2006 10th International Workshop on Cellular Neural Networks and Their Applications, Istanbul, Turkey, 28-30 August 2006

3-Layer CNN Chip for Focal-Plane Complex Dynamics with Adaptive Image Capture

C. M. Domínguez-Matas, R. Carmona-Galán, F. J. Sánchez-Fernández, A. Rodríguez-Vázquez
Instituto de Microelectrónica de Sevilla-CNM- CSIC
Av. Reina Mercedes s/n, 41012 Sevilla, Spain
e-mail: cmanuel@imse.cnm.es

Abstract— **This paper presents a CMOS implementation of a layered CNN concurrent with 32×32 photosensors with locally programmable integration time for adaptive image capture. The network is arranged in two layers containing feedback and control templates, inter-layer connections and programmable ratio of time constants. There are also feedforward connections to a third layer, which is faster, and devoted exclusively for combining the outputs of the other two. A more robust and linear multiplier block has been employed to reduce irregular analog wave propagation ought to asymmetric synapses. Global and local adaptation circuits are included on-chip. The predicted computing power per power consumption, 240MOPS/mW, is amongst the largest reported, what renders this kind of devices as especially adequate for portable applications of artificial vision.**

Index Terms— **Vision chips, CNN, parallel processing.**

I. Introduction

For the most of us humans, vision is the dominant sensory modality in the acquisition of information from the environment. For this to be possible, nature has developed one of the most efficient devices intended for adaptive image capture and real-time image processing: the retina [1]. Meanwhile, the struggle to bring artificial vision to fairly inaccessible places continues. This is mainly ought to the difficulties to handle extraordinary amount of data contained in the visual stimuli with the help of conventional microprocessors. Even if such data flow can be managed, it is done at the expense of considerable physical profile and energy consumption. This concern might not be a problem in machine vision applications in industrial environments. However, in applications like robotic vision [2], sensor networks for ambient intelligence [3] or retinal prosthesis for the blind [4], power efficient computation and the use of the simplest and the least hardware possible are mandatory. Here is where conventional digital processors, with a serial processing scheme, fail to meet the specifications. As can be seen in Fig. 1, general purpose processors are not very energy efficient. DSP's and hardware accelerated processors perform better, but the real boost in performance is obtained by the adaptation of the architecture to the nature of the stimuli. This is quite common in biological sensory organs, that exploit the high level of parallelism present in aggregates of neural cells. In order to realize an efficient VLSI implementation of array processing, analog and mixed-signal circuits represent a good alternative. The number of operations per second in analog chips has been calculated assuming peak performance is during a convolution, what renders the formula in [7]:

$$\mathrm{OPS}_{\mathrm{conv}} = \frac{(N_{\mathrm{add}} + N_{\mathrm{prod}})N_{\mathrm{cells}}}{\tau_{\mathrm{conv}}(N_{\mathrm{bits}} + 1)\ln 2} \qquad (1)$$

Therefore, the number of OPS is the ratio between the total number of additions and products realized in parallel in the chip, and the time it takes to the chip to settle to the final result of the convolution within the required accuracy. Back to Fig. 1, using analog circuits at the elementary processing units avoids A/D conversion at the pixel level, and, for moderate accuracy requirements, they occupy less area and consume less power than their digital counterparts.

In the following sections the CACE2 vision chip, the details of the elementary processing unit, and the extended features of the chip are explained.

II. CACE2 System Description

The CACE2 system architecture is intended to be implemented in single chip, constituting a complete vision

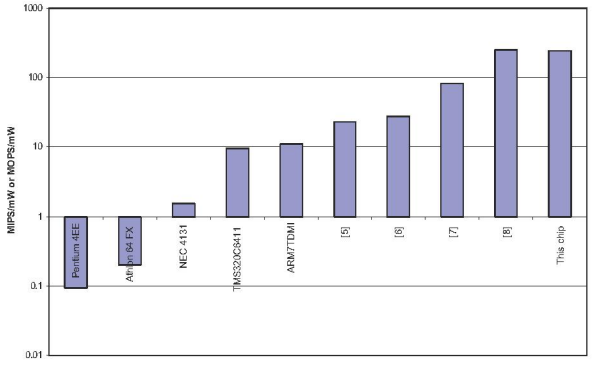


Fig. 1. Computing power per mW for different processors.

