Programmable retinal dynamics in a CMOS mixed-signal array processor chip

R. Carmona, F. Jiménez-Garrido[a], R. Domíguez-Castro, S. Espejo, A. Rodríguez-Vázquez
Instituto de Microelectrónica de Sevilla. IMSE-CNM-CSIC
Avda. Reina Mercedes s/n 41012 Sevilla (SPAIN)

ABSTRACT

The low-level image processing that takes place in the retina is intended to compress the relevant visual information to a manageable size. The behavior of the external layers of the biological retina has been successfully modelled by a Cellular Neural Network, whose evolution can be described by a set of coupled nonlinear differential equations. A mixed-signal VLSI implementation of the focal-plane low-level image processing based upon this biological model constitutes a feasible and cost effective alternative to conventional digital processing in real-time applications. For these reasons, a programmable array processor prototype chip has been designed and fabricated in a standard $0.5\mu m$ CMOS technology. The integrated system consists of a network of two coupled layers, containing 32×32 elementary processors, running at different time constants. Involved image processing algorithms can be programmed on this chip by tuning the appropriate interconnections weights. Propagative, active wave phenomena and retina-like effects can be observed in this chip. Design challenges, trade-offs, the buildings blocks and some test results are presented in this paper.

1. INTRODUCTION

Due to the vast amount of information contained in the visual stimuli, nature has developed a specialized part of the nervous system to handle it: the retina. On one side, the neuronal impulses conveying information along the nerves do not support such a large data rate. On the other side, because of the high correlation found between the elements of the image —most of the energy of the signal, in images displaying natural scenes, is concentrated in the lower spatial and temporal frequencies—, not every bit of information has to be passed to the brain to accomplish vision. Therefore, the retina, brought to the sensory periphery instead of being integrated in the central nervous system, processes the visual information at the focal plane, realizing what is called early vision[1 2 3]. This low-level processing reduces the enormous amount of information associated to the visual flow into data set of manageable size. Although retinas are not yet fully understood, and defines a challenging basic research area, the construction of vision processing devices with retina-like features shows large potential to overcome the limitations of conventional vision technologies. In that sense, during the last few years, several neuromorphic vision chips have been developed and reported in literature[4 5 6].

Recently, the behavior of the more external strata of the multi-layered structure of vertebrate retina has been successfully modelled by using Cellular Neural Network (CNN) framework[7]. Such model has been based on studies and observations about the mammalian retina which have been recently published in Nature[3]. In this model, interactions between cells in the retinal fabric are realized on a local basis; each cell interacts with its nearest neighbors. Also, every cell belonging to the same layer has the same interconnection pattern. For each retinal layer, the same set of interconnection weights is applied to each and everyone of its cells; i.e. layers are spatially-invariant. In addition to this, the signals supporting intra- and inter-layer interactions are continuous in magnitude and time. The phenomena observed in the mammalian retina[3] are modelled by two coupled sets of 2-D nonlinear differential equations[7]. Because of the local interactions and the spatial-invariance, the behavior of such a model is fully described by some 25 parameters. This set of controlling parameters include interaction strengths, time constants and bias terms. By properly setting their values complex, interacting waves are generated which emulates the phenomena observed in the mammalian retina.

This paper presents a fully-programmable mixed-signal implementation of this model[7] on a silicon chip. It is organized as follows. Section II is dedicated to the bio-inspired network models, the foundations of the mathematical network model in a sketch of the biological retina. Section III describes the architecture of the APAP chip and its main compo-

a.E-mail: garrido@imse.cnm.es, Tel.: +34955056666, Fax: +34955056686.

Bioengineered and Bioinspired Systems, Angel Rodríguez-Vázquez, Derek Abbott, Ricardo Carmona, Editors, Proceedings of SPIE Vol. 5119 (2003) © 2003 SPIE · 0277-786X/03/$15.00

nents. Section IV explains roughly the analog building blocks of the basic processing units. Experimental results obtained from testing a prototype chip are shown in Section V. Finally, Section VI displays some conclusions.

2. BIO-INSPIRED MODEL

2.1. Vertebrate retina.

The vertebrate retina has the structure displayed[8] in Fig.1.A first layer of photodetectors at the outermost layer of the retina, the cone cells —a different type of cell, the rods, are specialized in sensing in very dim light conditions and saturate very easily—, captures light and converts it to activation signals. Bipolar cells carry these signals across the retina layers to the ganglion cells that interface the retina with the optical nerve, in a trip of several micrometers[3]. The ganglion cells convert the continuous activation signals, proper of the retina, to spike-coded signals that can be transmitted over longer distances by the nervous system. In the way to the ganglion cells, the information carried by bipolar cells is affected by the operation of the horizontal and amacrine cells. They form layers in which activation signals are weighted and promediated in order to, first, bias photodetectors and, second, to account for inhibition on the vertical pathway. Patterns of activity are formed dynamically by the presence or absence of visual stimuli. The four main transformations that take place in this structure are: the photoreceptor gain control, the gain control of the bipolar cells, the generation of transient activity and the transmission of transient inhibition.

