A current-mode computational engine for stereo disparity and early vision tasks

G.M. Bisio[b], G.M. Bo[b], M. Confalone[a], L. Raffo[a], S.P. Sabatini[b], M.P. Zizola[a]

[a] DIEE-Dept. of Electrical and Electronic Engineering University of Cagliari
Piazza d'armi - 09123 Cagliari

[b] PSPC Research Group-DIBE University of Genova Via all'Opera Pia 11A - 16145 Genova
E-mail: luigi@diee.unica.it

Abstract

The paper reports on a long-term project aimed at developing analog integrated circuits that combine sensorial and computational capabilities adequate to perform in real time early vision perceptual tasks (e.g., texture segregation, depth from stereo, motion analysis, etc.). Such microsystems, which combine analog processing and neuromorphic computational paradigm, can satisfy the physical and economical constraints of low-end components. A programmable circuit for low-level visual tasks realized using a standard CMOS $0.8\mu m$ *technology is presented. The chip allows 1-D convolutions with Gabor functions of any phase through the interaction of an array of cells. The feasibility of the approach has been verified experimentally over a network of various sizes, up to 36 cells, obtained combining different chips. The area of the single processing cell is* $83\mu m \times 70\mu m$.

1 Introduction

Though the feasibility of a machine vision approach to early-vision tasks has been proven, their effective and low-cost implementation have not yet. Despite continuous advances in microelectronics technology, hardware simulation of machine vision algorithms still remain well below the performance level expected. To bridge this gap, a growing attention has been paid to analog perceptual solutions in which the system architecture integrates the image sensor with the preprocessing circuitry (cf. silicon retinas) [2] [22] [10]. This approach is particularly attractive for those early-vision applications in which flexibility of computation (otherwise, essential in a general-purpose vision system) is less important than the speed to go from signals to action. In such devices, information, indeed, is really distributed, being mapped directly in the electrical variables, and computation is carried out massively in parallel with high efficiency and speed. Information processing occurs at physical level, i.e., sensorial signals are transduced into new physical variables suitable for the system and it is the evolution of the system that performs computation.

Effective collective computation (small area and low-power consumption) on real-time signals is only possible when taking full advantage of the CMOS technology at transistor and/or circuit levels. Several examples are present in the literature illustrating how powerful this approach can be [16] [9] [14] [5] [22].

Our research intends to demonstrate how a variety of computational visual tasks can be mapped on the same "medium" which can be implemented efficiently through linear CMOS circuits. Early vision perceptual processes can be interpreted as a "measuring" operation on the visual signal [1]. Local relational operators are introduced to extract structural properties of visual data (orientation, texture, motion, etc.) by analyzing a single point of the visual field in terms of *how* it is related to what is around it. A proper formalization of these concepts is presented introducing lattice networks and analyzing their behavior in the Z-transform domain. In this way, it can be evidenced how each type of lattice network can be sustained by a unified entity, or computational *engine*, on which the various perceptual algorithms can be mapped by acting properly on the pattern of afferent connections. The functionality of 2nd-order lattice networks is discussed in detail and their VLSI implementation is verified. considering a programmable circuit realization of such 1-D processing arrays. The main design targets are: an operability over at least two decades of signal current, a pixel extension less than $100\mu m \times 100\mu m$, stability also for

low-spatial frequency, modularity. The use of this basic cooperative block in relevant perceptual problems such as stereo disparity estimation, or motion analysis is discussed.

2 The computational medium

2.1 Lattice models for early vision

A great deal of visual tasks is solved by animals and humans in a preattentive phase. In this phase, vast arrays of intercommunicating, identical processes carry out early vision tasks such as perceiving edges by texture differences, depth mapping, computing optical flow, recovering local surface structure, segmenting images, etc. These visual tasks can be formalized in terms of spatially continuous linear operators specified in differential and/or integral forms [7] [17] [11]. The discrete forms of these operators (i.e., lattice network) make it possible to map them on a layered network of interacting nodes. The network gains its system functionality through the cooperative action of ensembles of nodes topologically organized:

$$\sum_{k,l} a_{kl} e(m-k, n-l) = \sum_{k,l} b_{kl} s(m-k, n-l) \quad (1)$$

where $s(m,n)$ and $e(m,n)$ are the input and the output signals of the processing node at location (m,n), and the sets I and O determine the (rectangular) neighborhood domains of the feedforward excitation and cooperative interactions, respectively. Eq. 1 can be interpreted as a 2D iterative filter that detects specific characteristics present in the input pattern of excitation. The spatial domain on which these characteristics are detected is possibly larger than that of physical interconnections; this occurs both directly, by local interactions within the local neighborhood O, and indirectly, through the *propagation property* of iterative filters, which asserts that the output values of the filter can be recurrently affected by a larger neighbor region on the lattice network. As the system is linear and space-invariant, the output distribution (e) can be determined by the convolution between the input distribution (s) and the impulse response $h(\cdot,\cdot)$:

$$e(m,n) = \sum_k \sum_l h(m-k, n-l) s(k,l). \quad (2)$$

The shape and size of the kernel $h(\cdot,\cdot)$ depend on the sets of parameters a's and b's as can be explicited by introducing the Z transform $H(z_1, z_2)$ of the impulse response:

$$H(z_1, z_2) = \frac{\sum_{k=-I_m}^{I_m} \sum_{l=-I_n}^{I_n} b_{kl} z_1^{-k} z_2^{-l}}{\sum_{k=-O_m}^{O_m} \sum_{l=-O_n}^{O_n} a_{kl} z_1^{-k} z_2^{-l}}. \quad (3)$$

In this paper, we restrict our analysis to filters for which the kernels are separable:

$$e(m,n) = \sum_k h_1(m-k) \sum_l h_2(n-l) s(k,l). \quad (4)$$

Under these conditions, the 2D problem can be recast as the cascade of two 1D problems. As shown in the following, this assumption does not impact adversely on the solutions of the problems we are addressing.

We intend to investigate the functionality of 1-D lattice networks with respect to the degree of locality of interactions both excitatory and inhibitory. In order to better single out the essence of cooperative computation, we consider only the recurrent contributions, and assume the stimulus to be applied with just a one-to-one feedforward mapping. This will allow us to introduce the concept of "perceptual engine", thus pointing out how the same computational model can be used as a basic building block for a whole variety of filtering tasks.

In the following, we can assume that the perceptual engine is symmetric since, afterwards, we can regain generality of forms for the kernel acting upon the b's coefficients. Given the difference equation describing the nearest neighbor cooperation among nodes:

$$a_{-1} e(n-1) + a_0 e(n) + a_1 e(n+1) = b_0 s(n) \quad , \quad (5)$$

one derives the Z-domain transfer function and obtains the spatial impulse response [19]:

$$h^{(1)}(n) = C^{(1)} \mathrm{e}^{-\lambda |n|} \qquad \lambda > 0 \quad (6)$$

where $b_0 = C^{(1)}(1 - \mathrm{e}^{-2\lambda})$, $a_{-1} = a_1 = -\mathrm{e}^{-\lambda}$, and $a_0 = \mathrm{e}^{-2\lambda} + 1$.

Given the difference equation describing the nearest and next nearest neighbor cooperation among nodes:

$$\begin{aligned} a_{-2} e(n-2) + a_{-1} e(n-1) + a_0 e(n) \quad &+ \\ + a_1 e(n+1) + a_2 e(n+2) \quad &= \quad b_0 s(n) \end{aligned} \quad (7)$$

one derives the Z-domain transfer function and obtains the spatial impulse response [19]:

$$\begin{aligned} h^{(2)}(n) \quad &= \quad C^{(2)} \mathrm{e}^{-\lambda |n|} ((1 + \mathrm{e}^{-2\lambda}) \sin(\omega_0) \cos(\omega_0 n) \\ &+ \quad (1 - \mathrm{e}^{-2\lambda}) \cos(\omega_0) \sin(\omega_0 |n|)) \end{aligned} \quad (8)$$

where $b_0 = C^{(2)}[(1 - \mathrm{e}^{-4\lambda})(1 + \mathrm{e}^{-2\lambda}) \sin(\omega_0) - 4\mathrm{e}^{-2\lambda} (1 - \mathrm{e}^{-2\lambda}) \sin(\omega_0) \cos^2(\omega_0)](1 - \mathrm{e}^{-4\lambda})$, $a_{-2} = a_2 =$

$e^{-2\lambda}$, $a_{-1} = a_1 = -2e^{-\lambda}(1 + e^{-2\lambda})\cos\omega_0$, $a_0 = 1 + 4e^{-2\lambda}\cos^2\omega_0 + e^{-4\lambda}$. In the discussion that follows we shall consider the $h(\cdot)$ function as a primary characteristic of a network rather than a derived quantity to be obtained by the solution of a difference equation. These functions ($h^{(1)}$ and $h^{(2)}$) can be interpreted as the "perceptual engine" of our systems, providing computational primitives that can be composed to obtain more powerful image descriptors.