This work was partially supported by projects TIC2003-09817-C02-01 of the Spanish MCyT, and N-00014-02-1-0884 of the ONR. F. J. Sánchez is supported by a grant of the Spanish MEC.

system on a chip (VSoC). Its operation will be controlled by an embedded microprocessor, the CACE2 MCU (Fig. 2). At this point, this architecture has been implemented in 2 chips: image capture and early vision tasks are realized by a specialized peripheral, the CACE2 APAP, which is the chip reported here, and an FPGA containing the CACE2 MCU. The APAP consists in an analog/mixed-signal parallel array processor of 32×32 cells. Each cell is equipped with programmable spatio-temporal dynamics, local support for analog and logic in-pixel arithmetic operations, local analog and logic memories and a photosensor with extensions for adaptive image capture (Fig. 3). It follows the architecture of the CNN-UM [9]: an analog programmable array controlled by signals common to all of them and stored in its internal switch configuration registers (SCR's). From the point of view of the network topology, each cell incorporates two nodes of a CNN, belonging to layers of different time constants, and a third node for combining their outputs. This network supports complex dynamic phenomena expressed by a set of coupled reaction-diffusion equations [10].

In this first prototype of the system, the CACE2 MCU has been implemented in a FPGA, together with the necessary peripherals for booting up the system, program and data storage and communication. The SCR's of the APAP are allocated within the address space of the MCU. The programmability and control of the network dynamics, the calibration and biasing of the analog and mixed-signal building blocks, the adaptive image capture mechanisms, are controlled by the MCU via the signals stored in the SCR's. Access to them is directed by the address bus (A-bus). The access mode, either reading or writing, is indicated by the control bus (C-bus). The data bus (D-bus) is employed for sending or receiving parameters. Each SCR must be properly updated to manage the intra- and inter-cell connectivity of the processing elements, the sequences of signals that control the array operation and the codes of the internally generated analog references employed for mixed-signal circuits operation. The array processor counts also with a secondary data bus (8-bits wide) employed for image I/O. This bus is directly connected to the system memory via an access controller (DMA) which is also activated by the MCU when required by the software program.

III. CNN Processing Unit

The type of signal processing realized by the CACE2 APAP is based on the dynamic evolution of a 3×32×32 CNN. This behavior is described in terms of the input ($\mathbf{u}_k$), state ($\mathbf{x}_k$) and output ($\mathbf{y}_k$) variables. Each layer, k, of the array follows the evolution law expressed by:

$$\tau_k \frac{d\mathbf{x}_k(t)}{dt} = -\mathbf{g}[\mathbf{x}_k(t)] + \sum_n [\mathbf{A}_{kn} \otimes \mathbf{y}_n + \mathbf{B}_{kn} \otimes \mathbf{u}_n] + \mathbf{z}_k \qquad (2)$$

The symbol ⊗ stands for the linear convolution between the feedback and feedforward templates, with the output and input matrices of layer, n, where n can be 1, 2 or 3:

$$\begin{aligned} [\mathbf{A}_{kn} \otimes \mathbf{y}_n](i,j) &= \sum_{l=-r}^{r} \sum_{m=-r}^{r} \mathbf{A}_{kn}(l,m)\mathbf{y}_n(i+l,j+m) \\ [\mathbf{B}_{kn} \otimes \mathbf{u}_n](i,j) &= \sum_{l=-r}^{r} \sum_{m=-r}^{r} \mathbf{B}_{kn}(l,m)\mathbf{u}_n(i+l,j+m) \end{aligned} \qquad (3)$$

where r is the neighbourhood radius. In this particular implementation, τ_1 and τ_2 are comparable while τ_3 is much smaller than the others. If the full-signal-range CNN model is employed [11], the output and state variables can be identified. In this conditions, each matrix element in Eq. (3) is

