A 590,000 transistor 48,000 pixel, contrast sensitive, edge enhancing, CMOS imager –silicon retina–

Andreas G. Andreou
Electrical and Computer Engineering
The Johns Hopkins University, Baltimore, Maryland 21218 USA
Kwabena A. Boahen
CNS Program, Caltech, Pasadena, California 91125 USA
E-mail: andreou@jhunix.hcf.jhu.edu

Abstract

We present an experimental analog VLSI focal plane processor for the phototransduction, local gain control and edge enhancement of natural images. The single chip system incorporates 590,000 transistors in 48,000 pixels, and it has been fabricated on a 9.5×9.3 *mm die in a* $1.2\mu m$ *n–well double metal, double poly, digital oriented CMOS technology. The organization of the system abstracts from the structure and function of the vertebrate distal retina. The adopted design style, current–mode subthreshold CMOS using circuits of minimal complexity offers the possibility of ultra low power dissipation and area efficiency, commensurate with VLSI integration.*

1 Introduction

Over the last few years, new emerging opportunities in information technologies point towards markets for portable systems where battery operation, light weight and small factor will be in demand. From an economic perspective, miniaturization and high levels of system integration, with an implicit potential for large markets are predicted to be the technology drivers in the decades ahead [1]. From a technology and engineering perspective, the development of these systems will be done with *energetic efficiency* as the prime engineering constraint, taking a lead over other considerations. The *cost and reliability* of these portable systems are also important factors.

A distinct characteristic of these information technologies is their direct interface to people and real world environments. Undoubtly, portable operation and battery operation imposes severe constraints and a tight energy budget. However, with the widespread deployment of such technologies, there is another issue that is becoming increasingly more important. **Mobile operation implies that the speech and vision interfaces, much like other communication interfaces, must be capable of operating under highly variable environmental conditions**. To make this point clear we consider an example from vision.

0-8186-7047-9/95 $04.00

Image acquisition and early vision processing under naturally occuring illumination conditions is a task typical in the fields of robotics, prosthetic devices for the blind, and motor-vehicle navigation. Today this task is accomplished in two separate steps. First the light intensity is recorded through a standard imager such as a CCD camera. This 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 steps such as edge extraction or encoding may follow. However, even though the precision necessary for these tasks rarely exceeds 8 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 in the scene varies dramatically within a single frame (see Figure 1). The detrimental effects of non-uniform illumination in the performance of a face recognition system have been experimentally investigated by Buhman, Lades and Eeckman [2].

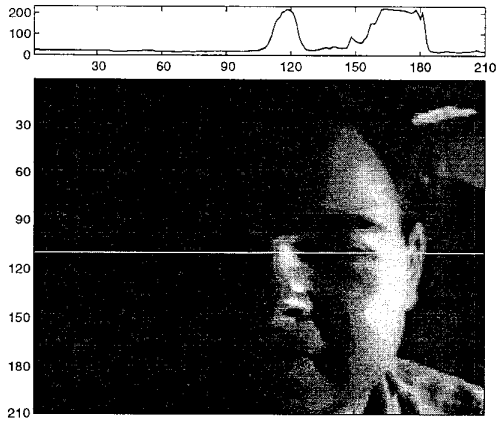

Figure 1. (Bottom) "Mark" as captured by a conventional camera. (Top) Intensity histogram at image line 110 (white line). The light source is positioned to the right side of the image and it introduces a large gradient in illumination within a single frame. This is clearly shown in the intensity histogram. 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.

1.1 The Biological Paradigm

Yet, biological organisms excel at solving problems in sensory perception and motor control, by sustaining high computational throughput with minimal energy dissipation. These are "real" physical systems, highly mobile and thus constraint by size, weight, and

the availability of energetic resources. They are also required to operate at temperatures near 300K where favorable conditions exist for the development of life.

Unlike most known forms of computing/calculating activity, 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 are useful for the ultimate goal of the organism, the interaction with the environment –in a closed loop configuration where the environment is integral part of a loop–. This interaction may involve only lower level functions such as early pereptual processing and sensorimotor coordination or it may include the higher level task of communication through speech and a natural language which is a unique characteristic to the human species.

