision and speech systems have been pursued by the research group at Caltech as a means of exchanging information between different chips and minimizing communication costs [please see Lazzaro et al. (1993) and references therein]. Information thus can be transferred in a multiresolution fashion on an as-needed basis. Depending on the particular problem, asynchronous communication can be done within a single chip, between single chip systems, or as part of a *client-server* computing paradigm.

3.3. Computational Sensors/Actuators

At this level the role of these subsystems is to provide the essential processing at the interface between the machine and the environment. Computational actuators is an emerging field of research since only recently the technologies for the manufacture of truly integrated micro-electro-mechanical systems have become available. These mechanical elements will not be further discussed here.

In the field of computer vision, computational sensors can be associated with what are commonly known as *silicon retinas*, *early vision processors* or *signal preprocessors*. In the field of speech recognition, one could view the processing at this level as representing the *silicon cochleas* or the *acoustic processors* of the system. Essential components of this interface are:

- adaptive transduction/amplification,
- adaptive quantization (encoding).

Transduction is where the physical stimuli are converted into electrical signals and amplified to the required signal to noise ratio levels. This conversion is determined by the physical structure of the transducer, frequency response, dynamic range, and noise characteristics. Adaptation at this level is necessary for two reasons, both related to sources of variability.

First we consider "endogenous" sources, i.e., the structural variability in the components. During fabrication, the different steps required to produce a functional device are themselves random processes (for example, ion implantation), and therefore no two devices can be fabricated with identical characteristics. The problem becomes more important in transducer arrays, especially those fabricated in technologies not as advanced as silicon VLSI, for example infra-red detector technology (Scribner et al., 1993, 1994).

The second source of variability that requires adaptation is exogenous to the system and is problem related. If the physical signals from the environment have a dynamic range larger than the transducer can handle, some form of adaptation (**gain control**) is necessary to match the properties of the stimulus to the characteristics of the transducer themselves.

Gain control is a form of encoding aimed at throwing away what is redundant; therefore, it is a process of data reduction. This is an essential requirement for alleviating the burden on both communication resources and subsequent processing stages. A well established technique for data encoding at this level is that of *predictive coding* (Shrinivasan et al., 1982).

A fixed encoding scheme, often having the goal of reducing redundancy, has the disadvantage that properties of the signal that are context-dependent may be lost. An adaptive encoding scheme is thus necessary to avoid data loss.

It is clear that computing for AOR systems poses some challenging problems. Substantial computational resources are necessary for robust operation, while at the same time signal and information processing must be done on a tight power budget and severe time limitations. We believe that robust, miniature, and energetically efficient hardware VLSI systems for AOR can ultimately be achieved by following a methodology which optimizes the design at and between all levels of system integration. This approach is applicable from the device and circuit technique levels all the way to algorithmic and architectural level considerations. This is the subject of discussion in the next section.

4. TECHNOLOGY AND DESIGN METHODOLOGY

At the most basic level, VLSI technology and analog models offer the possibility of experimentally

exploring computation by truly complex, *real systems* which lie beyond digital computing and the symbolic processing paradigm. In other words, there is no fundamental reason to believe that the systems that will ultimately solve the AOR problems will be entirely digital. It is likely that successful AOR systems will involve both CMOS analog and digital components as well as non-traditional forms of analog processing blocks such as integrated micro-electromechanical parts.

The adopted design style, **current-mode subthreshold CMOS** using circuits of minimal complexity (Boahen et al., 1989; Andreou et al., 1991a; Andreou & Boahen, 1994b) offers the possibility of ultra low power dissipation and area density commensurate with the high level of VLSI system integration.

CMOS technology and, in particular, *subthreshold* MOS operation has long been recognized as the technology of choice for implementing digital VLSI and analog LSI circuits that are constrained by power dissipation requirements (Vittoz & Fellrath, 1977; Vittoz, 1985a, 1985b). The advantages of using standard digital CMOS processes for cost-effective engineering solutions to analog signal processing problems are surveyed in Vittoz (1985a) and are also discussed in Tsividis (1987). CMOS has the highest integration density attainable today, making it especially attractive for analog VLSI models of neural computation (Mead, 1989; Vittoz, 1994).

It is appropriate at this point to ask the question: what kind of computational primitives does one have? In CMOS silicon, these are continuous functions (analog) of *time*, *space*, *voltage*, *current* and *charge*. To help manage the complexity in VLSI systems, these functions will be considered at three hierarchical levels: the *device level*, the *circuit level*, and the *architectural level*.

The understanding of complex information processing in neural systems through a discussion at different levels is an approach that was first introduced by Marr and Poggio (1977) and also discussed extensively in (Marr, 1982).

4.1. Device Level

The current in an MOS transistor operating in a subthreshold ohmic regime is an *exact difference* of exponential functions of the drain and source voltages (Vittoz & Fellrath, 1977; Vittoz, 1985b; Mead, 1989; Andreou & Boahen, 1994b) so that for an NMOS the current is given by:

$$I_D \equiv I_{DS} = I_{n_o} S \exp(\kappa_n v_{GB})[\exp(-v_{SB}) - \exp(-v_{DB})] \quad (1)$$

and for a PMOS

$$I_D \equiv I_{SD} = I_{p0} S \exp(-\kappa_p v_{GB})[\exp(v_{SB}) - \exp(v_{DB})]. \quad (2)$$

The terminal voltages v_{GB}, v_{SB}, v_{DB} are reference to the substrate and are normalized to the thermal voltage $V_t \equiv (kT/q)$. The constants I_{p0} and I_{n0} depend on mobility (μ) and other silicon physical properties; S is a geometry factor, the width W to length L ratio of the device. The constant κ takes values between 0.6 and 0.9. The above equations do not model drain conductance modulation or other short channel phenomena.

For NMOS devices that are biased with $v_{DS} \geqslant 4$ (saturation), the drain current is given by:

$$I_{DS} = SI_{n0} \exp(1 - \kappa_n v_{BS}) \exp(\kappa_n v_{GS}). \quad (3)$$

Equation (3) explicitly shows the current's dependence on v_{BS} and the role of the bulk as a *back-gate* that underlies this relationship. This equation, having only the dependence on v_{GS} and v_{BS}, is used for circuit designs where devices operate in saturation as a transconductance amplifier. A similar expression can be written for PMOS transistors.

Channel-length modulation (Early effect), which we have ignored completely, becomes significant in saturation. Thus, the device equations must be augmented with terms that model the Early effect to accurately predict the output conductance.

The transfer characteristics of MOS transistors are plotted in Figure 3 for both the above- and subthreshold regime (Pavasovic & Andreou, 1994). The transconductance per unit current increases as the current decreases through-out the above-threshold and transition regions and reaches a maximum in the subthreshold region. The MOS transistor has excellent circuit properties as a voltage-input, current-output device (transconductance amplifier)

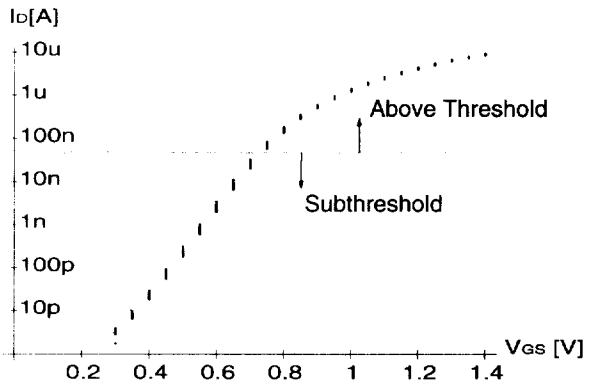


FIGURE 3. Measured drain current I_D versus gate-source voltage V_{GS} for 32 small geometry NMOS transistors (4×4 μm) fabricated in a 2 μm) *n*-well CMOS process; drain-source voltage of $V_{DS} = 1.5$ V. The fuzziness in the current (mismatch between devices), is constant in subthreshold [on a log(I) scale] and decreases as the device enters the transition and above threshold regime.

with good fan-out capabilities (high transconductance $g_m \equiv \partial I_{DS}/\partial V_{GS}$ and good fan-in capability (almost zero conductance at the input).

In highly integrated VLSI systems, small geometry transistors must be used, typically $4\,\mu m \times 4\,\mu m$ or $6\,\mu m \times 6\,\mu m$, to achieve high densities. Furthermore, as we have seen, it is preferable to operate the devices in the region where the transconductance per unit current is highest, i.e., the subthreshold and transition regions. Unfortunately, small device geometries and high transconductance per unit current also make the drain current strongly dependent on spatial variations of process-dependent parameters. The effect is especially true for I_{p0} and I_{n0}, which is the source of the variability observed in the drain currents of Figure 4. The apparent improvement in device matching for higher values of gate-source voltage is simply a manifestation of reduced transconductance per unit current as the device enters the above-threshold regime.

Our preference for subthreshold/transition region operation (despite what seem to be worse matching characteristics) is based on the observation that: *Active devices should be used in the region where their transconductance per unit current is maximized.* In this way, one can minimize the energy per operation and maximize the speed per unit power consumed, i.e., minimize the power-delay product:

$$\frac{\text{speed}}{\text{power}} = \frac{1/\tau}{I\Delta V} = \frac{g_m/C}{I^2/g_m} = \frac{1}{C}\left(\frac{g_m}{I}\right)^2, \tag{4}$$

where C is the load capacitance.

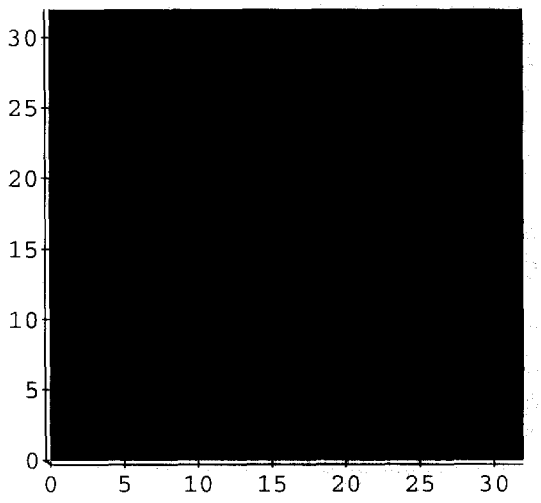


FIGURE 4. Density plots of currents in a 32×32 array of $4\mu m \times 4\mu m$—NMOS transistors. Each transistor is represented by a square pixel. Current level is coded by the shade of gray, where the minimum and maximum values are represented by black and white, respectively. The current is measured at a nominal level of 100 nA by setting V_{GS} to be the same for all transistors in the array. The devices are biased in saturation.

A squared factor is obtained because both voltage swings (ΔV) and propagation delays (τ) are inversely proportional to the transconductance for a given current level. However, only a linear factor is realized if the power supply voltage is not reduced to match the voltage swings $\Delta V \sim I/g_m$. When the device is operated in subthreshold, the drain-source conductance saturates at a few (kT/q) (see Figure 5). Power supplies of a few (kT/q) are also possible and thus power supplies can theoretically match the voltage swing levels.

The gate capacitance is also lowest when the device is operating at subthreshold. The effective mobility in the channel of MOS transistors exhibits a broad peak in the transition region (Yang et al., 1994). Since mobility is higher, losses in the channel are small, the channel conductance is high, and thus the noise of the transistors is low.

The maximum useful frequency of operation possible with an MOS transistor, when operating at subthreshold is determined by its transition frequency f_T which has an upper limit $f_{T\max}$ of:

$$f_{T\max} < \frac{\mu(kT/q)}{\pi L^2}, \tag{5}$$

where μ is the effective carrier mobility and L is the device channel length. The transition frequency of a

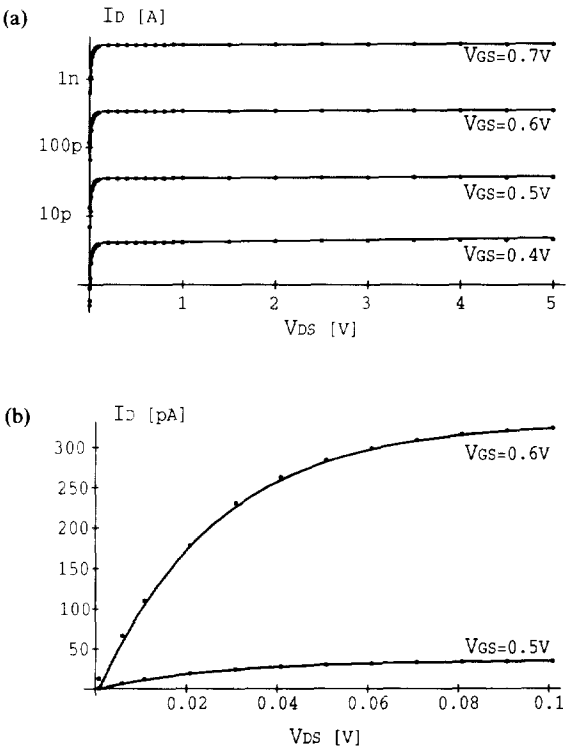


FIGURE 5. Measured output characteristics of NMOS transistors in subthreshold. Device dimensions are $W = L = 16\,\mu m$. The two sets of data emphasize the good saturation characteristics (a) and the non-linear ohmic behavior at low drain source voltages (b).

device is essentially the bandwidth (as determined by the internal gain and parasitic capacitances of the transistor). For 6–10 μm length devices (typical in analog VLSI today), functional systems in the hundreds of kHz processing rate per node are possible while for submicron devices, the limit extends to the MHz range.