In general, we can state that any convolution kernel, whose two-sided Z transform has at the denominator a fourth-order polynomial with symmetric coefficients, can be obtained by a weighted local sum over the output of the 1-D 2nd-order cooperative network. Typical convolution kernels useful for visual tasks (difference of Gaussians (DOG) [13], difference of offset Gaussians (DOOG) [12], Gabor functions [4] can be implemented by a 2nd order network [19]. Specifically, it can be proven that a local weighted sum over the perceptual engines leads to periodic functions as the result of an amplitude modulation of a sine plane wave with an exponential envelope:

$$\begin{aligned} g(n) &= \alpha h(n-1) + \beta h(n) + \gamma h(n+1) \\ &= Ce^{(-\lambda|n|)}\cos(\omega_0 n + \varphi). \end{aligned} \tag{9}$$

where $\alpha = Ce^{-\lambda}(e^{-2\lambda} - 1)\cos(\omega_0 - \varphi)$ $\beta = C(1 - e^{-4\lambda})\cos(\varphi)$ $\gamma = Ce^{-\lambda}(e^{-2\lambda} - 1)\cos(\omega_0 + \varphi)$. These functions closely resemble in space the Gabor functions which are extensively used in computational vision due to their optimal conjoint spatial/spatial-frequency representation property.

Section 3 presents an overview of this approach for various machine vision tasks, evidencing how the functionality of the perceptual engines can provide powerful computational resources.

2.2 Distributed computation on analog physical systems

Spatial filters can be implemented by using linear resistive networks with positive and negative resistors [14] [8] [2] [18], or by current generators driven by the voltage at other nodes (VCCS) [3]. Since our main interest is to investigate how complex functionalities can be achieved through cooperative computations, we choose to relate the lattice networks' equations to a circuit architecture of one-way-interacting elements implemented as CCCS. One advantage of this interpretation shows up at transistor level, where it is shown how to implement these interactions with "high" precision through current mirrors.

Circuit level Let us consider a regular grid of nodes, in which each node is fed by a current generator whose value is proportional to the incident light at that point (feedforward contribution). By using a current mode technique, in which all signals are encoded by currents, the excitation e of the network is mapped on a current, the interaction can be implemented by current controlled current sources (CCCSs) that feed or sink currents according to the current values of neighboring nodes. In this way, the analog signal processing of our circuit architecture will be based on ratios of matched component values [21], thus eliminating the dependence on the performances of single devices.

Assuming $a_0 \neq 0$, $a_1 = a_{-1}$ and, $a_2 = a_{-2}$, thus we can define $G = -a_1/a_0$ and $K = a_2/a_0$ and rewrite Eq. 7 as:

$$\begin{aligned} I_e(n) &= I_s(n) + GI_e(n-1) + GI_e(n+1) - \\ &- KI_e(n-2) - KI_e(n+2) \end{aligned} \tag{10}$$

$$e(n) = \frac{I_e(n)}{I_{norm}}, \qquad \frac{b_0}{a_0}s(n) = \frac{I_s(n)}{I_{norm}}$$

where $I_s(n)$ is the normalized input current (i.e., stimulus) and $I_e(n)$ is the normalized output current (i.e. excitation); the coefficients G and K are always positive. Eq. 10 can be interpreted in architectural terms as an array of circuital computational nodes (see Fig. 1), with local interconnections both at the nearest and the next nearest neighbors. Each computational node has two output currents $GI_e(n)$ toward the nearest nodes, two (negative) output currents $KI_e(n)$ toward the next nearest nodes, and receives the corresponding contributions from its neighbors, besides its input excitation $I_s(n)$. The circuit representation of a node is based on the use of CCCSs with the desired current gain G and K.