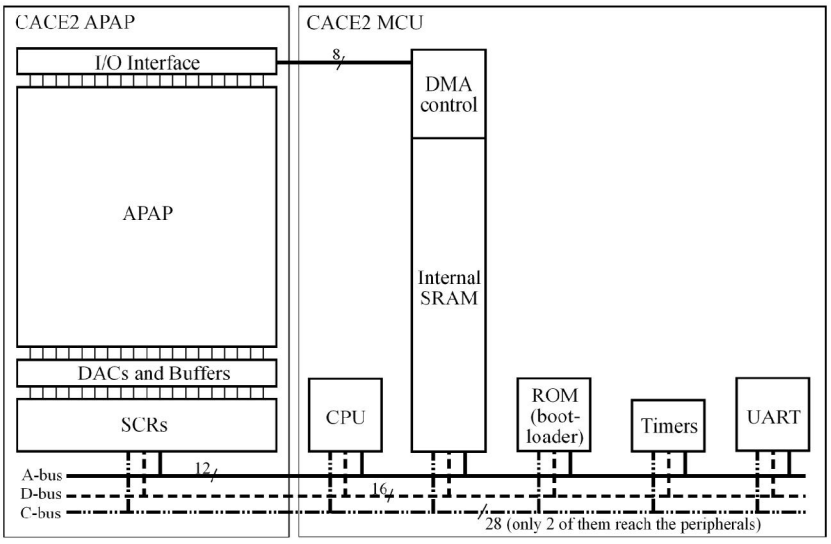


Fig. 2. Functional diagram of the CACE2 system

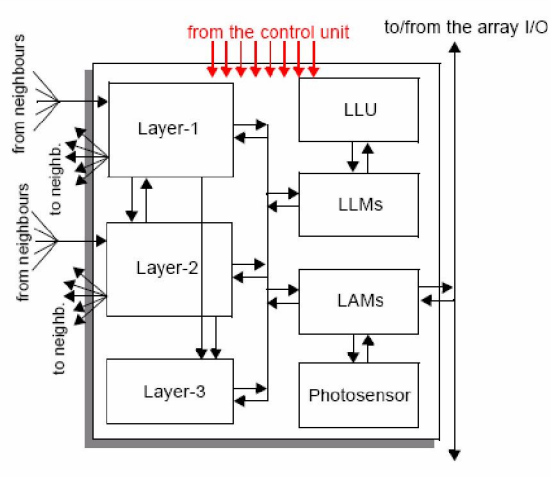


Fig. 3. Conceptual diagram of the basic cell.

obtained from the multiplication of the state (or input) variable by a programmable weight. These operators, responsible for multiplying the state (or input) variable by a programmable weight, are termed synapses or synaptic blocks in this context. They are basically four quadrants multipliers in which linearity with the state (or input) variable and a symmetric characteristic are strongly desired. The effect of varying the programmed weights is to modify the network dynamics, and thus, changing the type of processing realized by the array.

Continuing with the evolution law, there is a losses term:

$$g[\mathbf{x}_k(i,j)] = \lim_{m_o \to \infty} \begin{cases} m_o[\mathbf{x}_k(i,j)-1]+m_c & \text{if} \quad \mathbf{x}_k(i,j) > 1 \\ m_c \mathbf{x}_k(i,j) & \text{if} \quad |\mathbf{x}_k(i,j)| \le 1 \\ m_o[\mathbf{x}_k(i,j)+1]-m_c & \text{if} \quad \mathbf{x}_k(i,j) < -1 \end{cases} \tag{4}$$

and the activation function, to generate the output:

$$\mathbf{y}_k(i,j) = f[\mathbf{x}_k(i,j)] = \frac{1}{2} \lim_{m \to m_c} \left\{ \frac{1}{m} \left[|\mathbf{x}_k(i,j)+m| - |\mathbf{x}_k(i,j)-m| \right] \right\} \tag{5}$$

In both equations, m_c can be 0 or 1 for hard or sigmoidal type nonlinearity, respectively.