4.2. Circuit Level

The synthesis of computational structures begins at the circuit level, the focus of this section, and manifests itself as the emergence of *networks*. At the circuit level, conservation laws, i.e., the conservation of charge (Kirchhoff's current law), $\sum_i I_i = 0$ and the conservation of energy (Kirchhoff's voltage law), $\sum_i V_i = 0$, are used to realize simple constrain equations. The important concept of *negative feedback* is also exploited to trade the gain in the active elements for precision and speed in circuits.

Aside from the benefits of a device with a large gain, the exponential relationships between the controlling voltages and the current depicted in eqns (1) and (2) endow the MOS transistor with some interesting circuit properties. There exists a powerful synthesis (and analysis) procedure which can be used to generate a wide variety of circuits that perform linear and non-linear operations in the current domain. The methodology relies on the exponential form of current–voltage non-linearities. This procedure is based on what is known as the *translinear principle* (Gilbert, 1975, 1984), which was originally used in the context of bipolar transistors. The synthesized circuits are called *translinear* and may involve operations of one or more variables, such as products, quotients, power terms with fixed exponents, and scalar normalization of a vector quantity.

The application of the translinear principle to circuits implemented with MOS devices operating in subthreshold saturation, and an extension to the subthreshold ohmic regime, can be found in Andreou and Boahen (1994b, 1996). One fascinating aspect of translinear circuits is that, while the currents in its constitutive elements (the transistors) are exponentially dependent on temperature, the overall input/output relationship is insensitive to isothermal temperature variations. The effect of small local variations in fabrication parameters can also be shown to be temperature independent.

As another example of how computational primitives emerge at the network level from the device physics of the underlying technology, let us consider a summing operation, *local aggregation*. The linear addition of signals over a confined region of space occurs throughout the nervous system. Aggregation was discussed in Chapter 6 of Mead (1989), [also in Koch (1989)], and is the basis for many neuromorphic silicon VLSI systems described therein. Here we take a close look at *diffusion*, the physical process that underlies local aggregation in the nervous system, contrast the neural mechanism with the process of diffusion in MOS transistors, and come up with a novel network design technique.

4.2.1. *Linear MOS Transistor-only Networks.* The exponential functions of voltage in the square brackets of eqns (1) and (2) correspond to Boltzmann distributed charges at the source and drain diffusing through the channel. The exponentials can be conveniently represented as dimensionless quantities of charge $\mathcal{Q}_{(.)} \equiv \exp[-v_{(.)}]$ and diffusivity

$$\mathcal{D}(v_{GB}) \equiv S\exp(v_{GB}) \tag{6}$$

so that eqn (1) becomes:

$$I_{DS} = I_{n0}\mathcal{D}^{\kappa}[\mathcal{Q}_S - \mathcal{Q}_D]. \tag{7}$$

The charge-based representation depicted in eqn (7) suggests that the MOS transistor in subthreshold is a highly linear device in the charge domain; a property that has been exploited not only in neuromorphic analog VLSI (Boahen & Andreou, 1992) but also in more traditional forms of analog circuit design (Furth & Andreou, 1995). Viewing an MOS transistor in subthreshold as a basic diffusive element (Boahen & Andreou, 1992) allows for the effective implementation of systems that exploit properties of elliptic partial differential equations. The same idea was more recently revisited by Vittoz and Arreguit (1993).

We will now contrast the operation of traditional voltage/conductance based networks with diffusive networks. The network depicted in Figure 6a employs voltages and currents. Its node equation is

$$I_i = G(V_j + V_k + V_l + V_m - 4V_i) = G\nabla^2 V. \tag{8}$$

Note that the RHS term can be recognised as a first-order approximation to the Laplacian operator $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$ with the internode distance normalized to unity. This model is not amenable to VLSI integration because conductances (G) with a large linear range consume large amounts of area and power.

The second network uses charges (positive) and currents through NMOS transistors operating in subthreshold (Figure 6b). Its node equation is

$$I_i = I_{n0}\mathcal{D}^{\kappa}(\mathcal{Q}_j + \mathcal{Q}_k + \mathcal{Q}_l + \mathcal{Q}_m - 4\mathcal{Q}_i) = I_{n0}\mathcal{D}^{\kappa}\nabla^2\mathcal{Q}. \tag{9}$$

Note that I_i is the same as the current supplied to node i by the network. By using devices with identical

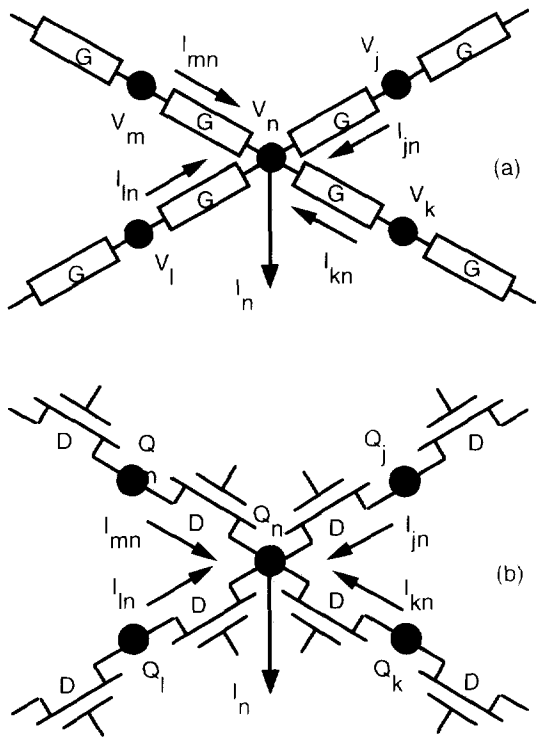


FIGURE 6. Discrete approximation to the Laplacian operator with (a) conductances and voltage/current variables or (b) diffusors and charge/current variables.

$S \equiv (W/L)$ ratios, the transistors are guaranteed to have the same diffusivity while using identical gate voltages guarantees that the charge concentrations at all the source/drains connected to node i are the same and equal $\mathcal{Q}_i$.

In both of these networks, the boundary conditions may be set up by injecting current into the appropriate nodes. In the voltage-mode network, the solution is the node voltages. They are easily read without disturbing the network. On the other hand, the second (MOS) network represents the solution by charge concentrations $\mathcal{Q}_S$ and $\mathcal{Q}_D$ at source/drains—not the charge on the actual nodes. The source/drain charge cannot be measured directly without disturbing the network. The solution charge may be inferred from the node voltage and measured with the appropriate "sense amplifiers".

Diffusive media in biological cell syncytia are hardly ever isotropic like our simple initial example. Nerve cells make gap junctions of varying area. Neuromodulators like dopamine can vary the pore permeability. Thus, nerve cells can control the diffusivity to neighboring cells or the extracellular fluid. We can control the diffusivity of the MOS network by exploiting the factorization property for the drain and source charges, with the result that they can be written as explicit products of two terms, one of which includes only the gate voltage and the source or drain voltages [see eqns (6) and (7)].

To summarize, the properties of the MOS transistor-only network depend on:

1. some actual physical parameter I_{n0} that is material dependent and is related to the diffusion coefficient of the carriers in silicon;
2. a design parameter S that is fixed *prior* to fabrication;
3. a variable parameter $\mathcal{Q}$ which is a function of the gate voltage and can be controlled during circuit operation.

4.2.2. *Linear MOS Transistor-only Loaded Networks.* Often, models of neural computation necessitate the realization of loaded networks. We begin with networks that employ linear conductances, voltages and currents and then compare them with translinear current-mode (Andreou & Boahen, 1994b) networks.

A voltage-mode circuit model for a loaded network is shown in Figure 7(a) for which:

$$I_{PQ} = (G_1/G_2)(I_Q - I_P).$$

This is a lumped parameter model where G_1 and G_2 correspond to resistances per unit length. The voltages on nodes P and Q are referenced to ground and represent the state of the network which can be read out using a differential amplifier with the negative input grounded.

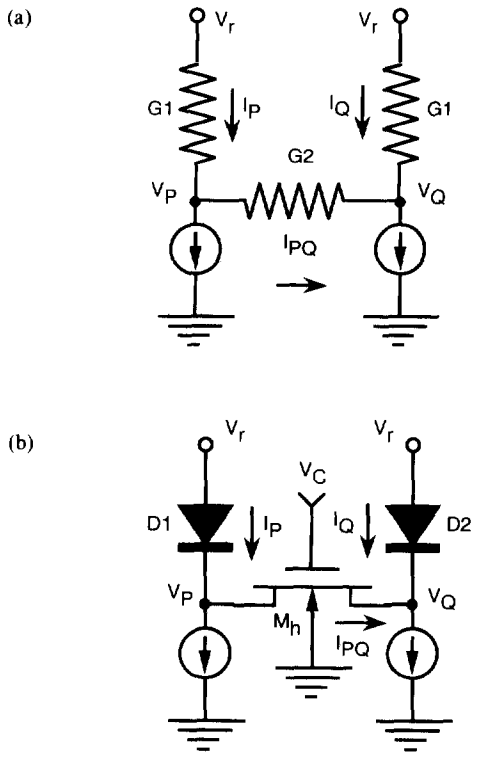


FIGURE 7. Building blocks for linear networks. Using segments that employ ideal (a) linear and (b) non-linear elements.

The equivalent circuit using idealized non-linear conductances is shown in Figure 7(b). The difference in currents through the diodes D_1 and D_2 is linearly related to the current through the diffusor MOS transistor. This relationship can be derived from eqn (1), describing subthreshold conduction, and the ideal diode characteristics where $I_D = I_S \exp(v_D)$. An expression can be derived for the current I_{PQ} in terms of the currents I_P and I_Q, the reference voltage v_r, and the bias voltage v_C, where:

$$I_{PQ} = \left(\frac{SI_{n0}}{I_S}\right) \exp(\kappa_n v_C - v_r)(I_Q - I_P). \qquad (10)$$

The current I_{n0} and S are the zero intercept current and geometry factor respectively for the diffusor transistor M_h; I_S is the reverse saturation current for the diode that is assumed to be ideal. The currents in these circuits are identical if

$$\frac{G_1}{G_2} = \left(\frac{SI_{n0}}{I_S}\right) \exp(\kappa_n v_C - v_r).$$

Increasing v_C or reducing v_r has the same effect as increasing G_1 or reducing G_2. The state of this network is represented by the charge at the nodes P and Q. Since the anode of a diode is the reference level (zero negative charge), the currents I_P and I_Q represent the result.

When diodes are not explicitly available in the process, diode connected PMOS or NMOS transistors can be used as shown in Figure 8. When the loads are PMOS, the current current I_{PQ} is given by:

$$I_{PQ} = \left(\frac{S_h I_{n0}}{S_v I_{p0}}\right) \exp(\kappa_n v_C - \kappa_p v_r)\left(I_Q{}^{1/\kappa_p} - I_P{}^{1/\kappa_p}\right). \qquad (11)$$

Unfortunately, the anode of a diode or a diode connected transistor is not a good current source. When NMOS transistors are used as loads, there is the additional benefit of exploiting the current conveying properties of a single transistor (Andreou & Boahen, 1994b). In so doing, we can obtain the current outputs I_P and I_Q on nodes that are low conductance (the drain terminals are now excellent outputs for the currents). Using eqn (8.45) in Andreou and Boahen (1994b), the current I_{PQ} is given as:

$$I_{PQ} = \left(\frac{S_h}{S_v}\right) \exp(\kappa_n v_C - \kappa_p v_r)(I_Q - I_P), \qquad (12)$$

where S_h and S_v are geometry parameters for transistors M_h and M_v, respectively.

In summary, we have provided a comprehensive overview of the *current-mode* approach for analyzing and synthesizing subthreshold MOS transistor-only linear networks. The essence of this approach is the representation of variables and parameters by charge, current, and diffusivity. Voltages and conductances are not used explicitly.