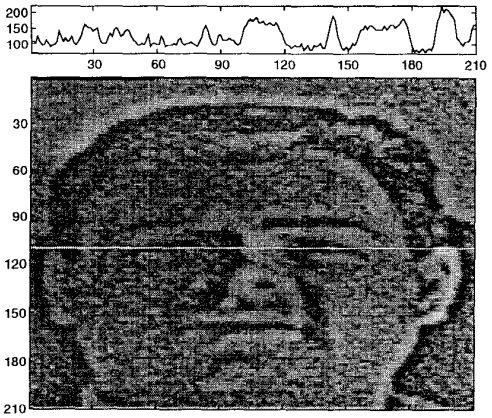

Figure 2. (Bottom) "Mark" as captured by our imager, a silicon model of the vertebrate distal retina. (Top) Intensity histogram. The light source is again positioned to the right side of the image and it introduces a large gradient in illumination within a single frame. The image captured by the silicon retina discards absolute illumination and preserves only local contrast information through local gain control at the pixel level. Unlike the image in Figure 1, the presence of a large illumination gradient does not degrade image acquisition here.

As such, neural computation is 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. One could argue, that the effectiveness of biological systems stems from their ability to deal effectively with the sources of *variability*. This suggest that an understanding of the *organizing* and *functionial* principles in biological information processing systems may be beneficial in the design of human engineered systems for performing similar tasks [3, 4]. This has been pursued over the last ten years by Carver Mead and his colleagues at Caltech [3]. Our work follows similar lines of thought [5].

In this paper, we discuss how such an approach has led to the development of an *analog* VLSI silicon system, a second generation, contrast sensitive, silicon retina. The architecture [6] is inspired by the processing performed at the outer plexiform layer of the vertebrate retina. It is mapped onto silicon using minimal complexity *current-mode* circuits that exploit *native* properties of subthreshold MOS transistors. High computational throughput at low levels of energy dissipation is achieved by employing analog processing in a massively parallel architecture. We begin with a discussion of technology and circuit techniques that are central to the implementation of the system.

2 Technology and Circuit Techniques

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 [8, 7]. The advantages of using standard digital CMOS processes for cost-effective engineering solutions to analog signal processing problems are surveyed in [8] and are also discussed in [9]. CMOS has the highest integration density attainable today, making it especially attractive for *analog VLSI* models of neural computation [3]. Moreover, the physical properties of silicon and its native oxides, together with recent advances in micromachining of electromechanical elements make silicon-based technologies the prime candidate for highly integrated, truly complex, analog computational systems. A final advantage is that CMOS silicon technologies are readily available for experimentation and rapid prototyping through foundry services at relatively low cost. This advantage of CMOS technologies accelerates the research and evolution of complex systems.

2.1 Current–Mode MOS Circuits

Computation when must be performed at the focal plane of an imager using circuits that are constrained to exist in essentially two dimensions, poses a challenging problem:

- First, 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.
- Second, active circuits other than phototransducers, produce heat which can increase the dark current of the phototransducing devices and the system performance can again be compromised.

However, we believe that even with a two dimensional implementation medium, it is possible to tradeoff light sensitivity and spatial resolution, for some essential processing at the focal plane. This is exactly what it has been done in the system discussed in this paper and the processing performed is that of robust local gain control.

The adopted design style, is current–mode subthreshold CMOS, using circuits of minimal complexity [10] offers the possibility of ultra low power dissipation with minimal complexity circuitry commensurate with VLSI integration. Subthreshold operation offers the highest processing rates per unit power. Current-mode (CM) operation yields large dynamic range,

simple and elegant implementations of both linear and nonlinear computations, and low power dissipation without sacrificing speed.

Our choice of subthreshold operation is based on the principle: **"Active devices should be used in the region where their transconductance per unit current is maximized."**. Because, this is the way to minimize the energy per operation and maximize the speed per unit power consumed:

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

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 $\sim I/g_m$. The transconductance per unit current increases as the current decreases—throughout the above-threshold and transistion regions—and reaches a maximum in the subthreshold region.

By taking advantage of the high subthreshold transconductance per unit current, voltage swings are kept to a few thermal voltages, and reasonable processing bandwidths are achieved. Dynamic power dissipation and supply noise are also reduced as a result of the smaller voltage swings. Smaller voltage swings eliminate the current that is wasted in charging and discharging parasitic capacitances, thereby allow us to use smaller current signals and cut quiescent power dissipation as well. Thus, this approach yields relatively fast analog circuits with power dissipation levels compatible with future trends in system integration. Fast digital circuits can also be designed using source-coupled logic gates and current steering (ECL-like circuits).