2.2. CNN analogy of the biological retina.

There are, in this description, some interesting aspects of the retinal layers that markedly resemble the characteristics of a CNN: the 2D aggregation of continuous signals, the local connectivity between elementary nonlinear processors, the analog weighted interactions between them. Motivated by these coincidences, and based on physiological and pharmacological studies[2], a CNN model has been developed that approximates the observed behavior of the vertebrate retina[9]. The outer plexiform layer of the retina, OPL, is responsible for the image capture. It has been characterized by experimental measurements[10], leading to a model with three different layers of cells. The first one, the photosensing layer, consists in an aggregation of cone cells. It is assumed here that the retina is adapted to lighting conditions and so the rods are saturated and remain silent. In addition to the layer containing the cones, there is a second layer composed of horizontal cells and a third one composed of bipolar cells. Each of these layers has the structure of a 2D CNN itself. each of them has its own interaction patterns (CNN templates) and its particular time constant. Cell dynamics are sustained by a first or a sec-

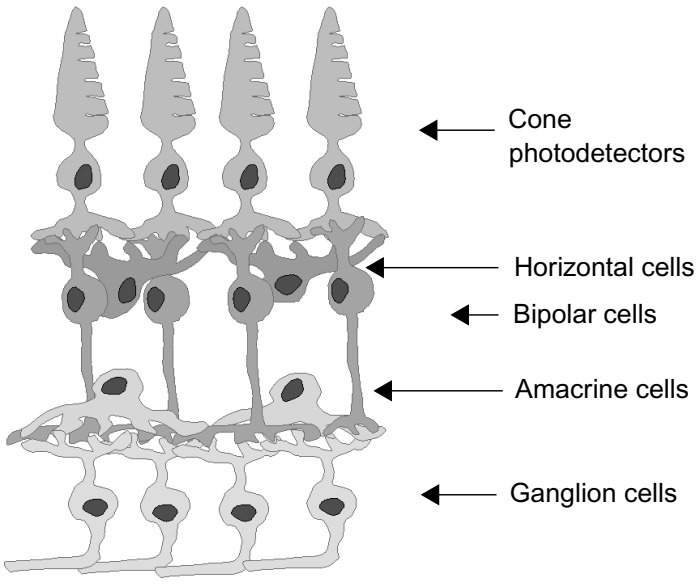


Figure 1: Schematic diagram of the vertebrate retina.

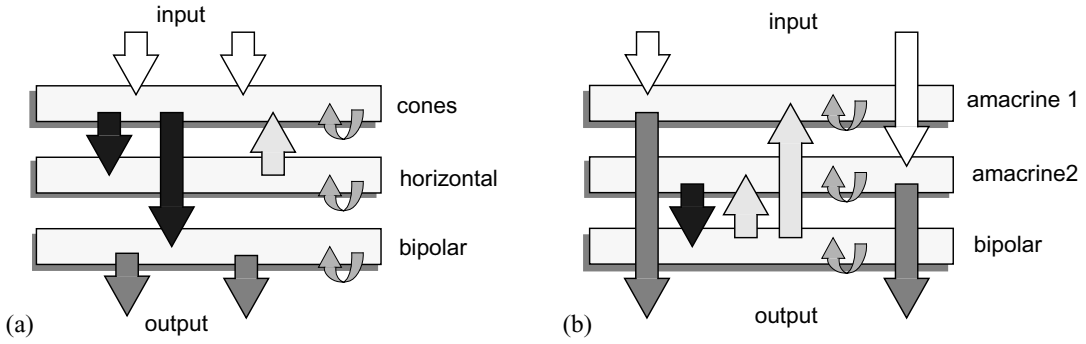


Figure 2: Conceptual diagram of the (a) OPL of the retina, and (b) the wide-field activity in the IPL.

ond order dynamic core. The structure of the OPL is depicted in Fig.2(a), where interactions between layers of cells are represented by arrows.Regarding the inner plexiform layer, IPL, it is responsible for the generation of the retinal output. A simplified model of the IPL[10] consists in two layers of wide field amacrine cells excited by the input signal, which in this occasion is the output of the bipolar cells, and a third layer that controls the dynamic of the previous layers by means of feedback signals. As before, the three layers are supposed to be 2D CNNs with their own internal coupling and their own time constant, Fig.2.

Because of the relative simplicity of these models, a programmable CNN chip has been proposed[11]. The programmable array processor of the chip consists in 2 coupled CNN layers, and a third layer, of a much faster dynamics ($\tau_3 \ll \tau_1, \tau_2$) that supports analog arithmetic (Fig.3). Each elementary processor contains the nodes for both CNN layers. The third layer is inherently implemented by these analog cores, with the local facilities for analog signal storage. The evolution of the coupled CNN nodes of a specific cell $C(i, j)$ is described by these coupled differential equations:

$$\tau_1 \frac{dx_{1,ij}}{dt} = -g[x_{1,ij}] + \sum_{k=-r_1}^{r_1} \sum_{l=-r_1}^{r_1} a_{11,kl} y_{1,(i+k)(j+l)} + b_{11,00} u_{1,ij} + a_{12} y_{2,ij} + z_{1,ij} \tag{1}$$

$$\tau_2 \frac{dx_{2,ij}}{dt} = -g[x_{2,ij}] + \sum_{k=-r_2}^{r_2} \sum_{l=-r_2}^{r_2} a_{22,kl} y_{2,(i+k)(j+l)} + b_{22,00} u_{2,ij} + a_{21} y_{1ij} + z_{2,ij} \tag{2}$$

where the nonlinear losses term and the output function in each layer are those of the FSR CNN model[12]:

$$g(x_{n,ij}) = \lim_{m \to \infty} \begin{cases} m(x_{n,ij} - 1) + 1 & \text{if} \quad x_{n,ij} > 1 \\ x_{n,ij} & \text{if} \quad |x_{n,ij}| \le 1 \\ m(x_{n,ij} + 1) - 1 & \text{if} \quad x_{n,ij} < -1 \end{cases} \tag{3}$$

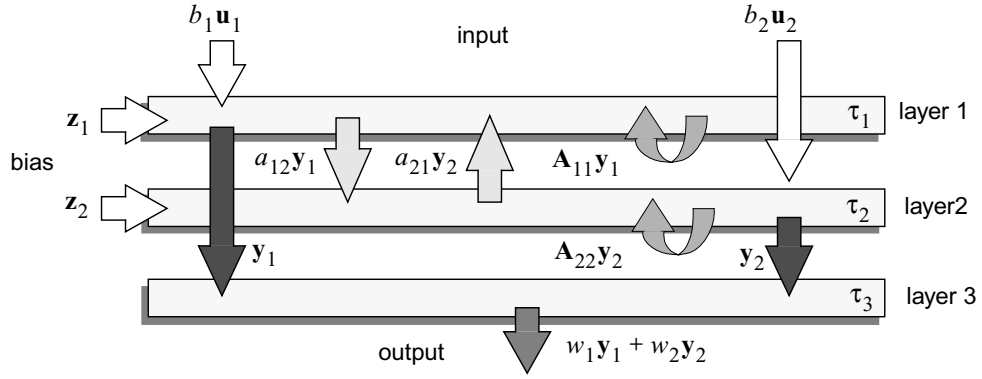


Figure 3: Diagram of the 2nd-order 3-layer CNN.

