Pulse-Width-Modulation (PWM) Alternatives for the Implementation of Programmable Analog Processing Arrays (PAPAs)

Servando Espejo*, Rafael Domínguez-Castro*, Gustavo Liñan*, Ricardo Carmona* and Angel Rodríguez-Vázquez*

Abstract ** **--- This paper presents some alternatives for the implementation of Programmable Analog Processing Arrays (PAPAs) based on Pulse-Width-Modulation (PWM) techniques. PAPAs, when used in the Focal Plane (FPPAPAs) with on-chip photosensors (one per processor) constitute a well defined class of versatile vision chips. Forecasted advantages and drawbacks of PWM alternatives are discussed in connection with the usual trends, obstacles and trade-offs of this kind of systems.**

1 Introduction

It is nowadays generally accepted that massively parallel analog array processors can overcome the limitations of conventional digital processing systems in many vision related applications [1], [2].

The strength of this new approach to real-time image processing relies on the exploitation of image-wide parallelism at the computational level, and also at the sensorial, memorization, and decision taking levels. Concurrent spatial distributions along the chip surface of the circuitry dedicated to these functionalities results in a new class of systems in which image acquisition and preprocessing tasks can be performed at very high speeds and, perhaps even more important, with very low cost, small-size, and power efficient single-chip electronic systems. Many vision applications require processing results in the form of a few data values, as opposed to complete images. This means that data throughput bottlenecks can also be avoided (input bottle-neck is solved by the on-chip photosensors), and therefore, that fully operative, single-chip, end-use oriented vision devices can be produced in the short term. The incorporation of full digital control of the programmable analog array, combined with a simple on-chip digital microcontroller and a reduced amount of memory will result in fully autonomous, low cost vision machines.

This paper presents some basic alternatives to the analog circuit techniques employed in recent prototypes [3] for the realization of the programmable analog convolutors that constitute the fundamental, underlying image-processing operator.

2 Background

The underlying common characteristic of a vast majority of programmable, spatially-uniform***, massively-parallel analog array processors employed in vision applications is the computation, at each elementary unit, of a weighted sum of contributions from a reduced set of neighboring elementary units,

$$y^c = \sum_d w^d x^d \qquad (1)$$

where index d sweeps the prescribed set of neighbor positions, relative to each cell c. Letter x denotes some analog value associated to each element in the array (normally, pixel values of some specific image), letter w the weighting factor associated to the corresponding nearby location, and letter y the process result (which constitutes a new image).

Weighting factors w constitute a reduced set of as many elements as defined by the local connectivity pattern. This set of values applies equally to all elementary units in the array, which is therefore termed as spatially invariant. The values of these weighting factors must be electrically programmable for versatility and to allow the realization of more complex, sequential and/or bifurcation type algorithms.

Processing algorithms differ mainly in the specific nature of some simple dynamic and/or static operator applied to the outcome of the basic local convolution described by (1). Some examples include continuous- or discrete-time integration, and/or a non-linear sigmoid, thresholding, or any other nonlinear static function [7]. Leaving aside this issue, this paper focuses on alternatives for the elec-

* Instituto de Microelectrónica de Sevilla, Centro Nacional de Microelectrónica, IMSE-CNM-CSIC. Avda. Reina Mercedes, s/n. Edif. CICA, 41012 Sevilla, SPAIN. E-mail: espejo@imse.cnm.es, Tel: +34-955-056666, Fax: +34-955-056686

** This work has been partially funded by ONR-NICOP N68171-98-C-9004 and DICTAM IST-1999-19007.

*** Spatial uniformity is a common characteristic of image processing systems, in parallelism with the common time-invariance of systems used to process time-dependent signals.

tronic implementation of the common, underlying processing core: the electrically-programmable convolutor of (1). However, as shall be seen, the inherent use of time increments as a mean of signal representation derived from PWM restricts the use of an (eventual) subsequent dynamic operator to those of discrete-time nature.

In recent prototypes, the convolution operator has been implemented by means of programmable analog synapses (multipliers*) with voltage inputs (V_x and V_w) and current-form output (I_{wx}) [8]. The convenience of this forms of signal representation is clear: easy distribution of weight values to all cells in the array and of pixel value to all synapses in every cell, and easy addition of neighboring contributions at each cell. The output current must then be transformed some how (e.g. by integration in a capacitor) into a voltage form output in order to close the (discrete- or continuous-time) loop, this is, to have both the input and output signals in the same (voltage) representation form, allowing iterative and/or subsequent processing.