There also exists a powerful synthesis (and analysis) procedure which can be used to generate a wide variety of circuits that perform linear and non–linear analog operations in the current domain, and relies on the exponential form of current-voltage non-linearities. This procedure is based on what is known as the *Translinear Principle* [14] 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, as well as 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 [5]. 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.

To demonstrate how computational primitives emerge at the network level from device physics of the underlying technology, let us consider an example of a summing operation, *local aggregation*. We take a close look at *diffusion*, the physical process that underlies local aggregation in the nervous system, contrast it with the process of diffusion in MOS transistors and come up with a novel network design technique [6].

2.2 The Diffusor

The current in an MOS transistor operating in subthreshold ohmic regime is an exact difference of exponential functions of the drain and source voltages [7, 3, 5] and for an NMOS the current is given by:

$$I = I_0 \, S \, exp(\kappa \, V_{GB}) \, [exp(-V_{SB}) - exp(-V_{DB})] \tag{1}$$

substrate and are normalized to the thermal voltage (kT/q). The constant I_0 depends on mobility (μ) and other silicon physical properties. S is a geometry factor, the width W to length L ratio the device. The constant κ takes values between 0.6 and 0.9.

The exponential functions of voltage in the square brackets of Equation 1, correspond to Boltzmann distributed charges at the source and drain. A charge based representation of the current in an MOS transistor was presented by Maher and Mead in [11] and in [3]. Our Equation 2 corresponds to the second term (diffusion term) in the R.H.S of Equation 11 in Reference [11].

$$I \propto [\mathcal{Q}_S - \mathcal{Q}_D] \tag{2}$$

The charge-based representation depicted in Equation 2, suggests that the MOS transistor in subthreshold is a highly linear device in the charge domain; a property that finds many uses in analog circuit design. Such view of an MOS transistor in subthreshold as a basic diffusive element allows for the effective implementation of systems that exploit properties of elliptic partial differential equations. With the appropriate logarithmic loads connected to the source and drain, linear networks can be obtained. The same idea was more recently revisited by Vittoz and Arreguit [12].

2.3 Local Aggregation Networks

Local gain control, an ultimate goal in the design of our system, necessiates the computation of a current representing a local spatial average of the incident illumination. Local aggregation (averaging), the addition of signals over a confined region of space occurs throughout the nervous system. Aggregation was discussed in Chapter 6 of [3], (also in [13]), and it is the basis for many neuromorphic silicon VLSI systems described therein.

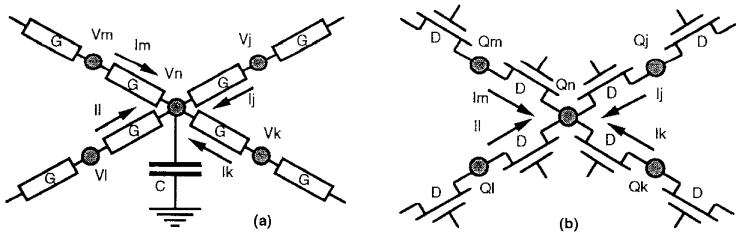

Figure 3. Local aggregation using conductances and voltage/current variables (left) or diffusors and charge/current variables (right).

Two alternative analog implementations of this process on a discrete grid are shown in Figure 3. The first network uses voltages and currents (Figure 3a). Its node equation is

$$\frac{dV_n}{dt} = \frac{4G}{C}\left(\frac{1}{4}(V_j + V_k + V_l + V_m) - V_n\right) \tag{3}$$

Note that term in large parenthesis is a first-order approximation to the Laplacian. However, this solution is not amenable to VLSI integration because transconductances (G) with a large linear range consume large amounts of area and power.

The second network uses charges (positive) and currents (Figure 3b). Its node equation is

$$\frac{dQ_n}{dt} = 4D\left(\frac{1}{4}(Q_j + Q_k + Q_l + Q_m) - Q_n\right) \tag{4}$$

Note that dQ_n/dt is the same as the current supplied to node n by the network. This solution is easily realized by exploiting diffusion in subthreshold MOS transistors. It was shown earlier that in the MOS transistor, the current is linearly proportional to the charge difference across the channel (See Equation 2). Therefore, the diffusion process may be modeled using devices with identical geometry S and identical gate voltages. The former guarantees they have the same diffusivity and the latter guarantees that the charge concentrations at all the source/drains connected to node n are the same and equal Q_n.