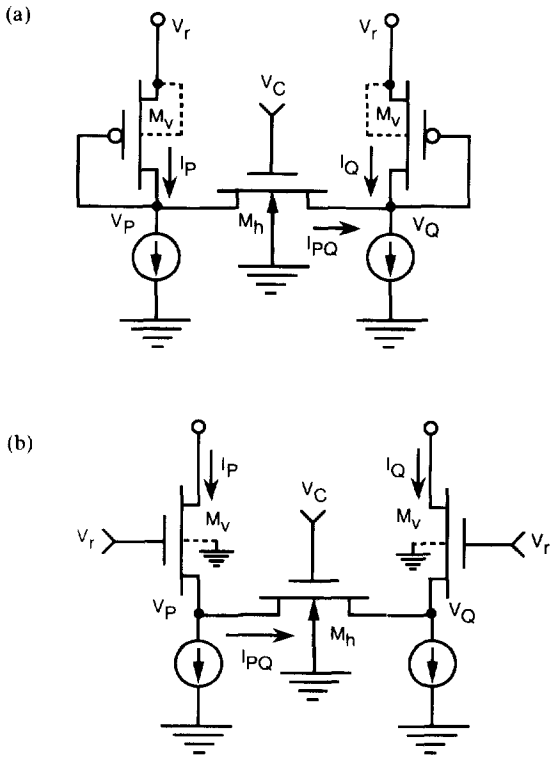


FIGURE 8. Current-mode building blocks for linear MOS transistor-only networks using (a) PMOS transistor implementation, (b) NMOS single transistor current-conveyor implementation.

4.3. Architectural Level

At this level, differential equations from mathematical physics can be employed to implement useful signal processing functions which nonetheless retain the form of constraint equations. For example, the *biharmonic* equation

$$\lambda \nabla^2 \nabla^2 \Phi + \Phi = \Phi_{in}, \qquad (13)$$

where $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplacian operator, constrains the sum of the fourth derivative, of Φ, and Φ itself to be equal to a fixed input Φ_{in}. From a *statistical* signal processing view-point, solutions to this equation could represent an *optimal estimation* Φ of the underlying smooth continuous function given a set of noisy, spatially sampled observations Φ_{in}. The solution is optimal in the sense that it simultaneously minimizes the squared error and the energy in the second derivative. Here, the parameter λ is the relative cost associated with the

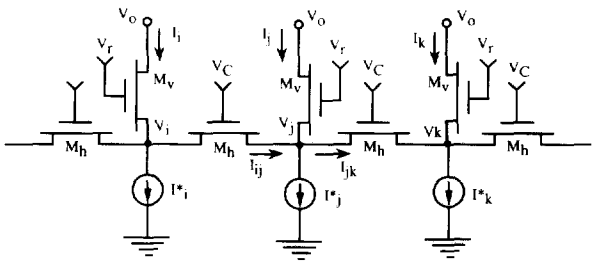


FIGURE 9. Discrete approximation to the Helmholtz equation using MOS transistor-only network to perform the local spatial averaging.

derivative term. A large value of λ favors smooth solutions while a small value favors a closer fit.

As an example of small system synthesis using the design methodology presented in this section, we will create a minimal, MOS transistors-only realization of the classical silicon retina architecture proposed by Mead and Mahowald (1988) and Mead (1989), and analyzed in detail by Taylor (1990).

The one dimensional MOS transistor-only network corresponding to the Helmholtz equation is shown in Figure 9 and models the averaging that occurs at the horizontal cell layer of the outer plexiform.

Summing the currents at node j we get:

$$I_{j^*} = I_{ij} - I_{jk} + I_j. \tag{14}$$

The results from the previous section for the currents I_{ij} and I_{jk}, which are given by eqn (12), can be substituted in eqn (14) to obtain:

$$I_{j^*} = I_j + \left(\frac{S_h}{S_v}\right) \exp(\kappa_n v_C + \kappa_n v_r)(2I_j - I_i - I_k). \tag{15}$$

By normalizing internode distances to unity, the above equation can be approximated on the continuum as:

$$I^*(x) = I(x) + \lambda \frac{\mathrm{d}^2 I(x)}{\mathrm{d}x^2}.$$

This equation yields the solution to the following optimization problem: find the smooth function $I(x)$ that best fits the data $I^*(x)$ with the minimum energy in its first derivative. The inputs are the currents $I^*(x)$ and the outputs are the currents $I(x)$. Since input signals and outputs (network average) are currents flowing in the same direction, they can be subtracted by first inverting one through a mirror and then adding the output of a mirrored signal to a replica of the input. The latter circuitry is not shown in Figure 9. The parameter

$$\lambda \equiv \left(\frac{S_h}{S_v}\right) \exp(\kappa_n v_C - \kappa_n v_r)$$

is the cost associated with the derivative energy—relative to the squared-error of the fit.

We have already seen how a diffusive grid can be used to compute a discrete approximation of the Laplacian and of the Helmholtz equation. In a subsequent section, we show how a model of early visual processing is related to the biharmonic equation and how to realize the model using diffusive networks.

5. NEUROMORPHIC FOCAL PLANE PROCESSORS

Computational sensors, i.e., information processing at the focal plane by intimately coupling detection and signal processing, as well as the use of novel sensing modalities, are areas that could ultimately contribute to the performance, compactness, and miniaturization of AOR systems. The visual systems of many biological species are existing proof that sensor specialization and integration of signal processing with sensing may be beneficial. The vertebrate regina (Dowling, 1987) is an even more impressive example of system integration and efficient processing.

However when computation must be performed at the focal plane of an imager using hardware that are constrained to exist in essentially two dimensions, two issues must be addressed first:

1. A portion of the area that otherwise could be used to collect light is now used by associated processing circuitry and therefore the spatial resolution, as well as the light sensitivity, of the system is compromised.
2. Active circuits other than phototransducers produce heat which can increase the dark current of the phototransducing devices and the system performance will be compromised.

Yet, we believe that even with a two dimensional implementation medium (such as single-plane VLSI circuitry), it is possible to trade off light sensitivity and spatial resolution for some essential processing at the focal plane.

In this section, we present a detailed discussion of two focal plane processing systems. The first is a contrast sensitive silicon retina, an edge enhancing imager that includes a rudimentary, yet effective *local gain control* mechanism at the transducer level. The second addresses the problem of *feature extraction*, where an embedded algorithm extracts position and motion information at the focal plane. These local signals are useful for tracking and robust image acquisition. A brief overview of other analog VLSI systems that provide local gain control through *temporal adaptation* will also be discussed.

5.1. A Contrast-Sensitive Silicon Retina

Image acquisition and early vision processing under naturally occurring illumination conditions is a common task in the fields of robotics, prosthetic devices for the blind, and motor-vehicle navigation and thus relevant to AOR. Today this task is accomplished in two separate steps (Rosenfeld, 1988). First the light intensity is recorded through a standard imager such as a CCD camera. The intensity field is subsequently processed outside the camera to discard any absolute luminance information and form a representation where only relative illumination, i.e., *contrast*, is retained. Additional processing such as edge extraction and/or low bit-rate encoding may follow. However, even though the precision necessary for these tasks rarely exceeds eight bits, the signal itself has a very large dynamic range–many orders of magnitude–which makes the problem difficult. This issue becomes acute when the illumination varies widely within a single frame, a common occurrence in natural scenes (see Figure 10). The detrimental effects of non-uniform illumination in the performance of a face recognition system have been investigated experimentally by Buhmann et al. (1994).

In contrast to the above approach, the same problem can be addressed by abstracting from the known organization and function of the distal retina (Dowling, 1987). Our method not only attempts to account for the physics of image formation in natural scenes but also attempts to address the problem of limited dynamic range in the actual hardware. This is an important issue since the performance of "real" physical computing systems are ultimately limited by fundamental limitations of the computational substrate.

During the design of the silicon retina, the *architectural* decision can be made to integrate some *low-precision* analog processing with the transducer elements and thus extract contrast information at the focal plane (Boahen & Andreou, 1992; Andreou & Boahen, 1994a). The resultant image that is captured with such a system is shown in Figure 11. The biologically motivated solution is attractive from a computational perspective because *contrast*, an invariant representation of the visual world, has been obtained with a front-end that is robust, small,

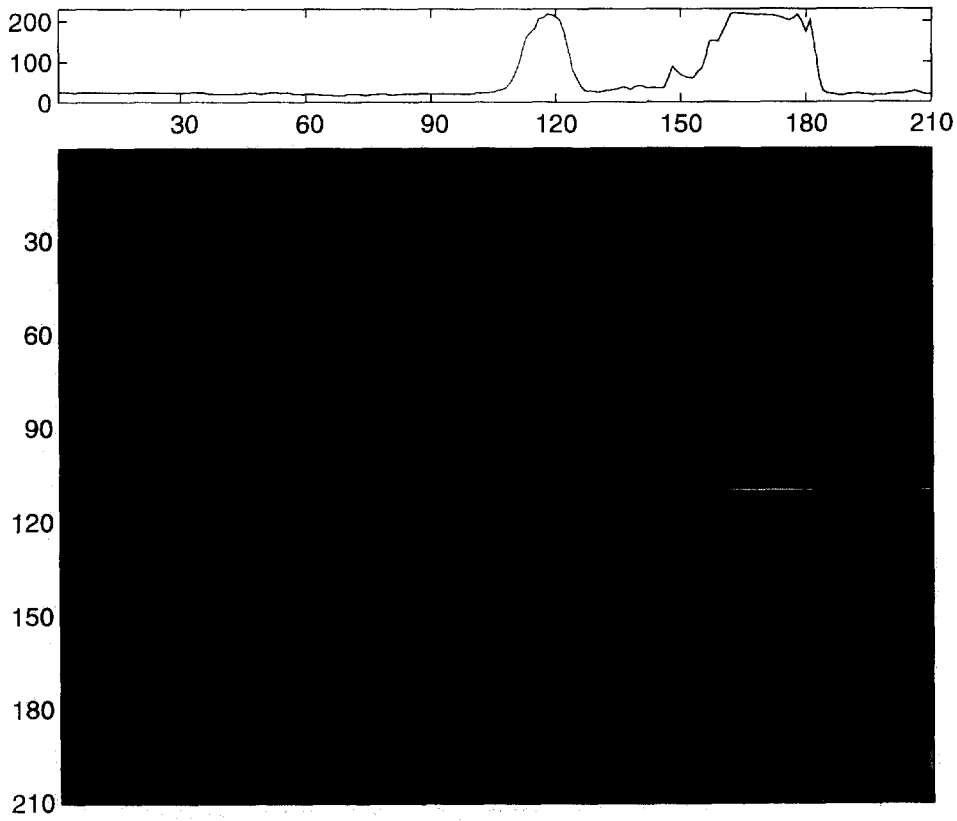


FIGURE 10. (Bottom) "Mark" as captured by a conventional camera. (Top) Intensity histogram at image line 110 (white line). The light source is positioned to the right side of the image and it introduces a large gradient in illumination within a single frame. This is clearly shown in the intensity histogram. The dynamic range of the scene exceeds the dynamic range of the camera. Aperture control on the camera provides a rudimentary global gain control mechanism. Information in this image is lost at this very first step because there is no gain control (adaptation) at the pixel level.

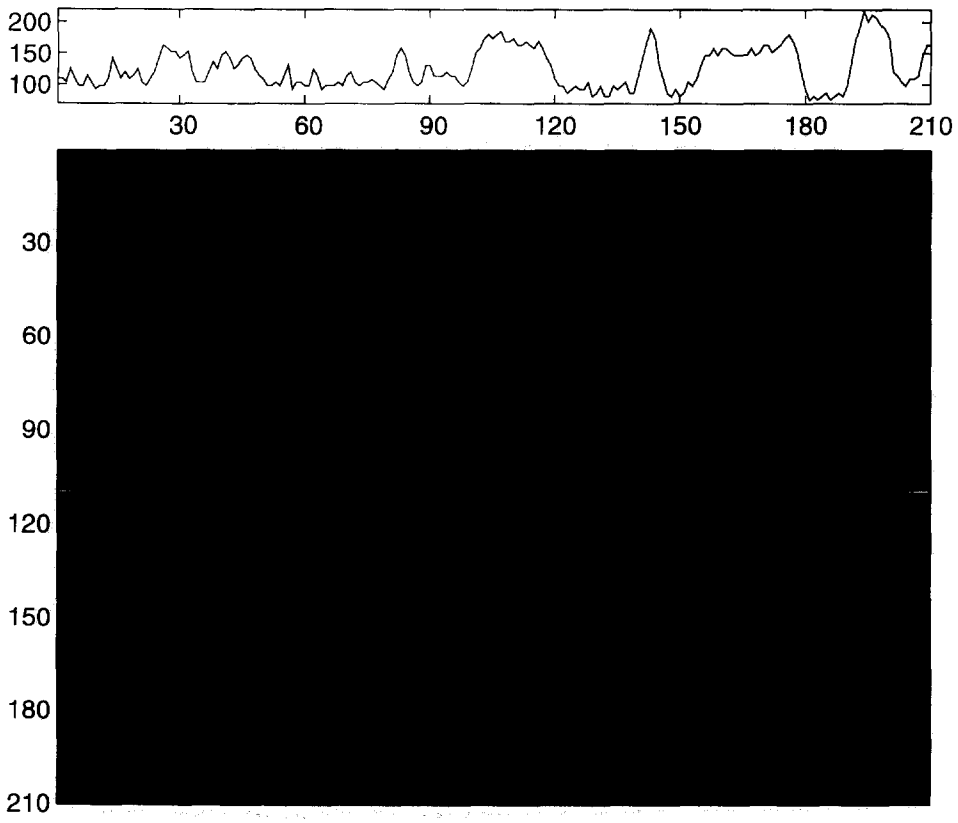


FIGURE 11. (Bottom) "Mark" as captured by a experimental imager, an analog VLSI model of the vertebrate distal retina. The light source is positioned to the right side of the image and it introduces a large gradient in illumination within a single frame. The image captured by the silicon retina discards absolute illumination and preserves only local contrast information through local gain control at the pixel level. Unlike the image in Figure 10, the presence of a large illumination gradient does not degrade image acquisition here.

and extremely low power (a few mW). There is also an engineering benefit; the output of the system is "encoded" in a representation with limited range and therefore subsequent processing/communication stages are not burdened with handling and processing signals of wide dynamic range. A performance comparison between a contrast sensitive silicon retina front end (Boahen & Andreou, 1992) and a conventional camera in a face recognition experiment is reported in Buhmann et al. (1994).