The CCCSs are implemented by using cascoded current mirrors (see Fig. 2). With respect to a simple current mirror the output resistance is increased allowing a reduction of the area of the transistors. Unfortunately, the increasing in the output resistance is at expense of the output voltage swing hence a smaller limit to the input current have to be set. To improve the matching between the currents on the branches of the current mirrors, keeping small the size of the transistors, we operate the transistors in the strong-moderate inversion limit. These conflicting requirements are satisfied when the input current is limited in the range $[10nA - 1\mu A]$. Furthermore, G and K are set to a rational number (e.g. p/q) where p and q are the number of identical transistors in the two branches of the current mirrors of G and K. The two CCCSs with gain K are obtained through the current mirror transistors M1-Ma(Mb); the two

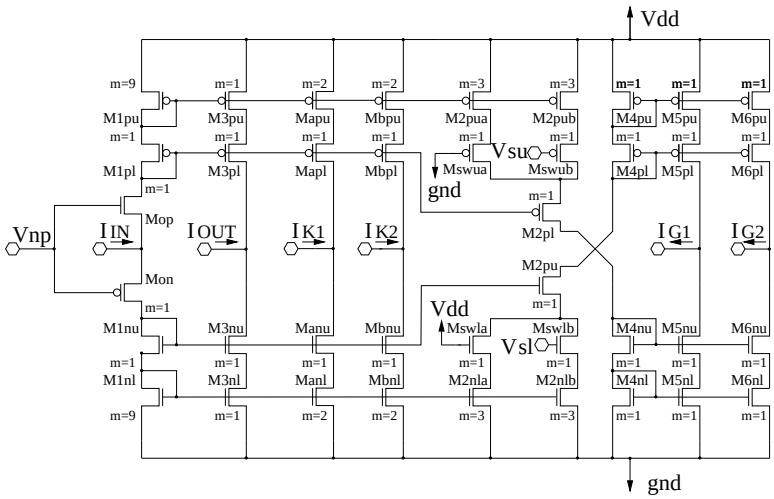


Figure 2: Circuital schema of one node of the network

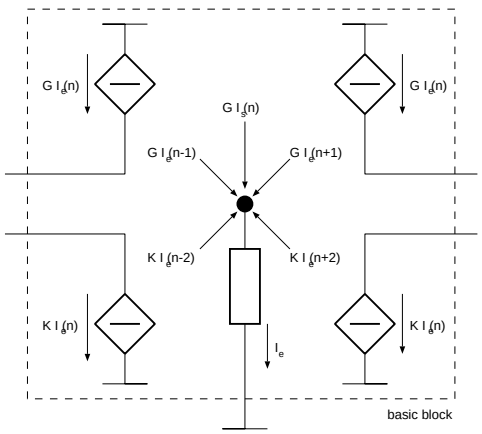


Figure 1: The node basic block and its internal circuital structure: the black dot represents the node which is stimulated by the photocurrent $I_s(n)$ and receives inhibitory (K) and excitatory (G) contributions from neighboring nodes. The sum of the current contributions from other nodes is performed by Kirchhoff current law at the node. The resulting excitation current, $I_e(n)$, acts, in turn, on neighboring nodes controlling the currents of four current generators.

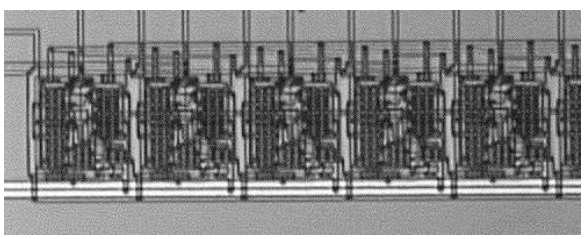


Figure 3: Micro photo of the 9-cell array chip.

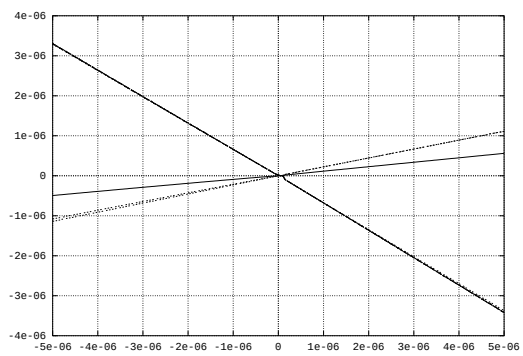
CCCSs with gain G are obtained through the current mirror transistors M1-M2-M4-M5(M6). The values of gains K and G are restricted to rational numbers, in relation to all possible combinations of the set of identical transistors connected in parallel in each branch of the current mirrors. In the present realization two combination of (G,K) are possible and can be selected through the switches Mswub, Mswlb: (3/9, 2/9), (6/9, 2/9). The two transistors Mop and Mon at the input of the cell are introduced to cancel the current flowing through M1p and M1n when the input current I_{in} is null. In this way only one of the two complementary parts of the cell is active when an input current is injected in or sunk from the cell.