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 otherhand, the network in Figure 4 represents the solution by charge concentrations $\mathcal{Q}_S$ and $\mathcal{Q}_D$ at source/drains—not the charge on the node capacitance. The source/drain charge cannot be measured directly, it can be inferred from the node voltage.

2.4 Loaded Networks

The silicon implementation of a retina model also requires circuits for "loaded" networks. This is the case were high conductances are attached to the nodes where currents are injected. To observe the behaviour of a loaded network, we begin with a small segment of a one dimensional network (Figure 4).

A voltage mode circuit model for a loaded network is shown in Figure 4(left) 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 referenced to ground, represent the state of the network and can be read out using a differential amplifier with the negative input grounded.

The equivalent circuit using idealized non-linear conductances is shown in Figure 4(right). The difference in currents through the diodes D_1 and D_2 are linearly related to the current through the diffusor MOS transistor. This relationship can be derived from Equation 1 describing subthreshold conduction, and the ideal diode characteristics where $I_D = I_S \exp[qV_D/(kT)]$. 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:

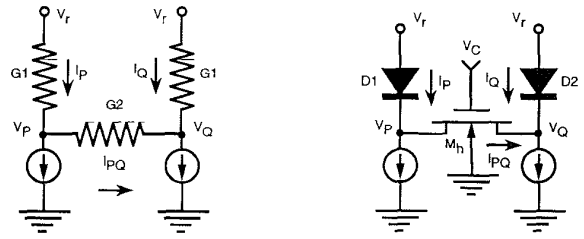

Figure 4. Building blocks for linear networks using ideal (left) linear and (right) non-linear elements.

$$I_{PQ} = \left(\frac{SI_{on}}{I_S}\right) \exp\left[\frac{\kappa V_C - V_r}{(kT/q)}\right](I_Q - I_P) \tag{5}$$

The current I_{on} and S is the zero intercept current and geometry factor respectively for the

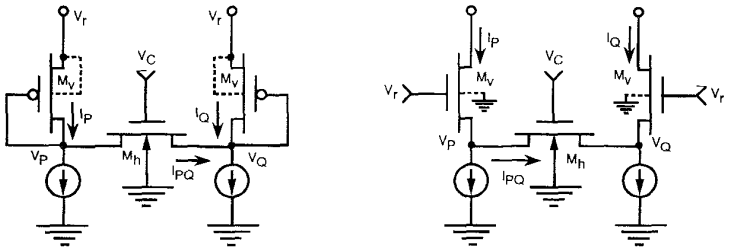

Figure 5. Current–mode CMOS building blocks for linear networks using (left) PMOS transistor implementation, (right) NMOS single transistor current-conveyor implementation.

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

$$\frac{G_1}{G_2} = \left(\frac{SI_{on}}{I_S}\right) \exp\left[\frac{\kappa V_C - V_r}{(kT/q)}\right]$$

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 explicitely available in the process, diode connected PMOS or NMOS transistors can be used as shown in Figure 5. With PMOS loads, the current current I_{PQ} is:

$$I_{PQ} = \left(\frac{S_h I_{onh}}{S_v I_{onv}}\right) \exp\left[\frac{\kappa_h V_C - \kappa_v V_r}{(kT/q)}\right]({I_Q}^{1/\kappa_v} - {I_P}^{1/\kappa_v}) \tag{6}$$

Unfortunately the anode of a diode or the drain terminal of a diode connected transistor is not a good current source. When NMOS transistors are used as loads, there is the additional

benefit, that of exploiting the current conveying properties of a single transistor [5], to obtain the current outputs I_P and I_Q, on nodes that are low conductance (the drain terminal are now excellent outputs for the currents). Using Equations 8.45 in [5] for the single transistor current conveyor, the current I_{PQ} is given as:

$$I_{PQ} = \left(\frac{S_h I_{onh}}{S_v I_{onv}}\right) \exp\left[\frac{\kappa_h V_C - \kappa_v V_r}{(kT/q)}\right](I_Q - I_P) \tag{7}$$

where S_h, I_{onh}, κ_h and S_v, I_{onv}, κ_v are geometry, zero bias intercept current and subthreshold slope parameters for transistors M_h and M_v, respectively.

