2nd-Order Neural Core for Bioinspired Focal-Plane Dynamic Image Processing in CMOS

Ricardo Carmona [1], *Francisco Jiménez-Garrido* [1], *Carlos M. Domínguez-Mata* [1], *Rafael Domínguez-Castro* [1], *Servando Espejo* [1], *Istvan Petras* [2] *and Ángel Rodríguez-Vázquez* [1]

[1] Instituto de Microelectrónica de Sevilla-CNM-CSIC
Campus de la Universidad de Sevilla. Avda. Reina Mercedes s/n, Edificio CICA
41012-Sevilla, Spain. Tel.: +34955056666, Fax: +34955056686
E-mail: rcarmona@imse.cnm.es

[2] Analogic and Neural Computing Laboratory
Hungarian Academy of Sciences

ABSTRACT

Based on studies of the mammalian retina, a bioinspired model for mixed-signal array processing has been implemented on silicon. This model mimics the way in which images are processed at the front-end of natural visual pathways, by means of programmable complex spatio-temporal dynamic. When embedded into a focal-plane processing chip, such model allows for on-line parallel filtering of the captured image; the outcome of such processing can be used to develop control feedback actions to adapt the response of photoreceptors to local image features. Beyond simple resistive grid filtering, it is possible to program other spatio-temporal processing operators into the model core, such as nonlinear and anisotropic diffusion, among others. This paper presents analog and mixed-signal VLSI building blocks to implement this model, and illustrates their operation through experimental results taken from a prototype chip fabricated in a $0.5\mu m$ CMOS technology.

I. INTRODUCTION

Physiological and pharmacological studies of the mammalian retina show that this amazing piece of wetware is not a simple phototransducer, but is responsible for very complex signal processing. The retina operates on the captured visual stimuli at early stages in the process of vision. Complex spatio-temporal processing encodes visual information into a reduced set of channels [1]. The visual information flow is compressed into a data set of a manageable size, to be delivered to the brain by the optic nerve. Although the mapping is retinotopic, it is not the raw image brightness that is sent to the visual cortex, but a specific set of image features (closely related with the spatial and temporal characteristics of the visual stimulus) which are obtained and codified in the retina. The purpose of this early vision processing is to alleviate the work of the central nervous system. The application of a highly regular computational task onto a large set of simple data (e.g. picture brightness samples) is transferred to the retina, while the cortex

activity is dedicated to higher level operations on more complex data structures. The massive parallelism of this model inspires a feasible alternative to conventional digital image processing. The limited bandwidth available for transferring signals between the camera array and the processor, and the limited computing speed achievable in a serial, or timidly parallel, processing architecture, make these systems fail to match the tight requirements found in real-time image processing.

We are interested in local monitoring and control of the photosensing devices for contrast enhancement. This capability improves the perceived sensation by extracting the reflectance information from the acquired luminance matrix [2]. Data bottlenecks, arising mostly in transferring image samples from the camera to the processor, and in delivering the appropriate control signals to each photosensor, and the enormous amount of data to be processed, make it hardly realizable at a practical frame rate by a conventional digital processing system. Yet, this task is gracefully implemented in the biological retina. Concurrent processing and sensing eliminate data bottlenecks in the forward and feedback paths, and massively parallel processing provides enough computing power. Mixed-signal VLSI permits the implementation of massively parallel multidimensional signal processing without serious area and power penalties. These chips are called neuromorphic [3] as they mimic the way in which the layers of neurons in the biological retina realize early vision.

An image acquisition and focal-plane processor chip must have, at every pixel, a reliable, locally adaptive photosensing device (the opto-electronic interface) plus the analog and/or mixed-signal core which realizes signal processing at the pixel-level. Concerning the distributed processing facilities, the CNN universal machine architecture [4] has several advantages. It has an analog front-end, which is compatible with the nature of the signals coming from the photosensors, it is general-purpose and fully programmable, it has a distributed memory to store intermediate results, and it has been proven to realize the type of processing required for sensor control [5]. In addition, retinal features have been successfully modelled and simulated within the CNN framework [6].

This paper presents, in the first place, a network model inspired on the layered structure of the mammalian retina. Then the implementation of a fully-programmable 2nd-order neural core to provide active wave computing at the focal-plane is shown. By setting the appropriate parameters: such as interaction strengths, time constants and bias terms, an array of such processing elements can emulate some phenomena observed in the mammalian retina. At the end of the paper, experiments in a $0.5\mu m$ CMOS prototype of 32×32 cells, each one containing a 2nd-order neural core, are displayed.

II. BIOINSPIRED NETWORK MODEL

A) A sketch of the mammalian retina

The retina is a peripheral component of the central nervous system responsible of acquiring and coding the information contained in the visual stimuli. Specialized neurons develop a particular kind of massively parallel processing of raw sensory information. Visual stimuli trigger patterns of activation in the layered structure of the retina, which are processed as they advance

towards the optic nerve. These patterns of activation are analog waves supported by continuous-time signals, contrarily to the spike-like coding of neural information found elsewhere in the nervous system [7]. The biological motivation for this peculiarity can be found in the lack of bandwidth offered by the spike-like neural impulses to handle the vast amount of data contained in the visual stimuli. Fig. 1 displays a conceptual diagram of the functional architecture of the mammalian retina [8]. In this scheme, light comes through the inner retina, all the way across the eye, crosses the transparent layers of cells and is captured by the photosensors in the outer retina. At the outermost end of the layered structure, the retinal pigment epithelium (RPE) is found. This is a non-neuronal layer of cells that surrounds the outer segments (OS) of the photoreceptors. It is the source for the regeneration of the pigment chromofore after its isomerization by light. The following layer is composed of specialized photoreceptive cells of two types: rods and cones. Rods are more light sensitive and responsible for scotopic vision. Cones are less sensitive, more numerous, and are responsible forcolour vision. Their OS contain stacks of discs with rhodopsin, the visual pigment. Rods and cones capture light and convert it into activation signals. Their inner segments (IS) contain the rest of the cellular organelles. The next visible layer is the outer nuclear layer (ONL), which contains the cell bodies of the rods and cones. The outer plexiform layer (OPL) contains the axons from the horizontal cells and the dendritic trees of bipolar cells. They receive synaptic inputs from the rods and cones. Bipolar cells carry the activation signals across the retinal layers to the ganglion cells that interface the retina with the optical nerve, in a trip of several micrometers [1]. The inner nuclear layer (INL) contains the cell bodies of bipolar, horizontal and amacrine cells. The inner plexiform layer (IPL) contains the axons of the bipolar and amacrine cells, and the dendritic trees of the retinal ganglion cells. The ganglion cell layer (GCL) contains the bodies of the ganglion and displaced amacrine cells. The optic nerve fibre (ONF) is built from the axons of the retinal ganglion cells.

The ganglion cells convert the continuous activation signals, proper of the retina, into spike-

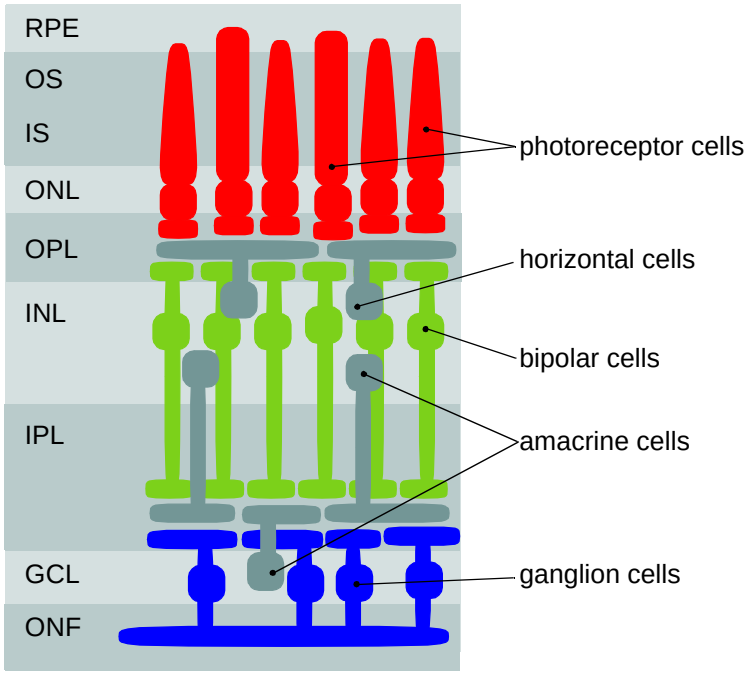

FIGURE 1. Schematic diagram of the functional architecture of the mammalian retina [8].

coded signals which can be transmitted over longer distances by the nervous system. On its 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. 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 [1]. Briefly, captured stimuli are promediated and the high-gain characteristics of the cones and the bipolar cells are shifted to adapt to the particular light conditions. These operations have a local scope and depend on the recent history of the cells. Once adaptation is achieved, patterns of activity are formed dynamically by the presence or absence of visual stimuli. Also inhibition is generated and transmitted laterally through the layers of horizontal and amacrine cells. As a result of these transformations, the patterns of activity reach the layer of ganglion cells. At this point, the patterns are converted into pulse-coded signals that are sent to the brain to be interpreted. In a sense, the layered structure of the retina translates the visual stimuli into a compressed language which can be understood by the brain in recreating vision.