This approach has resulted in highly efficient and fully operative systems [3], [4], [5], [6] with new ones currently under manufacturation [9].

This paper presents a class of different alternatives based on Pulse-Width Modulation (PWM), and discusses some forecasted advantages and drawbacks without reaching a final conclusion. A substantial amount of thorough analysis and design effort, as well as some silicon demonstrators, would be needed to unveil the real significance of these alternatives on the basis of a fair comparison with the presently mature approaches. Still, it is hoped that these early-stage proposal will be of interest to the affected community.

3 Pulse-Width Multiplication

The operations reflected in (1) are only *sums* and *scalings* (multiplication). Pulse-width modulation is a well known technique in the electronic design community. It has also been proposed for the implementation of several types of Artificial Neural Networks (ANNs). However, up to our knowledge, it has not been explored yet in the field of programmable vision chips. It can be briefly described in connection with Fig. 1.

The capacitor is first precharged to some reference level V_r. Then, a current is integrated in it during a prescribed time interval:

$$V_o(t) = V_r + \frac{1}{C}\int_{t_0}^{t} I_i(\tau)d\tau \qquad t \in [t_0, t_0 + \Delta t] \qquad (2)$$

* Multipliers and synapses differ in the nature of their input signals. In general, multipliers process two time-variant signals, while synapses have one time-invariant signal (the weight) and one time-variant signal (properly called "the signal").

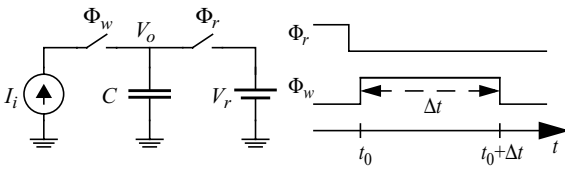

Fig. 1: Basic PWM concept.

Assuming I_i constant during the integration time,

$$\Delta V_o = V_o(t_0 + \Delta t) - V_o(t_0) = \frac{1}{C} \cdot I_i \cdot \Delta t \qquad (3)$$

or, alternatively

$$\Delta Q_c = C\Delta V_o = I_i \cdot \Delta t \qquad (4)$$

Equations (3) and (4) reflect a simple form of achieving multiplication. Summation is also straight forward by accumulating different charge-packages/ voltage-increments in the same capacitor, either simultaneously (hardware replication) or sequentially (time-multiplexing).

4 PAPAs realization: PWM Alternatives

The convenience of having voltage representations of the magnitudes to be multiplied and a current-form output has been mentioned earlier. Neither representation is followed in (3) or (4).

Therefore, we need to represent the variable to be scaled (x) and the scaling factor (w) as a current or as a time increment, which in turn must be defined some how from voltage waveforms.

Table 1 shows the four combination alternatives, which we will now discuss in more detail.

Alt.	I_i	Δt
A	*prop. to* $\boldsymbol{x}$	*prop. to* $\boldsymbol{w}$
B	*prop. to* $\boldsymbol{w}$	*prop. to* $\boldsymbol{x}$
C	*constant*	*prop. to* $\boldsymbol{x \bullet w}$
D	*prop. to* $\boldsymbol{x \bullet w}$	*constant*

Table 1: Alternatives for the obtention of a product w•x based on equation (4).

4.1 Alternative A: $I_i \propto x$, $\Delta t \propto w$

In this alternative, the state variable is represented by a current I_i, while the weight signal is represented by the width Δt of a pulse. The generation of a bipolar current proportional to a voltage-form variable is achieved conceptually with an OTA, as shown in Fig. 2. Voltage waveform pulse' widths, however, are always positive, and therefore, the obtention of bipolar weights requires some additional mechanism. A straight forward approach is to swap the input pins of the OTA, if needed, in order to account for the weight sign. In this manner, the pulse-width represents the absolute value of the weight, while its sign is transmitted by a digital signal controlling which of

the OTA input pins is driven by the signal, the other being driven by the zero-signal reference voltage.

4.2 Alternative B: $I_i \propto w$, $\Delta t \propto x$

In this alternative, the state variable is represented by a pulse width Δt, while the weight signal is represented by a time-invariant current level I_i. Fig. 3 shows a conceptual (one quadrant) implementation. Current source I_B is globally controlled by a voltage node V_B used to achieve a current value proportional to the desired weight (assumed positive). The comparator, driven by a global voltage ramp V_{SR} and by the state variable, generates a pulse with a width proportional to the state variable value (assumed positive). The voltage ramp must start at the reference (zero) level for the state variable.

