Invited Paper

Mixed-Signal CNN Array Chips for Image Processing

A. Rodríguez-Vázquez, S. Espejo, R. Domínguez-Castro, R. Carmona and E. Roca

Dept. of Analog and Mixed-Signal Circuit Design
Instituto de Microelectronica de Sevilla-Universidad de Sevilla
Edificio CICA, C/Tarfia s/n, 41012-Sevilla, SPAIN
FAX:: 34 5 4231832; Phone:: 34 5 3623811; email: angel@cnm.us.es

ABSTRACT

Due to their local connectivity and wide functional capabilities, Cellular Nonlinear Networks (CNN) are excellent candidates for the implementation of image processing algorithms using VLSI analog parallel arrays. However, the design of general purpose, programmable CNN chips with dimensions required for practical applications raises many challenging problems to analog designers. This is basically due to the fact that large silicon area means large development cost, large spatial deviations of design parameters and low production yield. CNN designers must face different issues to keep reasonable enough accuracy level and production yield together with reasonably low development cost in their design of large CNN chips. This paper outlines some of these major issues and their solutions.

1. INTRODUCTION

Vision machines intend to obtain a representation of their environment through the analysis of a flow of optical images. They are targeted to complete in real-time the processing steps required to pass from the constantly changing sensor data to a semantic description of the scene. This is realized through a sequence of spatio-temporal operations where many nonlinear dynamic interactions are made among the bi-dimensional signal values to: first, determine the local image properties (smoothing, thresholding, edges, motion, color, texture, motion, etc.); second, obtain the generic scene attributes (boundaries, regions, surfaces, objects, etc.); and, third, build a semantic description from these attributes. It results in a continuous compression of the information, from the raw bi-dimensional input signals to the symbolic representation of the scene [1].

Traditional vision machines use a CCD camera for parallel acquisition of the input image, and serial transmission of a digitalized version of the input data to a separate computer. It results in huge data rates which conventional computers are not capable to analyze in real-time. For instance, a 3-colour@512 x 512 pixel camera delivers about F million byte/second, where F is the frame rate. Conventional computers and DSP´s are able to manage such a huge rate for auto-focus, image stabilization, control of the luminance/chrominance, etc. However, real-time completion of the spatio-temporal operators required for understanding images requires much more sophisticated digital processors. Consequently, traditional vision machines with real-time capabilities are bulky, expensive and extremely power-hungry. This is in contrast to living beings, where even very tiny and power-efficient brains (such as that of insects) are capable to analyze complex time-varying scenes in real-time.

This contrast between the performance of artificial vision machines and "natural" vision system is due to the inherent *parallelism* of the former. In particular, the retina (the "camera" of human vision system) combines image sensing and parallel processing to reduce the amount of data transmitted for subsequent processing by the following stages of the human vision system. Based on this, universities and industries have focused their efforts on the development of new generations of vision artifacts capable to overcome the drawbacks of traditional ones through the incorporation of distributed parallel processing, and by making this processing to act concurrently with the signal acquisition. One possible strategy uses flip-chip bonding of physically isolated sensing and processing devices [2] . Other possibility is to incorporate the sensory and the processing circuitry on the same semiconductor substrate. Silicon retinas, smart-pixel chips and focal plane arrays are members of this class of vision chips [3]. CMOS technologies offer unique features for their development due to the availability of good phototransduction devices and the possibility to realize a large catalog of linear and nonlinear processing functions by using the different operating regions of the MOS transistor.

Previously reported CMOS vision chips are typically intended for fixed function. However, industrial applications demand chips capable of flexible operation, with programmable features and standard interfacing to conventional equipment. Some efforts in the direction of building these programmable devices have been reported elsewhere [4,5,6] . Also, a powerful methodological framework for their systematic development has been set-up recently by researchers of the University of California at Berkeley and the Hungarian Academy o Sciences[7,8] . It is based on the observation that most image processing operations can be formulated as well-defined tasks on signal values placed over regular 2-D spatial distributions, and with direct interactions among signals limited to local receptive fields. Consequently, they are directly mappable onto *C*ellular *N*onlinear (or Neural) *N*etworks (*CNN*s), which are *arrays* of *nonlinear dynamic analog* processing units (cells), arranged on regular grids where direct interactions among cells are limited to finite *local* neighborhoods. By enabling these interactions to be programmable, and incorporating the possibility to sequentially realize a stored program over an 2x*N-D* image memory, the CNN have evolved into the CNN *U*niversal *M*achine (CNN-UM) [8] . Since CNNs realize the vast majority of image processing tasks through proper selection of the interaction strengths and/or task sequencing, the CNN-UM can be considered as a general-purpose image processing computer on a chip (or chip set).

2. THE CNN PARADIGM

CNNs are [7]:

- multi-dimensional arrays defined on a grid and composed of,
- mainly identical nonlinear, dynamical processing units (cells), one per grid vertex, which satisfy two properties:
- all significant variables are continuous-valued, and,
- physical interconnections among cells are mostly local, i.e. most cells are physically connected only to other located within finite radii of the grid.

The operation of CNNs becomes described by using three variables per cell:

- Cell *state*: x_c, which conveys cell energy information as a function of time.
- Cell *output*: $y_c(t)$, obtained from the cell state through a nonlinear transformation,

$$y = f(x_c) \tag{1}$$

Fig.1(a) shows some useful nonlinear output functions [7].

- Cell *input*: u_c , representing external excitations.