5.1.1. *Biological Organization.* The analog silicon system in the core of the array is modeled after neurocircuitry called the outer-plexiform layer, which is located in the distal part of the vertebrate retina. Figure 12 illustrates interactions between cells in this layer (Dowling, 1987). The well-known center/surround receptive field emerges from this simple structure that consists of just two types of neurons. Unlike the ganglion cells in the inner retina and the majority of neurons in the nervous system, the neurons that we model here have graded responses (they do not spike); thus this system is well-suited to analog VLSI. The original architecture (Boahen & Andreou, 1992) was inspired by the linear model and numerical simulations of Yagi et al. (1989).

The photoreceptors are activated by light: they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells, in turn, suppress the activity of the receptors through inhibitory chemical synapses. The receptors and horizontal cells are electrically coupled to their neighbors by electrical synapses. These allow ionic currents to flow from one cell to another, and are characterized by a certain conductance per unit area.

In the biological system, contrast sensitivity—the normalized output that is proportional to a local measure of contrast–is obtained by shunting inhibition. The horizontal cells compute the local average

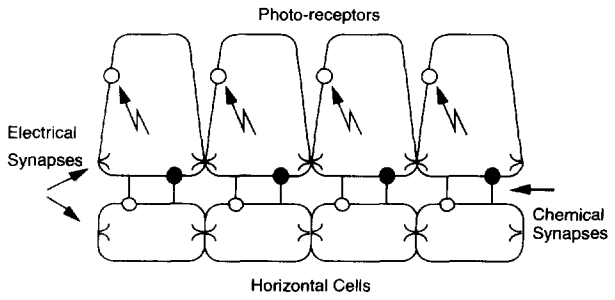


FIGURE 12. One-dimensional model of neurons and synapses in the outer-plexiform layer. Based on the red-cone system of the turtle retina.

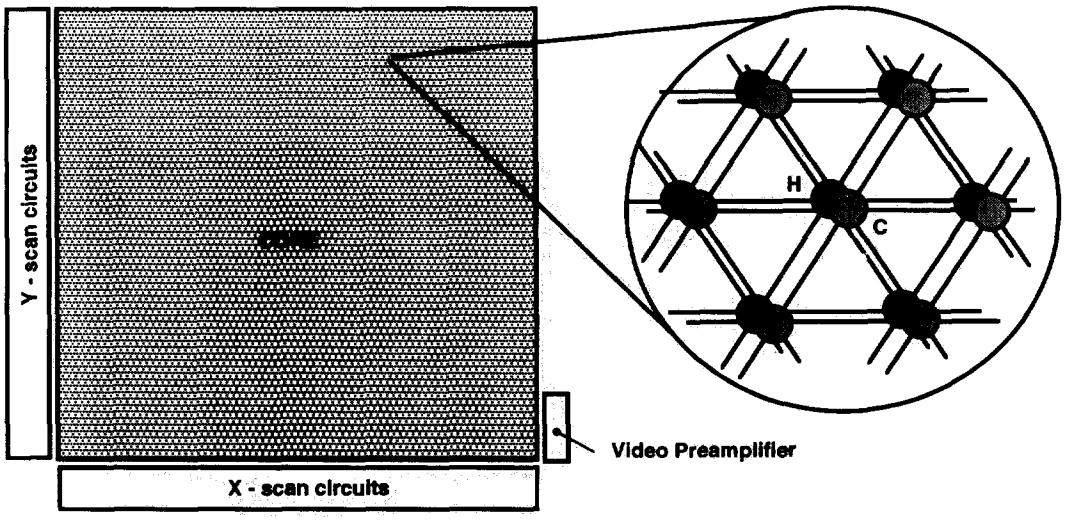


FIGURE 13. Floorplan and system organization. It comprises two functional components, the core, and the support circuitry. Focal plane processing is performed in the core area.

intensity and modulate a conductance in the cone membrane proportionately. Since the current supplied by the cone outer-segment is divided by this conductance to produce the membrane voltage, the cone's response will be proportional to the ratio between its photoinput and the local average, i.e., to contrast. This is a very simplified abstraction of the complex ion-channel dynamics involved.

5.1.2. *Silicon System.* The analog architecture that attempts to abstract the processing performed at the outer plexiform layer of the vertebrate regina is shown in Figure 13. The architecture is mapped onto silicon using circuits of minimal complexity that exploit native properties of subthreshold MOS transistors. High computational throughput at low levels of power dissipation is achieved by employing *low precision* analog processing in a massively parallel architecture.

Support circuitry in the periphery extracts the data from the core and interfaces with the display. The chip incorporates a video pre-amplifier and some digital logic for scanning the processed images out of the array. This circuitry is discussed in detail in the paper by Mead and Delbriick (1991). Standard NTSC video is produced off-chip using an FPGA controller and a video amplifier.

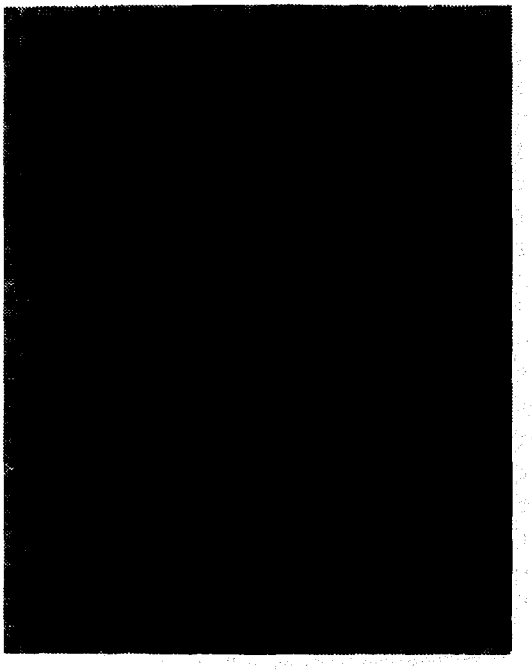
The core of the silicon retina is an array of pixels with a six-neighbor connectivity (see Figure 13). The wiring is included in the layout of the cell (see Figure 14) so that they may be tiled in a hexagonal tesselation to form the focal plane processor. This is a mesh processor architecture where two layers of processors, C and H, use both intra- and inter-layer communication through local paths. This parallel processing scheme features the locality of reference and thus minimizes communication costs.

The basic analog MOS circuity for a one dimensional pixel with two neighbor connectivity is shown in Figure 15. We begin with the non-linear aspect of the system's operation, its *contrast sensitivity*. The non-linear operation that leads to a local gain-control mechanism in the silicon system is achieved through a mechanism that is qualitatively similar to its biological counterpart, but quantitatively different [see discussion in Boahen and Andreou (1992)]. Refering to Figure 15, the output current $I_c(x_m, y_n)$ at each pixel, can be given (approximately) in terms of the input photocurrent $I(x_m, y_n)$ and a local average of this photocurrent in a pixel neighborhood (M, N). This region may extend beyond the nearest neighbor. The fixed current I_u supplied by transistor M_3 normalizes the result and Ψ is a parameter.

$$I_c(x_m, y_n) \approx I_u \frac{I(x_m, y_n)}{\left(I(x_m, y_n) + \Psi \sum_{M,N} I(x_i, y_i)\right)}. \tag{16}$$

At any particular intensity level, the outer-plexiform behaves like a linear system that realizes a powerful second-order regularization algorithm (Poggio et al., 1985) for edge detection. This operation can be seen by performing an analysis of the circuit about a fixed operating point. To simplify the equations, we first assume that $\hat{g} = \langle I_h \rangle g$, where $\langle I_h \rangle$ is the local average. Now we treat the diffusors (devices M_4) between nodes C and C' as if they had a fixed diffusivity $\hat{g}$. The diffusivity of the devices M_5 between nodes H and H' in the horizontal network is denoted by h. Thus, the simplified equations describing the full two

(a)



(b)

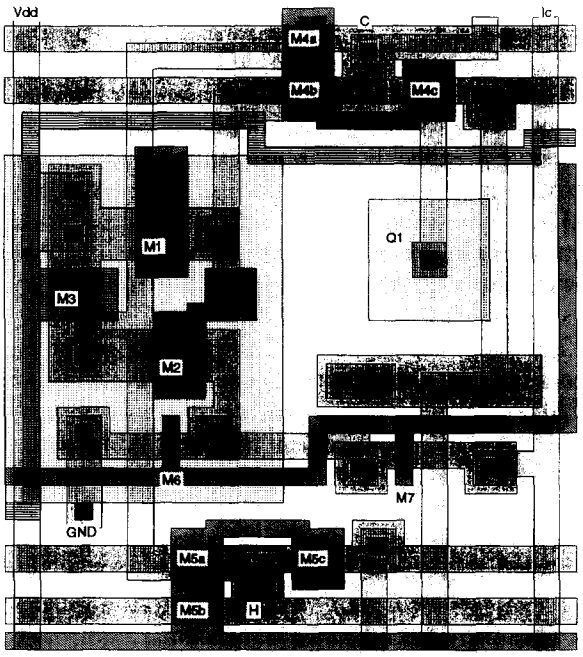


FIGURE 14. (a) Photomicrograph of the chip. The surface is covered by second metal except where there are openings for the phototransistors (the dark square areas). Note the hexagonal connectivity between the pixels. (b) Layout of the basic cell.

dimensional circuit on a square grid are:

$$I_h(x_m, y_n) = I(x_m, y_n) + \hat{g} \sum_{\substack{i=m\pm 1 \\ j=n\pm 1}} \{I_c(x_i, y_j) - I_c(x_m, y_n)\}$$

$$I_c(x_m, y_n) = I_u + h \sum_{\substack{i=m\pm 1 \\ j=n\pm 1}} \{I_h(x_m, y_n) - I_h(x_i, y_j)\}$$

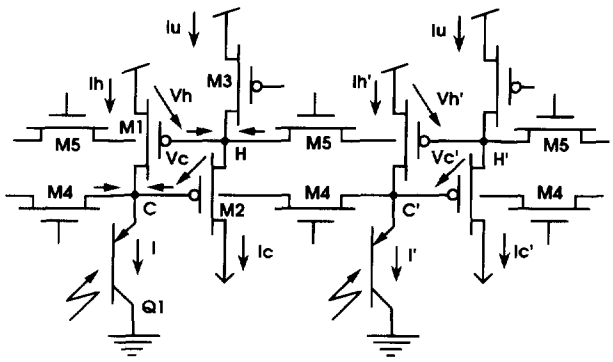


FIGURE 15. One dimensional implementation of outer-plexiform retinal processing. There are two diffusive networks implemented by transistors M_4 and M_5, which model electrical synapses. These are coupled together by controlled current-sources (devices M_1 and M_2) that model chemical synapses. Nodes H in the upper layer correspond to horizontal cells while those in the lower layer (C) correspond to cones. The bipolar phototransistor Q_1 models the outer segment of the cone and M_3 models a leak in the horizontal cell membrane. Note that the actual system has a six neighbor connectivity.

Using the second-difference approximation for the Laplacian, we obtain the continuous versions of these equations:

$$I_h(x, y) = I(x, y) + \hat{g}\nabla^2 I_c(x, y) \tag{17}$$

$$I_c(x, y) = I_u - h\nabla^2 I_h(x, y) \tag{18}$$

with the internode distance normalized to unity. Solving for $I_h(x, y)$, we find:

$$\hat{g}h\nabla^2\nabla^2 I_h(x, y) + I_h(x, y) = I(x_i, y_j). \tag{19}$$

This is the *biharmonic* equation used in computer vision to find an optimally smooth interpolating function $I_h(x, y)$ for the noisy, spatially sampled data $I(x_i, y_j)$; it yields the function with minimum energy in its second derivative (Poggio et al., 1985). The coefficient $\lambda = \hat{g}h$ is called the regularizing parameter which determines the trade-off between smoothing and fitting the data.