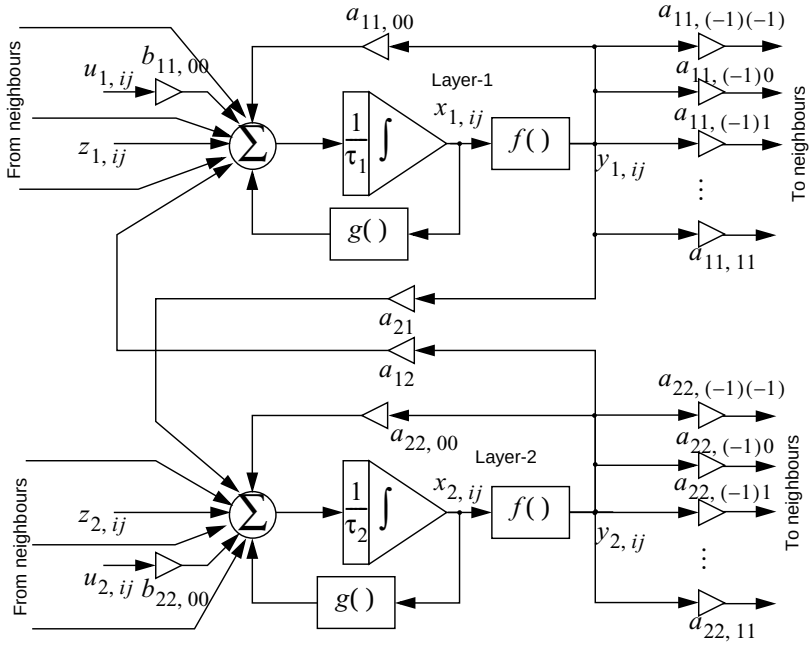


Figure 4: Model of the two coupled CNN layer nodes.

and:

$$y_{n,\,ij} = f(x_{n,\,ij}) = \frac{1}{2}(|x_{n,\,ij}+1| - |x_{n,\,ij}-1|) \tag{4}$$

Fig.4 depicts the block diagram of the vertically coupled CNN nodes. Synaptic connections between cells are linear. Each CNN layer incorporates feedback connections, by means of which the output of each cell contributes to the state of its neighbors, weighted by the elements $\{a_{nn,\,kl}\}$; a feedforward connection, weighted by $b_{nn,\,00}$, that regulates the contribution of the cell's input; a bias term $z_{n,\,ij}$, that can be different for each cell; and finally coupling connections between both layers, weighted by a_{21} and a_{12}. Each layer has its own time-constant τ_n. Programming different dynamics in this CNN model is possible by adjusting the template elements and the time-constants of the layers. The total number of synapses to be implemented on each cell is 22, plus the 2 bias maps multipliers, which can be treated as a second input image for each layer.

3. PROTOTYPE ARCHITECTURE

3.1. Analog programmable array processor.

The proposed chip consists in an Analog Programmable Array Processor (APAP) of 32×32 identical cells (Fig.5). It is surrounded by the circuits implementing the boundary conditions for the CNN dynamics. The I/O interface consists in a serializing-deserializing analog multiplexor. The timing and control unit is composed by a micro-instruction decoder, generating the appropriate signals to configure the network, and an internal clock/counter with a set of finite state machines that generate the internal signals that enable program memory accesses and other data transfers. The operation control unit constitutes the interface between the program memory and the processing array. In this program memory, the algorithm to be implemented is stored in several digital memory banks through a digital interface. A program instruction contains the control bits for chip operation and binary codes for analog voltages. Thus, a bank of D/A converters interfaces these memory blocks with the processing array.

3.2. Basic cell.

The basic cell of the CNN-based array processor has a similar architecture to that of the CNN universal machine cells[13].However, in this occasion, the prototype includes two different continuous-time CNN layers. As depicted in

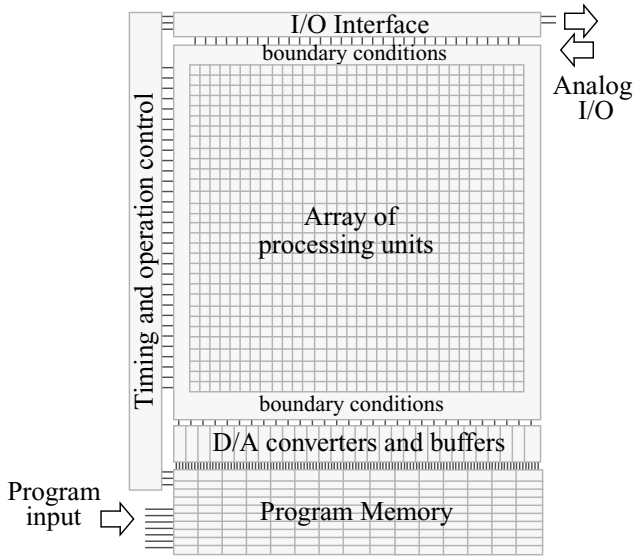


Figure 5: Chip Architecture.