CNNs are multidimensional signal processing devices whose inputs are the input vector $\mathbf{u} = \{u_c, \ \forall c\}$ and the vector of initial states $\mathbf{x}(0) = \{x_c(0), \ \forall c\}$, and whose outcome is represented by the vector of output variables $\mathbf{y} = \{y_c, \ \forall c\}$. For given input and topology, the processing performed by CNNs is determined by the following:

- An *evolution law,* described by ordinary differential equations (ODEs), finite-difference equations (FDEs), or a mixture of both [7,9]. For instance, in a case where ODEs are involved,

$$\tau_c \frac{dx^c}{dt} = -g\left[x^c(t)\right] + d_c + \sum_{d \in N_r(c)} \{a_{cd}(y_d, t) + b_{cd}(u_d, t)\} \qquad \forall c \tag{2}$$

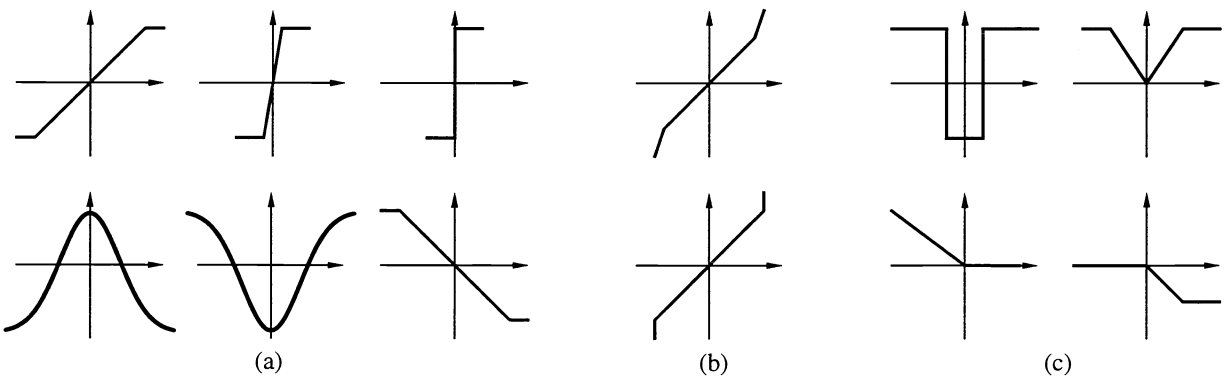

Fig. 1: Some Common nonlinearities found in CNNs: a) output funstion, b) dissipative term, c) templates.

where $N_r(c)$ denotes the neighborhood of the cell. Similar to the output function, the dissipative term above can be nonlinear [10]. Fig.1(b) shows some typical nonlinear dissipative terms.

- The states, $\mathbf{x}_s$, and inputs, $\mathbf{u}_s$, of the cells in the grid surrounding.

The design parameters that control the operation of CNNs are the shapes of the dissipative and the output functions (i.e. $f(.)$ [7] and $g(.)$ [10]), the offset values, and the nature and values of the interactions within each cell neighborhood, which are represented by the feedback ($a_{cd}(.)$) and control ($b_{cd}(.)$) contributions. In the more general case, these latter may be nonlinear. Fig.1(c) shows some common nonlinearities typically encountered at the interactions among cells [11].

An important class of CNNs have all inner cell identical, arranged on a 2-D grid, and with the same pattern of interactions for all of them (*translation-invariant*) . A subclass which is of special interest for VLSI realization has *linear, time-invariant* interactions among cells. Their inner control and feedback parameters are univocally represented by two matrices, called respectively control (**B**) and feedback (**A**). They are *templates* which repeat across the net, and provide direct insight into the topology and the strength of the cell interactions.

The input-output mapping performed by these linear, time- and translation-invariant CNNs is basically determined by the offset term and the feedback and control templates -- up to $2(2r + 1)^2 + 1$ different real parameters, where r denotes the radius of the neighborhood. This limitation in the number of parameters to be realized and controlled enables larger number of cells per unit silicon area (cell density). Also, although the computational power increases with r, the use of small values of this parameter does not constraint the functional capabilities of the CNN-UM. This is due to the propagation phenomena, which provoke interaction among cells located far away in the network. Thus, the CNN-UM compensates the use of small values of r (in particular, $r = 1$) through proper programing [8,11].

The design of general purpose, programmable CNN chips with dimensions required for practical applications raises many challenging problems to analog designers. This is basically due to the fact that large silicon area means large development cost, large spatial deviations of design parameters and low production yield. CNN designers must face different issues to keep reasonable enough accuracy level and production yield together with reasonably low development cost in their design of large CNN chips. Some of the major issues are outlined below.

3. AREA

Most practical CNN applications require relatively large number of cells (from 10^3 to 10^5). To keep the chip area reasonably low, large design expertise and optimization is required at the cell and the system levels. In any case, however, total silicon area will remain large. Large chip areas have several dramatic consequences. Process yield decreases exponentially with system area [18, 20] . Although the cost of silicon is, a priori, linear with area, the exponential yield reduction results in an exponential increase of the cost per good sample. Large chips are also affected by variations of the process parameters over the chip area (long distance mismatch [17]), which represent a challenge for the design of what are assumed to be spatially-regular processing systems. Temperature gradients, originated either by periphery miscellaneous on-chip circuitry, or simply by the die boundary conditions, represent an additional contribution to spatial variations on the behavior of elementary processing units. Cell area minimization has several consequences as well, first of which can be an additional reduction on process yield originated by compact layout styles. Also, short-distance matching among devices is proportional to device areas, which means that small devices exhibit high mismatch levels. Power supply lines should be wide enough to prevent electromigration phenomena [19]. Also, due to metal resistivity, even moderate current densities result in non flat power supply levels over the cell array. This is particularly critical if reference signals are defined with respect to power supply levels (low PSRR). Scaling down the technology does not directly imply a proportional reduction in the cell area. This is again due to several reasons: loss of accuracy of small devices, larger current densities on narrow lines, increased resistances, etc. These effects are particularly relevant on analog circuitry. An additional problem for circuits with optical input capability originates from the direct relationship between sensor sensitivity and its area.

4. YIELD

Large silicon area means large number of faults and, hence, low process yield. Some faults result in severe malfunctioning (catastrophic faults: i.e. shorts and open circuits). Others result in deviations of some specified behavior from the allowed tolerance windows (parametric faults: e.g. accuracy or speed) [21]. The yield factor due to catastrophic faults is inversely exponential with total system area, while the factor due to parametric faults decreases with the inverse of the square root of individual devices area. Many additional factors influence the final yield figure, like layout style, circuit architecture and topology, human manipulation, etc. The many factors involved, and the almost general lack of technology characterization in this matters, make it very difficult to predict yield figures, even with moderate accuracy. In CNN chips, yield is always a crucial factor with drastic consequences on the cost and even on the viability of the design for a particular technology. For this reason, large efforts dedicated to yield characterization and to the study and validation of layout techniques oriented to catastrophic-yield improvement should be undertaken. Although in the last years new specific CAD tools have evolved in this field [23], the obtention of technology-specific parameters is in general tremendously expensive in terms of manufacturing costs and iteration time. The lack of technology characterization is particularly frequent in the newest, submicron processes.