Again the obtention of a four quadrants behavior presents some practical problems. The generation of a current proportional to a bipolar weight can be achieved by several means, for instance using a differential transconductor, or two current sources (one "source" and one "sink") with only one active at any time depending on the sign of the weight. A more complex problem exists with the representation of a bipolar state variable by means of a pulse width. One alternative is to perform a level-shift in order to obtain unipolar representations of the state variable [10]. Another alternative is the replication of the architecture to obtain a "differential" approach: two ramps (both starting at zero, with same slopes of opposite sign), two comparators, two current sources. The swapping of the slope signs of the two global signals would allow an easy way of implementing the weight sign. Another alternative would be the multiplexation in time of the positive and the negative parts of each contribution.

4.3 Alternative C: I_i independent, $\Delta t \propto w \bullet x$

Alternative C can be considered a variation of alternative B. Indeed, the structure shown in Fig. 3 results in a pulse width inversely proportional to (the absolute value of) the slope of the global ramp voltage. If the current level I_B is left constant (independent of the weight), the weighting factor can be controlled by the slope of $V_{SR}(w)$. In practice, both the current value and the slope could be used for weight control. Alternatives B and C are extreme cases of this general one, with fixed slope and fixed current respectively.

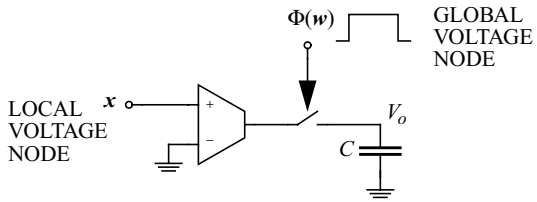

Fig. 2: Conceptual implementation following alternative A

4.4 Alternative D: $I_i \propto w \bullet x$, Δt independent

In a similar form, alternative D can be considered a variation of alternative A (see Fig. 2) in which "the transconductance of the OTA" is controlled by the weight signal. Actually, this means that a (voltage inputs, current output) analog multiplier is needed, the programmable-OTA being just an specific example (as a matter of fact, with significant practical problems [11]). This approach, is specially suited for continuous-time dynamical systems, and is the alternative employed in most reported PAPA implementations (both continuous- and discrete-time). The pulse width is normally not used for multiplication, although it is some times used to stop the (continuous-time) transient after a prescribed amount of time. In fact, this is alternative is not based on PWM. Alternatives A and D could also be considered as extreme cases of a more general one.

5 Advantages and drawbacks

The monolithic implementation of PAPAs has the same general objectives of other electronic devices: accuracy, speed, area efficiency, low power consumption, etc. However, these systems should be composed of as many cells (processors/sensors) as possible in order to achieve image resolutions in the range of common actual standards. This is not easy in general with present manufacturation technologies, although resolutions of 128 x 128 have been obtained [9] and larger ones are likely to be reported in the near future.

In this scenario, the objectives of area efficiency and low power dissipation are specially important, in particular if we have in mind the natural trend to increase the functionality of each array element in terms of processing power, memory capacity, sensor sensibility, and other special functionalities. Therefore, the usual conflict between area efficiency and accuracy of analog systems becomes specially relevant. Fortunately, the accuracy required for most image processing applications is not high in general. Still, substantial design efforts must be made to optimize the different compromises between area and power efficiency, functionality, accuracy and speed.

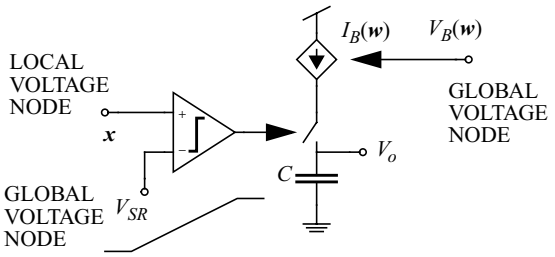

Fig. 3: Conceptual implementation following alternative B

In particular, two aspects seem to dominate the accuracy requirements in this type of systems. One is the proper proportionality of the effective weight values applied to the contributions added at each cell (local proportionality or uniformity). This has a significant effect on the sign of the result of the summation, for any prescribed set of state variable values in the neighborhood. Another is the spatial uniformity of weight values along the cell array (global uniformity), required for an effective spatial uniformity in the processing tasks. Derived from accuracy requisites, it is convenient that null weights be effectively null, this is, if a weight is to be zero, there should be a mean of making it really zero (independently of mismatch or other accuracy degradation sources). These objectives enter in direct conflict with the well-known inverse relationship between mismatch-errors and device area [12].