Physical level and performance The proposed circuit has been designed in a 0.8 μm standard CMOS, N-well, double poly, double metal technology. The area of the single cell is $83\mu m \times 70\mu m$ without the sensor. The test chip contains a stripe of 9 cells array module that can be connected to others to form a longer chain (Fig. 3).

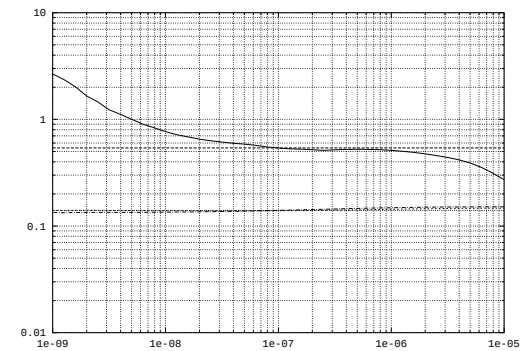
Fig. 4a shows the IG1, IG2, IK1, IK2, IOUT of a cell for input currents ranging from $-5\mu A$ to $5\mu A$ and for voltage at nodes ranging from 1.5V to 3.5V. A perfect match between the expected values and measures is evidenced.

The ratio between the output and the input current at the central node is depicted in Fig. 4b versus the input current (logarithmic scale). It can be noted that analytical and measured linearity shows a good matching in the range [$1nA$, $10\mu A$] for G=3/9 (lower curves) and only for two decades for G=6/9 (upper curves).

Fig. 5a shows the spatial inpulse response of a single

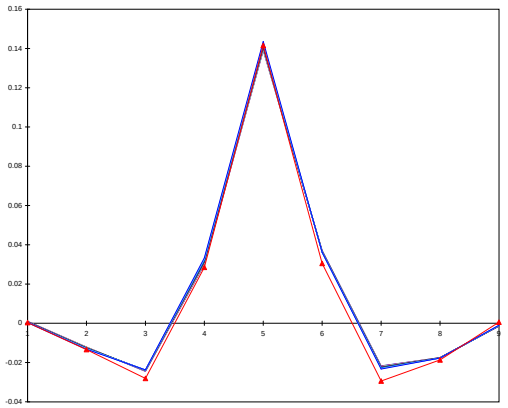


(a)

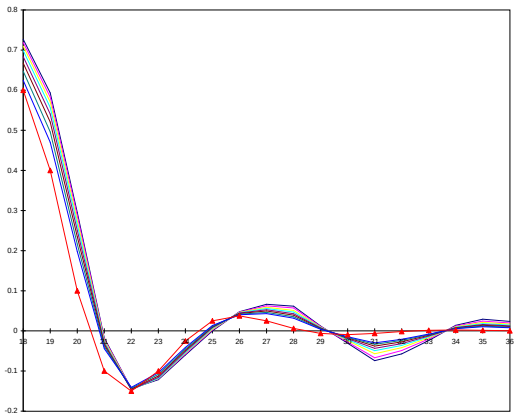


(b)

Figure 4: (a) Output currents from a cell (IG,IK,IOUT) for input currents ranging from $-5\mu A$ to $5\mu A$ for different voltage at nodes, (b) O/I transfer function at the central node versus input current (logarithmic scale) for G=3/9 (upper lines) and for G=6/9 (lower lines)



(a)



(b)

Figure 5: (a) Output of a 9 nodes network stimulated in the central node and ideal response (triangular dotted line), (b) Zoom on the output of a 36 nodes network obtained from the interconnection on four test chips (right half nodes)

chip network with Msw transistors open (G=3/9) when stimulated at the central node with a constant current ranging from $100nA$ to $1\mu A$. The output is normalized to the input current and compared to the theoretical response (the line with triangular dots).