The two node network segments discussed above can be used to construct models of electrical conduction in biological networks. One such model for a loaded network is shown in Figure 6.

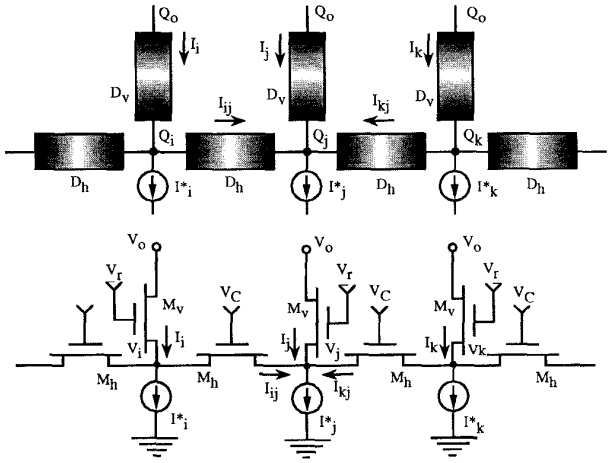

Figure 6. One-dimensional loaded network. The lateral diffusors model gap junctions between cells and the vertical ones model the membrane leakage. (a) Schematic representation. (b) MOS transistor implementation.

The network shown in Figure 6 may be analyzed with the help of a diffusor model. The node equation is

$$I_j^* = I_j + I_{ij} + I_{kj} = D_v(Q_0 - Q_j) + D_h(Q_i + Q_k - 2Q_j) \tag{8}$$

where (D_h and D_v) are the effective diffusivities. This is a discrete approximation to the differential equation:

$$I^*(x) = D_v(Q_0 - Q(x)) + D_h \frac{d^2Q(x)}{dx^2}$$

Compared to the Laplacian, there is an extra term that arises from the vertical diffusors which shunt charge to ground. This equation yields the solution to the following

optimization problem: Find the smooth function $Q(x)$ that best fits the data $I^*(x)$ with the minimum energy in its first derivative. The ratio D_h/D_v is the cost associated with the derivative energy—relative to the squared-error of the fit. Note that in this example $Q_0 > Q(x)$ and thus the extra term is positive.

The vertical elements afford us the opportunity to read the solution, i. e. the source/drain charge Q_j. This is directly proportional to the vertical current I_j if Q_0 is zero. Equation 1 predicts that Q_0 is negligibly small if V_0 is a few thermal voltages above V_j (saturation). Therefore, we can measure Q_j by setting V_0 at a high voltage and measuring the current there. The actual value used does not matter because Q_0 only serves to set the quiescent (zero input) drain/source charge.

For circuit analysis, it is convenient to use the Translinear Principle together with channel current decomposition [5] to analyze these diffusor circuits:

$$I_{Q_j} = I_j e^{\frac{\kappa(V_c - V_r)}{kT/q}}$$

assuming I_j's drain-driven component is zero, i. e. M_v is in saturation. Similar results are obtained for I_{Q_i} and I_{Q_k}; simply replace I_j with I_i and I_k, respectively.

This is immediately obvious if we observe that M_h and M_v are a differential pair operating subthreshold. These devices act as a current-divider for current driven by the charge at their common node. The divider ratio is set by their effective widths which depend on the geometrical width as well as the surface potential. Here, we have used the κ approximation to relate the surface potential to the gate-bulk voltage. The surface potential is constant as long as the gate and bulk voltages are fixed—assuming the mobile charge is negligible. Therefore, the divider ratio is constant and linear division occurs. However, as we enter the transistion and above-threshold regions this assumption fails and the surface potential starts to follow the source voltage. Consequently, the divider-ratio is no longer independent of the current level. This limits the dynamic range of the diffusor.

