CMOS REALIZATION OF A 2-LAYER CNN UNIVERSAL MACHINE CHIP

R. CARMONA, F. JIMÉNEZ-GARRIDO, R. DOMÍNGUEZ-CASTRO,
S. ESPEJO AND A. RODRÍGUEZ-VÁZQUEZ

Instituto de Microelectrónica de Sevilla-CNM-CSIC. Avda. Reina Mercedes s/n
41012 Sevilla (SPAIN). Tel.: +34 955056666, Fax: +34 955056686.
E-mail: rcarmona@imse.cnm.es

Some of the features of the biological retina can be modelled by a cellular neural network (CNN) composed of two dynamically coupled layers of locally connected elementary nonlinear processors. In order to explore the possibilities of these complex spatio-temporal dynamics in image processing, a prototype chip has been developed implementing this CNN model with analog signal processing blocks. This chip has been designed in a 0.5μm CMOS technology. Design challenges, trade-offs and the building blocks of such a high-complexity system (0.5×10^6 transistors, most of them operating in analog mode) are presented in this paper[a].

1 CNN-UM chip architecture

1.1 CNN-based analogy of the biological retina

The vertebrate retina is composed of several layers of horizontal and amacrine cells [1]. These layers, coupled by means of bipolar cells, end, on one side, in a layer of photodetectors and, on the other, in a layer of ganglion cells. The photodetectors capture the visual stimuli and translate it into activation patterns. The ganglion cells, at the other end of the retina, convert the continuous activation signals into pulse-like action potential signals that can be transmitted over longer distances by the nervous system. The activation signals in the retina are weighted and promediated to bias photodetectors and to inhibit the vertical pathway. Patterns of activity are formed dynamically by the presence or absence of visual stimuli. In this description, similarities can be found with the CNNs [2]: not only in the topology, but also in that we have 2D aggregations of continuous signals, local connectivity between elementary nonlinear processors and analog weighted interactions between them. Motivated by these coincidences, a model for the operations of the biological retina based on CNNs has been developed [3]. It contains two coupled CNN layers plus an additional layer incorporating analog arithmetics to combine the outputs of the dynamically linked layers. This can be realized by a CNN Universal Machine (CNN-UM) architecture [4] in which each cell contains two first-order cores, common local analog and logic memories (LAMs and LLMs) and common logic and communication units (LLU and LAOU). The evolution law of each cell, $C(i, j)$, is given by two coupled equations:

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

a. This work has been supported by ESPRIT V Project IST-1999-19007 and by ONR/NICOP Grant N-00014-00-1-0429, and the Spanish CICYT Project TIC-1999-0826.

where the nonlinear losses term and the output function in each layer are those of the Full-Signal-Range (FSR) CNN model [5], which, having a limitation on the cell state voltage allows for identifying state and output:

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

and

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

1.2 Prototype chip floorplan

The proposed chip consists in an analog parallel array processor (APAP) of 32×32 identical cells (Fig. 4). It is surrounded by the circuits implementing the boundary conditions for the CNN dynamics. There is also an I/O interface, a timing and control unit and a program memory. The I/O interface consists in a serializing-deserializing analog multiplexor. The program memory is composed of 24 blocks of SRAM of 64 bytes of capacity, 1kB dedicated to the analog program, and 0.5kB to the logic program. In addition, the analog instructions and reference signals need to be transmitted to every cell in the network in the form of analog voltages. Thus, a bank of D/A converters interfaces the analog program memory with the processing array. Finally, the timing unit is composed by an internal clock/counter and a set of finite-state- machines that generate the internal signals that enable the processes of image up/downloading and program memory accesses.

1.3 Basic cell scheme

The elementary processor of the CNN includes two coupled continuous-time cores (Fig. 1(a)). Each one belongs to one of the two different layers of the network. The synaptic connections between processing elements of the same or different layer are represented by arrows in the diagram. The basic processor contains also the LLU, and the LAMs and LLMs to store intermediate results. All the blocks in the cell communicate via an intra-cell data bus, which is multiplexed to the array I/O interface. Control bits and switch configuration are passed to the cell directly from the global programming unit.

The internal structure of each CNN core is depicted in the diagram of Fig. 1(b). Each core receives contributions from the rest of the processing nodes in the neighbourhood, and these contributions 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 for implementing the FSR CNN model. The 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.