5. POWER DISSIPATION

Together with process yield, power dissipation is one of the fundamental limits for the realization of high complexity monolithic systems. Large power dissipation results in large chip temperature. Depending on packaging, environment conditions and other factors, a hard limit for power dissipation exists beyond which system failure and/or permanent damage will occur. Below this limit, high temperatures affect the behavior of the circuitry. Homogeneous temperature distribution may be easy to consider at the design stage. However this requires uniform power dissipation. Boundary conditions can also affect temperature uniformity. This is particularly true for large area chips operating at temperatures highly above the environment. The relationship between power dissipation and chip temperature is not trivial, an depends on factors like the package type, system board, external ventilation, environment, etc. Problems associated with power distribution within the chip result in additional effects on accuracy and area figures. Metal lines resistivity affect the actual supply voltages at the cell locations, resulting in general in accuracy degradation. In some cases, the use of voltage-regulators within the cell circuitry would solve the associated accuracy problem, at the expense of additional dissipation and area. Proper circuit techniques (high PSRR circuits) can also attenuate this effect. An additional limit, with significant effects on area efficiency, is imposed by metal electromigration effects, which forces the use of wide power supply lines. For similar reasons, a high number of power bonding pads is needed in general, and they should be distributed uniformly throughout the perimeter of the system (assuming uniform power dissipation). Substantial reduction on power dissipation has the general drawback of accuracy and speed degradation. As usual, different design techniques and device primitives have drastic effects on the final compromise. For instance, the use of weak inversion on MOS devices allows extremely low power dissipations, at the expense of high local mismatch and low speeds. The simultaneous availability of MOS and bipolar transistors (in BiCMOS technologies) improves the accuracy problem, leaving only the speed penalization. Some increase in production costs should be expected in this case. The use of reduced power supply voltages, decreases the power consumption of the system. The decrease is linear with voltage reduction (for constant current), and thus, drastic improvements are not possible. Also, analog circuits accuracy is generally deteriorated by voltage range reductions.

6. ACCURACY

Accuracy specification of analog circuits is generally very dependent on the system and on the application. This is particularly true in the case of CNNs. Any real circuit used for the CNN cells can only be roughly modeled by the aimed CNN cell equation. Many additional parameters are generally needed, each of them with its corresponding deviation from nominal values (both systematic and random). Also, spurious dependencies of parameters (e.g. weights) and functions (e.g. the output non-linearity) on other variables (e.g. state variables, spatial location, time, etc.) should be taken into account. The high complexity and the dynamic nonlinear nature of CNNs makes it difficult to evaluate the permissible tolerances of each parameter and function. The problem is further increased on programmable networks, for which no predefined application is determined. A widely accepted approach to accuracy specification relays on the following simplification: Assume first that each individual cell behaves in fact like a first order dynamical nonlinear system, and proceed to characterize the static accuracy of the involved parameters and functions. Then, evaluate the dynamic dependencies of the previous parameters and functions on the dominant state variable (ideally the only one) and on other parasitic state variables, for which additional first-order nonlinear differential equations must be formulated. Usually, and due to the high complexity of the dynamic characterization, which must be built over the already complex static analysis, only the first step is undertaken. A generally accepted figure for maximum static accuracy of un-calibrated, typical analog circuits is about 9 bits [18, 17]. However, the obtention of such resolution levels requires large areas and moderate power consumptions, both of which are prohibitive in CNN design. For this reason, a lower figure of about 6 to 8 bits is commonly considered a good level of accuracy for analog CNN chips. This is still difficult to achieve.

6.1. Static Accuracy

Static accuracy is limited by a number of factors, and affects to parameters like template values, time constants, offset levels, nonlinear functions, initial state and input values, boundary conditions, etc. As in any other circuit, non-ideal deterministic properties of the employed devices (e.g. Early effect, parasitic resistances, leakage currents, etc.) establish a limit to the accuracy on the emulation of the CNN algorithm. The limit imposed by these systematic errors is generally easy to predict from commonly available technology data, (e.g. transistor models, sheet-resistances, etc.), and can be minimized by proper design techniques. Process parameter windows, usually provided by the manufacturer, must be considered during the minimization of systematic errors in order to ensure appropriate behavior even for worst-case process conditions. In many cases, a careful internal reference level generation is needed to ensure proper operation of all devices under large process parameter variations. Another problem is due to the unavoidable spatial variation of process parameters within the chip. These can be decomposed into long-distance and a short-distance components. The former is due to process gradients [22], which introduce a slowly-varying spatial-dependency on local parameter values. Although the phenomena is deterministic in nature, the exact value of the spatial function is usually unpredictable and variable from chip-sample to chip sample, since it depends, among many other things, on the location and orientation of the particular chip-sample within the wafer. For this reason, this component is commonly modeled as a random, time-invariant, low spatial-frequency function (wavelength much larger that typical device sizes) [17]. In general, this has the effect of smooth variations of the otherwise uniform cell parameters along the cell array. Additional sources of smooth spatial variations are found in temperature gradients and effective power supply values. Temperature variations can be originated by heat dissipation of peripheral on-chip circuits (placed around the cell array area for miscellaneous purposes), or simply by the boundary conditions of the die and the thermal properties of the package. Effective power supply variations are generally due to the non null resistivity of metal lines used for power distribution within the chip. Both error sources are particularly important in large systems of even moderate dissipation levels. Circuit and layout techniques can be used to attenuate or even eliminate most sources of smooth spatial variations. The short-distance component of process parameter variations represents a more serious challenge. These can be modeled as random, time-invariant, high spatial-frequency function (wave lengths much shorter that typical device sizes). The effect of these component is a random contribution to key electrical parameters of every individual device. It can be seen [17] that standard deviations are inversely related with device sizes (e.g. area, width, length), which means that small "identical" devices present significant random differences in practice. Clearly, this affects the accuracy of cell parameters, introducing a random spatial component which is added to the previous smooth dependency. This effect, usually known as local mismatch, represents one of the fundamental limits for both accuracy and cell density in CNN chips. It can also be seen that local mismatch in MOS circuits deteriorates (in general) as bias currents are decreased. This results in a compromise with the power reduction objective. Final accuracy figures are strongly dependent on circuit techniques (e.g. gm-C, MOSFET-C, etc.) [14], basic building blocks and global architectures, device operation regions (e.g. in MOS devices, weak or strong inversion and triode or pinch-off regions) [24], available design primitives (e.g. MOSFETs, BJTs, linear/floating/accurate capacitors, high resistivity layers, etc.) [13, 12], and of course, on area and power efficiency.