A one dimensional solution to this equation can be obtained (see eqn 20) using Green's functions valid for vanishing boundary conditions at plus and minus infinity; this has the characteristic mexican hat shape (see Figure 16).

$$I_h(x, \lambda) = \frac{1}{2\lambda^{1/4}} \exp\left(-|x| \Big/ \sqrt{2}\lambda^{1/4}\right) \cos\left(\frac{|x|}{\sqrt{2}\lambda^{1/4}} - \frac{\pi}{4}\right). \tag{20}$$

5.1.3. *Layout Considerations.* The two-layer architecture for the silicon retina can be accommodated in a cell area of $80\lambda \times 94\lambda$ using a single poly, two metal technology. In the implementation reported in

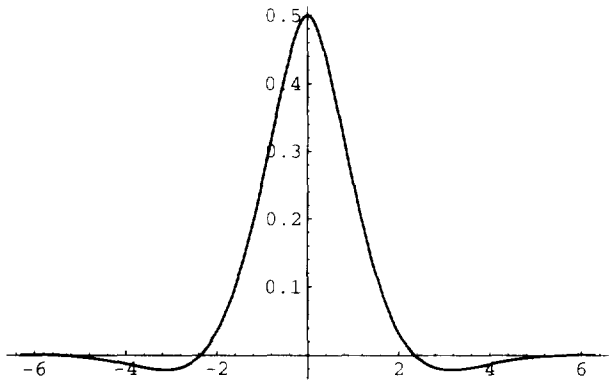


FIGURE 16. Plot for the one dimensional solution of the biharmonic equation; $\lambda = 1$, and input is a delta function at the origin.

Boahen and Andreou (1992) and here, a double poly, double metal technology is used and the cell area is $66\lambda \times 73\lambda$. First metal and polysilicon wires are used for interconnects; second metal is used to cover the entire array, shielding the substrate from undesirable photogenerated carriers. Transistors are implemented using both polysilicon layers.

The system has been fabricated with 230×210 pixels on a 9.5×9.3 mm die in a 1.2 μm n-well double metal, double poly, digital oriented CMOS technology. The chip incorporates 590,000 transistors in the 48,000 pixels and support circuitry, with the core operating in subthreshold/transition region.

A conservative estimate for the energetic efficiency can be obtained by assuming that a total of 18 low precision operations (OP) are performed per pixel. Six operations are necessary for the convolution with bandpass kernel of eqn (20), six for the Laplacian operator [eqn (18)], and six for the local gain control computation [eqn (16)]. If the system is biased so that at the pixel level the frequency response is 100 kHz, approximately 1×10^{11} low precision calculations per second are performed in the 210×230) pixels. The power dissipation under the above biasing conditions is about 50 mW when operating from 5 V power supplies. This is equivalent to 0.5 pW/OP.

5.2. A Silicon Retina with Embedded 2-D Motion and Position Estimation

Spatiotemporal processing, such as local gain control implemented either spatially or temporally, is only a first step, albeit a vital one, in extracting meaningful information from a scene. For tracking and for providing temporal cues, image motion information is important. In the case of tracking, the motion signals can be used either to direct the "attention" of the focal plane imager, or for motion compensation within the object/target recognition algorithm. It is important to note that we will be discussing global image motion rather than determining a collection of local velocity vectors, i.e., an optical flow field. Computation of a complex optical flow field (Hildreth, 1983), while potentially of great use, is a difficult focal plane task. As such the computation that will be performed on a chip will be employed locally to compensate for the motion of the image acquiring device.

A number of global motion computing chips have been fabricated to data [see, for example, Delbrück (1993), Andreou et al. (1991b) and Etienne-Cummings et al. (1992)] and a comparison of several approaches is contained in Horiuchi et al. (1992). The common characteristic underlying the aforementioned implementations is the use of correlation to compute the image's velocity or displacement. This general class of algorithms is biologically inspired, being based on Reichardt's original work on the visual system of the fly (Reichardt, 1961). Since both biological systems and analog VLSI rely on parallel computations with low-precision components, it is not surprising that these correlation algorithms map very well onto current-mode analog VLSI circuitry. This was the motivation for pursuing this approach by Andreou and co-workers (1989). In contrast, chips using local gradient algorithms such as those of Tanner and Mead (1986), were hampered by the necessity of explicitly computing spatial and temporal derivatives and as a result suffered in terms of pixel size and robustness. A search of the literature reveals that virtually all recent chip designs have utilized correlation algorithms, indicating the suitability of this approach for analog VLSI.

The following sections will describe the design and operation of a chip which capitalizes on analog VLSI's effectiveness to provide concurrent image centroid and displacement motion information (Meitzler et al., 1995). The centroid and displacement computations are each performed by two one dimensional linear arrays [earlier versions of which are detailed in Andreou et al. (1991b) and Meitzler et al. (1993)] oriented at 90° angles to each other. Thus, the chip provides fully two dimensional information on the image's global movement and centroid. Furthermore, the output of the 50×50 pixel contrast sensitive retina (Boahen & Andreou, 1992) (described earlier) can be viewed on an NTSC-compatible monitor with the addition of an external resistor, transistor, and oscillator. Thus, this highly integrated system provides three forms of information (normalized image, centroid, motion) with a minimum of external support circuitry.

5.2.1. *System Architecture and Design.* Figure 17 shows the architecture for the entire chip. Incident photons are first detected using a vertical pnp

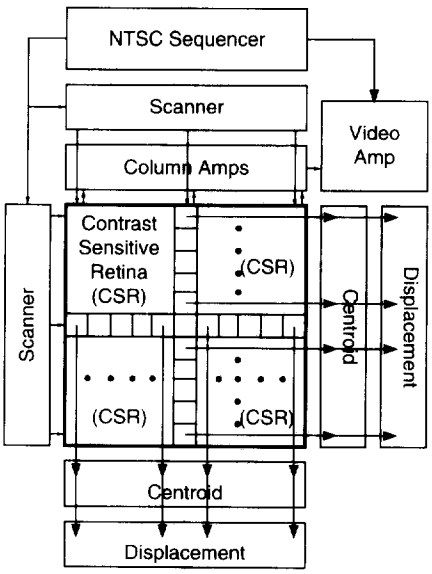


FIGURE 17. Functional organization of the motion/centroid/retina chip.

phototransistors. A corresponding microphotograph is shown in Figure 18. Subsequently, the raw image is passed through the contrast sensitive retina circuit described in the previous section and denoted by "CSR" in the diagram. By enhancing edges and removing gradients, the retina computes a version of the image which yields a more robust motion signal for tracking. In addition, removal of average illumination information greatly improves the invariance of the motion output to changes in ambient light.

Prior to any other computation, each of the retina pixel outputs is replicated and fed through a bank of amplifiers (one per column as shown in Figure 17) to an output driver. An on-chip state machine (NTSC sequencer) then generates the appropriate scanner clocks, synchronization, and blanking signals to produce an NTSC-compatible signal. The sequencer is a digital circuit that was compiled from a high level description of its function. The video amplifier circuit combines the blanking, sync, and pixel data and drives the composite video signal off the chip. As mentioned previously, a single transistor and resistor are all that is required to drive a standard NTSC monitor. In addition to the potential of possibly using this information as part of a larger, multi-chip system, the ability to view what the chip is "seeing" is an invaluable aid in aligning and focusing the

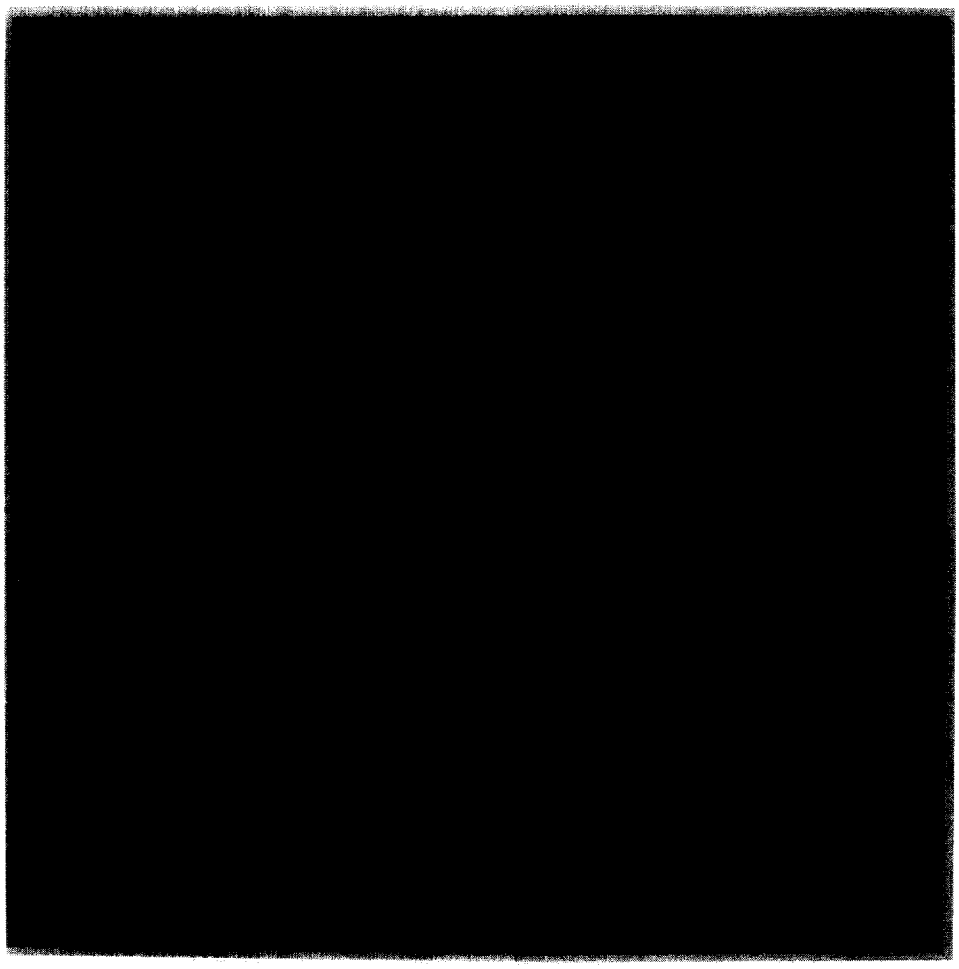


FIGURE 18. Photomicrograph of the motion/centroid/retina chip.

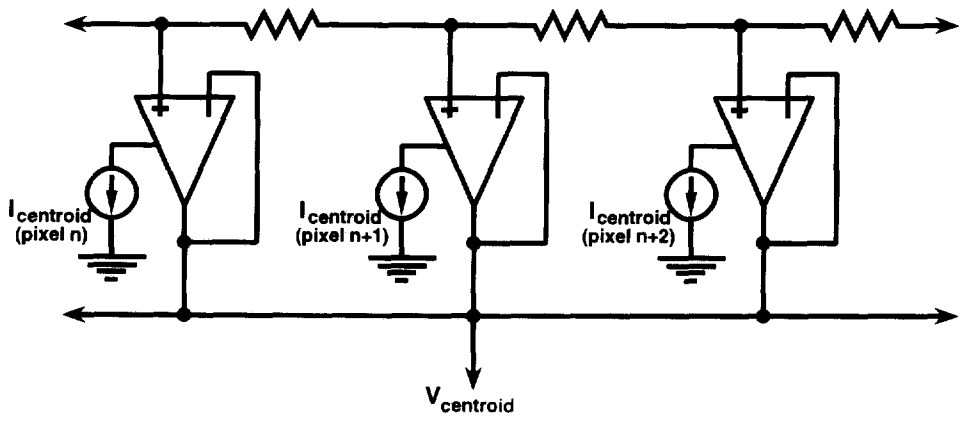


FIGURE 19. Centroid computation circuit.

image. Also, adjusting the retina's spatial bandpass characteristic to optimize the motion and centroid response is greatly simplified.

The pixel outputs of the retina's center row and column are also fed to the periphery of the array and are used for the motion and centroid computations. The image centroid is calculated using a scheme by DeWeerth (1992) which is shown in Figure 19. In this circuit, the centroid is computed using a series of transconductance amplifiers with the position encoded by a resistive divider across the array. The amplifier tail current, I_{centroid}, is proportional to the retina's output. Interestingly, the raw retina output, being contrast sensitive and having a fixed average value, does not contain adequate DC information to generate a strong centroid signal. Thus, an intermediate stage (see Figure 20) was added which effectively computes the absolute difference of a retina pixel's output and twice the average retina output current, I_{avg} (set by $V_{\text{retina\_bias}}$). In addition, an offset current controlled by V_{offset} can also be added for a total amplifier tail current of:

$$I_{\text{centroid}} = |I_{\text{pixel}} - 2I_{avg}| + I_{\text{offset}}. \tag{21}$$

This change effectively reintroduces DC information and allows the centroid circuit to operate properly when using the output of the contrast sensitive retina.