B) CNN analogy of the inner and outer plexiform layers

In the above description there are some aspects of the retinal layers that markedly resemble the features of a cellular neural network (CNN) [9]: the 2D aggregation of continuous signals, the local connectivity between elementary nonlinear processors, and the analog weighted interactions between them. Also, the complete signal pathway in the retina has the topology of a 3D network, or, more properly $2\frac{1}{2}$D network, a pile of 2D layers connected vertically. Motivated by these coincidences, a CNN model has been developed which approximates the observed behaviour of different parts of the mammalian retina. For instance, the outer plexiform layer (OPL). The OPL is responsible for the generation of the first activation patterns immediately after image capture. It has been characterized by experimental measurements, leading to a model with three different layers [10]. These layers stand for the contribution of photoreceptors, horizontal and bipolar cells. Each of them 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 at each layer are supported by a first or a second order continuous-time core.

The inner plexiform layer (IPL) has been also modelled within the CNN framework. The IPL is responsible for the generation of the retinal output. A simplified model of the IPL has three layers. Two of them represent the influence of the wide field amacrine cells excited by the input signal, which in this case is the output of the bipolar cells, and there is a third layer that controls the dynamic of the previous layers by means of feedback. As before, the three layers can be seen as 2D CNNs with their own internal coupling and their own time constant [10].

Because of the relative simplicity of these models, a programmable CNN chip has been proposed [11]. The programmable array processor consists of 2 coupled CNN layers. Each elementary processor contains the nodes for both CNN layers. The third layer, supporting analog arithmetics, is implemented off-line by these analog cores, with the help of 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:

$$\tau_1 \frac{dx_{1,ij}}{dt} = -g[x_{1,ij}] + \sum_{k=-r_1}^{r_1} \sum_{l=-r_1}^{r_1} a_{11,kl} y_{1,(i+k)(j+l)} + b_{11,00} u_{1,ij} + a_{12} y_{2,ij} + z_{1,ij}$$
$$\tau_2 \frac{dx_{2,ij}}{dt} = -g[x_{2,ij}] + \sum_{k=-r_2}^{r_2} \sum_{l=-r_2}^{r_2} a_{22,kl} y_{2,(i+k)(j+l)} + b_{22,00} u_{2,ij} + a_{21} y_{1ij} + z_{2,ij} \quad (1)$$

where the loss term and the activation function are those of the FSR CNN model [12]:

$$g(x_{n,ij}) = \lim_{m \to \infty} \begin{cases} m(x_{n,ij} - 1) + 1 & \text{if} \quad x_{n,ij} > 1 \\ x_{n,ij} & \text{if} \quad |x_{n,ij}| \le 1 \\ m(x_{n,ij} + 1) - 1 & \text{if} \quad x_{n,ij} < -1 \end{cases} \quad (2)$$

and:

$$y_{n,ij} = f(x_{n,ij}) = \frac{1}{2}(|x_{n,ij} + 1| - |x_{n,ij} - 1|) \quad (3)$$

Fig. 2 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 neighbour, weighted by the elements $\{a_{nn,kl}\}$; a feedforward connection, weighted by $b_{nn,00}$, which regulates the contribution of the cell's input; a bias term $z_{n,ij}$, which 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 will be treated as a second input image for each layer.

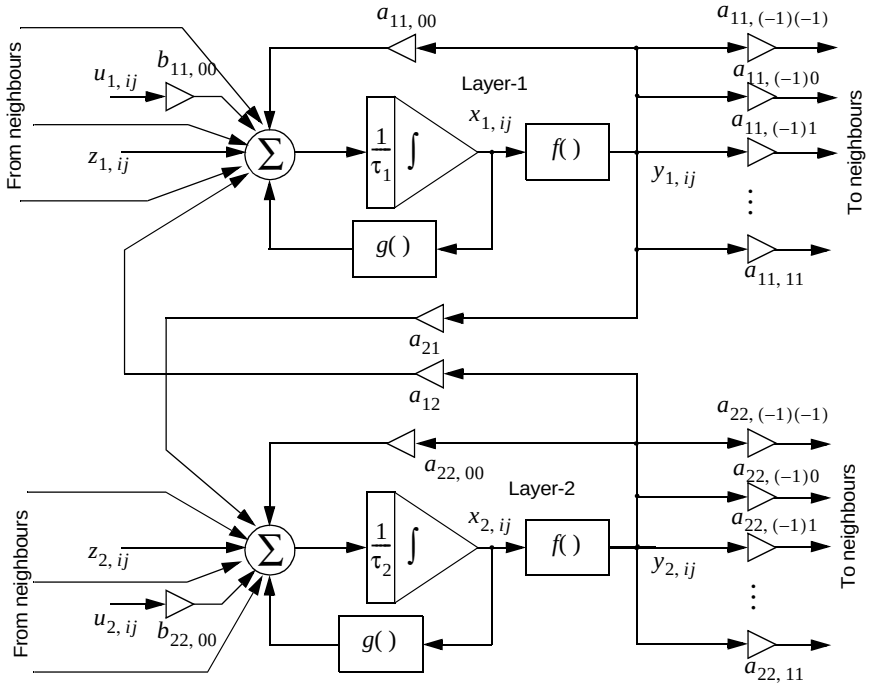

FIGURE 2. Block diagram of the two coupled CNN layer nodes

III. 2ND-ORDER CORE IMPLEMENTATION

A) 2nd-order cell structure

The internal architecture of the basic processing cell presented here is similar to the structure of the cells in the CNN universal machine [4]. However, in this case, the prototype cell includes two different continuous-time CNN layers, as described in the conceptual diagram of Fig. 2. Together with the two different analog CNN core blocks (Fig. 3(a)), local analog and logic memories (4 LAMs and 4 LLMs) are provided at the pixel-level for the storage of intermediate results, and a local logic unit (LLU) is built as well for pixel-level logic operations. The synaptic connections between the analog processing nodes of the same layer are built around the cell core, as shown, while inter-layer coupling, kept within the pixel scope in this model, is placed inside the cell (represented by arrows between the processing layers 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, located outside the array processor.