6.2. Dynamic Accuracy

As stated previously, dynamic accuracy is much harder to characterize, mainly due to the nonlinear nature of the system and its large complexity. A general rule of thumb is that every circuit block should be "much faster" than the integrator block (dominant time constant). However, the dynamic characterization of nonlinear circuits is not an easy matter, and most modeling attempts result in high-order nonlinear differential equations, even for very simple circuit blocks. For this reason, the traditional practice reduces to measuring some delays and frequency responses under very specific situations for each particular circuit block (e.g. nonlinear operators, switches, current mirrors and sources, multipliers, multiplexers, etc.). Note that simple circuit blocks commonly used in linear applications, when employed in CNN circuitry, operate both in their linear and nonlinear regions at different time instants. Dynamic inaccuracy is affected by a large number of factors, most of them associated with the unavoidable existence of additional state variables within every cell due to parasitic capacitors. Among these factors, charge redistribution due to turn-on of the analog switches (e.g. initialization) and feedthrough [26] errors due to switches turn-off represent an important error source, in particular under strong area constraints (small capacitors). Mismatch (both long and short-range) produces variations in dynamically relevant components. Long-range variations have the effect of smooth spatial variations of time constants (as discussed previously), while local mismatch has effects on cell dynamics, in particular in fully-differential architectures, in which dynamical dissymmetries introduce additional parasitic state variables. A different class of error sources is related to the DC stability of power supplies and intermediate reference levels. These global metal lines, in large area systems, suffer appreciable capacitive coupling from global control lines, which together with the average large distances to source locations result in general in appreciable variations on what are assumed to be DC voltage levels. Power supply lines suffer as well from current spikes produced by local digital circuitry (the usual separation of analog and digital power supply lines is costly in terms of area consumption). These problems are increased by the area efficiency constraint. For instance, the use of large capacitors (relative to parasitics) to get a clearly dominant time constant is tremendously expensive, and hence, regardless the speed (also related to power consumption), parasitic and nominal dynamics are generally difficult to separate. Also, because low current levels deteriorate the static accuracy, high speed (several MHz dominant frequency) is generally a must. An important additional consideration is needed when a multi-chip-modules architecture is aimed. In this case, external capacitive loads introduce external time constants that may be drastically different from internal ones. This problem must generally be addressed using high-speed buffering techniques. The high number of interconnections among chips requires a high number of buffers, and thus, large area and power consumptions.

7. SPEED

As already discussed in previous sections, speed is highly correlated to power consumption, area dynamic accuracy. These correlations highly dependent on the circuit primitives available in the technology (e.g. MOSFETs and/or BJTs). An important consideration is whether the speed of the analog CNN process is really significative within the global processing throughput of the system: in many cases, the throughput bottle-neck will be defined by the image I/O processes, rather than by the processing time itself. I/O throughput is also compromised with area and power efficiency, and in the case of analog images (e.g. gray scale), with dynamic and static accuracy. Multi-chip modules design presents additional challenges regarding speed due to the dynamic effects of external connections.

8. POWER SUPPLY LEVELS

Nowadays, the 5V power supply range standard is rapidly becoming obsolete, in particular for high complexity digital systems (e.g. CPUs). A new standard of 3.3V is widely extended in commercial products and lower ranges are gaining acceptance [27]. This tendency is basically imposed by fundamental limits of device physics, which have become relevant on modern scaled down technologies [28]. Secondary arguments are the associated power reduction and speed improvement (this last effect is due to technology scaling). Technology scaling down (and the necessary supply reduction) is mainly driven by the digital circuitry. Analog circuit design, although adversely affected, must find solutions within the new environment due to several reasons, among which is its smaller relevance (in terms of area occupation) in the increasingly important field of mixed-signal applications. In the case of CNN chips, the use of scaled down technologies is clearly convenient to area efficiency and cell-count maximization, although the improvement is not linear with scaling due many reasons, some of which have already been commented (e.g. local mismatch dependency with device area). Supply range reduction directly affects the available voltage swing of analog signals, and thereof, their total signal to noise ratio [27]. The effect is less clear on current-mode analog signals, due to the nonlinear relationship between voltages and current in active devices. Effects are also drastically different depending on particular design styles and employed circuit primitives (e.g. strongly and weakly inverted MOSFETS, BJTs, etc.). As in many other aspects, circuit techniques appear as a crucial factor on the final performance of analog circuits under scaled down voltage supplies.

9. INTERFACING

Control-specific signals are normally digital (binary), wile weight programing require, in principle, the specification of values within some continuous range (using analog signals). The desired weight range can also be discretized, resulting in a digital representation of weight values. Both approaches present advantages and drawbacks with drastic effects on the final performance of the system. A hybrid approach (externally digital, internally analog), which combines the best of the two alternatives has been described in [29]. This hybrid alternative results in robust behavior, ease of program storage, and simple external control by means of conventional computers. At the same time, it allow the use of analog synapses within the cells, and global analog weight signals throughout the cell array, which results in higher area and power efficiencies.