The physical realization of the elementary processing unit of the CNN starts with the selection of the appropriate format for the representation of the signals. On one side, voltages can be easily delivered to neighbouring areas by connecting wires to high-impedance nodes. Therefore, input, output and state variables are chosen to be represented by the matrices of voltages $\mathbf{V}_u$, $\mathbf{V}_y$ and $\mathbf{V}_x$, respectively. On the other side, signal addition can be easily realized in the form of currents wired together to a virtual ground. Hence, the summands in the second member of Eq. (2) should be represented by currents. And then, this sum of currents will be integrated in the state capacitor to obtain the instantaneous value of the state variable voltage:

$$C_k \frac{d\mathbf{V}_{x,k}(t)}{dt} = -\mathbf{G}_g[\mathbf{V}_{x,k}(t)] + \sum_n [\mathbf{G}_{A,kn} \otimes \mathbf{V}_{y,n} + \mathbf{G}_{B,kn} \otimes \mathbf{V}_{u,n}] + \mathbf{I}_{z,k} \tag{6}$$

As can be seen, the elements of the feedback and feedforward templates, $\mathbf{A}_{kn}(i,j)$ and $\mathbf{B}_{kn}(i,j)$, are now programmable linear transconductances, $\mathbf{G}_{A,kn}(i,j)$ and $\mathbf{G}_{B,kn}(i,j)$, that multiplied by input and output voltages render the neighbourhood contributions in the form of currents. Thus, the synaptic block is a transconductor whose output current is proportional, in the ideal case, to the product of the state (or input) variable and the weight. The double transformation implicit in Eq. (6), V-I and then I-V, allows for a compact realization of the processing node, achieving higher cell densities, meaning an array size of practical interest and, besides, a tolerable fill factor.

The accuracy of these terms is very important to accomplish a correct operation of the network, since the synapse offsets, as well as every mismatch on ideally symmetric weights, are integrated in the state capacitor. Precisely, in the implementation of four-quadrant multipliers, one of the common difficulties is to maintain the symmetry with respect to the origin of the weights. A mismatch in weights having the same absolute value but opposite signs can modify the dynamic routes of the cells in the network, ending in displaced equilibrium points, and thus, distorting the prescribed processing. The main linearity concerns are found in the V-I conversion, as linear current integration, and thus I-V transformation, can be provided by available highly linear double-poly capacitors. In this design, we have employed a linearized OTA in order to generate the unitary current contribution. Though the elementary transconductor achieving V-I conversion has a larger number of transistors than the single-transistor synapse in [12], advantages in the linearity with the state (or input) variable and symmetry of the V-I characteristic justify its use. In addition, the supporting circuitry can be simplified resulting in a more robust implementation finally without any area penalty.

The schematics in Fig. 4 represent the core of the elementary dynamic processor. Operating in closed loop (when the switch controlled by 'Loop' is on), it implements the evolution law described by Eq. (6). The weighted V-I conversion of the state voltage is carried at several stages. The single-to-differential V-to-I conversion is realized by a linearized transconductor (left-side of the schematics), these current signals are replicated and scaled by several programmable current mirrors to generate the contributions towards the neighbors and itself (at the center and right sides) and the current signals from the neighbors and self-feedback are added and integrated in the state capacitor, when feedback loop is closed (by the block at the center).

The transconductor responsible of transforming the state capacitor voltage V_x into a differential current is a source degenerated differential pair with diode-connected loads. It is based on a linearized OTA [13]. The operation of this circuit