The node equation may be rewritten as

$$I_j{}^* = I_{ij} + I_{kj} + I_j = I_{Q_i} - I_{Q_j} + I_{Q_k} - I_{Q_j} + I_j$$

and substituting the expressions for the current components, we get

$$I_j{}^* = I_j + e^{\frac{\kappa(V_c - V_r)}{kT/q}}(I_i + I_k - 2I_j) \tag{9}$$

This is identical to Equation 8 if we replace I_j with $D_v Q_j$, I_k with $D_v Q_k$, etc, and $\exp(\kappa q(V_c - V_r)/kT)$ by D_h/D_v. The area efficiency and controlled coupling strength available using the diffusor this circuit particularly attractive for implementing the local aggregation.

The network in Figure 6 was recently described in terms of "pseudo-conductances" [12]. We prefer the charge/current mode description as this provides an intuitive understanding of the device and yielded the insight that enabled us to extend the translinear principle to subthreshold MOS transistors in the ohmic region and diffusors.

In this section, we have provided a comprehensive view for *current-mode* approach in subthreshold MOS circuits. The essence of this approach is the representation of variables and parameters by charge, current, and diffusivity. Voltages and conductances are not used explicitly. In the next section, we show how these techniques have been used at the system level.

3 System Organization

We will now discuss the organization of the contrast sensitive silicon retina. The architecture of the system is shown in Figure 7. There are two functional components in this organization. The core, and the support circuitry.

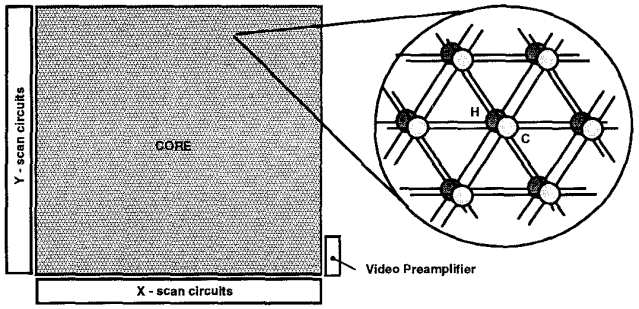

Figure 7. Floorplan and system organization. Focal plane processing is performed in the core area.

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 Delbrück [17]. 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-neighbour connectivity (see Figure 7). The wiring is included in the layout of the cell (see Figure 10) 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, communicate both intra and inter layer through local paths. This parallel processing scheme features locality of reference and thus minimizes communication costs. We now proceed with the discussion of the core circuitry.

3.1 Biological Organization

The analog silicon system in the core of the array is modeled after neurocircuitry in the distal part of the vertebrate retina—called the outer-plexiform layer. Figure 8 illustrates interactions between cells in this layer [15]. The well-known center/surround receptive field emerges from this simple structure, consisting 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 photoreceptors are activated by light; they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells, in turn, suppress the activity

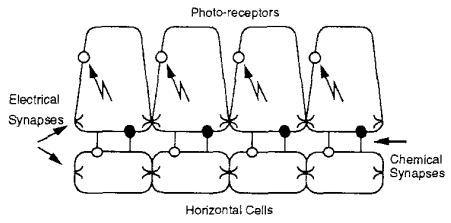

Figure 8. One-dimensional model of neurons and synapses in the outer-plexiform layer. Based on the red-cone system in the turtle retina.

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 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. The advantage of performing this complex operation at the focal plane is that the dynamic range is extended (local automatic gain control).

3.2 Silicon Implementation

The basic analog MOS circuitry for a one dimensional pixel with two neighbor connectivity is shown in Figure 9. We begin with the non-linear aspects of system operation, its *contrast sensitivity*. The non-linear operation that leads to a local gain-control mechanism in the silicon system is acheived through a mechanism that is qualitatively similar to the biological counterpart, but quantitatively different (see discussion in [6]). Refering to Figure 9, 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.

$$I_c(x_m, y_n) = I_u \frac{I(x_m, y_n)}{\left(I(x_m, y_n) + \sum_{\mathrm{M,N}} I(x_i, y_j)\right)} \tag{10}$$

At any particular intensity level, the outer-plexiform behaves like a linear system that realizes a powerful second-order regularization algorithm [16] for edge detection. This 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 diffusitivity $\hat{g}$.