The combination of local sensing and processing capabilities over large array of identical elementary units is cornerstone for vision CNM-UM ships. Two problems arise in connections with this strategy. First, the realization of high performance photosensors (fast, accurate, high sensitivity, etc.) and their interface with the local processing circuitry requires in general large areas, in particular when analog input image area to be processed. For instance, the sensitivity of typical photoactive elements in CMOS processes (diodes, vertical BJTs,...) is directly related to their areas, and thus, technology scaling will not have the expect area reductions effect on this part of the circuitry. Second, if special protective measures are not taken, image projection over the surface of the chip results in undesired effects of light on the processing circuitry, including drastic increase of leakage currents, and additional short and long-range time-variant mismatch due to light-intensity variations over the surface of the chip.

10. CIRCUIT TECHNIQUES

Digital CMOS ICs use NMOS and PMOS transistors as the only primitives, and exploit mostly their operation as switches. On the contrary, analog CMOS ICs, analog CMOS CNN-UM chips, uses a much larger set of CMOS primitive components:

- CMOS-BJT transistors: vertical and lateral [12].
- Passive components: resistors and capacitors [13,14].
- CMOS photodevices [15,16].

Although the reduced number of available components is a drawback for IC design, this can be palliated with the full exploitation of the functional features of the primitives. In particular, the MOS transistor can be used as a voltage-controlled transconductor, a voltage-controlled resistor, an analog switch, a capacitor, a current rectifier, etc. [14, 25]. Obviously, proper exploitation of these functional features require detailed knowledge of the intrinsic limitations of the primitives. Some general considerations for analog IC design include the following:

- Reduced spreading. It is a consequence of the fact that monolithic element values are determined by sizes and shapes of corresponding physical components, so that, as a general rule of thumb, large element values implies large area occupation and large parasitics. Consequently, the spreading of component values must be kept as small as possible to reduce total area occupation and equalize parasitics.
- Poor absolute accuracy. Typical tolerance margins are about 20%, with large influence of temperature and aging, and in cases, large nonlinearities.
- Good relative accuracy. Tolerance of ratios between similar components are much smaller (as low as 0.1% [13]) and the same is encountered concerning on the influence of temperature, aging, or even nonlinearities (in analog circuit design there are many examples of nonlinearity cancellation in a pair of highly nonlinear components, for instance in a current mirror). Consequently, monolithic circuit design techniques should rely on component ratios (capacitor, resistor, or transistor ratios), instead of on absolute component values, and should try to exploit simple schemes for nonlinearity cancellation, or the tuning of absolute component values. It drives the last consideration to be made here,
- Matching depends on geometry and biasing. As a general rule of thumb, matching increases by increasing component sizes, decreasing component distance and by making ratioed components have homogeneous shapes [17]. On the other hand, the matching of MOS transistors degrades as their drain current decreases.

Within the general framework and constraints imposed by technological limits and design specifications, the designer must choose the proper circuit design technique. Different design styles have, in general, drastic effects on final performance and complexity figures. The selection of the particular circuit technique must be basically oriented to the achievement of high area efficiency, i.e., to the maximization of the number of cells in a single chip, while simultaneously maintaining the required system specifications. At the same time, the many challenges, problems, and technological constraints described below in this paper must be taken into account at the design stage, and when possible, specific circuit techniques should be used to overcome their effects (e.g. internal supply voltage regulators, self-tuning and/or self-calibration schemes, low mismatch-sensitivity circuit blocks, adaptive reference level generation, etc.). Typical design alternatives include whether to use or not weakly inverted MOS transistors, to select a BiCMOS technology, current or voltage (or hybrid) signal representation, triode or pinch-of operation region of synapse transistors, etc.Although the general design flow goes from the algorithmic to the circuit level, the achievement of maximal efficiency justifies, in some cases, the realization of slight modifications at the algorithmic level with appreciable effects on implementation efficiency, as long as functionality is not substantially affected [9].Technological characteristics (available circuit primitives, mismatch figures, power supply range, etc...) have a strong influence on the final style selection. Thus, technology rticular, selection is a previous critical choice. In pyield and mismatch characterization data (seldom completely available to the designer), are needed at the design stage if the many challenges are to be undertaken with some probability of success. This means that, in many cases, previous technology characterization will be needed, when pushing the limits of the designs. Technology scaling down, although clearly advantageous, should not be foreseen as a proportional reduction in system area (alternatively an increase in system complexity) for many reasons already commented (direct dependency with area of some crucial properties like, matching, maximum current density, resistivity, photosensors sensitivity and accuracy, etc.).

11. CAD

It is a well known fact that analog VLSI CAD tools are well behind their digital counterparts. The design of large CNN chips is also affected by the lack of advanced CAD tools, specially concerning accuracy an yield prediction. Both problems, although common to other analog VLSI systems, are increased in the case of CNNs due to the nonlinear nature of the processing, and to the high area efficiency requirement (an thus usual compact layout style). Future developments in these fields will probably reduce the design time (and thus the final cost), as well as improve the final performance.

12. MULTI-CHIP MODULES

Many fundamental limits, like yield and power dissipation, can be avoided by designing smaller, cheaper and more accurate chips capable of being interconnected with other identical units to compound a highly complex system. This however presents some additional challenges, like the high number of bonding pads required (at least two per border cell), speed and dynamic accuracy problems due to the influence of external loads on border cells, and the need of complex optical systems if optical inputs are to be used. Accuracy can also be affected if special tuning schemes are not used to ensure proper static an dynamic matching among different chips.

13. A 0.8μm CMOS CNN-UM VISION CHIP

We present here a 0.8μm CMOS CNN-UM vision-chip which has been designed following previous guidelines. To the best of our knowledge, this chip is the first fully operational CNN vision-chip reported in literature which combines the capabilities of image-transduction, programmable image-processing and algorithmic control on a common silicon substrate.

13.1. System Architecture and Functionality

Architecture: The prototype contains 20 x 22 identical cells arranged in a a rectangular grid, each of them with local transduction, processing, control, and storage capabilities. In addition, global control, interfacing and storage circuitry is placed on the surrounding of the cell array, yielding a total silicon area of 30mm^2, the maximum allowed by the foundry in MPW runs.

Transduction: Image acquisition relies on photogenerated currents at floating-base vertical BJTs, available on standard CMOS technologies. An automatic adaptive scheme is used to ensure appropriate contrast levels by shifting the observed scene to obtain a zero-mean distribution of pixel values. Optional external circuitry can be employed to adjust the mean of the distribution when needed, for instance for highly regular images with dominant background.