Simultaneous with the centroid computation, image motion in the form of displacement with respect to a fixed reference is also computed on the chip. The motion computation is performed using a rather direct implementation of the Reichardt detector (Andreou et al., 1989) except that the usual delay is replaced with a sample and hold circuit. The modified architecture for a single detector unit is shown in Figure 21. To summarize the operation, one can calculate that if the filtered image $I(x)$ is moving with a time varying displacement $s(t)$, then the detector's response can be found:

$$s(t)\left[\left(\frac{\partial I}{\partial x}\right)^2 - I\frac{\partial^2 I}{\partial x^2}\right]\mathrm{d}x \tag{22}$$

where the pixel separation, dx, approaches 0 (Meitzler et al., 1993). The output can be seen to be proportional to $s(t)$, which is the distance from the point at which the image is sampled. In the one-dimensional arrays

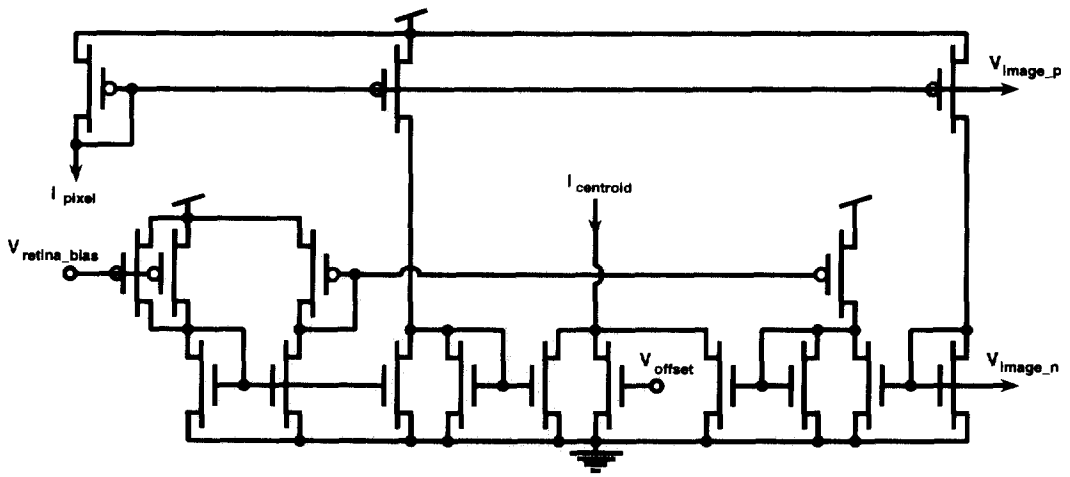


FIGURE 20. Centroid DC-restoring bias circuit.

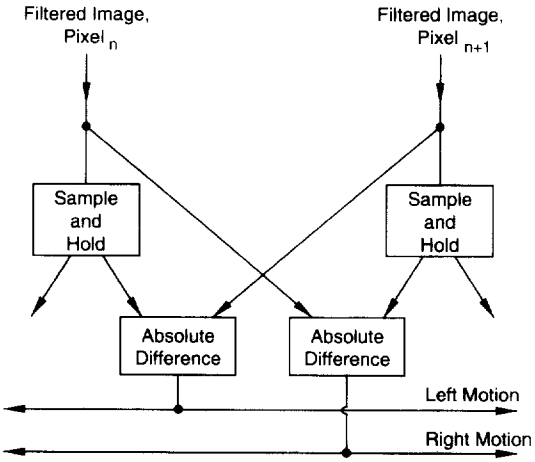


FIGURE 21. Modified Reichardt–Hassenstein motion computation architecture. This sub-block employs discrete time circuits and sampled data.

used on the chip, each pixel is shared by two detectors, one to the left and one to the right. Since only nearest neighbor interactions are taken into account, the displacement is assumed to always be less than one pixel. Thus, this architecture is best suited for use in a closed-loop tracking system in which the image's range of movement will be limited. Finally, note that in reality, the motion computation is performed by a large number of these detectors whose current outputs are aggregated and normalized to yield a robust estimate of the motion signal.

The sample and hold circuit is taken from a design by Vittoz et al. (1991) and is shown in

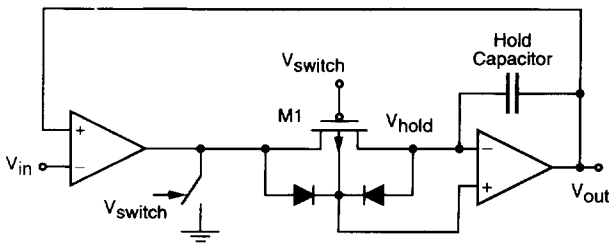


FIGURE 22. Sample and hold circuit (analog memory).

Figure 22. Instead of a simple capacitor and switch, two transconductance amplifiers are used to obtain a much slower decay of the held voltage, V_{hold}. This improvement is accomplished by using the amplifiers to form a virtual ground across the leakage diode formed by the bulk and right source/drain terminal of the switch transistor M_1. Although the circuit is more complex, the longer hold time is necessary for this architecture. With a faster decay of the held image, it would be necessary to resample frequently; therefore, since image motion cannot be calculated during resampling, the effective duty cycle of the system would be greatly diminished.

From Figure 21, it can be seen that the standard Reichardt architecture has been altered in another way. The multiplication operation originally proposed by Reichardt and used in earlier motion chips (Andreou et al., 1991b) has been replaced with a simpler absolute difference circuit, shown in Figure 23. In this circuit, although the inputs are voltages,

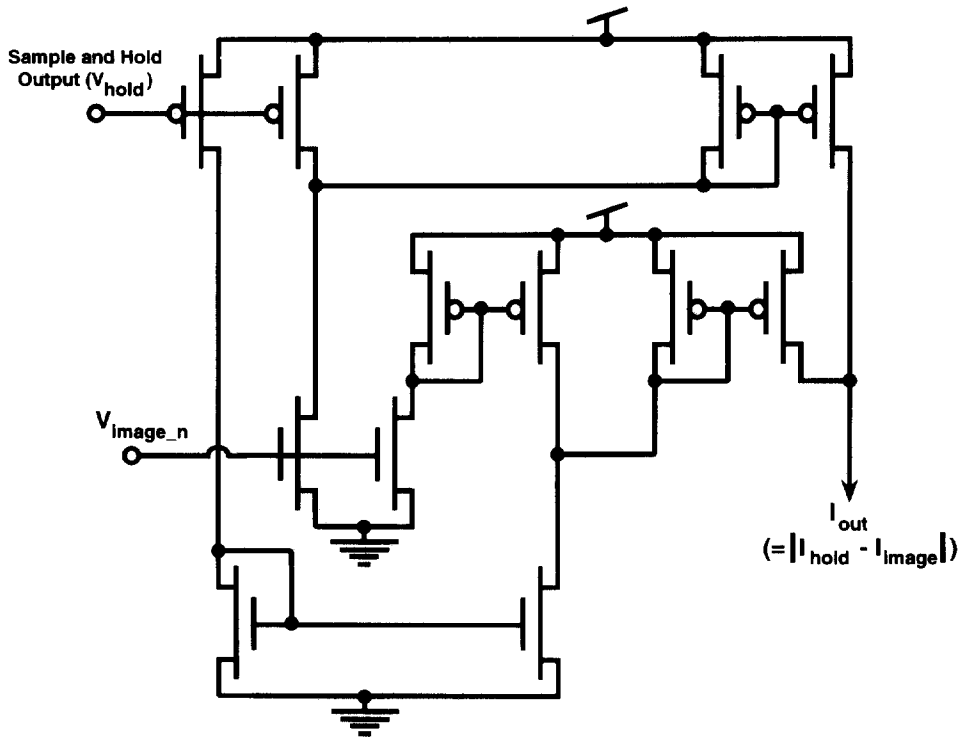


FIGURE 23. Absolute difference circuit.

they nonetheless encode the instantaneous and held currents from the retina array. Since the remainder of the circuit is composed of current mirrors, the absolute difference is a current-mode computation. The opponent currents are then normalized and passed off-chip where they are converted to voltages and subsequently differenced and amplified with an instrumentation amplifier.

5.2.2. *Chip Performance.* The chip was fabricated in a standard 2 μm double-poly CMOS process on a 6.8 mm $\times$ 6.9 mm die. Note that the NTSC sequencer and video amplifier can be disabled along with the on-chip digital circuits to reduce power consumption when video output is not needed. Even with the internal video external support circuitry, this system's power dissipation is tens of mW, which is orders of magnitude below what would be required to perform similar computations on a more traditional digital computer vision system.

Some results replicated from Meitzler et al. (1995) are shown in the following figures. The displacement computation circuitry was tested using an input image of six, three pixel wide stripes perpendicular to the motion-sensitive axis. The x- and y-axes DC response curves for a total image motion of one pixel are shown in Figure 24. Note that the response is normalized to partially compensate for variations caused by the image dependent terms in eqn (1); however, variation in the retina's phototransistor and bias transistor and in the output mirrors causes additional distortion in signal fed to the x- and y-axis displacement circuitry. It is therefore not surprising that the response curves do not have identical shapes. In addition, non-idealities in the circuitry and the nature of the algorithm make a linear response unlikely as well. The displacement circuit's AC response, shown in Figure 25, has a 3-dB point at approximately 70 Hz. Thus, the chip seems to have adequate bandwidth for a variety of image tracking applications.

Another factor affecting the utility of the motion circuitry is the performance of the on-chip analog

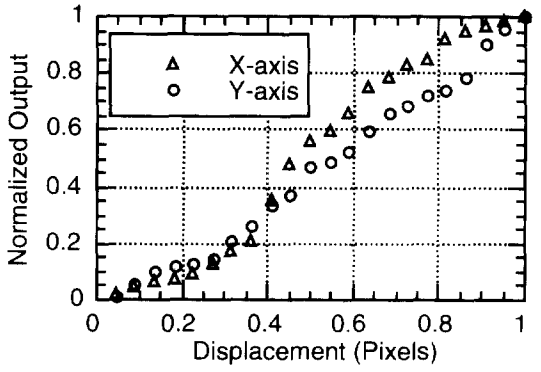


FIGURE 24. Displacement circuitry DC response.

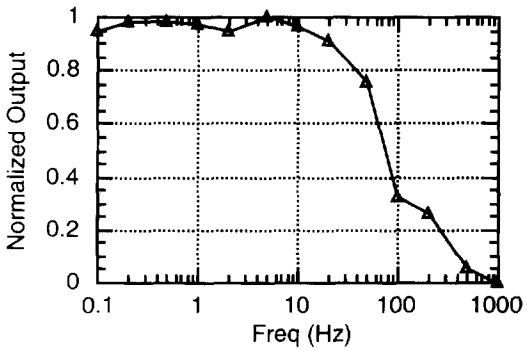


FIGURE 25. Displacement circuitry AC response.

memory (sample and hold circuit). Measurements of the chip's output, taken by sampling an image and then shifting it one pixel, are shown in Figure 26 for an average over ten trials. The output decay rate is approximately 19%/min and therefore should not be a problem so long as the image does not need to be continuously held for over a minute. Note that the displacement circuit's output decay and frequency response were almost completely independent of the sample and hold's overall biasing levels. Thus, the sample and hold could be biased in subthreshold for increased power savings.

In summary, from the standpoint of autonomous image acquisition and pre-processing, the integration of both centroid and image displacement computations on a single chip resulted in a compact and low power solution to several problems. The chip could potentially find application in a large number of tracking/location systems, especially those in which power consumption and physical size are limiting factors. When stabilizing an image, displacement is a potentially more useful control variable than velocity because offset errors may not be integrated [see Meitzler et al. (1993)].

As an indicator of the practicality of these types of processors, please note that the design of this system was motivated by the need to stabilize an image of the sun in the Flare Genesis project being conducted in part by the Johns Hopkins Applied Physics Laboratory. Because the instrument platform will be carried

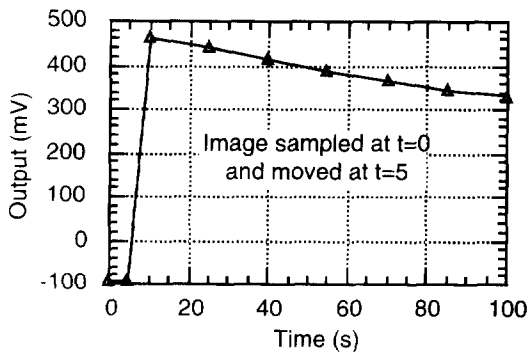


FIGURE 26. Displacement circuitry output decay.

by balloon to an altitude of 35,000 m, size and weight are of critical importance. Thus, the highly integrated analog VLSI system presented in this paper is an ideal solution. The chip is inserted in a control loop to compensate for image motion, resulting in higher resolution photographs showing the fine structure of the sun's surface and is in place for the upcoming scientific mission.

5.3. Other Analog VLSI Systems

One of the most fundamental functions that is found in the early vision of biological systems, is *temporal adaptation* to changes in the image level that are below the bandwidth of interest. This capability can also be built into subthreshold focal-plane circuitry to effectively generate a temporally high-passed version of the input. One such circuit is the adaptive photoreceptor design by Delbrück, shown in Figure 27 (Delbrück & Mead, 1994a, b).