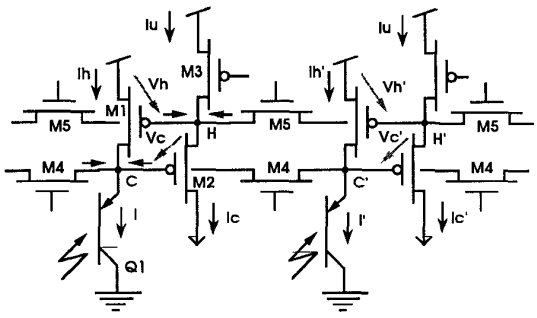

Figure 9. One-dimensional implementation of outer-plexiform retinal processing. There are two diffusive networks implemented by transistors M_4 and M_5, which model electrical synapses. These are coupled together by controlled current-sources (devices M_1 and M_2) that model chemical synapses. Nodes 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.

The diffusitivity of the devices M_5 between nodes H and H' in the horizontal network is denoted by h. Then the simplified equations describing the full two-dimensional circuit on a square grid are:

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

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

$$I_h(x,y) = I(x,y) + \hat{g}\nabla^2 I_c(x,y) \tag{11}$$

$$I_c(x,y) = I_u - h\nabla^2 I_h(x,y) \tag{12}$$

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

$$\hat{g}h\nabla^2\nabla^2 I_h(x,y) + I_h(x,y) = I(x_i, y_j) \tag{13}$$

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 [16]. The coefficient $\lambda = \hat{g}h$ is called the regularizing parameter; it determines the trade–off between smoothing and fitting the data.

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

$$I_h(x,\lambda) = \frac{1}{2\lambda^{1/4}} \exp(-|x|/\sqrt{2}\lambda^{1/4}) \cos\left(\frac{|x|}{\sqrt{2}\lambda^{1/4}} - \frac{\pi}{4}\right) \tag{14}$$

3.3 Layout Considerations

The two-layer architecture for the silicon retina can be accomodated in a cell area of $80\lambda \times 94\lambda$ using a single poly two metal technology. In the implementation reported in [6] 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.

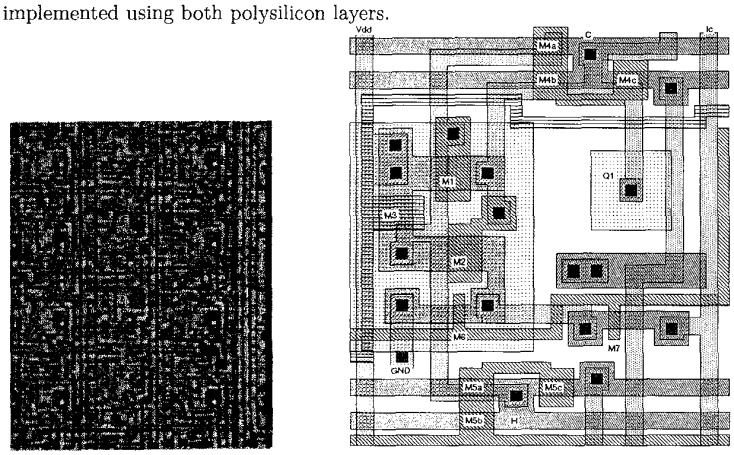

Figure 10. (Left) Photomicrograph of the chip. The surface is covered by second metal except where there are openings for the phototransistors (the dark square areas). Note the hexagonal connectivity between the pixels. (Right) Layout of the basic cell.

The system has been fabricated with 230×210 pixels on a 9.5×9.3 mm die in a 1.2μm 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.

4 Discussion

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 is the operating point of the system; that is also how the automatic gain control is achieved. At an 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 [14]. 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. For a detail discussion on Translinear circuits in subthreshold MOS please refer to [5].

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 with bandpass kernel of Equation 14, six for the Laplacian operator (Equation 12) and six for the local gain control computation (Equation 10). If the system is biased so that at the pixel level the frequency response is 100Khz, approximately 1×10^{12} low precision calculations per second are performed in the (210×230) pixels. The power dissipation under the above biasing conditions is about 50mW when operating from 5 Volt power supplies. This is equivalent to 0.05 pW/OP.

This performance is a result of an optimization done at the system level, rather than trying to optimize the energetic efficiency of an individual gate. The biological inspired architecture resulted a system capable of dealing well with the both the variability in the problem that is solving (data acquisition under variable illumination conditions) [2] and at the same time enabling robust operation in the presence of structural variability and mismatch present in MOS transistors [18].