Dedicated calibration circuitry within the array elements is a natural alternative to overcome the above mentioned limitations. This, however, conflicts with the area efficiency requirements in general, and therefore should be used only when worth it (i.e., at critical points). If the contributions from the different cells in the neighborhood of a given cell could be generated by the same circuit elements (using multiplexation in time), the local uniformity would be warranted. Also, area efficiency could be greatly increased. In addition, since the number of circuit blocks would be small, calibration could be affordable, resulting in global uniformity. The number of global control lines, often a source of area consumption beyond the need of active circuitry area, could also be reduced. Finally, real zero weights would be straight forward by simply skipping the corresponding time slots. This would also result in the corresponding instant-power saving, another typical problem related with the resistivity of large metal lines in large systems with moderate or large static power dissipation. The penalty to be paid is speed, since time multiplexation is required. However, most real application cases employ templates with a relatively high number of zero elements. This means that the speed penalty may not be severe, specially if the approach allows the integration of a larger number of cells as a compensation (thus increasing the system-level effective computation speed).

The PWM alternatives presented in this paper can be considered as initial proposals. It is accepted that these, as any implementation approach, will show practical problems related with the described optimization goals. More specific decisions, like the selection of some point along the hardware/time multiplexation alternatives, the amount and type of calibration circuitry, the mechanisms to achieve four-quadrants behavior, etc., provide a large space for design optimization. It is certainly possible that careful PAPA designs following the proposed alternatives result in some advantages as compared to presently reported devices.

References

[1] Gupta, M.M. and Knopf, G.K. (Ed.): *"Neuro-Vision Systems, Principles and Applications"*. IEEE Press, 1994.

[2] C. Koch, H. Li (Eds.), *Vision Chips, Implementing Vision Algorithms with Analog VLSI Circuits*, IEEE Press, 1995. ISBN: 0-8186-6492-4

[3] G. Liñán, P. Foldesy, S. Espejo, R. Domínguez-Castro and A. Rodríguez-Vázquez. *"A 0.5μm CMOS 106 Transistors Analog Programmable Array Processor for Real-Time Image Processing"*, Proc. of the 25th European Solid-State Circuits Conference, pp. 358-36, Duisburg-Germany, Sept. 1999.

[4] R. Domínguez-Castro, S. Espejo, A. Rodríguez-Vázquez, R. Carmona, P. Foldesy, A. Zarándy, P. Szolgay, T. Sziranyi and T. Roska, *"A 0.8 μm CMOS Programmable Mixed-Signal Focal-Plane Array Processor with On-Chip Binary Imaging and Instructions Storage"*, IEEE Journal of Solid State Circuits, Vol. 32, No. 7, pp. 1013-1026, July 1997.

[5] A. Paasio, A. Dawidziuk, K. Halonen and V. Porra, *"Minimum Size 0.5μm CMOS Programmable 48 x 48 CNN Test Chip"*, Proc. of the 1997 European Conference on Circuit Theory and Design, pp. 154-156, Budapest, Hungary, September 1997.

[6] P. Kinget and M. Steyaert. *"Analog VLSI Integration of Massive Parallel Processing Systems"*, Ed.Kluwer Academic Publishers, 1996.

[7] T. Roska, *"Computer-Sensors: Spatio-Temporal Computers for Analog Array Signals, Dynamically Integrated with Sensors"*. Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, Vol. 23, pp. 221-238, Kluwer Academics November/December 1999.

[8] G. Liñán, R. Domínguez-Castro, S. Espejo and A. Rodríguez-Vázquez, *"Design of a Large-Complexity Analog I/O CNNUC"*. Proc. 1999 Design Automation Day on Cellular Visual Microprocessor, pp. 15-41, Stressa, August 1999.

[9] http://www.imse.cnm.es/Proyectos/dictam

[10] Ari Paasio: *"Integration of Cellular Nonlinear Network Universal Machine"*, Ph.D. Thesis, Helsinki University of Technology, December 1998.

[11] S. Espejo: *"VLSI Design and Modeling of CNNs"*. Ph. Dissertation, University of Sevilla, March 1994.

[12] M.J.M Pelgrom, A.C.J. Duinmaijer and A.P.G. Welbers: *"Matching Properties of MOS Transistors"*. IEEE J. Solid-State Circuits, Vol. 24, pp 1433-1440, October 1989.