The internal structure of each CNN core is depicted in the diagram of Fig. 3(b). Each one receives contributions from the rest of the processing nodes in the neighbourhood 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 neighbours' 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 which con-

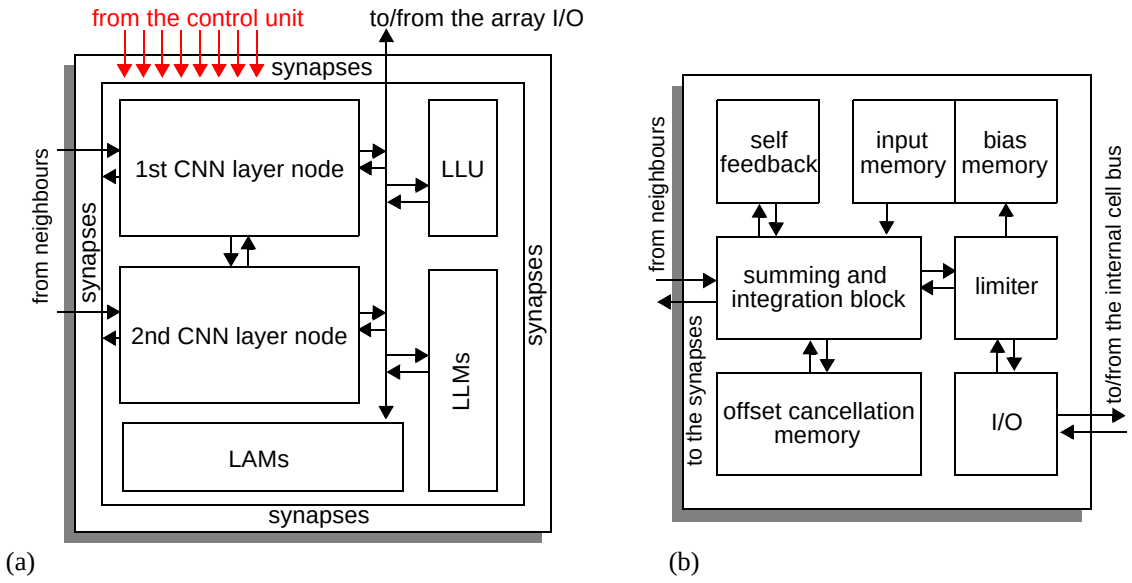

FIGURE 3. Conceptual diagram of the (a) basic cell and the (b) internal structure of each CNN layer node.

nect to the cell's internal data bus.

Running complex spatio-temporal dynamics in this network requires following several initialization and calibration steps. First of all, acquisition of the input image and auxiliary masks and/or patterns. To this purpose, the array I/O interface is directed to specific LAM locations in a row-by-row basis. After that, the analog instruction, i. e. the set of synaptic weights required for a specific operation, is selected and transmitted to all the cells in the array. Then, the offset of the critical OPAMPs is extracted in a calibration step. After that, the time-invariant offsets of the synaptic blocks are computed and stored in the current memories. Now the network is almost ready to operate. The state capacitors and the feedforward synapses are then initialized by means of the appropriate switch configuration, and the network evolution is run by closing the feedback loop in each processing element. Before stopping the network evolution, the final state is stored in a LAM register for further operation.

B) Single-transistor synapse

One of the most important blocks in the cell is the synaptic block. The synapse is, *simply*, a four-quadrant analog multiplier. Its inputs are the cell state, V_x, or input, V_u, variables and the corresponding weight signal, V_w, while the output is the cell's contribution to a specific neighbouring cell. The multiplier is required to have voltage inputs, which can be easily conveyed to any high-impedance node by a simple wire, and current output, which may be easily summed by wiring all current contributions concurrently to a low-impedance node. Two important facts for the implementation of the synaptic blocks are, first, that there is no need to have a strictly linear relation between the weight signal, V_w, and the output current, I_o, and second, that the weight signal does not change during the evolution of the network. Thus, any deviation depending on V_w is not a gain error, but an offset error, i.e. an error which can be cancelled by auto-zeroing in a pre-processing calibration step.

Direct multiplication can be achieved by a MOS transistor operating in the ohmic region. Its low-frequency large-signal characteristic is found in the first-order approach by (if n-type):

$$I_{DS} = \beta_n \left[V_{GS} - V_T(V_{SB}) - \frac{V_{DS}}{2} \right] V_{DS} \tag{4}$$

where $\beta_n = \mu_o C_{ox}'(W/L)$. A multiplication can be realized with this device as long as $V_{DS} \ll 2[V_{GS} - V_T(V_{SB})]$ holds [13]. This alternative has several advantages, compared with multipliers built with MOS transistors in weak inversion or in strong inversion saturation [14]: it requires a reduced amount of area, because four-quadrant behaviour is achieved with one single transistor. In addition, it has a better relation between bias power and signal power, thus leading to higher accuracy at lower power consumption, while in the saturation region the information is carried by a small fraction of the actual currents flowing through the devices. Third, the use of the ohmic region shows better mismatch figures than any other region [15].

The one-transistor synapse works as follows. Consider a p-type MOS transistor operating in the ohmic region (Fig. 4). The transistor selected is of type p because the more resistive p-type channel requires smaller currents (hence smaller power consumption) for the same transistor lengths. Alternatively, for the same current levels, the required p-channel MOS is shorter than its n-type counterpart. The source-to-drain current of a PMOS transistor in the ohmic region is

given by:

$$I_o = -\beta_p(V_A - V_L)V_G - \beta_p(V_A - V_L)\left(\left|\hat{V}_{T_p}\right| - \frac{V_A + V_L}{2}\right) \tag{5}$$

where the threshold adopts one of these two analogue forms:

$$\hat{V}_{T_p} = \begin{cases} -\left|V_{T0_p}\right| - \gamma(\sqrt{\phi_B + V_{DD} - V_A} - \sqrt{\phi_B}) & \text{if } V_A \geq V_L \\ -\left|V_{T0_p}\right| - \gamma(\sqrt{\phi_B + V_{DD} - V_L} - \sqrt{\phi_B}) & \text{if } V_A \leq V_L \end{cases} \tag{6}$$

V_L must be kept fixed in order to use V_A and V_G as single-ended input voltages, and to sense I_o as the output of the synapse. For this purpose we can employ a current conveyor [16] at the current input node of each cell. The current conveyor permits current sensing while maintaining a virtual reference at node (L). All the synapses contributing to the same cell can be connected to the same virtual reference. The only objection is that the impedance at this node must be well below the parallel of the output impedances of all the synaptic blocks.

Back to Eq. (5), notice that the second term on the right side of the equation does not depend on V_G, therefore node (G) is a strong candidate to hold the cell state variable voltage. But V_G must always be positive for the MOS transistor to operate above threshold, thus let V_G be composed of a reference voltage V_{x_0}, sufficiently high, and a superposed cell state signal V_x:

$$V_G \equiv V_X = V_{x_0} + V_x \tag{7}$$

And, in order to achieve four-quadrant multiplication, V_A must be permitted to go above and below V_L. Let us select V_L as the reference for the weight signal, V_{w_0}, being:

$$V_A \equiv V_W = V_{w_0} + V_w \tag{8}$$

Eq. (5) can then be rewritten as:

$$I_o = -\beta_p V_w V_x - \beta_p V_w\left(V_{x_0} + \left|\hat{V}_{T_p}\right| - V_{w_0} - \frac{V_w}{2}\right) \tag{9}$$

which is a four-quadrant multiplier with an offset term which is time-invariant (at least during the transient evolution of the network) and does not depend on the cell state. Therefore, we have arrived at a four-quadrant multiplier with single-ended voltage inputs and a current output, with a offset which can be eliminated by a calibration step, with the help of a current memory:

$$I_o = -\beta_p V_w V_x + I_{\text{offset}}(V_w) \tag{10}$$

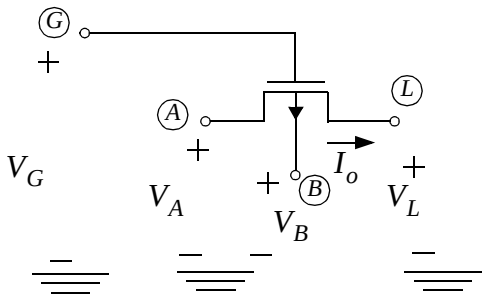

FIGURE 4. Multiplier using one single MOS transistor in the ohmic region.

The limitations found to this behaviour are the upper and lower boundaries of the ohmic region in strong inversion and the degradation of the mobility. The transversal electric field pushes the carriers towards the semiconductor surface where they suffer scattering, which renders a reduction in the speed of the carriers, thus degrading the mobility. This transversal electric field depends on the gate voltage, thus the first summand in Eq. (10) is no longer linear with V_x. Combining the two limiting factors:

$$V_{w_{\max}} + V_{x_{\max}} \leq \frac{1}{2} V_{GE_{\max}} - \frac{1}{2}\left[\left|V_{T_p}(V_{DD} - V_{w_0})\right| - \left|V_{T_p}(V_{DD} - V_{w_0} - V_{w_{\max}})\right|\right] \tag{11}$$

where $V_{GE_{\max}} = (V_{SG} - \left|V_{T_p}(V_{SB})\right|)\big|_{\max}$ is a maximum effective gate voltage, beyond which the distortion introduced by mobility degradation exceeds the linearity requirements.