The operation of this circuit centers on the amplification of the difference between the DC state of the system and the instantaneous value of the input. The pixel's bias point is stored on the gate of M_4, which can be considered fixed for small signal, high frequency changes due to C_1. Thus, when the input photocurrent moves from the bias point, the source voltage of M_4 will be forced to change (dropping for increasing photocurrent and vice versa) a very small amount to compensate. This change is then amplified by M_1, M_2, and M_3.

The key to the temporal differentiation and adaptation lies in the combination of M_5, C_1, and C_2. As shown in Figure 27, M_5 and C_1 operate as a low-pass filter; however, M_5 is configured as a non-linear conductance. The circuit uses the lateral bipolar mode for M_5 when the gate voltage of M_4 is greater than the drain of M_3. As a result, the M_5 structure is effectively two parallel opposing diodes. Thus, for small excursions from the DC bias point, the voltage across M_5 is small and the gate of M_4 is charged very slowly. A strong temporal high-pass response is the net result. For larger input changes, charging occurs more quickly, allowing the gate of M_4 to adapt to the new input level and set new bias conditions. A more thorough discussion of the circuit's subtleties is contained in Delbrück and Mead (1994b). For our purposes, it is most important to note that this design is a low power solution to the problem of detecting temporal transients.

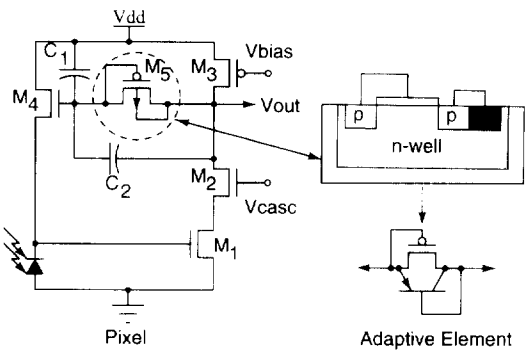


FIGURE 27. Delbrück's adaptive pixel circuit.

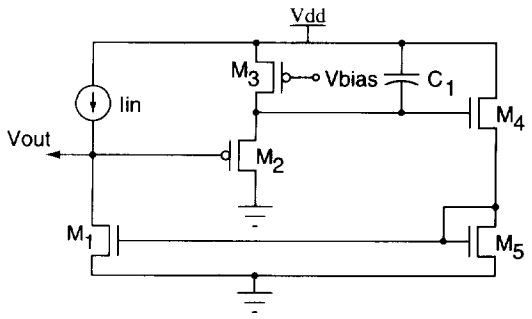


FIGURE 28. Chong's adaptive pixel circuit.

Another design is the temporal differentiator circuit by Chong et al. (1992). The underlying principle is much the same as in the previous case, i.e., delayed negative feedback caused by integrating output-supplied charge on a capacitor. In this particular case (see Figure 28 for the circuit) the delay is accomplished using the transconductance of M_2 and C_1 with M_3 as a bias. Since M_2's transconductance is effectively set by V_{bias}, the sensitivity of the circuit to temporal changes can be manually adjusted.

If one compares these two circuits, one can see that Chong et al.'s circuit is nearly a one-sided differentiator, i.e., there is a large response only to positive increases in the input current. This behavior arises from the fact that C_1 can be charged arbitrarily slowly through M_3 but is discharged very rapidly through M_2. Therefore, the negative feedback experiences only a very small delay for decreases in the input current which greatly reduces the temporal differentiation. Such one-sided behavior may, however, be desirable depending on the nature of subsequent processing.

Finally, note that the transistor count for Chong et al.'s pixel is at least as good as Delbrück's, even without a cascading operation to speed up the response.

6. DISCUSSION

On the approach: In the previous sections, we have seen how an analysis-by-synthesis methodology (Mead, 1989) using analog computation and VLSI technology has led to the development of energetically efficient analog VLSI systems for early vision. Crucial to the success of our endeavour is a hierarchical view of information processing as

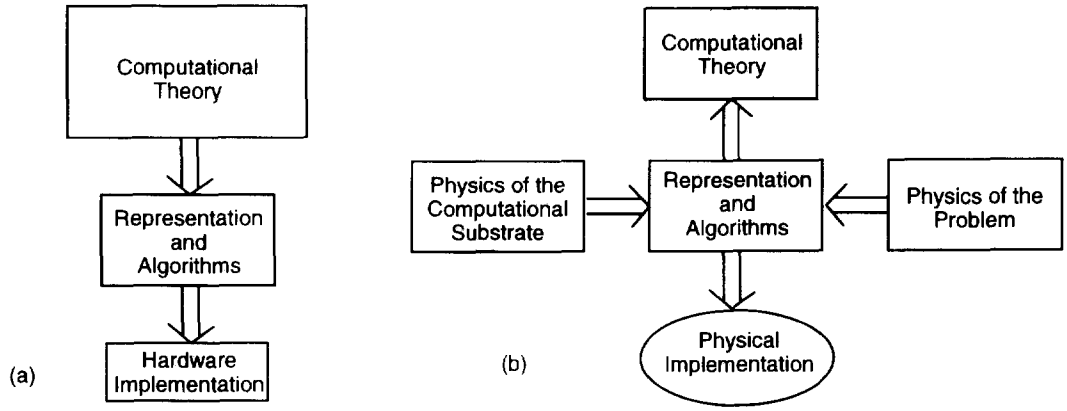


FIGURE 29. (a) Marr's three levels of looking at complex information processing systems. (b) An alternative view.

discussed in Chapters 1 and 7 of Marr (1982). Marr strongly believed, however, that computational theory should be at the top of the hierarchy and play the most important role, while the particulars of the implementation have only a peripheral role (see Figure 29a).

Our work suggests that it may be beneficial to view the different levels from a slightly different perspective, which is depicted in Figure 29b. We begin with the physics of the problem and the physical properties of the computational substrate. Good algorithms and representations emerge as a result of constraints imposed at this level.

This perspective is directly applicable to the problem at hand. It is well established that in the back-end of an ATR system, the physics electromagnetic radiation in the atmosphere (energy flow), and the physical properties of the imaged structures must all be considered to accurately model the scene, the target, and the process of signal degradation in the environment. A similar approach must be taken at the front end, where the physical properties of signal transducers (for example, the FLIR sensor or other infra-red sensing arrays) must be taken into account. Therefore, research aimed at understanding what are the ultimate performance limitations in AOR systems and their engineering must include a discussion of fundamental limitations in microelectronics technology (Keyes, 1987).

On algorithms and architectures: We have experimentally demonstrated that, in considering possible algorithms and architectures for solving sensory communication problems, one need not be restricted to a particular model of computation. The a priori assumption should be made that a structure exists (within a well defined set of "real" constraints consistent with the computational substrate). This approach incorporates the best possible model of computation. The particular mapping of a model to a computational substrate is thus guided by fundamental limitations of the basic elements, the properties that make the solution scalable, and the existence of a synthesis procedure that enables the emergence of a complex structure.

For example, it is easier today to write software that implements a filter function on a digital computer, than to implement the filter function using ASIC digital circuits, than to design and implement analog filters as analog integrated circuits, than to design and manufacture a filter based on the physical properties of some mechanical silicon micro-structure. Given the subject matter of this paper, there is no reason to believe that the last solution is not the preferred solution given adequate research resources to solve "algorithm" and technological problems.

On physical models: The charge-based formulation and analog VLSI implementation of the silicon retina presented here is an example of a physical model that could be cast in the dynamical systems framework (a relaxation network). Our method is mathematically interesting, and at the same time perhaps practical. Indeed, by judiciously employing "physical models" of computation such as a detailed biophysical model of a retina (Boahen & Andreou, 1992) the inherent parallelism and nature of physical laws (Hillis & Boghosian, 1993) is exploited in the computational process.

It can be argued that the analog VLSI retina model has an a priori internal model of the world—one that assumes that the intensity is either uniform or, in the case of non-uniform illumination, is a linear function of space. The output of the system is the difference between the input intensity field and the model. As such, the output is a measure of the second spatial derivatives (or the Laplacian) of the intensity field. In the field of computer vision, linear methods based on regularization theory are used to impose smoothness constraints (Poggio et al., 1985) on the discretely sampled and noisy real world data. These computationally demanding algorithms are run on general purpose digital hardware.

In the physical realization of a computational systems, the same "regularization" benefits could be beneficial in dealing with the "noise" introduced by the variability in gain of MOS transistors (see Figure 3).[1] Thus, we see how in the organization of the system one could account for the properties of the computational substrate at the architectural level. Such properties are irrelevant when implementing algorithms on general purpose, digital computers. In digital computers and symbolic processing machines, structural variability and noise in the basic elements is handled at a much lower level, at the gate level. Switching levels are chosen so that adequate noise margin is introduced for large scale reliable computation [see further discussion in Andreou (1994) and references therein].

In the context of the biological model, the function of the horizontal cells (corresponding to nodes H) is to "optimally" compute a smoothed version of the image (through a convolution with the kernel shown in Figure 20) while the cones (corresponding to nodes C) perform edge enhancement by taking the Laplacian of the smoothed image as given by eqn (18). The space constant of the solutions is $\lambda^{1/4}$ or $(\hat{g}h)^{1/4}$. The model suggests that specialized structures in biological systems could effect some type of "wet-ware regularization" to compensate for the inherent random variations in the neuronal characteristics. Such a property could in turn lead to robust performance in the presence of "noise". The latter statement is just a hypothesis subjected to experimental verification.

The notion of an "optimal" computation step has been introduced by Bialek and Owen (1990). They have considered the signal and noise characteristics of the photoreceptors in the outer retina and they have derived "optimal" temporal filters to further process the receptor signals. Our work (Boahen & Andreou, 1992) addresses a similar problem in the space domain where "noise" is introduced by the structural variability in the gain of the individual elements and spatial smoothing is needed to increase the information capacity of the system.

The contrast sensitive silicon retina is an architecture that yields the ON-center/OFF-surround response at the level of the cone (photoreceptor) network. Even though from an engineering perspective one can employ this function for edge enhancement (as we have done), the question of why such a structure exists in the neural system is still open. To put it more succinctly; is edge enhancement the goal or is it simply an emerging property from a computational function that is aimed at dealing with signals of large dynamic range using imprecise components?

On large scale analog computation: The analog VLSI system presented in this paper is essentially an *analog floating point* processor. As a first step, the system computes the range (the voltages in the horizontal cell synthycium correspond to the value of the exponent). This level is the operating point of the system. Note that the automatic gain control is also achieved via this computation. At a given operating point, sophisticated spatial filtering is performed to smooth the sampled data and enhance the edges. Having separated the problem of precision and dynamic range, the signal processing within the range can be done with low precision analog hardware. The issue of precision versus dynamic range in analog circuits was addressed by Barrie Gilbert 10 years ago with his elegant implementation of an "array normalizer" that used bipolar transistors and current-mode translinear circuits (Gilbert, 1984). The system presented here is similar in two ways to Gilbert"s array normalizer. First, there is *local* normalization of the input current signals. Second, all processing is done in the current domain where the translinear properties of MOS subthreshold devices are exploited to implement the required functions.

7. CONCLUSIONS

Our research was aimed at exploring different ideas on neuromorphic computations and their VLSI implementations for image acquisition and preprocessing in AOR. The results of our investigation are encouraging. It has been demonstrated that analog circuits of limited precision, when assembled in large networks following an appropriate design methodology, can successfully perform linear and non-linear computation with an energetic efficiency unmatched by their digital counterparts. The 590,000 transistor analog VLSI, contrast sensitive, silicon retina and the integrated tracking systems are small steps towards the engineering of truly intelligent machines.

NOTE: A small number of the two analog VLSI systems that were discussed in this paper can be made available for experimentation by researchers that are interested in autonomous object/target tracking/recognition systems.

REFERENCES

Andreou, A. G. (1990). Synthetic neural systems using current-mode circuits. In *Proc. 1990 IEEE Int. Symp. on Circuits and Systems* (pp. 2428–2432).

Andreou, A. G. (1991). Electronics art imitate life. *Nature*, **19**, 26.

Andreou, A. G. (1994). On physical models of neural computation and their analog VLSI implementation. *Proc. 1994 Workshop*

[1] Noise here denotes structural variability, as opposed to thermodynamic noise.

on Physics and Computation, IEEE Computer Society Press, Los Alamitos, CA, pp. 255–264.

Andreou, A. G., & Boahen, K. A. (1994a). A 48,000 pixel, 590,000 transistor silicon retina in current-mode subthreshold CMOS. In *Proc. 37th Midwest Symp. on Circuits and Systems.*

Andreou, A. G., & Boahen, K. A. (1994b). Neural information processing II: The current-mode approach. In M. Ismail, & T. Fiez (Eds.), *Analog VLSI* (Chapter 8). New York: McGraw-Hill.