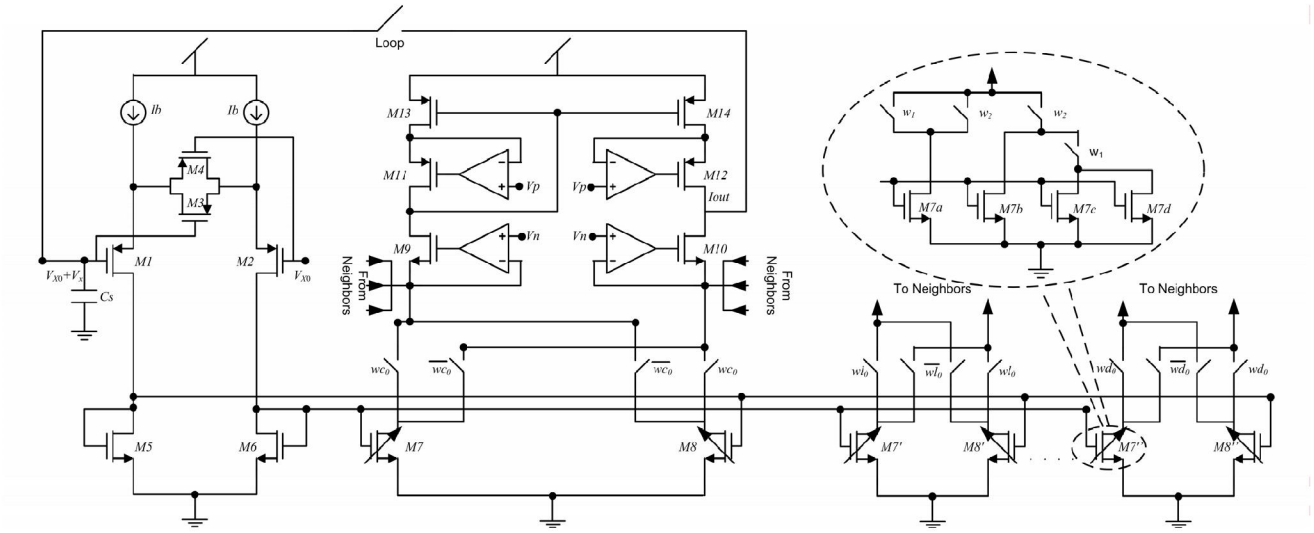


Fig. 4. Schematic of the linearized OTA and synaptic blocks

alone is inherently symmetric if working in fully-differential mode, representing an enhancement from what have been achieved by previous implementations. This symmetry though is broken by using a single-ended input voltage, but still the resulting V-I characteristic maintains symmetry levels beyond those of other implementations. The benefits of a differential representation were not significant to be worth handling with double capacitor area and a complex signal routing.

The implementation of the weights is based on geometrical relations between transistors. This has the advantage of being less influenced by process parameter variations both inter- and intra-die. It has also the drawback of only permitting the use of a discrete set of weight values, namely –4, –2, –1, 0, 1, 2 and 4. Opposite-sign contributions are obtained by crossing the wires conveying the currents to the collecting nodes, thus, achieving by architecture a symmetric operation.

Finally, the sum of all the currents coming from the neighborhood is injected into the target state capacitor. But

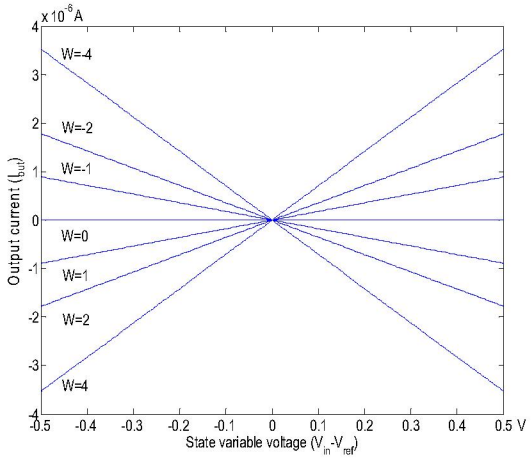


Fig. 5. Output current vs. state voltage of the OTA-based synapse.

before that, differential to single-ended current conversion is realized with the help of a current mirror. It is important to mention that the achievable output resistances using self-biased or externally biased Cascode current mirrors are not sufficient to ensure the necessary independence from the output voltage. In other words, the resulting error in the copied current, because of the finite output resistance of the mirror, was beyond the predicted errors due to parameters mismatch. Therefore, gain boosting of the Cascode devices is needed to reduce this effect. The accuracy of the current replication in this mirror is crucial for achieving the required linearity and symmetry in the V-I characteristic. Also, we have employed 0.5/2.0 transistors for the current mirrors, ensuring enough matching.