For moderate linearity requirements, in a typical CMOS technology, the right hand side of Eq. (11) becomes approximately equal to 1V. If V_x and V_w are assigned the same voltage ranges, $\pm 400\text{mV}$ around their reference values, then $V_{x_{\max}} = V_{w_{\max}} = 400\text{mV}$. With these values of $V_{x_{\max}}$, $V_{w_{\max}}$ and $\left|V_{T_p}\right| \approx 0.8\text{V}$, V_{x_0} must be kept 1.6V below V_{w_0}. Thus, V_{w_0} must be high enough to leave room for V_{x_0}, but not too large because the weight signal will progress up to $V_{w_{\max}}$ above V_{w_0}. In addition, we have to provide a range for the current conveyor circuitry to maintain a virtual reference precisely at V_{w_0}, and for the circuits generating the weight voltages, which will have a limited output swing. If we select $V_{w_0} = 2.55\text{V}$, then they are 0.75V above V_{w_0} before hitting the power rail at 3.3V, which means one $\left|V_{T_p}\right|$, approximately. With this value, V_{x_0} results in 0.95V. Finally, once the voltage ranges are fixed, a maximum current per synapse is selected for meeting power requirements, in this case it will be $1.4\mu\text{A}$. With these values, the synapse is dimensioned. In this chip, it will be $2\mu\text{m}$ wide and $25.9\mu\text{m}$ long.

C) Current conveyor

The current conveyor, required for creating a virtual reference node at which the synapses outputs can be sensed, is implemented in the circuit of Fig. 5. Any difference between the voltage

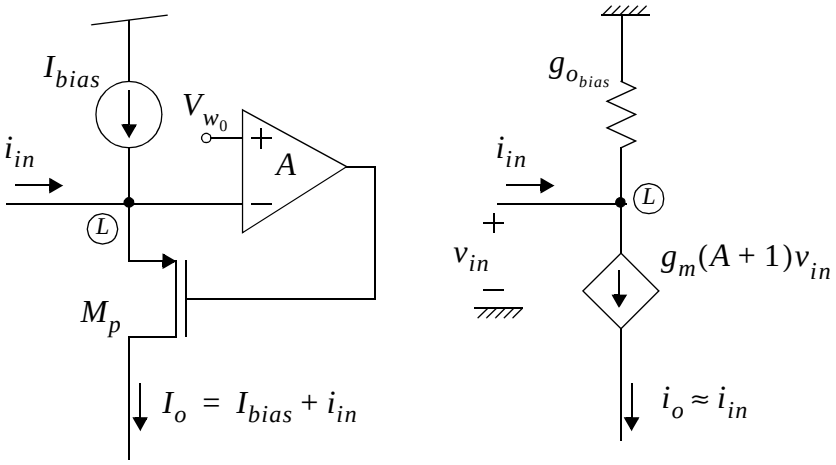

FIGURE 5. Current conveyor realization and small-signal equivalent.

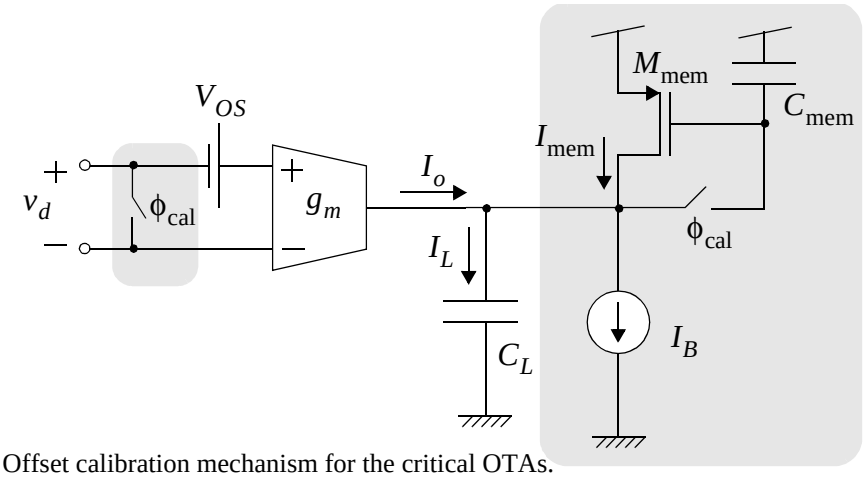

FIGURE 6. Offset calibration mechanism for the critical OTAs.

at node (L) and the reference V_{w_0} is amplified and the negative feedback corrects the deviation. The input impedance of this block is very low, which means that changes in the small-signal input current Δi_{in} do not disturb appreciably the virtual reference at node (L), thus: $\Delta v_{in} \approx 0$. The bias current is required to ensure that node (L) is always the source of transistor M_p. At the same time, this circuit permits the injection of a nearly exact copy of the input current at the state node, whose voltage range differs from that of the weight signals. The only drawback of using this circuit is that a voltage offset, V_{OS}, at the input of the differential amplifier —which can be implemented with a simple OTA as it drives a very high impedance node, the gate of M_p — results in an error of the same amount in the reference voltage implemented at node (L). Since the main contribution to the offset is random, this error will be distributed all along the array resulting in mismatched synaptic blocks which can degrade performance, e. g. anisotropic evolution of the network yielded by a symmetrical propagation template. As we are impelled to use small-size devices, in order to achieve the highest cell-packing density possible, the random offset can be quite large. In order to avoid this, an offset calibration mechanism has been implemented at the critical OTAs (Fig. 6). The input-referred offset voltage, V_{OS}, has been taken out of the OTA block symbol. Without the offset cancellation circuit (the shadowed area), at low frequencies, and considering a negligible output conductance, the output of the OTA is:

$$I_o = g_m(v_d + V_{OS}) \tag{12}$$

Considering the error cancellation mechanism, when ϕ_{cal} is ON, then the inputs are short-circuited, $v_d = 0$, and M_{mem} is connected as a diode,with its source-to-drain is in steady-state:

$$I_{mem} = I_B - g_m V_{OS} \tag{13}$$

After some time, ϕ_{cal} is turned off and, except for a remnant switching error, the current I_{mem} is memorized by means of the voltage stored in C_{mem}. Thus, the total current injected into the load is free of any offset:

$$I_L = I_o + I_{mem} - I_B = g_m v_d \tag{14}$$

D) Current memory

The offset term of the synapse current must be removed for the output current to accurately represent the result of a four-quadrant multiplication. To this purpose, before the CNN opera-

tion, but right after the new weights have been up-loaded, all the synapses are reset to $V_X = V_{x_0}$. 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 during the network evolution, resulting in a one-step cancellation of the errors of all the synapses. The validity of this method relies on the accuracy of the current memory. For instance, in this chip, the sum of all the contributions will range from $18\mu A$ to $46\mu A$. On the other hand, the maximum current signal of the synapse is:

$$I_{\max} = \beta_p V_{x_{\max}} V_{w_{\max}} \approx 0.5\mu A \tag{15}$$

which means a total current range of $1\mu A$. If an equivalent resolution of $8bits$ is intended, then, $(1/2)LSB = 2nA$. In these conditions, our current memory must be able to distinguish $2nA$ from the $46\mu A$. This represents an equivalent resolution of $14.5bits$. In order to achieve such accuracy levels, a so-called S^3I current memory will be employed [17]. As depicted in Fig. 7, it is composed of three stages, each one containing a switch, a capacitor and a transistor. At the beginning, while ϕ_1, ϕ_2 and ϕ_3 are ON, the current I_{in} is divided into I_1, I_2 and I_3, and:

$$V_1 = V_2 = V_3 = V_m = V_{T0_n} + \sqrt{\frac{I_{in}}{\beta_1 + \beta_2 + \beta_3}} \tag{16}$$