Fig.6(a), the basic processor contains together with the local analog and logic memories (LAMs and LLMs), committed to the storage of intermediate results, the local logic unit (LLU), responsible for pixel-level logic operations, and two different analog CNN core blocks, each one belonging to one of the two different CNN layers implemented. The synaptic connections between processing elements of the same or different layer are represented by arrows in the diagram. All the blocks in the cell communicate via an intra-cell data bus, which is multiplexed to the array I/O interface. Control and cell configuration bits are passed directly from the control unit.

The internal structure of each of the CNN cores of the cell is depicted in the diagram of Fig.6(b). Each core receives contributions from the rest of the processing nodes in the neighborhood which are summed and integrated in the state capacitor. The two layers differ in that the first layer has a scalable time constant, controlled by the appropriate binary code, while the second layer has a fixed time constant. The evolution of the state variable is also driven by self-feedback and by the feedforward action of the stored input and bias patterns. There is a voltage limiter which helps to implement the limitation on the state variable of the FSR CNN model. This state variable is transmitted in voltage form to the synaptic blocks, in the periphery of the cell, where weighted contributions to the neighbors' are generated. There is also a current memory that will be employed for cancellation of the offset of the synaptic blocks. Initialization of the state, input and/or bias voltages is done through a mesh of multiplexing analog switches that connect to the cell's internal data bus.

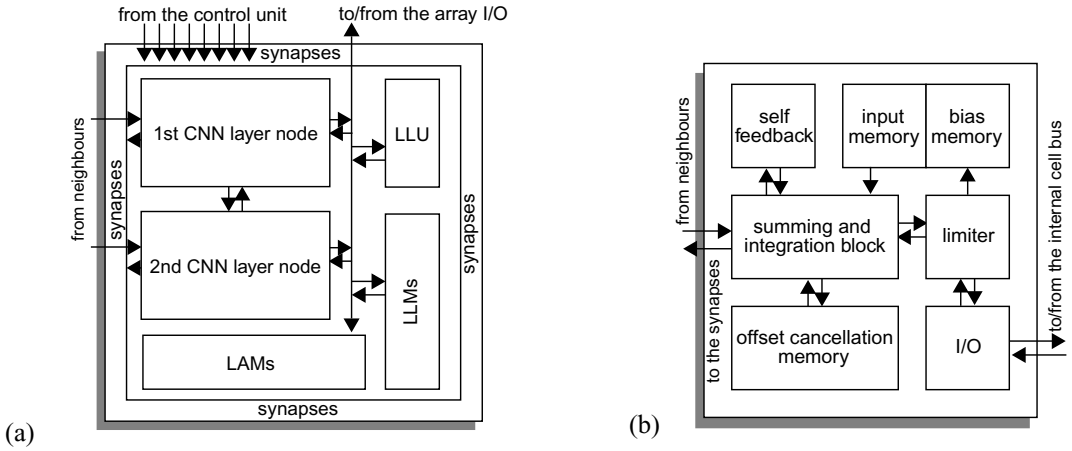


Figure 6: Conceptual diagram of the (a) basic cell and the (b) internal structure of each CNN layer node.

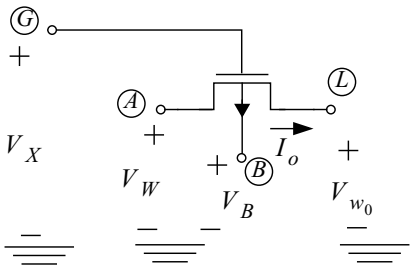


Figure 7: Multiplier using one single MOS transistor in ohmic region.

4. ANALOG BUILDING BLOCKS

4.1. Single transistor synapse.

The synapse is a four-quadrant analog multiplier. Their inputs will be the cell state V_X or input and the weight voltages V_W, while the output will be the cell's current contribution to a neighboring cell. It can be realized by a single transistor biased in the ohmic region[14]. For a PMOS with gate voltage $V_X = V_{x_0} + v_x$, and the p-diffusion terminals at $V_W = V_{w_0} + v_w$ and V_{w_0}, —where V_{x_0} and V_{w_0} are the reference central values for the state and weight voltages— ,the drain-to-source current is:

$$I_o \approx -\beta_p V_w V_x - \beta_p V_w \left(V_{x_0} + \left| \hat{V}_{T_p} \right| - V_{w_0} - \frac{V_w}{2} \right) \tag{5}$$

which is a four-quadrant multiplier with an offset term that is time-invariant —at least during the evolution of the network— and not depending on the state. This offset is eliminated in a calibration step, with a current memory.

For the synapse to operate properly, the input node of the CNN core, (L) in Fig.7, must be kept at a constant voltage. This is achieved by a current conveyor,(Fig.8(a)).Any difference between the voltage at node (L) and the reference V_{w_0} is amplified and the negative feedback corrects the deviation. Notice that a voltage offset in the amplifier results in an error of the same order. An offset cancellation mechanism is provided,(Fig.8(b)). Signal ϕ_{cal} shorts the OTA inputs and enables diode-mode operation of transistor M_{mem}, that will conduce a current I_{mem} such as to cancel out the current offset. Once ϕ_{cal} is turned off, the total current injected into the load capacitor is offset-free:

$$I_L = I_{o_{\text{OTA}}} + I_{\text{mem}} - I_{\text{b}} = g_m v_d \tag{6}$$

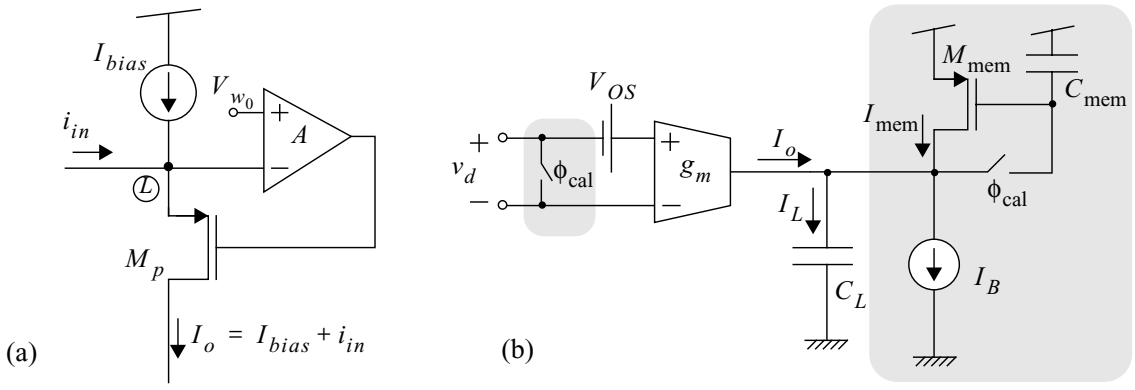


Figure 8: Current conveyor (a) and offset calibration mechanism (b).

4.2. Current memory.

As it has been referred, the offset term of the synapse current must be removed for its output current to represent the result of a four-quadrant multiplication. For this purpose all the synapses are reset to $V_X = V_{x_0}$. Then the resulting current, which is the sum of the offset currents of all the synapses concurrently connected to the same node, is memorized. This value will be subtracted on-line from the input current when the CNN loop is closed, resulting in a one-step cancellation of the errors of all the synapses. The validity of this method relies in the accuracy of the current memory. For instance, in this chip, the sum of all the contributions will range, for the applications for which it has been designed, from $18\mu A$ to $46\mu A$. On the other side, the maximum signal to be handled is $1\mu A$. If a signal resolution of $8b$ is pretended, then $0.5LSB = 2nA$. Thus, our current memory must be able to distinguish $2nA$ out of $46\mu A$. This represents an equivalent resolution of $14.5b$. In order to achieve such accuracy level, a S^3I current memory is used. It is composed by three stages, Fig.9, each one consisting in a switch, a capacitor and a transistor. I_B is the current to be memorized. After memorization the only error left corresponds to the last stage.

4.3. Time-constant scaling.

The differential equation that governs the evolution of the network can be written as a sum of current contributions injected to the state capacitor. Scaling up/down this sum of currents is equivalent to scaling the capacitor and, thus, speeding up/down the network dynamics. Therefore, scaling the input current with the help of a current mirror, for instance, will have the effect of scaling the time-constant. A circuit for continuously adjusting the current gain of a mirror can be designed based on a regulated-Cascode current mirror in the ohmic region. But the strong dependence of the ohmic-region biased transistors on the power rail voltage causes mismatches in τ between cells in the same layer. An alternative to this is a digitally programmable current mirror. It trades resolution in τ for robustness, hence, the mismatch between the time constants of the different cells is now fairly attenuated.

A new problem arises, though, because of current scaling. If the input current can be reshaped to a 16-times smaller waveform, then the current memory has to operate over the largest and the smallest signals. But, if designed to operate on large currents, the current memory will not work for the tiny currents of the scaled version of the input. If it is designed to run on small input currents, long transistors will be needed, and the operation will be unreliable for the larger currents. One way of avoiding this situation is to make the S^3I memory to work on the original unscaled version of the input cur-

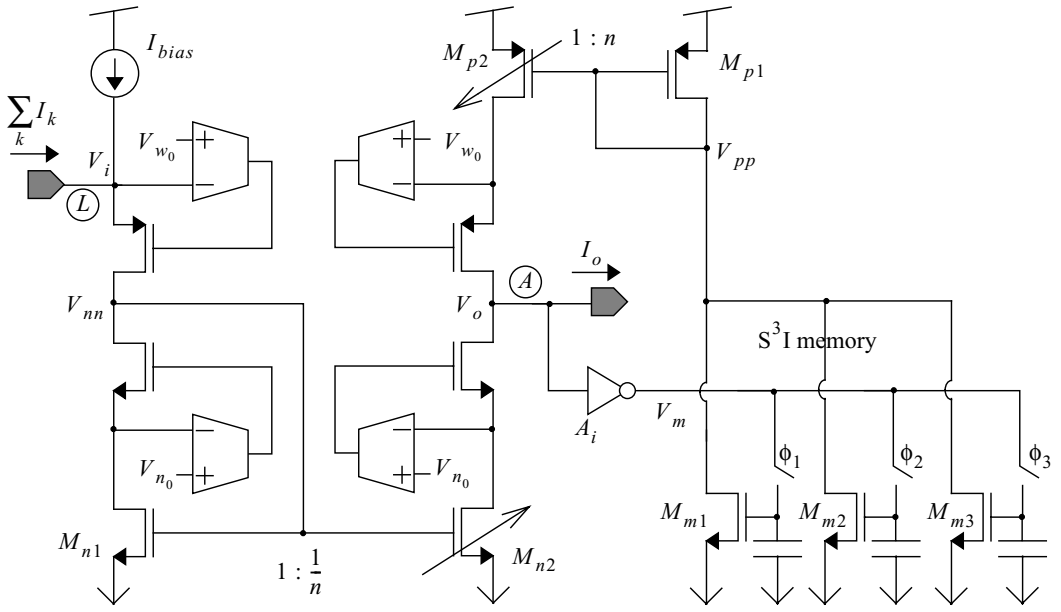


Figure 9: Input block with current scaling and S^3I memory.

rent. Therefore, the adjustable-time-constant CNN core will be a current conveyor, followed by the S^3I current memory and then the binary weighted current mirror. The problem now is that the offsets introduced by the scaling block add up to the signal and the required accuracy levels can be lost. Our proposal is depicted in (Fig.9). It consists in placing the scaling block (programmable mirror) between the current conveyor and the current memory. In this way, any offset error will be cancelled in the auto-zeroing phase. In the picture, the voltage reference generated with the current conveyor, the regulated-Cascode current mirrors and the S^3I memory can be easily identified. The inverter, A_i, driving the gates of the transistors of the current memory is required for stability.