As a result, the output current has a high linearity. The transconductance relative error in large signal is kept below 0.7%. Concerning the symmetry of the characteristic, the difference of the output currents corresponding to weights with the same absolute value but opposite sign is zero on average because offset cancellation, the standard deviation, obtained by Monte Carlo simulation, being 2% of the absolute value of the individual currents. Compared to previous implementations, in terms of linearity of the V-I characteristic of the multipliers, this circuit performs one order of magnitude better for comparable area and power consumption. This is not always required for the correct operation, i. e. convergence of the network dynamics to the correct equilibrium points, but decisive for linear diffusion.

Functional operation of the complete cell, including the local memories and the switching tree that allows communication with the outside of the array, has been verified by simulation. A small network composed of 3x3 cells —actually a 5x5 network if the boundary cells are considered— has been programmed to implement different image processing templates. Fig. 6(a) displays the state

variables of the cells of the 3x3 network when programmed to realize connected component detection in the horizontal direction. From the simulations, it can be seen that the time constant of the cells is designed to be under 100ns.

IV. Extended Chip Features

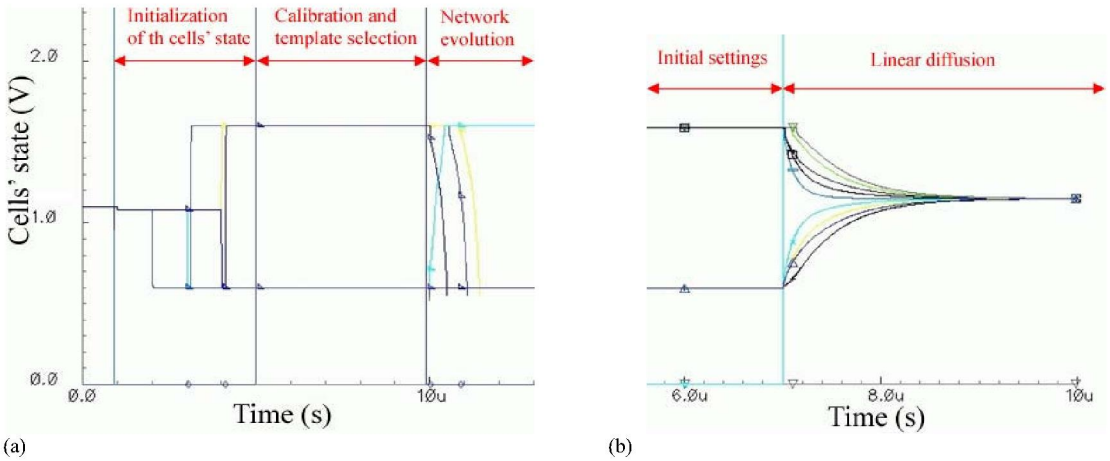
A. True resistive grid

One of the most useful tools for focal-plane image processing is the diffusive propagation of the pixel values. This is achieved by a Gaussian lowpass filter, which for a discretized image grid, can be realized by the convolution of the original image with a spatial mask. The evolution law implemented at every node of the CNN can support this diffusive dynamics by appropriately setting the correct interconnection weights in the feedback template. The main drawback of implementing this operator and others using symmetric weights, in a VLSI structure designed for fully-programmable CNN dynamics is mismatch in generating current contributions. Because of the local computation of the contributions to the neighborhood, the amount of current being injected from cell $C(i,j)$ into neighboring cell $C(i+1,j)$, for instance, does not match in absolute value the current being injected from cell $C(i+1,j)$ back into the state capacitor of cell $C(i,j)$. The consequence of this is easy to derive, the supposedly symmetric diffusion is converted into an unruled propagation of the pixel values.