Switches controlled by ϕ_1, ϕ_2 and ϕ_3 are successively turned off. Each time that one of these switches turns off, the voltage stored in its associated capacitor changes, e. g. V_1 changes from V_m to $V_m + \Delta V_1$, because of charge injection. The other transistors have to accommodate to absorb the error, as the sum of currents is still forced to be I_{in}, and thus V_2 and V_3 change to:

$$V_m' = V_{T0_n} + \sqrt{\frac{I_{in}}{\beta_1 + \beta_2 + \beta_3} + \frac{g_{m_1}\Delta V_1}{\beta_2 + \beta_3}} \tag{17}$$

when ϕ_1 turns off. Correspondingly, V_3 changes to:

$$V_m'' = V_{T0_n} + \sqrt{\frac{I_{in}}{\beta_1 + \beta_2 + \beta_3} + \frac{g_{m_1}\Delta V_1}{\beta_2 + \beta_3} + \frac{g_{m_2}\Delta V_2}{\beta_3}} \tag{18}$$

when ϕ_2 falls. Finally ϕ_3 is turned off, and V_3 ends in $V_m'' + \Delta V_3$. The final current, I_{out}, is:

$$\begin{aligned} I_{out} = {} & \beta_1 (V_m - V_{T0_n})^2 - g_{m_1}\Delta V_1 + \beta_2 (V_m' - V_{T0_n})^2 - g_{m_2}\Delta V_2 + \\ & + \beta_3 (V_m'' - V_{T0_n})^2 - g_{m_3}\Delta V_3 \end{aligned} \tag{19}$$

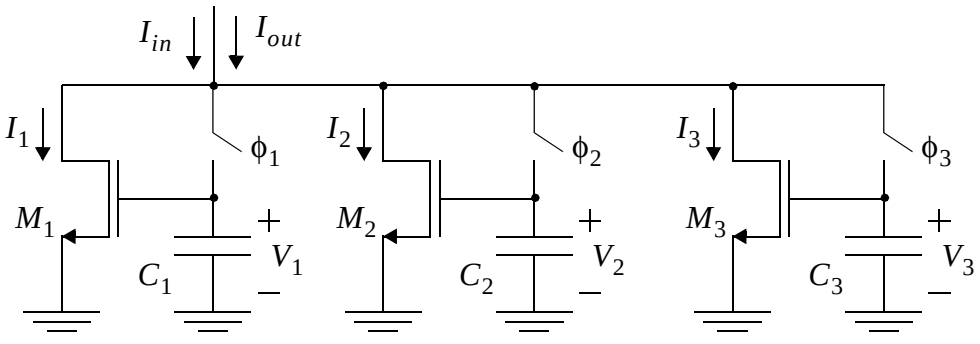

FIGURE 7. S^3I current memory schematics and timing

and substituting here the values of V_m, V_m' and V_m'', we find that:

$$I_{out} = I_{in} - g_{m_3}\Delta V_3 \tag{20}$$

the only error left is that corresponding to the last stage. The former stages do not contribute to the error in the memorized current. If the S^3I block is designed to store the most significant bits in the first capacitor, and the less significant bits in the last one, then the error in the memorized current can be made quite small. Consider that the total resolution of the current memory is N. Let us assume that M_1 is conducting the $N/3$ most significant bits of the current I_{in}, then M_2 conducts the next $N/3$, and M_3 conducts the rest. Thus, for the last stage an effective resolution can be defined:

$$I_3 = \left(\frac{I_{in}}{2^N}\right)\sum_{k=\frac{2N}{3}+1}^{N} 2^{N-k} = \left(\frac{I_{in}}{2^N}\right)2^{N_{\text{eff}}} \tag{21}$$

If the error in the memorized current has to be kept below 0.5LSB, and $g_{m_3} = 2\sqrt{\beta_3 I_3}$, then:

$$\Delta V_3 \le \sqrt{\frac{I_3}{\beta_3}} \cdot 2^{-\left(\frac{N_{\text{eff}}}{2}+\frac{N}{2}+2\right)} \tag{22}$$

And this is the design equation which relates the geometric aspect of transistor M_3, through β_3, with the magnitude of the storage capacitor, via ΔV_3. Once we have β_3, β_1 and β_2 it may be easily derived that:

$$\beta_1 = \left(\frac{\beta_3}{2^{N_{\text{eff}}}}\right)\sum_{k=1}^{N/3} 2^{N-k} \qquad \beta_2 = \left(\frac{\beta_3}{2^{N_{\text{eff}}}}\right)\sum_{k=N/3+1}^{2N/3} 2^{N-k} \tag{23}$$

One might think that adding more stages to the current memory will endlessly increase accuracy. However, there is one factor that has not been addressed yet. As the order of the memory increases, the smaller the currents become which have to be sensed by the last stages. There comes a point in which the leakages from the capacitors of the first stages are of the size of the current to be memorized by the last stages, thus making it impossible to reach a steady state current which corrects the previous errors. This problem worsens as temperature rises. For instance, at 70C leakages can introduce changes in the memorized current in the order of $0.2\text{nA}/\mu\text{s}$. If the dynamics of the current memory require several μs to settle (because of the use of large capacitors and due to the tiny currents involved) the memorized current will display an error that is quite above the initial estimation.

E) Time constant scaling block

The time constant of the CNN layer is defined as $\tau = C_c/G_c$, the ratio between the state capacitor and the transconductance G_c obtained by multiplying the current factor of the synapse, $\beta_p = 3.13\mu\text{A/V}^2$, times the weight signal voltage V_w. This time constant depends on the specific set of templates being implemented in the CNN. The state capacitor is composed by the gate capacitances of the 11 synapses driven by the cell's state. As $C_{ox}' = 3.45\text{fF}/\mu\text{m}^2$ in this technology, this makes a total of 1.97pF. In the most favourable case, when every neighbour,

even the cell itself, is contributing the maximum amount of current to the cell state, a parallel stack of 18 synapses, with a transconductance of $22.5\mu A/V$ is found. This represents a minimum CNN time constant of $87.4ns$.

Scaling the time constant of one of the CNN layers involves either modifying the value of the state capacitor or of the synapse transconductance. For the first alternative, it will be necessary to implement a regulable capacitor. If a continuously regulable capacitor is pretended, it does not seem to be easy to realize. If a capacitor with a discrete set of capacitances is adequate, an area of 16 times $2 \times 25.9 = 51.8\mu m^2$ will be required to implement a $1:16$ time constant ratio.

The second alternative, scaling the transconductances of every synapse contributing to the cell, can be achieved with a current mirror. Scaling up/down the sum of currents entering the cell is equivalent to scaling up/down the transconductances of the synapses, and thus, to scaling down/up the time constant of the CNN core. A circuit for continuously adjusting the gain of a mirror can be designed based on the active-input regulated-Cascode current mirror [18]. The major disadvantage of using this circuit is its strong dependence on the power rail voltage. The power rail voltage can deviate further more than 1% in a densely packed 32×32-cell parallel array processor chip. This will cause a large mismatch in the time-constants of the different cells in the layer. An alternative to this is a binary programmable current mirror. Its output current is given by:

$$I_o = (1 + b_0 + 2b_1 + 4b_2 + 8b_3)\frac{I_{in}}{16} \tag{24}$$

where b_0, b_1, b_2 and b_3 are the decimal values of the control bits. In this case, 4 bits will be more than enough to program the required relations between τ_1 and τ_2. The mismatch between the time constants of the different cells is now fairly attenuated by design.

A new problem arises related with the placement of the scaling block in the signal path. There are several alternatives. First, the scaling block, the binary weighted current mirror, can be placed after the offset cancellation memory, as in Fig. 8(a). The problem is that any offset introduced by the scaling block is incorporated to the signal path without possible cancellation. The second alternative (Fig. 8(b)) is to place the scaling block before the offset cancellation memory. This means that the S^3I memory will have to operate over a wider range of currents, thus complicating its design and surely degrading its performance. Our choice, depicted in Fig. 8(c) has been to place the scaling block in the memorization loop. The current memory will operate on the unscaled version of the input current, and any offsets associated with the scaling blocks will be sensed and memorized to be cancelled on-line during the network evolution.