Andreou, A. G., & Boahen, K. A. (1996). Translinear circuits in subthreshold CMOS. *J. Analog Integrated Circuits and Signal Proc.*, **8**.

Andreou, A. G., Strohbehn, K., & Jenkins, R. E. (1989). A hardware implementation of the Reichardt motion detector. In *Neural Networks for Computing* (Abstract).

Andreou, A. G., Boahen, K. A., Pouliquen, P. O., Pavasović, A., Jenkins, R. E., & Strohbehn, K. (1991a). Current-mode subthreshold MOS circuits for analog VLSI neural systems. *IEEE Trans. Neural Networks*, **2**(2), 205–213.

Andreou, A. G., Strohbehn, K., & Jenkins, R. E. (1991b). Silicon retina for motion computation. In *Proc. 1991 IEEE Int. Symp. on Circuits and Systems* (pp. 1373–1376).

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, **1**, 295–311.

Bialek, W., & Owen, W. G. (1990). Temporal filtering in retinal bipolar cells: Elements of an optimal computation? *Biophys. J.*, **58**, 1227–1233.

Boahen, K. A., & Andreou, A. G. (1992). A contrast sensitive silicon retina with reciprocal synapses. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems* (Vol. 4, pp. 764–772). San Mateo, CA: Morgan Kaufmann.

Boahen, K. A., & Andreou, A. G. (1993). Design of a bidirectional associative memory chip. In M. Hassoun (Ed.), *Associative neural memories: theory and implementation* (Chapter 17). New York: Oxford University Press.

Boahen, K. A., Andreou, A. G., Pouliquen, P. O., & Pavasović, A. (1989). Architectures for associative memories using current-mode analog MOS circuits. In C. Seitz (Ed.), *Proc. Decennial Cal Tech Conf. on VLSI.* Cambridge, MA: MIT Press.

Boser, B. E., Säckinger, E., Bromley, J., Cun, Y. L., & Jackel, L. (1991). An analog neural network processor with programmable topology. *IEEE J. Solid-State Circuits*, **26**, 2017–2025.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proc. R. Soc. (London) B*, **220**, 89–113.

Buhmann, J. M., Lades, M., & Eeckman, R. (1994). Illumination-invariant face recognition with a contrast sensitive silicon retina. In *Advances in neural information processing systems 6.* San Mateo, CA: Morgan Kaufmann.

Cauwenberghs, G., Neugebauer, C. F., & Yariv, A. (1992). Analysis and verification of an analog VLSI incremental outer-product learning systems. *IEEE Trans. Neural Networks*, **3**(3).

Chandrakasan, A., Burstein, A., & Brodersen, R. W. (1992). A low power chipset for portable multimedia applications. In *Proc. IEEE Int. Solid-State Circuits Conf.*

Chong, C. P., Salama, C. A. T., & Smith, K. C. (1992). Image-motion detection using analog VLSI. *IEEE J. Solid-State Circuits*, **27**(1), 93–96.

Delbrück, T. (1993). Silicon retina with correlation-based, velocity-tune pixels. *IEEE Trans. Neural Networks*, **4**(3), 529–541.

Delbrück, T., & Mead, C. A. (1994a). Adaptive photoreceptor with wide dynamic range. In *Proc. Int. Symp. Circuits and Systems (ISCAS)* (Vol. 4, pp. 339–342). Piscataway, NJ: IEEE.

Delbrück, T., & Mead, C. A. (1994b). Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits. Technical Report CNS Memo No. 30, California Institute of Technology, Pasadena, CA.

Deweerth, S. P. (1992). Analog VLSI circuits for stimulus localization and centroid computation. *Int. J. Computer Vision*, **8**(3), 191–202.

Dowling, J. E. (1987). *The retina: an approachable part of the brain.* Cambridge, MA: Belknap Press of Harvard University Press.

Duda, R. O. & Hart, P. E. (1973). Pattern classification and scene analysis. New York: John Wiley and Sons.

Etienne-Cummings, R. R., Fernando, S. A., der Spiegel, J. V., & Mueller, P. (1992). Real-time 2-d analog motion detector VLSI circuit. In *Proc. Int. Joint Conf. on Neural Networks* (pp. 426–431). Baltimore, MD: IEEE.

Furth, P. M., & Andreou, A. G. (1995). Linearized transconductors in subthreshold CMOS. *Electronics Letters*, **31**(7), 545–547.

Gilbert, B. (1975). Translinear circuits: A proposed classification. *Electronics Letters*, **11**(1), 14–16, 136.

Gilbert, B. (1984). A monolithic 16-channel analog array normalizer. *IEEE J. Solid-State Circuits*, **19**(6).

Gorin, A. L., Levinson, S., Gertner, A., & Goldman, E. (1991). Adaptive acquisition of language. *Comput. Speech Language*, **5**(2), 101–132.

Grossberg, S. (Ed.) (1988). *Neural networks and natural intelligence.* Boston, MA: MIT Press.

Haykin, S. (1994). *Neural networks: a comprehensive foundation.* New York: McMillan College Pub.

He, Y., Cilingiroglu, U., & Sánchez-Sinencio, E. (1993). A high density and low-power charge-based hamming network. *IEEE Trans. VLSI Systems*, **1**(1), 55–62.

Hildreth, E. C. (1983). *The measurement of visual motion.* ACM Distinguished Dissertation Series, Cambridge, MA: MIT Press.

Hillis, D. W., & Boghosian, B. M. (1993). Parallel scientific computation. *Science*, **261**, 856.

Holler, M., Tam, S., Castro, H., & Benson, R. (1989). An electrically trainable artificial neural network ETANN with 10420 floating gate synapses. In *Proc. Int. Joint Conf. Neural Networks* (Vol. II, pp. 191–196).

Horio, Y., & Nakamura, S. (1992). Analog memories for VLSI neurocomputing. In E. Sánchez-Sinencio, & C. Lau (Eds.), *Artificial neural networks: paradigms, applications, and hardware implementations* (pp. 344–366). Piscataway, NJ: IEEE Press.

Horiuchi, T., Bair, W., Bishofberger, B., Moore, A., Koch, C., & Lazzaro, J. (1992). Computing motion using analog VLSI chips: An experimental comparison among different approaches. *Int. J. Computer Vision*, **8**(3), 203–216.

Keyes, R. W. (1987). *The physics of VLSI systems*, chapter 1. Wokingham, UK: Addison-Wesley.

Koch, C. (1989). Seeing chips: Analog VLSI circuits for computer vision. *Neural Computation*, **1**, 184–200.

Kohonen, T. (1988). *Self-organization and associative memory*,2nd ed. Berlin: Springer-Verlag.

Lazzaro, J., Wawrzynek, J., Mahowald, M., Sivilotti, M., & Gillespie, D. (1993). Silicon auditory processors as computer peripherals. *IEEE Trans. Neural Networks*, **4**(3), 523–528.

Li, Z., & Atick, J. J. (1994). Toward a theory of the striate cortex. *Neural Computation*, **6**(1), 127–146.

Linsker, R. (1986). Self-organization in a perceptual network. *IEEE Computer*, 105–117.

Mahowald, M., & Douglas, R. (1991). A. silicon neuron. *Nature*, **354**, 515–518.

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

Marr, D., & Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosci. Res. Program Bull.*, **15**, 470–488.

Mead, C. A. (1989). *Analog VLSI and neural systems.* Reading, MA: Addison-Wesley.

Mead, C. A. (1990). Neuromorphic electronic systems. *Proc. IEEE*, **78**(10), 1629–1636.

Mead, C. A. & Delbrück, T. (1991). Scanners for visualizing activity of analog VLSI circuitry. *Analog Integrated Circuits and Signal Processing*, **1**(2), 93–106.

Mead, C. A., & Ismail, M. (Eds.), (1989). *Analog VLSI implementation of neural systems*. Norwell, MA: Kluwer Academic Publishers.

Mead, C. A., & Mahowald, M. A. (1988). A silicon model of early visual processing. *Neural Networks*, **1**, 91–97.

Meitzler, R. C., Andreou, A. G., Strohbehn, K., & Jenkins, R. E. (1993). A sampled-data motion chip. In *36th Midwest Symp. Circuits and Systems* (Vol. 1, pp. 288–291). Detroit, MI: IEEE.

Meitzler, R. C., Strohbehn, K., & Andreou, A. G. (1995). A silicon retina for 2-D position and motion computation. In *Proc. ISCAS* (pp. 2096–2099). Seattle, WA: IEEE.

Pavasović, A., & Andreou, A. G. (1994). Characterization of subthreshold MOS mismatch in transistors for VLSI systems. *J. VLSI Signal Processing*, **8**, 75–85.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, **317**, 314–319.

Pouliquen, P. O., Andreou, A. G., Strohbehn, K., & Jenkins, R. E. (1993). An associative memory integrated system for character recognition. In *Proc. 36th Midwest Symp. on Circuits and Systems* (pp. 762–765).

Ramacher, U., & Ruckert, U. (Eds.) (1991). *VLSI design of neural networks*. Boston, MA: Kluwer Academic Publishers.

Reichardt, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. A. Rosenbluth (Ed.), *Sensory communication* (pp. 303–317). Cambridge, MA: MIT Press.

Roe, D. B., & Wilpon, J. G. (Eds.) (1994). *Voice communication between humans and machines*. Washington, D.C.: National Academy Press.

Rosenfeld, A. (1988). Computer vision: Basic principles. *Proc. IEEE*, **76**(8), 857–1056.

Sánchez-Sinencio, E., & Lau, C. (Eds.) (1992). *Artificial neural networks: paradigms, applications, and hardware implementations*. Piscataway, NJ: IEEE Press.

Scribner, D. A., Sarkady, K. A., Kruer, 'M. R., Caulfield, J. T., Hunt, J. D., Colbert, M., & Descour, M. (1993). Adaptive retina-like preprocessing for imaging detector arrays. In *Proc. Int. Conf. Neural Networks*.

Scribner, D. A., Fisher, J., Caulfield, J. T., Colbert, M., Sarkady, K. A., Zadnik, J., Satyshur, M. P., Kruer, M. R., & Brooks, J. (1994). On-chip image processing for staring infrared focal plane arrays using retina-like techniques. In *Proc. IRIS Detector Meeting*.

Sheng, S., Chandrakasan, A., & Brodersen, R. W. (1992). A portable multimedia terminal. *IEEE Communic. Mag.*, **30**(12), 64–75.

Srinivasan, M. V., Laughline, S., & Bubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proc. R. Soc. (London) B*, **216**, 427–459.

Tanner, J. E., & Mead, C. A. (1986). An integrated analog optical motion sensor. In S. Y. Kung, R. E. Owen, & J. G. Nash (Eds.), *VLSI signal processing II* (pp. 59–76). Piscataway, NJ: IEEE Press.

Taylor, J. G. (1990). A silicon model of vertebrate retinal processing. *Neural Networks*, **3**, 171–178.

Tsividis, Y. P. (1987). Analog MOS integrated circuits–certain new ideas. *IEEE J. Solid State Circuits*, **22**, 317–321.

U.S. Army ATR Report (1994). Report of the Working Group on Automatic Target Recognition. U.S. Army Research Office.

Vittoz, E. A. (1985a). The design of high-performance analog circuits on digital cmos chips. *IEEE J. Solid State Circuits*, **SC-20**(3), 657–665.

Vittoz, E. A. (1985b). Micropower techniques. In Y. P. Tsividis (Ed.), *VLSI for telecommunications*. Englewood Cliffs: Prentice Hall.

Vittoz, E. (1994). Analog VLSI signal processing: Why, where and how? *J. Analog Integrated Circuits and Signal Processing*, **6**, 27–44.

Vittoz, E., & Arreguit, X. (1993). Linear networks based on transistors. *Electronics Letters*, **29**, 297–299.

Vittoz, E. A., & Fellrath, J. (1977). CMOS analog integrated circuits based on weak inversion operation. *IEEE J. Solid-State Circuits*, **12**(3), 224–231.

Vittoz, E. A., Oguey, H., Maher, M. A., Nys, O., Dijkstra, E., & Chevroulet, M. (1991). Analog storage of adjustable synaptic weights. In U. Ramacher, & U. Ruckert (Eds.), *VLSI design of neural networks* (pp. 47–63). Boston, MA: Kluwer Academic Publishers.

Yagi, T., Funahasi, Y., & Ariki, F. (1989). Dynamic model of dual layer neural network for vertebrate retina. In *Proc. Int. Joint. Conf. Neural Networks* (pp. 1787–1789). Washington, D.C.: IEEE.

Yang, H., Sheu, B. J., & Lee, J.-C. (1992). A nonvolatile analog neural memory using floating-gate MOS transistors. *Analog Integrated Circuits and Signal Processing*, **2**(1).

Yang, K., Meitzler, R. C., & Andreou, A. G. (1994). A model for MOS effective channel mobility with emphasis in the subthreshold and transition regions. In *Proc. Int. Symp. Circuits and Systems* (Vol. 1, pp. 429–432). London: IEEE.