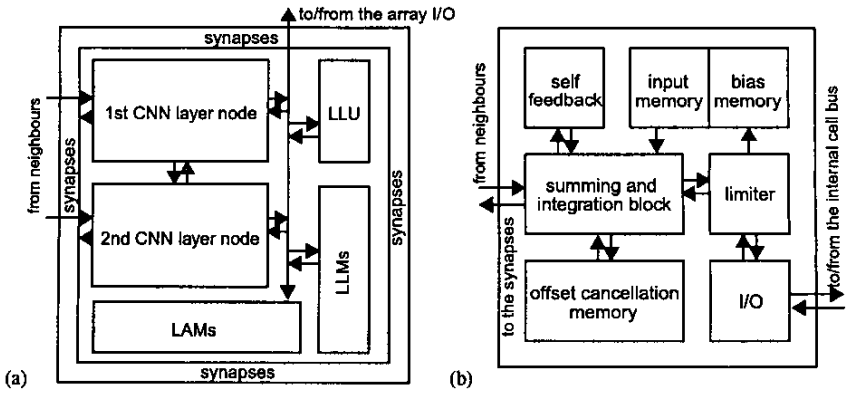


Figure 1. (a) Conceptual diagram of the basic cell and (b) the CNN layers' nodes.

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.

2 Analog building blocks for the basic cell

2.1 Single-transistor synapse

The synapse is a four-quadrant analog multiplier. Their inputs will be the cell state (V_x)— identified with the cell output in the FSR model— or input and the weight voltages (V_w), while the output (I_{ds}) will be the cell's current contribution to a neighbouring cell. It can be achieved by a single transistor biased in the ohmic region [6]. 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, that allow signals V_x and V_w to have either sign— the drain-to-source current is:

$$I_{ds} = -\beta_p V_w V_x - \beta_p V_w \left(V_{x_0} + \left| \hat{V}_{T_p} \right| - V_{w_0} - \frac{V_w}{2} \right) \tag{4}$$

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 cell state. This offset that can be eliminated by a calibration step, with the help of a current memory.

2.2 Current conveyor and level shifting

For the synapse to operate properly, the input node of the CNN core must be kept at constant voltage, independently of what current is entered. This is achieved by a current conveyor (Fig. 2(a)). Any difference between the voltage at node Ⓛ and the reference V_{w_0} is amplified and the negative feedback corrects the deviation. Notice that a voltage offset in the amplifier will result in an error of the same order. An offset cancellation mechanism is provided (Fig. 2(b)). Signal ϕ_{cal} shorts the Operational Transconductance Amplifier (OTA) inputs and enables diode-mode operation of transistor M_{mem}, that will conduce a

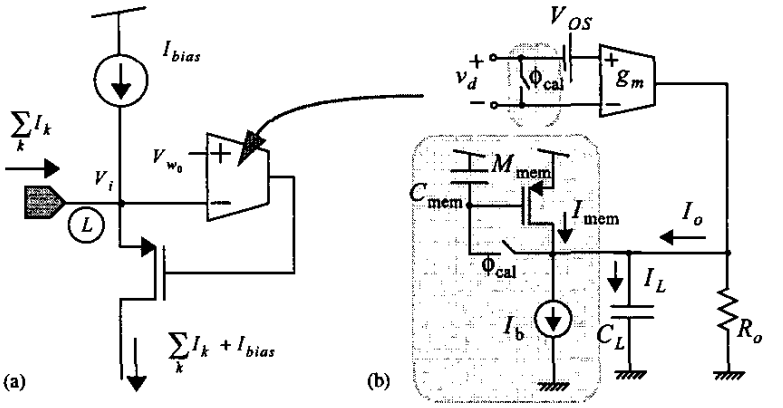


Figure 2. (a) Current conveyor and (b) OTA realization with offset-correction mechanism.

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:

$$I_L = I_o + I_{\text{mem}} - I_{\text{b}} = g_m v_d \tag{5}$$

2.3 *S^3I current memory*

As 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 substracted 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 $0.5\text{LSB} = 2\text{nA}$. 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. 3), 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. The former stages do not contribute to the error in the memorized current. If the S^3I block is designed so as to store the most significant bits in the first capacitor, and the less significant bits in the last one, this error can be made quite small.

2.4 *Time-constant scaling*

The differential equation that governs the evolution of the network, Eq. 1, 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 binary 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 is allowed to be reshaped to a 16-times smaller waveform, then the current memory is obliged to operate over a wider dynamic range. 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. On the contrary, 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 current. Therefore, the adjustable-time-constant CNN core consists in 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. 3. 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 at 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. Without it, the output node, (A), will diverge from the equilibrium.