The resulting CNN core is shown in Fig. 9 [19]. In this picture, the voltage reference generated with the current conveyor, the 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. Without it, the output node, (A), will diverge from the equilibrium. The operation of this circuit is as follows. Before running the CNN dynamics, the current offsets of all the synapses are injected to the virtual reference at node (L). This current is scaled down to one n-th of its value by means of the adjustable current mirror formed by M_{n1} and M_{n2}. The arrow over M_{n2} stands for the binary programmability of this device. The value of n is:

$$n = 1 + b_0 + 2b_1 + 4b_2 + 8b_3 \tag{25}$$

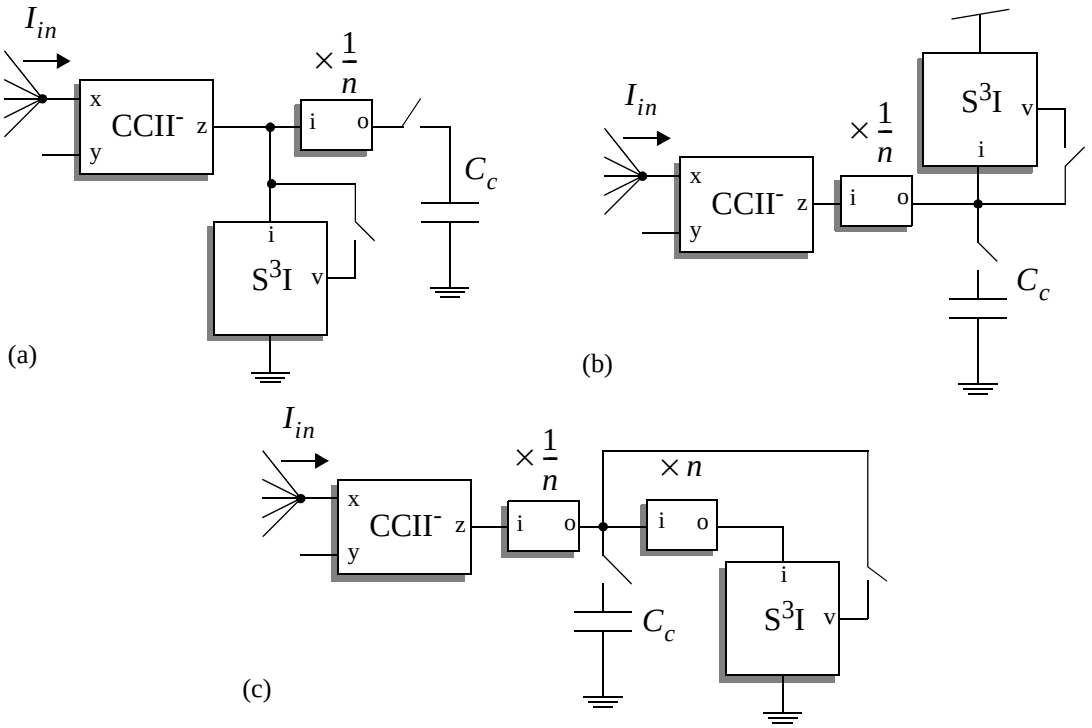

FIGURE 8. Alternatives for the placement of the scaling block.

Then, if all the transistors of the S^3I memory are conducting, that is, if ϕ_1, ϕ_2 and ϕ_3 are ON, then the negative feedback loop makes M_{p_2} conduct the same current as M_{n2}. M_{p2} is also adjustable so as to make M_{p1} and the current memory work with the same current ranges as in the input stage. The rest of the operation has been already described. The current memory stores successively the remaining most significant bits of the input current, plus the errors accumulated. When this is done, the CNN loop can be closed and the output current I_o represents the scaled sum of the contributions, with the state-independent errors substracted.

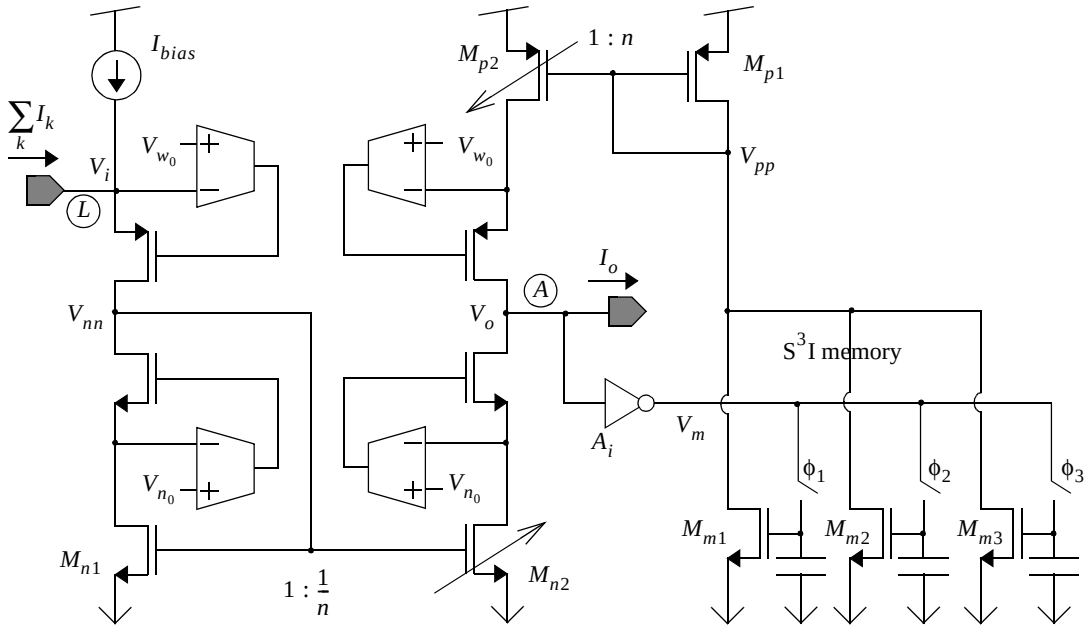

FIGURE 9. Input block with current scaling.

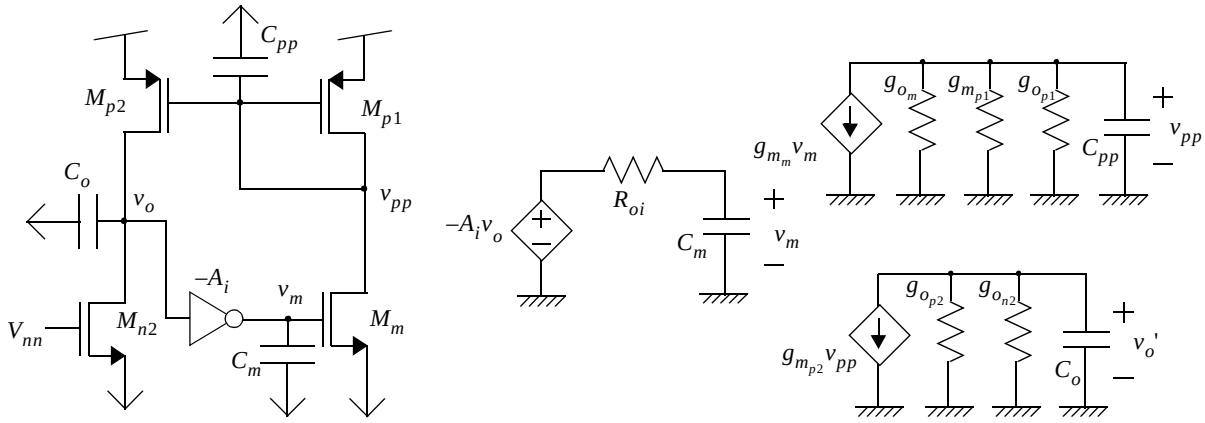

FIGURE 10. Simplified schematics of the feedback loop and its small signal equivalent.

The critical aspects of this circuit are related to the feedback loop formed by M_{p1}, M_{p2}, M_{n2}, inverting amplifier A_i and transistors M_m, when sensing the offset current. During this process output current I_o is zero because the current path to the state capacitor is open. Once the input current has been established, V_{nn} can be considered a bias voltage. First of all, it must be taken into account that during the three different phases in which the loop is closed (ϕ_1,ϕ_2 and ϕ_3 ON, ϕ_1 OFF and ϕ_2 and ϕ_3 ON, and, finally, ϕ_1 and ϕ_2 OFF and ϕ_3 ON) the values of g_{m_m} and C_m change, so the stability conditions must hold for any possible set of values. Considering the small-signal equivalent circuit for this loop, a three-pole system is found (Fig. 10), with pole frequencies: $p_1 = -g_{o_2}/C_o$, $p_2 = -1/C_m R_{oi}$ and $p_3 = -g_{m_{p1}}/C_{pp}$. The nearest pole, at node V_o, will be employed to compensate the loop for stability. As g_{m_m} and C_m decrease for the latest phases of the current memorization, the loop will be more stable because this causes the loop dc gain, T_0, to decrease and p_2 to grow, breaking away from p_1 and thus increasing the phase margin. Therefore, the worst situation will occur when ϕ_1,ϕ_2 and ϕ_3 are ON, and thus the circuit is designed to be stable in these conditions. It is also important that A_i is kept reasonably low, otherwise it will displace the unity-gain frequency, ω_u, towards the value of the inversion ω_{180}. This means a loss of phase margin, and can compromise the loop stability.