5. EXPERIMENTAL RESULTS

5.1. Prototype chip data.

A prototype chip has been designed and fabricated in a $0.5\mu m$ single-poly triple-metal CMOS technology. Its dimensions are $9.27 \times 8.45 mm^2$ (microphotograph in Fig.10). The cell density achieved is $29.24 cells/mm^2$, once the overhead circuitry is detracted from the total chip area —given that it does not scale linearly with the number of cells. The power consumption of the whole chip is around $300mW$. Data I/O rates are nominally $10MS/s$. In the first test results, with a non-optimized platform, I/O times of 220ns have been measured for a full-scale step. The time constant of the fastest layer (fixed time constant) is intended to be under $100ns$. The peak computing power of this chip is, therefore, $470GXPS$, what means $6.01 GXPS/mm^2$, and $1.56 GXPS/mW$. The chip handles analog data with an equivalent resolution of $7.5 bits$, as measured at the first tests.

5.2. Retinal behavior emulation.

Different image processing algorithms can be programmed on this chip by setting the corresponding switches configuration and by tuning the appropriate interconnection weights. Propagative and wave-like phenomena, similar to those found at the biological retina, can be observed in this chip. For instance, the wave-fronts generated at the slower layers can be employed to inhibit propagation in the faster layer, thus generating a trailing edge for the waves in the fast layer.

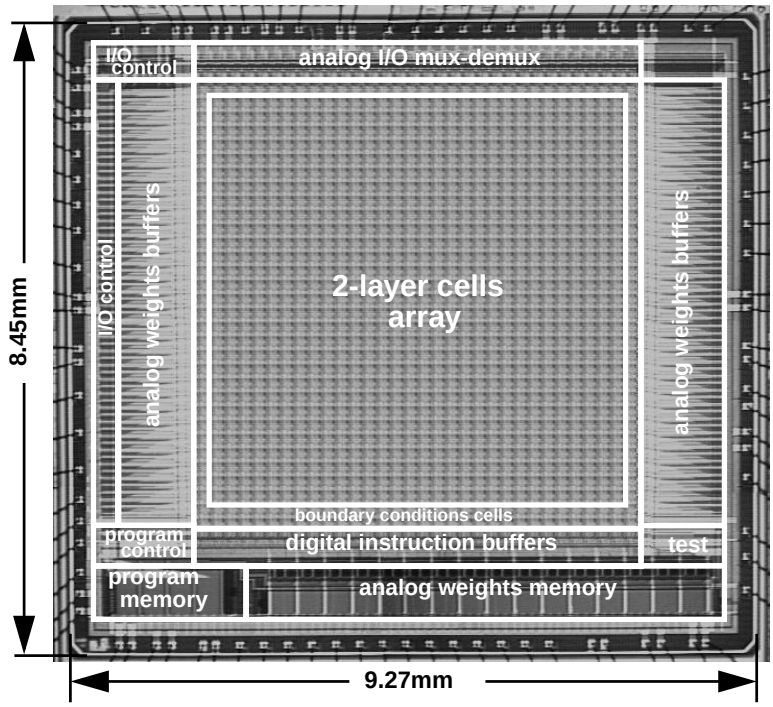


Figure 10: Microphotograph of the prototype chip.

This produces the similar results as the wide field erasure effect observed in the IPL of the retina (Fig.11**).** Another interesting effect observed in the OPL of the retina is the detection of spatio-temporal edges followed by de-activation of the patterns of activity. This phenomenon has been also programmed in the chip (Fig.12).

5.3. Active waves phenomena.

By setting the appropriate interconnection weights, active wave phenomena —the propagation of waves in an energetically active medium—, can be observed in the chip. For instance, the triggering of a travelling wave or the generation of spiral wave (Fig.13).

6. CONCLUSIONS

Based on the results obtained, we can state that the proposed approach supposes a promising alternative to conventional digital image processing for applications related with early-vision and low-level focal-plane image processing. Based on a simple but precise model of part of the real biological system, a feasible efficient implementation of an artificial vision device has been designed. The peak operation speed of the chip outperforms its digital counterparts due to the fully parallel nature of the processing. This especially so when comparing the computing power per silicon area unit and per watt.

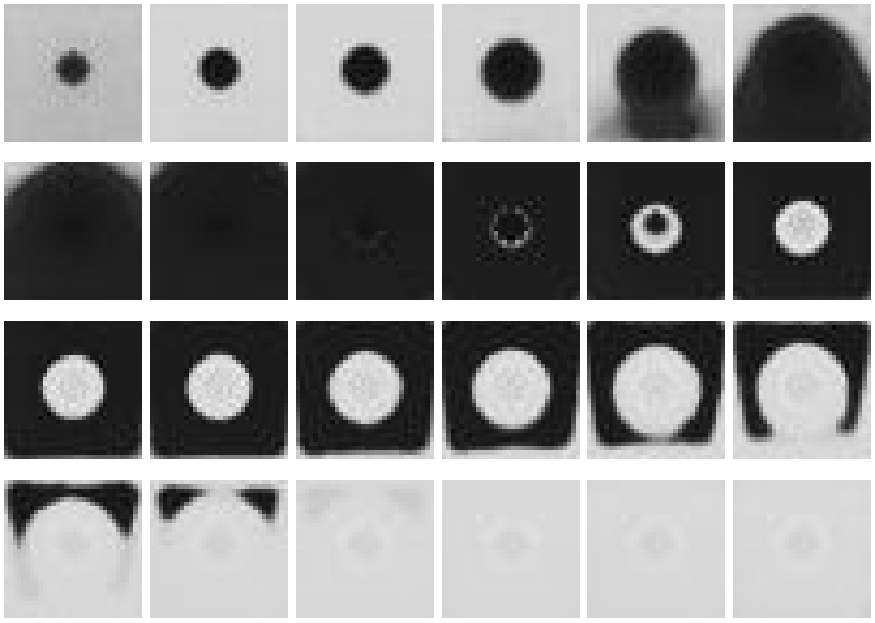
7. ACKNOWLEDGEMENTS

This work has been partially funded by ONR Project N00014-00-10429 (POAC), EC Project IST-1999-19007 (DICTAM) and the Spanish MCyT Project TIC1999-0826 (SIVA-2).