Special care has been put in counteracting the effects of mismatch by design. This is, resizing the transistors in order to avoid excessive deviation from the nominal in the generation of the unitary current contributions. Apart from this, the prototype chip includes a true resistive grid concurrent with the CNN array. Each cell contains two resistors, made of high-resistivity poly-Si, that can be connected to the state capacitor in order to form a rectangular resistive grid. The time constant of this grid is between 0.2-1.0μs, and it is not correlated to the CNN time constant, neither can be controlled by the user. The operation of the resistive network is illustrated in Fig. 6(b). Here the 3x3 array of cells is programmed to evolve with null templates, using only the grid of poly resistors. In less than 2μs, the state voltages of all cells converge to the global average. This is not fast, and it becomes worse when the size of the network increases, but it is convenient from the point of view of the control of the algorithm to count with a diffusion mechanism that runs slower than the switch configuration updating signals.

B. Adaptive image capture

Adaptive image capture in the CACE2 APAP is based in the local and global control of the photosensors' gain. Operating in photocurrent integration mode, the voltage representing the value of the pixel depends on the integration time, i. e. for the same power of the incident light over the sensor surface, a larger integration time will allow the same photogenerated current to discharge the sensing capacitance for a longer time, resulting in a larger voltage excursion form the reset value. In this chip, each pixel has a reset transistor governed either by a global signal —automatic adaptation of the integration time is off— or by a comparator driven by a local reset control voltage and a global time-evolving reference. The local support for this comparison is explained in [14]. Its main function is to adapt the local gain of the photosensor. As they are integrating sensors, this gain adjustment is achieved via the adaptation of the local integration time according to a locally derived voltage level. In order to do that, the global reference will be an inverse voltage ramp that is delivered to every sensor in the array. When the inverse ramp crosses —with negative slope— the



(a) (b)

Fig. 6. Simulation of the evolution of a reduced (3x3) array with local memory update, offset cancellation and reset processes.

local threshold, the integration of the photocurrent starts. In this way, the darker the pixel then the larger integration time that will be allocated for that pixel for the next capture —the algorithm relies in the correlation between the values for the same pixel in different frames in a sequence. Correspondingly, the brighter the pixel, the less time it will have in the next capture.

Concerning the global adaptation mechanism, the inverse ramp is centered in a predicted average integration time value. Therefore, if the previous image capture resulted in an over-exposed picture, the average voltage will be below the middle point of the pixels' voltage range. If the previous image is under-exposed, the average voltage will be above this point. The algorithm programmed into the chip corrects the time extent of the ramp accordingly in order to have smaller exposures for brightness saturated images and larger exposures for extremely dark pictures. This is achieved by comparing the average voltage of the pixels with upper and lower thresholds. If the resulting average falls between these thresholds, the only corrections introduced are due to local adaptation. If the voltage falls above/below the upper/lower limit, a digital circuit triggered by these comparators, corrects the frequency division realized onto the systems master clock, employed to generate the inverse ramp, in the proper sense. This ends in a wider/narrower ramp shape until the average pixel voltage falls between the two thresholds.

V. Chip Data and Conclusions

The prototype chip has been designed and fabricated in a CMOS 0.35µm. The die size is 7.6mm x 7.6mm. Fig. 7 displays a microphotograph of the chip. Table I shows a survey of chip data. These features are predicted from the simulation results. The chip in now under test, in order to confirm the expected performance.

TABLE I

CMOS process	0.35µm
No. of CNN cells	3x32x32
Die area	7.6mm x 7.6mm
Array area	6.2mm x 6.2mm
Processing cell size	176µm x 176µm
Current consumption per cell	300µA@3.3V
Weight resolution	7-8b
Image resolution	8b
I/O rates	10MHz
CNN time constant	below 100ns

References

[1] D. H. Hubel, *Eye, Brain and Vision*. Scientific American Library, No. 22. W. H. Freeman and Co., New York, 1995.

[2] T. Makimoto, T. T. Doi, "Chip Technologies for Entertainment Robots - Present and Future". *Int. Electron Devices Meeting*, pp. 9-16, Dec. 2002.