Leakage currents can degrade the S^3I memory operation especially as the operation temperature rises. Although the negative feedback moves the circuit towards the correction of the errors, it may be too slow to settle at a value before leakages modify the position of the equilibrium point. Therefore, compensation must be kept under a limit to avoid slowing down the loop dynamics in excess.

IV. EXPERIMENTAL RESULTS

A) Prototype chip data

A prototype chip has been designed and fabricated in a standard $0.5\mu m$ CMOS technology with single-poly and triple-metal layers. Fig. 11 displays a microphotograph of the chip. It contains a central array of 32×32 2nd-order cells of the type formerly described (this prototype does not incorporate the adaptive photosensors). Surrounding the array, a ring of boundary cells,

implementing the contour conditions for the CNN dynamics, is found, together with the necessary buffers to transmit digital instructions and analog references to the array. On the lower part of the chip, the program control and memory blocks can be found. The last major subsystem is the I/O interface including S/H batteries, decoders, counters and different sequential logic. The whole system fits in $9.27 \times 8.45\text{mm}^2$, including the ring of bonding pads. One single processing element occupies $188\mu\text{m} \times 186\mu\text{m}$. The resulting cell density is 29.24cells/mm^2. In order to cautiously handle this data, it is important to notice that the area occupied by the cell array scales linearly with the total number of cells, which is not the case with the overhead circuitry, which tends to be a smaller fraction of the total chip size as the number of cells rises. The power consumption of the whole chip has been estimated at 300mW. Data I/O rates are nominally 10Ms/s. The time constant of the fastest layer (fixed time constant) is designed to be under 100ns. The chip can handle analog data with an equivalent resolution of 7.5bits (measured). The peak computing power of this chip is of 470GOPS. Here, OPS means analog arithmetic operations per second. In a time constant, 100ns, each CNN core performs 12 multiplications and 11 additions. Thus, for each cell, with two cores, there are 46 operations within each cell in 100ns. Having 1024 processing cells, the chip can reach 470GOPS when running the network dynamics. The computing power per unit area —considering the main array alone— is 6.01GOPS/mm^2 and per unit power is 1.56GOPS/mW.

B) Retinal behaviour emulation

Image processing algorithms can be programmed on this chip by setting the configuration of switches and by tuning the appropriate interconnection weights — the programming interface is digital while the internal coding of the weights is analog. Propagative and wave-like phenomena, similar to those found at the biological retina, can be observed in this chip by just setting the proper coupling between cells in the same or in different layers. For instance, it can be pro-

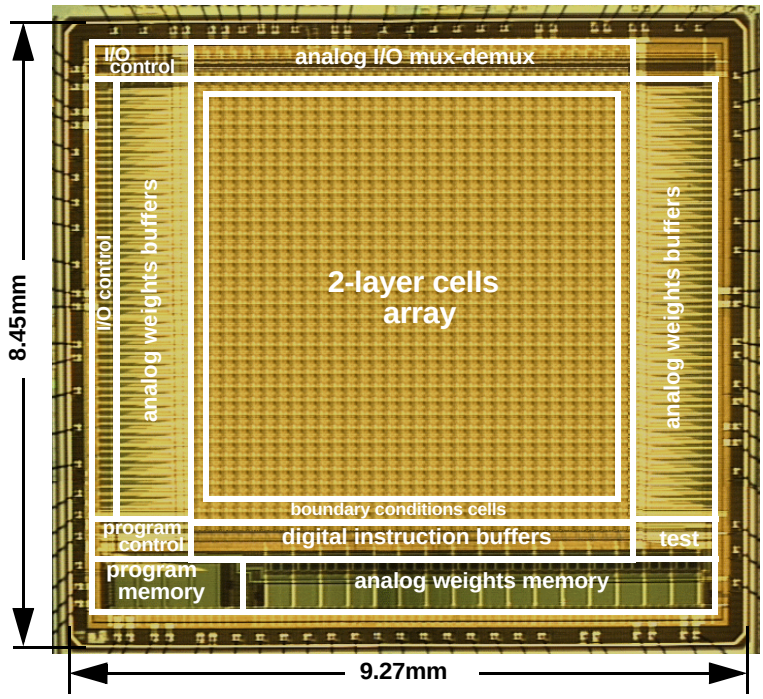

FIGURE 11. Microphotograph of the prototype chip.

grammed to propagate spots in the faster layer towards the border of the array. These spots trigger a slower set of waves in the first layer. The wavefronts 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. This produces similar results to the wide field erasure effect observed in the IPL of the retina. Fig. 12 displays a 3-D plot of this effect. These pictures have been generated with the prototype chip by running the network dynamics, from the same initial state, during successively larger periods of time. This permits the reconstruction of the actual evolution of the state of the cells during the CNN dynamics. Another interesting effect observed in the OPL of the retina [10] is the detection of spatio-temporal edges followed by de-activation of the patterns of activity. This phenomenon has also been programmed in the chip (Fig. 13).

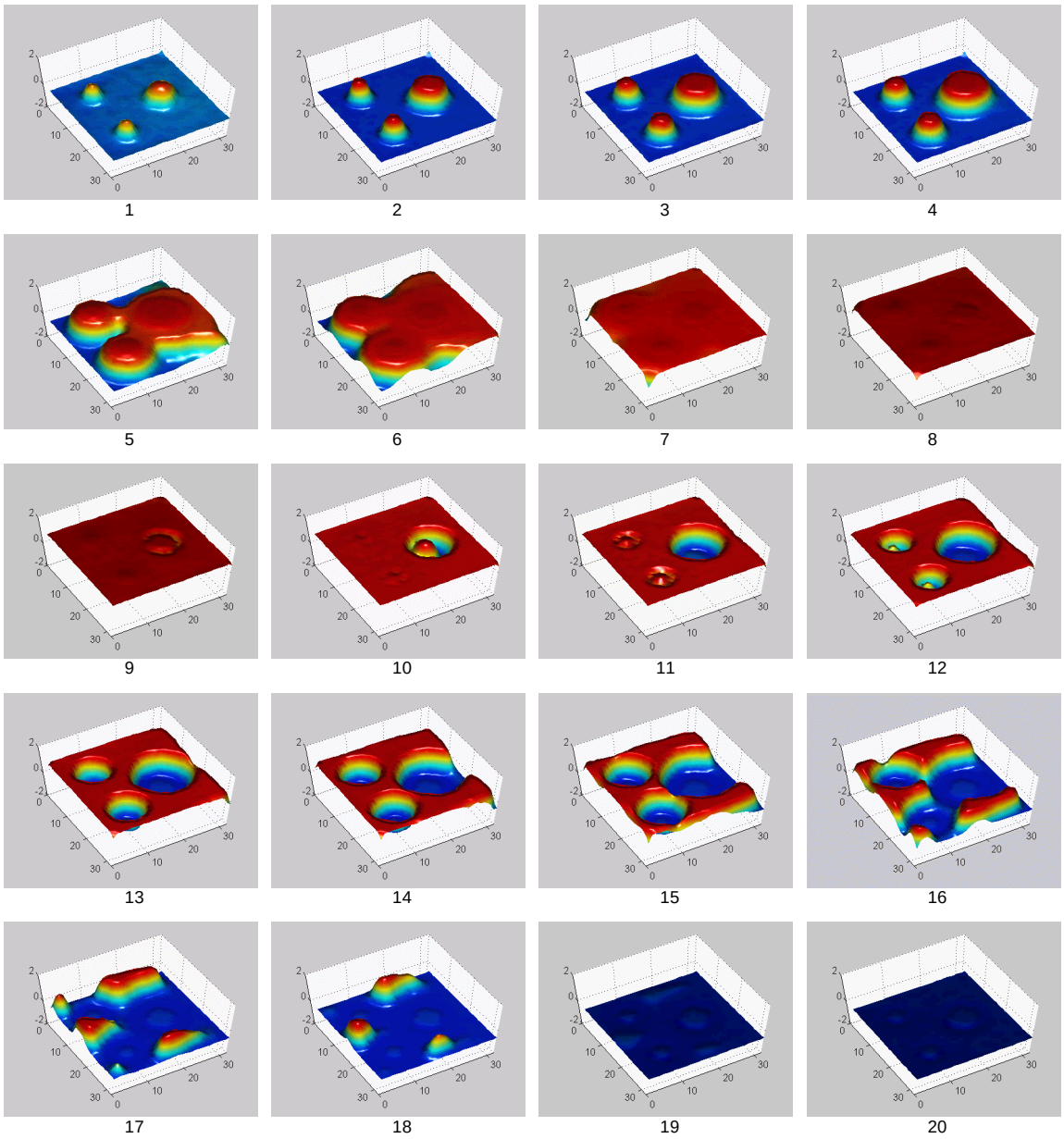

FIGURE 12. Wide field erasure effect, represented in 3-D.