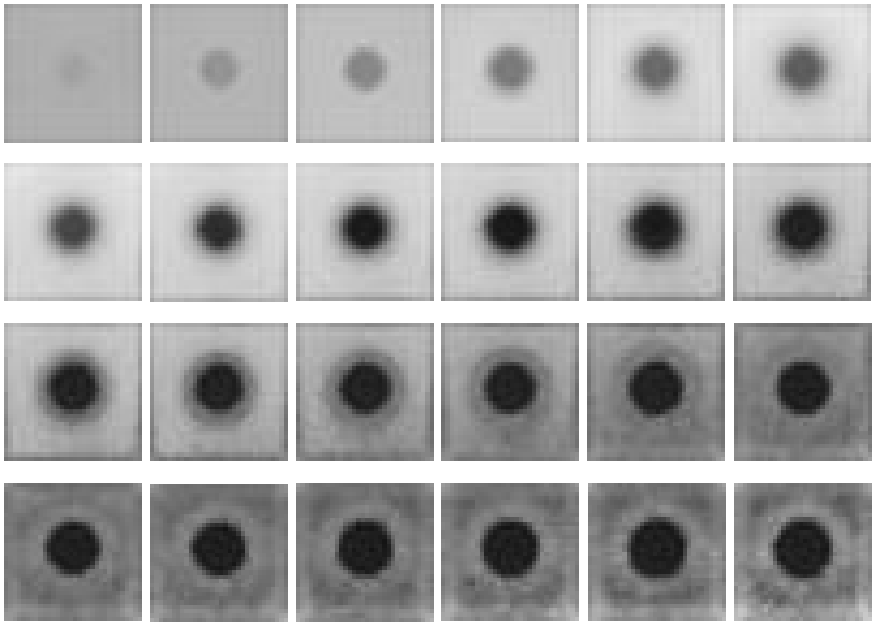
8. REFERENCES

1. D.H. Hubel, *Eye, Brain and Vision*, W.H. Freeman, New York,1988.
2. F. Werblin, “Synaptic Connections, Receptive Fields and Patterns of Activity in the Tiger Salamander Retina”, Investigative Ophthalmology and Visual Science, **Vol. 32**, No. 3, pp.459-483, March 1991.
3. B. Roska, and F.S. Werblin, “Vertical interactions across ten parallel, stacked representations in the mammalian retina“, Nature, **Vol. 410**, pp. 583-587, March 2001.
4. Carver Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA, 1989.
5. Christof Koch and Hua Li (Eds.), *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, IEEE Computer Society Press, Los Alamitos, 1995.
6. Alireza Moini, *Vision Chips*, Kluwer Academic Publishers, Boston, 1999.
7. D. Bálya, B. Roska, E. Nemeth, T. Roska, and F.S. Werblin, “A Qualitative Model Framework for Spatio-temporal Effects in Vertebrate Retina”, *Proc. of the 2000 IEEE Conf. on Cellular Neural Networks and their Applications*, pp. 165-170, 2000.
8. F. Werblin, T. Roska and L. O. Chua, “The Analogic Cellular Neural Network as a Bionic Eye”, International Journal of Circuit Theory and Applications, **Vol. 23**, No. 6, pp. 541-69, November-December 1995.
9. A. Jacobs, T. Roska and F. S. Werblin, “Methods for Constructing Physiologically Motivated Neuromorphic Models in CNNs”, International Journal of Circuit Theory and Applications, **Vol. 24**, No. 3, pp. 315-339, May-June 1996.
10. Cs. Rekeczky, B. Roska, E. Nemeth and F. Werblin, “Neuromorphic CNN Models for Spatio-Temporal Effects Measured in the Inner and Outer Retina of Tiger Salamander”, *Proc. of the Sixth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 15-20, Catania, Italy, May 2000.
11. Cs. Rekeczky, T. Serrano-Gotarredona, T. Roska and A. Rodríguez-Vázquez, “A Stored Program 2nd Order/3-Layer Complex Cell CNN-UM”, *Proc. of the Sixth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 219-224, Catania, Italy, May 2000.
12. S. Espejo, R. Carmona, R. Domínguez-Castro and A. Rodríguez-Vázquez, “A VLSI Oriented Continuous-Time CNN Model”, International Journal of Circuit Theory and Applications, John Wiley & Sons, **Vol. 24**, No. 3, pp. 341-356, May-June 1996.

13. T. Roska and L. O. Chua: “The CNN Universal Machine: An Analogic Array Computer”, IEEE Transactions on circuits and Systems-II: Analog and Digital Signal Processing, **Vol. 40**, No. 3, pp. 163-173, March 1993.
14. R. Domínguez-Castro, A. Rodríguez-Vázquez, S. Espejo and R. Carmona, “Four-Quadrant One-Transistor Synapse for High Density CNN Implementations”, *Proc. of the Fifth IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 243-248, London, UK, April 1998.



(a) The fastest layer



(b) The slowest layer

Figure 11: Wide field erasure effect.