Fig 5b shows the output for the right half nodes of a network obtained by connecting four chips to form a 36 nodes network with G=6/9 (i.e. Msw's closed). It is useful to note that even though the match between the expected response (the line with triangular dots) and the measures is not exact, the output corresponds in any case to a perceptual engine with slightly different parameters. This mismatching is due to the fact that the computational schema that generates a perceptual engine is very sensitive to variations of G,K parameters especially for low ω_0 [19]. The error detected in our measures corresponds to variations of G and K less than 1%, whilst one should keep in mind that the real risk for this kind of implementation is to drive the behavior of the network outside the convergence region (for the parameters considered errors larger than 4%). Therefore, from an application point of view, our results are really effective and show that the approach followed is feasible for the realization of an optical sensor with low-level processing capabilities.

3 System solutions and applications

Early vision tasks concern the analysis of visual information by the decomposition of the input image into coherent components that can be independently characterized [13]. Thus, the first stages of a vision system should be able to extract quickly and efficiently the *salient* information in the image, represented in natural scenes by lines, edges, textures, stereo cues, velocity fields, etc. Various computational vision models can make use of Gabor-like functions as localized spatial filters or as basic functions for image transformations [4] [6]. Accordigly, the perceptual engine can be used as an elementary operator on its own, or as a building block to obtain more elaborate visual operators.

In this section, we'll examine the application capabilities of our perceptual microsystem as a front-end module in a machine vision system, describing its potentialities in stereoscopic depth perception and in motion analysis.

Stereopsis Stereoscopic depth estimation is mainly based on disparity measurement. Disparity is a local property that can be associated to a stereo pair (left and right) of images or image sequences. The problem has an elegant formulation in the frequency domain, in relation to the so-called phase-based matching methods [20]. Basically, disparity estimation is computed as the shift necessary to align the phase values of the band-pass filtered version of the binocular stereo signals. One can assume that the disparity between the stereo image pair occurs only along the direction of the epipolar lines [?, HowardRogers95] In the simplest situation, in which optical axes are parallel, epipolar lines are parallel to the line joining the optical centers of the cameras. In such conditions, stereo phase-based matching can be formulated as a 1D problem. Hence, we can use a rank of 1D perceptual engines acting separately on different rows of the input image (see the module *perceptual array* in Fig. 6). The local

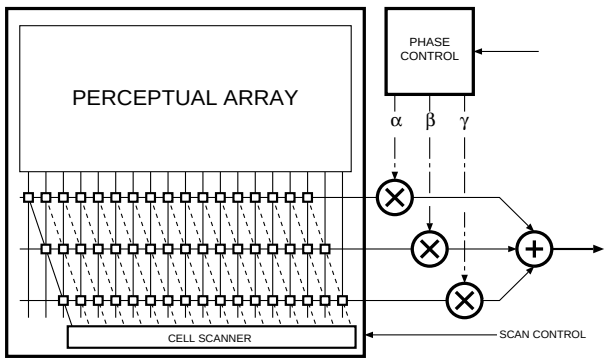


Figure 6: Overall architecture of the module for stereo depth perception. The convolution with Gabor-like functions of any phase can be obtained through the weighted sum of the output of three neighboring nodes in the perceptual array.

phase and thence the local disparity are computed following Sanger's algorithm [20], whose response can be obtained through the computational scheme evidenced in Fig. 6.

Motion analysis Several approaches to motion analysis rely upon correspondence measures in the spatio-temporal domain of a visual scene [15] i.e., upon space/time linear filtering of the visual input. It is possible to obtain a spatio-temporal kernel using the perceptual engine if one introduces a temporal dependence in the weighting coefficients α, β and γ of Eq. (9). Specifically, we can rewrite Eq. (9) as:

$$g(n,t) = C' \mathrm{e}^{-t/\tau} \mathrm{e}^{(-\lambda|n|)} \cos(\omega_0 n + \omega_0 v_f t). \qquad (11)$$

having posed $C(t) = C' \mathrm{e}^{-t/\tau}$ and $\varphi(t) = \omega_0 v_f t$, where the parameters τ and v_f characterize the temporal response of the second-order network generating $\alpha(t)$,