Acknowledgments: One of the authors (AGA) was supported in part by Nissan corporation and by a Research Initiation Award from NSF (MIP-9010364). Chip fabrication was provided by MOSIS. We thank Professor Carver Mead for making us believe in the power of low energy analog computation and for his encouragement over the last 6 years. We thank Mark Martin for helping with image acquisition and processing.

References

[1] L.A. Glasser, "Electronics Technology for Low-Power Computing and Wireless Communication," Proceedings of IEEE IEDM-93, Washington D.C., Dec, 1993.

[2] J.M. Buhmann, M. Lades and R. Eeckman, "Illumination-Invariant Face Recognition with a Contrast Sensitive Silicon Retina," *Advances in Neural Information Processing Systems 6,* J.D. Cowan, G. Tesauro, and J. Alspector. (eds.), Morgan Kaufmann Publishers, San Mateo, CA 1994.

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

[4] C.A. Mead, "Neuromorphic electronic systems," *Proceedings IEEE,* vol. 78, no. 10, pp. 1629-1636, Oct. 1990.

[5] A.G. Andreou and K.A. Boahen, "Neural Information Processing (II)," Chapter 8, in *Analog VLSI Signal and Information Processing,* M. Ismail and T. Fiez eds., McGraw-Hill, 1994.

[6] K.A. Boahen and A.G. Andreou, "A contrast sensitive silicon retina with reciprocal synapses," *Advances in Neural Information Processing Systems 4,* Moody, J.E., Hanson, S.J. and Lippmann, R.P. (eds.), Morgan Kaufmann Publishers, San Mateo, CA 1992.

[7] E.A. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation," *IEEE Journ. of Solid-State Circuits,* SC-12, No. 3, pp. 224-231, June 1977, E.A. Vittoz, "Micropower techniques," in *VLSI Circuits for Telecommunications,* edited by Y. P. Tsividis and P. Antognetti, Prentice Hall, 1985.

[8] E.A. Vittoz, "The design of high-performance analog circuits on digital CMOS chips," *IEEE Journ. of Solid-State Circuits,* SC-20, No. 3, pp. 657-665, June 1985.

[9] Y.P. Tsividis, "Analog MOS integrated circuits–certain new ideas, trends and obstacles," *IEEE J. Solid-State Circuits,* vol. SC-22, pp. 317-321, June 1987.

[10] K.A. Boahen, A.G. Andreou, P.O. Pouliquen and A. Pavasovič, "Architectures for associative memories using Current–Mode analog MOS circuits," Proceedings of the Decennial Caltech Conference on VLSI, C. Seitz editor, MIT Press, 1989, A. G. Andreou, et.al. "Current–mode subthreshold MOS circuits for analog VLSI neural systems," *IEEE Trans. on Neural Networks,* vol. 2, no. 2., pp. 205-213, March 1991.

[11] M.A. Maher and C.A. Mead, "A Physical Charge Controlled Model for MOS Transistors," Proceedings of the Stanford Conference on Advanced Research in VLSI, P. Losleben editor, MIT Press, 1987.

[12] E. Vittoz and X. Arreguit, "Linear networks based on transistors," Electronics Letters, vol. 29, pp. 297-299, Feb. 4th, 1993.

[13] C. Koch, "Seeing Chips: analog VLSI circuits for computer vision," *Neural Computation,* vol. 1, 2, pp. 184-200, 1989.

[14] B. Gilbert, "Translinear circuits: A proposed classification," *Electronics Letters,* vol. 11, No. 1, pp. 14-16, 1975, B. Gilbert, "A Monolithic 16-Channel Analog Array Normalizer," *IEEE Journ. of Solid-State Circuits,* vol. SC-19, No. 6, 1984.

[15] J. E. Dowling, "The retina: an approachable part of the brain," The Belknap Press of Harvard University, Cambridge, MA, 1987.

[16] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory," *Nature,* 317, pp. 314-319, 1985.

[17] C.A. Mead and T. Delbrück, "Scanners for visualizing activity in analog VLSI circuitry," *Analog Integrated Circuits and Signal Processing,* vol. 1, no. 2, Oct. 1991.

[18] A. Pavasović, A. G. Andreou, and C. R. Westgate, "Characterization of subthreshold MOS mismatch in transistors for VLSI systems," *Journal of Analog Integrated Circuits and Signal Processing,* 6, pp. 75-84, June 1994.