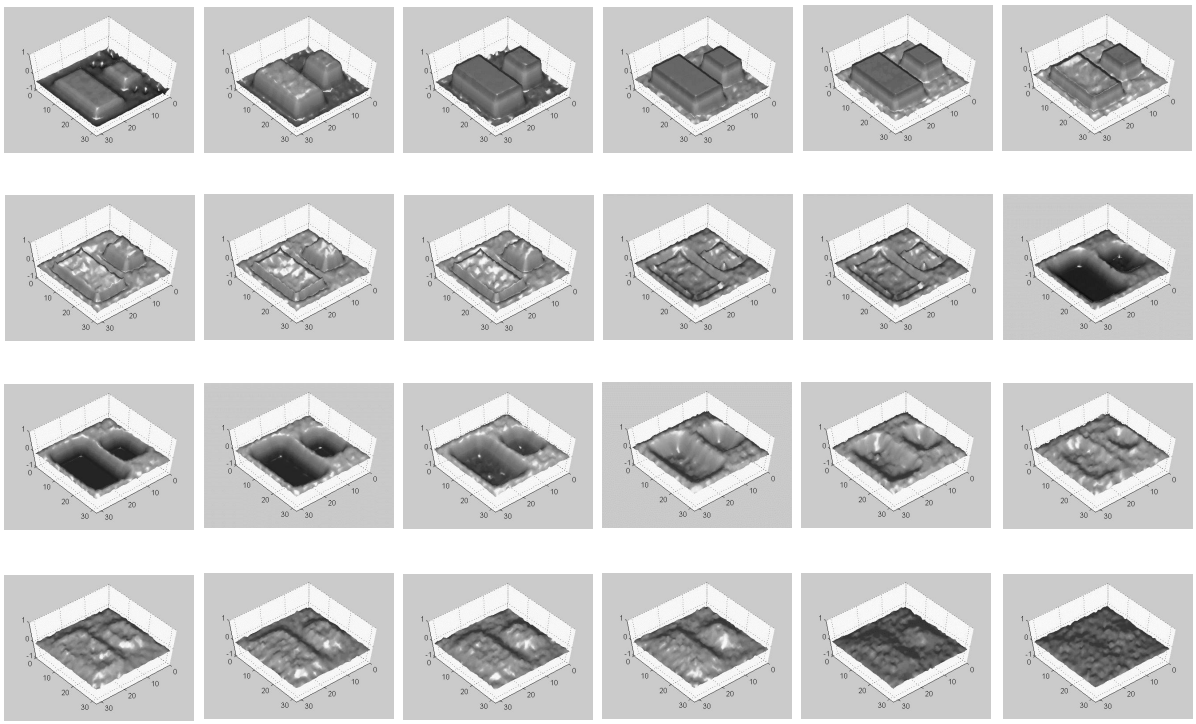


Figure 12: Spatio-temporal edge detection (fast layer).

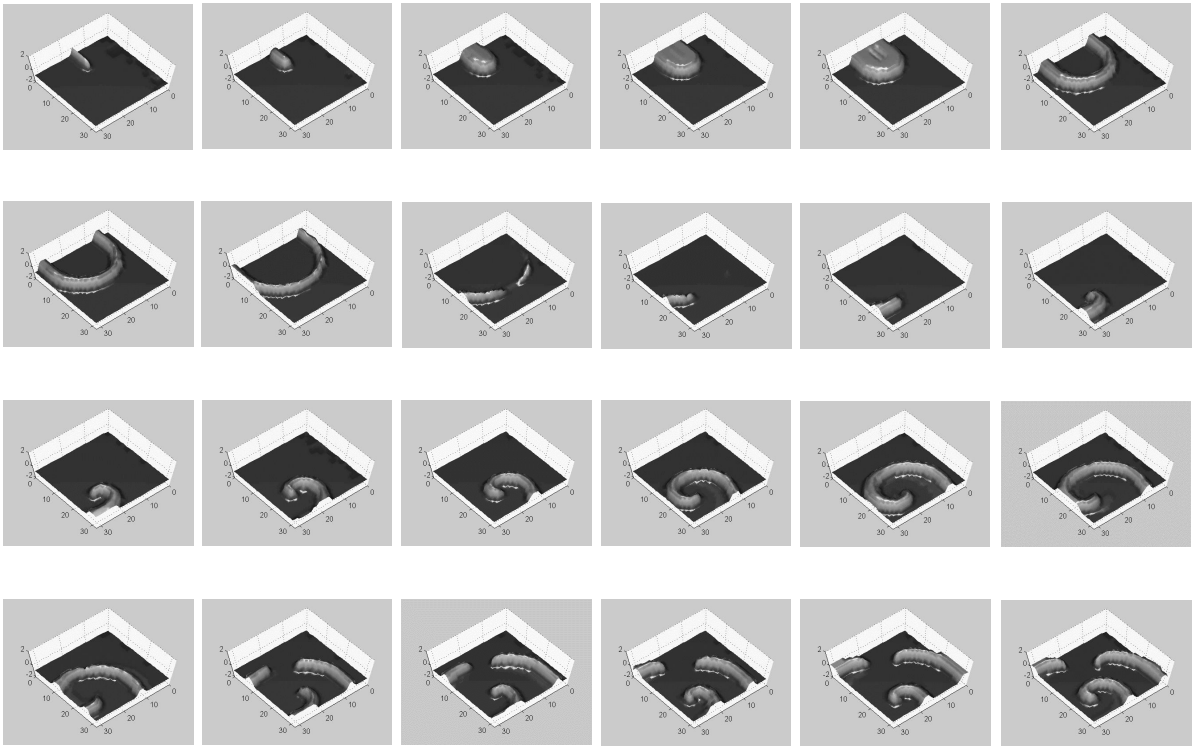


Figure 13: Spiral wave (fast layer).