Control and storage: The basic functions of image acquisition and processing are complemented in this prototype with storage, algorithmic control, and programmable boolean operations among images. Four images can be stored on-chip. Individual pixels are physically located within corresponding cells, facilitating parallel data-transference tasks. Image-memories can be loaded from the image-sensor, the output of the neural network, or the output of the programmable boolean operator. Any stored image can be used as any of the two input images of the network, or as input to the boolean operator. Any image can also be loaded or downloaded, on a row by row basis, through an external I/O bidirectional bus.

Global control and interface: Eight complete sets of CNN coefficients can be stored on chip. Although their internal control is analog, they are discretized and stored in digital form, with a resolution of 7bits+sign, more than enough for the expected accuracy of the analog processing circuitry [29]. These digital values are internally transformed into analog synapse-control signals by adaptive loops comprising linear D/A converters and a synapse identical to those used within the cells. This strategy results in simple and robust external control, virtually independent of the detailed relationship between the control-signal and the actual weight programmed in the synapses. The internal storage of images and processing coefficients, together with the possibility of using them in any order and any number of times, yields a highly flexible system usable for relatively complex and generic image-processing tasks including sequential and bifurcated-flow algorithms. Also, although the fundamental processing function is analog, the digital nature of the interface makes the prototype extremely easy to control with conventional computing systems.

13.2. Circuit Implementation

The major trend in the design has been the maximization of the number of cells in the array under the area limitation imposed by the foundry (30mm^2), while maintaining a reasonable degree of accuracy in the analog operations. The involved area-accuracy trade-off has been addressed through intensive structural and parametric optimization . The prototype contains 20 x 22 cells, each with an area of 180 x 180 μm^2; it means a cell density of 31 cells/mm^2.

Image Acquisition Circuitry: This is identical to that previously described by the authors in [30] for fixed-function CNN CMOS chips. Pixel sensors consist of two vertical BJTs arranged in Darlington configuration, as shown in Fig.2a. The photogenerated current at the base-collector junction (n-well/p-substrate) of transistor Q_1 is amplified by a factor $(\beta_F+1)^2$ by Q_1 and Q_2, yielding output current levels of about 0.8μA under an environmental laboratory lighting of 0.9W/m^2. The acquired gray-scale image is converted to binary by comparison to the spatial average of the image, thus ensuring proper contrast adjustment over a wide range of illumination conditions. The averaging and comparison circuitry, illustrated in Fig.2b, is included in every cell and globally interconnected through a common node SUM. A CMOS inverter transforms the shifted output current to digital levels, which can then be stored at internal memories. The area of the imaging circuitry (sensor+regulation) amounts up to 7% of the cell area.

CNN Processing Circuitry: Its fundamental building blocks are programmable synapses and nonlinear integrators (integrators with saturation). Fig.3a shows the structure selected for the former, consisting of a linear multiplier core with transistors operating in ohmic region [14, 25], and two source-follower buffers. Its selection is based on exhaustive analysis of the area-accuracy trade-off in alternative CMOS synapses [14]. Note that synapse inputs are voltages, which eases the intra-chip distribution of template coefficients, and the intra-cell distribution of the state-variable to all the cell synapses. On the other hand, the synapse output is a current, thus facilitating the summation of different contributions at integrators' input nodes. The nonlinear behavior of the output with respect to the differential weight signal $V_{wp}-V_{wn}$, mainly due to the resistively-loaded source-followers is not important, since it is taken into account by the weight-control adaptive circuitry. On the other hand, the behavior of the structure with respect to the state-variable signal $V_{xp}-V_{xn}$ is, as required, highly linear.

The differential nature of the synapse signals, and the necessity to reduce common-mode parasitics led us to the fully-differential integrator architecture of Fig.3b, realized with two current-conveyors and two grounded capacitors. The latter are realized through the gate capacitance of the MOS transistors tied to the state-variable terminals of the neuron synapses, which are indeed the natural load for the current conveyors in an integrator configuration. Integrator saturation is achieved by a nonlinear resistor with a sharp voltage-limiting characteristic, realized with two diode-connected transistors as shown in Fig.3c, and connected between the differential-rail signals.

Local Logic and Control Circuitry: The realization of the programmable boolean operator and the control circuitry is based on switches and conventional digital circuitry. The 4bit memory, based on charge storage, employs metal-1 shields over sensitive areas to avoid the adverse effect of light on reverse diode-currents, which could result in a significative reduction of storage-time. About 30% of the cell area is dedicated to these digital capabilities.

Interfacing Circuitry: About 50% of the prototype area is dedicated to miscellaneous circuitry placed at the periphery of the cell array. This includes the weight-control stages used to generate analog programming signals from their digitally stored values, boundary cells employed to establish spatial boundary conditions, biassing circuitry, and bonding pads. Also, some conventional digital blocks are needed for the control of the loading and downloading processes of images and CNN coefficients.

13.3. Experimental Results

The prototype, whose photograph is shown in Fig.4, has been designed and manufactured in a standard, digitally oriented, two-metals, one-poly, n-well, 0.8μm CMOS technology available through the EUROCHIP consortium. Two additional smaller chips with analog and digital parts were fabricated for testing and characterization purposes. The total area of the prototype is 30mm^2, the time constant of the analog network is 0.25μs, and the operation speed of the digital circuitry is 10MHz.

Global Performance: The prototype has been globally tested and its functionality verified by the authors and by an independent research group [31]. It has been successfully used for such diverse functions as low-pass image filtering, corners and borders extraction, hole filling, motion detection, and many other CNN applications reported in literature. Task sequencing and algorithmic control has been also verified, in particular for texture-detection applications. While the functionality of the prototype has been completely verified, the accuracy of the analog circuitry (around 6 bits) is slightly under the objective (7 bits). The source of this degradation has already been identified and could be easily avoided with minor changes in the design.