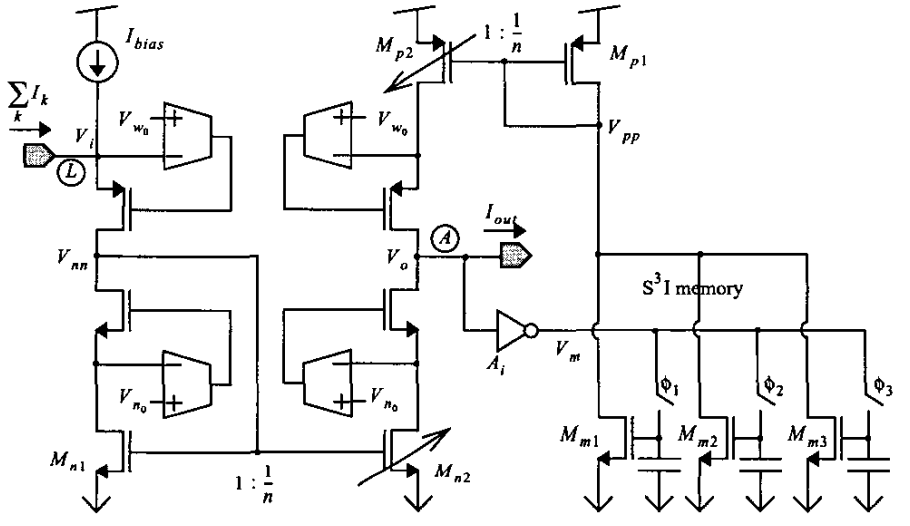


Figure 3. Input block with current scaling, S^3I memory and offset-corrected OTA schematic.