[3] E. Aarts, R. Roovers, "IC Design Challenges for Ambient Intelligence", *Design, Automation and Test in Europe (DATE)*, pp. 2-7, March 2003.

[4] Eyal Margalit et al. "Retinal Prosthesis for the Blind", *Survey of Ophthalmology*, Vol. 47, No. 4, pp. 335- 356, July-August 2002.

[5] T. Komuro, I. Ishii, M. Ishikawa, A. Yoshida, "A Digital Vision Chip Specialized for High-Speed Target Tracking". *IEEE Trans. on Electron Devices*, Vol. 50, Np. 1, pp. 191-199, Jan. 2003.

[6] P. Dudek and P. J. Hicks, "A General-Purpose Processor-per-Pixel Analog SIMD Vision Chip". *IEEE Transactions on Circuits and Systems-I*, Vol. 52, No. 1, pp. 13-20, Jan. 2005

[7] G. Liñan et al., "A 1000 Fps At 128*128 Vision Processor With 8-Bit Digitized I/O". *IEEE Journal of Solid-State Circuits,* Vol. 39, No. 7, pp. 1044-1055, Jul. 2004.

[8] R. Carmona et al., "A Bio-Inspired 2-Layer Mixed-Signal Flexible Programmable Chip for Early Vision", *IEEE Trans. on Neural Networks*, Vol. 14, No. 5, pp. 1313-1336, Sep. 2003.

[9] T. Roska and L. O. Chua: "The CNN Universal Machine: An Analogic Array Computer". *IEEE Trans. on Circuits and Systems-II,* Vol. 40, No. 3, pp. 163-173, March 1993.

[10] Cs. Rekeczky, T. Serrano, T. Roska and A. Rodríguez, "A Stored Program 2nd Order/3-Layer Complex Cell CNN-UM". *Int. W. on CNN's and Apps,* pp. 219-224, Catania, Italy, May 2000.

[11] S. Espejo et al., "A VLSI Oriented Continuous-Time CNN Model". Int. J. of Circuit Theory and Apps, Wiley. Vol. 24, No. 3, pp. 341-356, May-June 1996.

[12] R. Domínguez-Castro, S. Espejo, A. Rodríguez-Vázquez and R. Carmona, "Four-Quadrant One-Transistor-Synapse for High-Density CNN Implementations". Int. Workshop on CNNs and their Apps. (CNNA'98), pp. 243-248, London, UK, April 1998.

[13] F. Krummenacher and N. Joehl, "A 4-MHz CMOS Continuous-Time Filter with On-Chip Automatic Tuning". IEEE J. of Solid-State Circuits, Vol. 23, pp. 750-758, June 1988.

[14] R. Carmona, C. M. Domínguez-Matas, J. Cuadri, F. Jiménez-Garrido and A. Rodríguez-Vázquez, "A CNN-Driven Locally Adaptive CMOS Image Sensor". *Int. Symp. Circuits and Systems* (ISCAS'04), Vol. V, pp. 457-460, Vancouver, Canada, May 2004

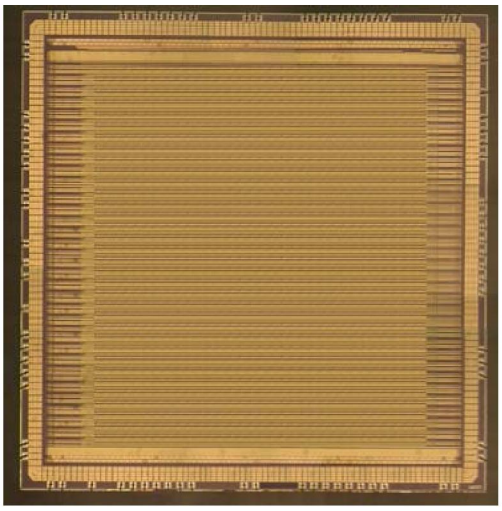


Fig. 7. Microphotograph of the CACE2 prototype.