Basic Blocks Performance: most of the basic analog blocks have been exhaustively tested and characterized. Fig.5a shows the response of the programmable synapse for different weight values. Integral linearity is better than 0.4% and total harmonic distortion of 0.2% within the required state-variable signal-range. Statistical characterization, based on 10 samples located on different chip-units, yield a standard deviation of the output current offset of 0.1% relative to the maximum output current; and for weight values it is 0.8% of its full-range. Fig.5b illustrates the large signal V-I characteristic observed at the low-impedance input node of current conveyors. Statistical characterization over 20 samples shows a standard deviation for the input offset-voltage of 3mV. Finally, Fig.5c contains the I-V characteristic of the voltage-limiter employed for integrators saturation. Saturated voltage levels exhibit a standard deviation of 2% of the full signal range.

14. SUMMARY

The realization of high complexity, accurate and programmable CNN systems presents appreciable difficulties related with technological constraints and system level specifications, in particular, with the high cell-density required for practical applications. In this sense, it present numerous interesting challenges for the design community, most of them common to other analog VLSI systems like general artificial neural networks. For this purpose, substantial efforts must be invested in yield improvement (both at the manufacturing and the design levels) including a careful characterization of failure probabilities, and from there on, possible technological improvements and yield-oriented design techniques. Also, some investment in CAD development, basically in yield estimation and functional circuit verification tools would be helpful, as well as a careful determination of static and dynamic accuracy requirements of CNNs. Some general guidelines for the future development of high performance, programmable CNN chips are, for instance, the use of multi-chip modules, micro-power circuit techniques, and design for low voltage supply operation.

15. REFERENCES

1 M.M. Gupta and G.K. Knopf, "Neuro-Vision Systems: A Tutorial", in *Neuro-Vision Systems: Principles and Applications* (M.M. Gupta and G.K. Knopf, eds.), pp.1-34, IEEE Press, New-York, 1994.

2 T. Duong, S. Kemeny, M. Tran, T. Daud, A. Thakoor, D. Ludwig, C. Saunders y J. Carson, "Low Power Analog Neurosynapse chips for a 3-D 'Sugarcube' Neuroprocessor", *Proceedings Conf. on Neural Networks*, pp. 1907-1911, 1994.

3 C. Koch and H. Li (eds.), *Vision Chips, Implementing Vision Algorithms Using Analog VLSI Circuits*, IEEE Press, New-York, 1994.

4 K. Kyuma, E. Lange, J. Ohta, A. Hermanns, B. Banish and M. Oita, "Artificial Retinas - Fast, Versatile Image Processors", *Nature*, Vol. 372, pp. 197-198, November 1994.

5 F. Werblin, A. Jacobs and J. Teeters, "The Computational Eye", *IEEE Spectrum*, pp. 30-37, May 1996.

6 S. Espejo, R. Carmona, R. Domínguez-Castro and A. Rodríguez-Vázquez, "A 0.8μm CMOS Programmable Analog-Array-Processing Vision-Chip with Local Logic and Image Memory", *Proc. of the 1996 European Solid-State Circuits Conference*, to appear, September 1996.

7 L.O. Chua and T. Roska, "The CNN Paradigm". *IEEE Trans. Circuits and Systems-I*, Vol. CAS-40, pp 147-156, March 1993.

8 T. Roska and L.O. Chua, "The CNN Universal Machine: An Analogic Array Computer", *IEEE Transactions on Circuits and Systems-II*, Vol., 40, No.-3, March 1993.

9 A. Rodríguez-Vázquez, S. Espejo, R. Domínguez-Castro, J.L. Huertas, and E. Sánchez-Sinencio, "Current-Mode Techniques for the Implementation of Continuous- and Discrete-Time Cellular Neural Networks", *IEEE Trans. on Circuits and Systems II*, Vol 40, No. 3, pp 132-146, March 1993

10 S. Espejo, R. Carmona, R. Domínguez-Castro and A. Rodríguez-Vázquez, "A VLSI-oriented Continuous-Time CNN Model", *Int. J. Circuit Theory and ApplicationsI*, Vol. 24, pp. 341-356, May 1996.

11 T. Roska and L. Kék, *Analogic CNN Program Library*, Analogical and Neural Computing Laboratory Memo. DNS-5-1994, Budapest, 1994.

12 X. Arreguit, *Compatible Lateral Bipolar Transistors in CMOS Technology*, PhD dissertation, Lausanne EPFL, 1989.

13 D.J. Allstot and W.C. Black, "Technological Design Considerations for Monolithic MOS Switched-Capacitor Filtering", *Proceedings of the IEEE*, Vol. 71, pp. 167-186, August 1983.

14 A. Rodríguez-Vázquez, S. Espejo, R. Domínguez-Castro and R. Carmona, "On the VLSI Design of CNNs", *in VLSI Implementation of the Cellular Neural Network Universal Machine* (T. Roska and A. Rodríguez-Vázquez, eds.), John Wiley, 1997.

15 E. Roca and A. Rodríguez-Vázquez, "CMOS Photosensitive Devices", *in VLSI Implementation of the Cellular Neural Network Universal Machine* (T. Roska and A. Rodríguez-Vázquez, eds.), John Wiley, 1997.

16 T. Delbrück and C.A. Mead, *Analog VLSI Phototransduction*, California Institute of Technology CNS Memo No. 34, Pasadena, 1994.

17 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.

18 P.R. Gray and R.G. Meyer, "*Analysis and Design of Analog Integrated Circuits*", John Wiley & Sons, 3rd Ed., 1993.

19 H.B. Bakoglu, *"Circuits, Interconnections and Packaging for VLSI"*, Reading: Addison Wesley, 1990.

20 S.W. Director, W. Maly and AJ. Strojwas, *"VLSI Design for Manufacturing: Yield Enhancement"*, Boston: Kluwer, 1990.

21 J.A. Abraham: *"Fault Modeling in VLSI"*. North-Holland, 1986.

22 S.K. Ghandi, "*VLSI Fabrication Principles: Silicon and Gallium Arsenide*", New York: John Wiley, 1994.

23 A. Jee and C. Bazeghi, "*CARAFE User's Manual. Release Alpha.4*", UCSC-CRL-94-20, June 13, 1994.

24 Y.P. Tsividis, *"Operation and Modeling of the MOS Transistors"*. New York: McGraw-Hill, 1987.

25 M. Ismail and T. Fiez, *"Analog VLSI: Signal and Information Processing"*. New York: McGraw-Hill, 1994.

26 C. Eichenberger and W. Guggenbuhl, "On Charge Injection in Analog MOS Switches and Dummy Switch Compensation Techniques", *IEEE Trans. Circuits and Systems*, Vol. CAS-37, pp 256-264, 1990.