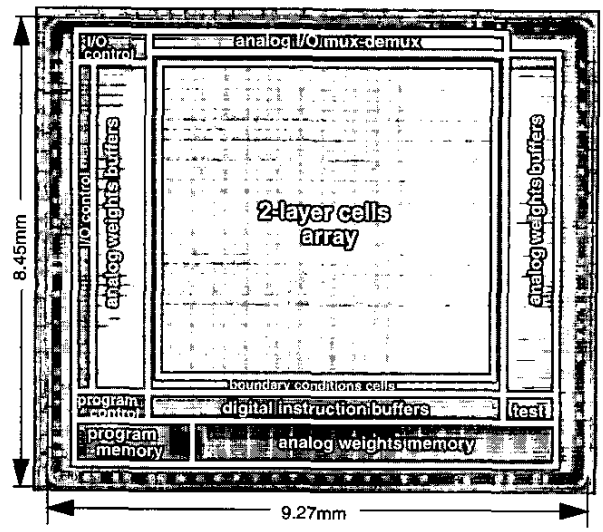


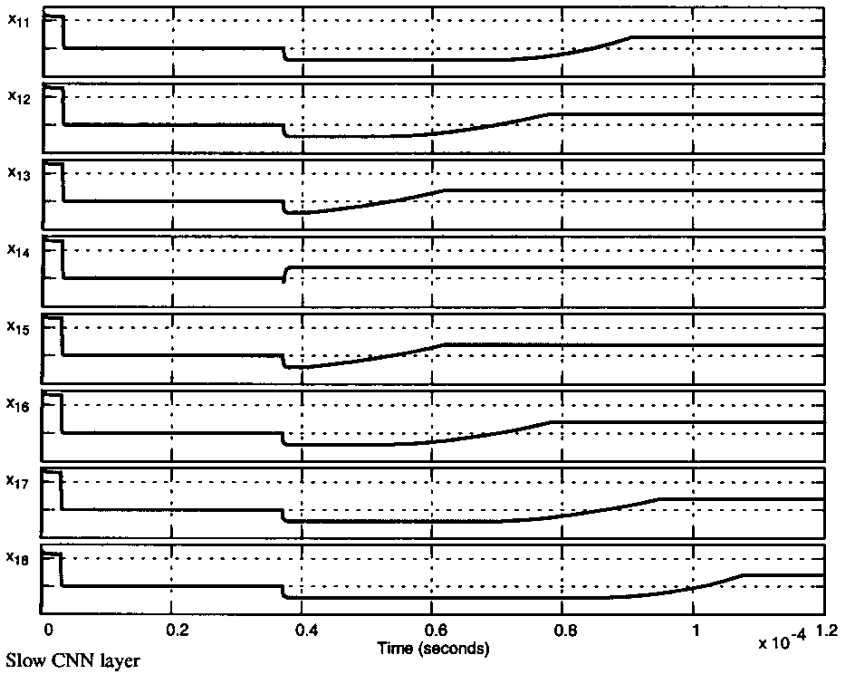
Figure 4. Prototype chip photograph

3 Chip data and simulations

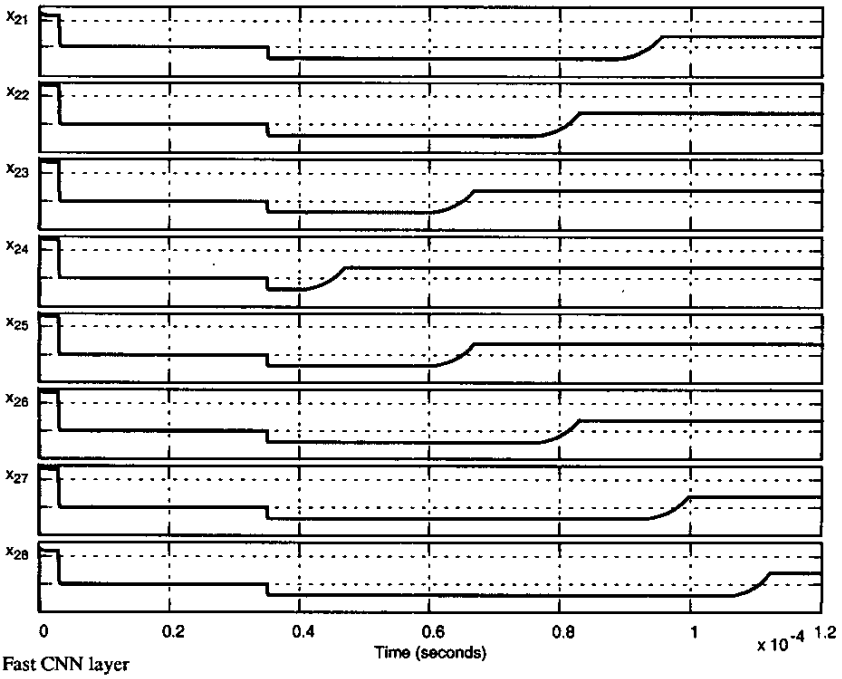
A prototype chip has been designed and fabricated in a $0.5\mu m$ single-poly triple-metal CMOS technology. Its dimensions are 9.27×8.45 sq. mm (photograph in Fig. 4). The cell density achieved is $29.24 cells/mm^2$. The programmable dynamics of the chip permit the observation of different phenomena of the type of propagation of waves, pattern generation, etc. Fig. 5 displays the evolution of the state variable in a reduced network, 1×8 cells, in which the propagation of a wave front in 1-D has been programmed. It is triggered by a marker in the first layer of cell C_{14} and induced in the second layer as can be seen. By controlling the network dynamics and combining the results with the help of the built-in local logic and arithmetic operators, rather involved image processing tasks can be programmed, for instance, grayscale contour detection, skeletonization, etc. [3].

4 Conclusions

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 will outdo its digital counterparts due to the fully parallel nature of the processing, which is, once more, based on the analogy not on the simulation.



(a) Slow CNN layer



(b) Fast CNN layer

Figure 5. 1-D wave propagation.

References

1. F. Werblin, Synaptic Connections, Receptive Fields and Patterns of Activity in the Tiger Salamander Retina, *Inv. Oph. and Vis. Sc.* **32** (3), 459 (1991).
2. F. Werblin, T. Roska and L. O. Chua, The Analogic Cellular Neural Network as a Bionic Eye. *Int. J. Circ. Theor. and App.* **23** (6), 541 (Wiley, Boston, 1995).
3. Cs. Rekeczky, T. Serrano-Gotarredona, T. Roska and A. Rodríguez-Vázquez, A Stored Program 2nd Order/3-Layer Complex Cell CNN-UM. *Proc. 6th Int. W. Cel. Neur. Net. Apps.*, 219 (Catania, 2000).
4. T. Roska and L. O. Chua, The CNN Universal Machine: An Analogic Array Computer. *IEEE Trans. Circ. Syst. II: Anal. Dig. Sign. Proc.*, **40** (3), 163 (1993).
5. S. Espejo, R. Carmona, R. Domínguez-Castro and A. Rodríguez-Vázquez, A VLSI Oriented Continuous-Time CNN Model, *Int. J. Circ. Theor. Apps.*, **24** (3) 341 (Wiley, Boston, 1996)
6. R. Domínguez-Castro, A. Rodríguez-Vázquez, S. Espejo and R. Carmona, Four-Quadrant One-Transistor Synapse for High Density CNN Implementations. *Proc. 5th Int. W. Cel. Neur. Net. and Apps.*, 243 (London, 1998).