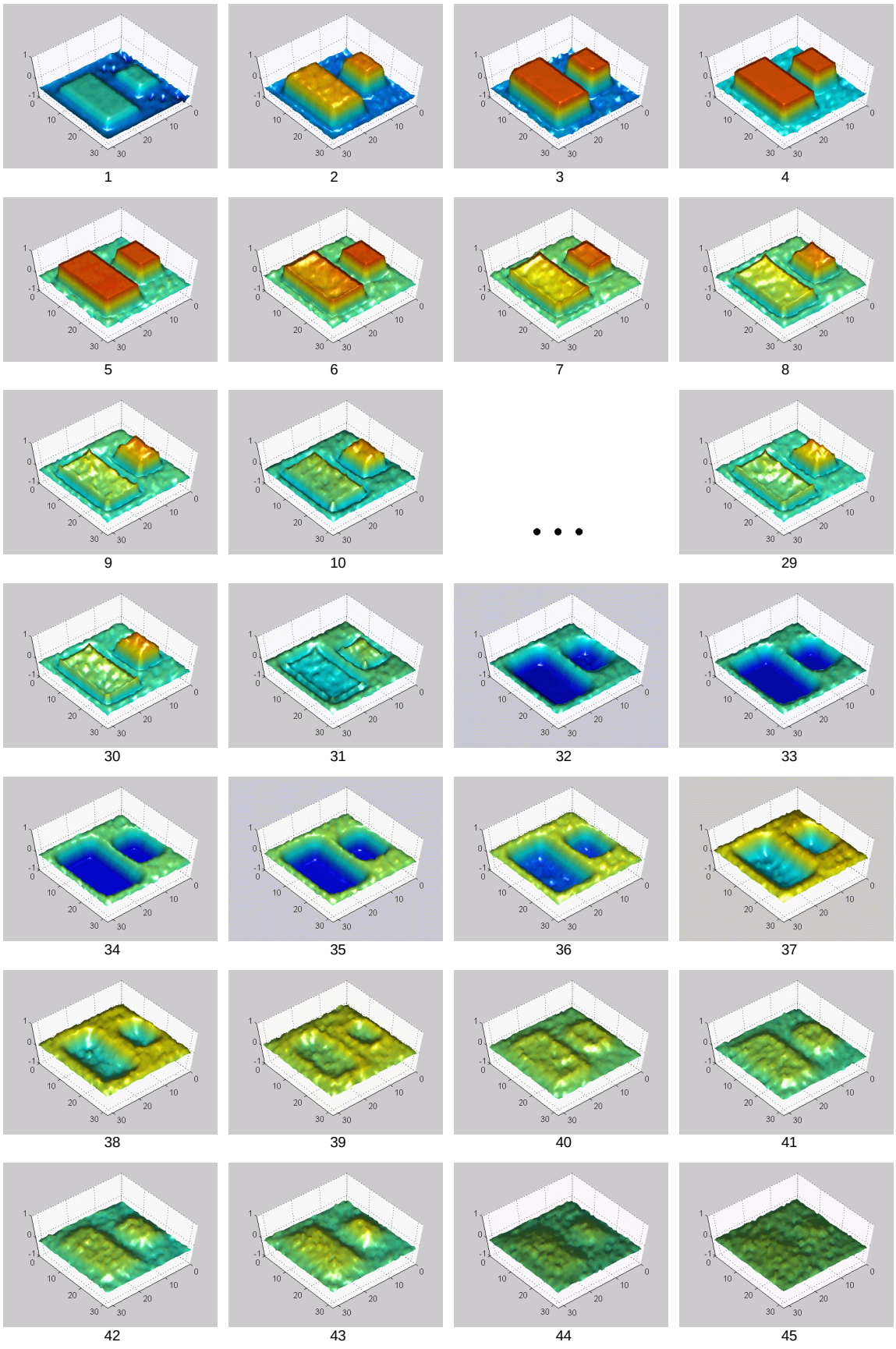

FIGURE 13. Spatio-temporal edge detection and de-activation (fast layer).

V. CONCLUSIONS

Based on a simple but precise model of the real biological system, a feasible efficient implementation of an artificial vision device has been designed. Tailored analog building blocks for fully programmable focal-plane image processing are provided. A prototype chip containing a network of $32 \times 32 \times 2$ CNN nodes have been designed, fabricated and successfully tested in standard CMOS technologies. Different wave-computing algorithms can be implemented in this chip by simply programming the network dynamics with only a few parameters: connection weights, time constant ratio, bias map and boundary conditions.

ACKNOWLEDGEMENTS

This work has been partially supported by projects IST2001 – 38097 (LOCUST), TIC2003 – 09817 - C02 – 01 (VISTA), and ONR-NICOP N000140210884. The authors deeply appreciate the many useful and fruitful discussions with G. Liñán related to chip architecture and circuit design, T. Serrano-Gotarredona regarding the implementation of programmable current mirrors and with D. Bálya, P. Földesy and I. Petrás related to the experiments.

REFERENCES

[1] 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.

[2] V. Brajovic, “A Model for Reflectance Perception in Vision”. *Bioengineered and Bioinspired Systems, Proceedings of SPIE*, Vol. 5119, pp. 307-315, May 2003.

[3] Alireza Moini, *Vision Chips*. Kluwer Academic Publishers, Boston, 1999.

[4] 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.

[5] Cs. Rekeczky, T. Roska, and A. Ushida, “CNN-based Difference-controlled Adaptive Nonlinear Image Filters”. *International Journal of Circuit Theory and Applications*, Vol. 26, pp. 375-423, July-August, 1998.

[6] D. Balya, Cs. Rekeczky and T. Roska, “Basic Mammalian Retinal Effects on the Prototype Complex Cell CNN Universal Machine”. *Proceedings of the 7th IEEE International Workshop on Cellular Neural Networks and their Applications*, pp. 251-258, 2002. Frankfurt am Main, Germany, July 2002.

[7] 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.

[8] H. Wassle and B.B Boycott, “Functional Architecture of the Mammalian Retina”. *Physiological Reviews*, Vol. 71, No. 2. pp. 447-447, 1991.

[9] 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.

[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] Y. P. Tsividis, "Integrated Continuous-Time Filter Design —An Overview". *IEEE Journal of Solid-State Circuits*, Vol. 29, No. 3, pp. 166-176, March 1994.

[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.

[15] A. Rodríguez-Vázquez, E. Roca, M. Delgado-Restituto, S. Espejo and R. Domínguez-Castro, "MOST-Based Design and Scaling of Synaptic Interconnections in VLSI Analog Array Processing Chips". *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, Vol. 23, pp. 239-266, Kluwer Academics, November/December 1999.

[16] K. C. Smith and A. S. Sedra: "The Current Conveyor —A New Circuit Building Block". *IEEE Proceedings*, Vol. 56, pp. 1368-1369, August 1968.

[17] C. Toumazou, J. B. Hughes and N. C. Battersby (Eds.), *Switched-Currents: An Analogue Technique for Digital Technology*. Peter Peregrinus, London, England, 1993.

[18] T. Serrano and B. Linares-Barranco: "The Active-Input Regulated-Cascode Current Mirror". *IEEE Transactions on circuits and Systems-I: Fundamental Theory and Applications*, Vol. 41, No. 6, pp. 464-467, June 1994.

[19] Ricardo Carmona, *Analysis and Design of CNN-based VLSI Hardware for Real-Time Image Processing*. Ph. D. Thesis, University of Seville, June 2002.