27 A. Rodríguez-Vázquez and E. Sánchez-Sinencio (editors), Special issue on "Low-Voltage and Low-Power Analog and Mixed-Signal Circuits and Systems", *IEEE Trans. on Circuits and Systems-I,* November 1995.

28 M. Nagata, "Limitations, Innovations and Challenges of Circuits and Devices into a Half Micrometer and Beyond", *IEEE Journal of Solid-State Circuits*, Vol. 27, pp. 465-472, 1992.

29 S. Espejo, R. Domínguez-Castro, A. Rodríguez- Vázquez, and R. Carmona, "Weight-Control Strategy for Programmable CNN Chips", *Proc. 3rd Int. Workshop on Cellular Neural Networks and their Applications*, pp. 405-410. Rome, December 1994.

30 S. Espejo, A. Rodríguez-Vázquez, R. Domínguez-Castro, J.L. Huertas, and E. Sánchez-Sinencio, "Smart-Pixel Cellular Neural Networks in Analog Current-Mode CMOS Technology", *IEEE Journal of Solid-State Circuits*, Vol. 29, pp. 895-905, August 1994.

31 P. Foldesy, A. Zarandy, P. Szolgay and T. Roska, *Measurement Results of the 20 x 22 CNNUM Chip*, Analogic and Neural Computing Laboratory, Hungarian Academy of Sciences, February 1996.

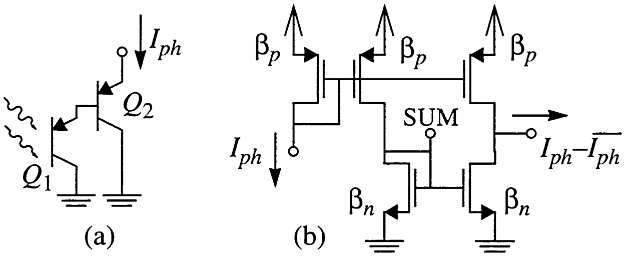

Fig. 2: a) Optical transducer implementation, b) Automatic threshold adjustment circuitry.

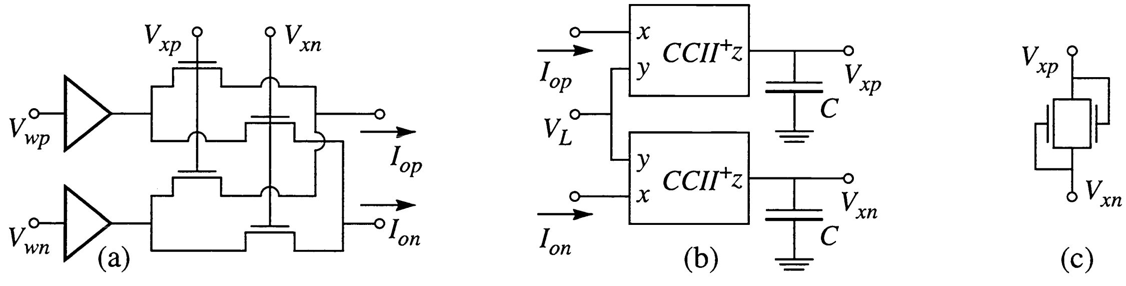

Fig. 3: Main CNN processing blocks, a) programmable synapse, b) integrator, c) voltage limiter.

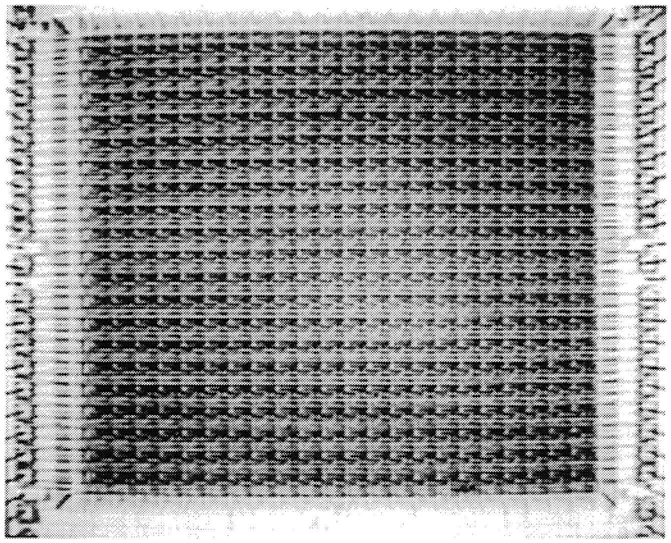

Fig. 4: Prototype's photograph.

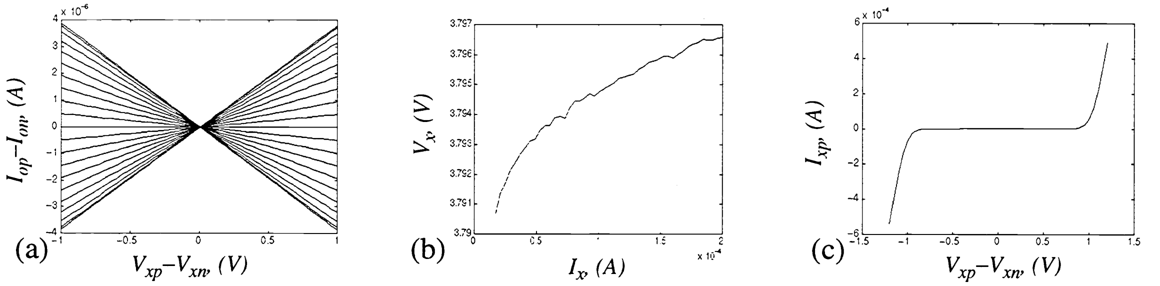

Fig. 5: Measured response of (a) the programmable synapse, (b) the low-impedance input node of CCII,(c) the voltage-limiter element.