$\beta(t)$, and $\gamma(t)$. In this way, we obtain a 1D spatio-temporal Gabor-like kernel with a linearly-variable phase interpretable by a dumped localized traveling wave with constant velocity v_f. Let us consider a system with a transfer function given by Eq. (11): if we stimulate the system with a moving spot of light, with speed v_p, its temporal response will be the envelope of the Gabor-like function *if and only if* $v_f = v_p$. Such a system will be able to reveal the presence of an impulsive input moving along a specific direction with speed $v_p = v_f$. With more complex inputs pre-processing step will be necessary (i.e. contrast and edge enhancement) in order to provide the system with impulsive inputs only. We simulated such a system with several inputs. Among others, a sequence of a moving electrical toy-train (see Fig. 7) was used as a "realistic" input to test the system; the speed of the train was $v_p = 1$ pixel/time step. For each frame, a single row was extracted to generate the 1D input sequence shown in Fig. 8a. The outputs of temporally modulated perceptual engines, tuned to different velocities, go through a bank of matched filters tuned to the envelope of the Gabor-like functions. As shown in Figs. 8b, the correct velocity of the train is revealed by the filter with maximum response, corresponding to $v_f = 1$ pixel/time step.

Acknowledgment This work is partially supported by CEC-Esprit-BRA Project CORMORANT, ASI Project *Microsistemi Ottici per la Robotica Spaziale* and MURST 40% *Sistemi microelettronici e microsistemi.*

References

[1] E.H. Adelson and J.R. Bergen. The plenoptic function and the elements of early vision. In M.S. Landy and J.A. Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[2] X. Arreguit and E.A. Vittoz. Perception systems implemented in analog VLSI for real-time applications. In *Proc. of the PerAc'94 Conference: From Perception to Action*, Lausanne, Sept.5-9, 1994, 1994.

[3] G.M. Bisio, M. Bruccoleri, P. Cusinato, L. Raffo, and S.P. Sabatini. An analog VLSI massively parallel module for low-level cortical processing in machine vision. In *Proc. IEEE-MICRONEURO94 - Torino (Italy)*, pages 207–211, 1994.

[4] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Amer.*, A/2:1160–1169, 1985.

[5] T. Delbrück. Silicon retina with correlation-based velocity-tuned pixels. *IEEE Trans. Neural Net.*, 4:529–541, 1993.

[6] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biol. Cybern.*, 61:103–113, 1989.

[7] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[8] H. Kobayashi, J.L. White, and A.A. Abidi. An active resistor network for Gaussian filtering of images. *IEEE J. on Solid State Circuits*, 26:738–748, May 1991.

[9] C. Koch, J. Marroquin, and A. Yuille. Analog "neuronal" networks in early vision. *Proc. Natl. Acad. Sci.*, 83:4263–4267, 1986.

[10] C. Koch and B. Mathur. Neuromorphic vision chips. *IEEE Spectrum*, 33(5):38–46, 1996.

[11] J.J. Koenderink. Scale-time. *Biol. Cybern.*, 58:159–162, 1988.

[12] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Amer.*, 7:923–932, 1990.

[13] D. Marr. *Vision.* New York: Freeman, 1982.

[14] C.A. Mead. *Analog VLSI and Neural Systems.* Addison-Wesley, Reading, 1989.

[15] K. Nakayama. Biological image motion processing: a review. *Vision Res.*, 25:625–660, 1985.

[16] T. Poggio. Early vision: from computational structure to algorithms and parallel hardware. *Computer Vision, Graphics, and Image Processing*, 31:139–155, 1985.

[17] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:638–643, 1985.

[18] L. Raffo. Resistive network implementing maps of Gabor functions of any phase. *Electronics Letters*, 31(22):1913–1914, 1995.

[19] L. Raffo. Analysis and synthesis of resistive networks for distributed visual elaborations. *Electronics Letters*, 32(8):743, April 1996.

[20] T.D. Sanger. Stereo disparity computation using Gabor filters. *Biol. Cybern.*, 59:405–418, 1988.

[21] E.A. Vittoz. Analog VLSI signal processing: Why, where, and how? *J. of VLSI Signal Processing*, 8:27–44, 1994.

[22] J.L. Wyatt, C. Keast, M. Seidel, D. Standley, B. Horn, Knight T., C. Sodini, H.-S. Lee, and T. Poggio. Analog VLSI systems for image acquisition and fast early vision processing. *Int. J. of Computer Vision*, 8:217–230, 1992.

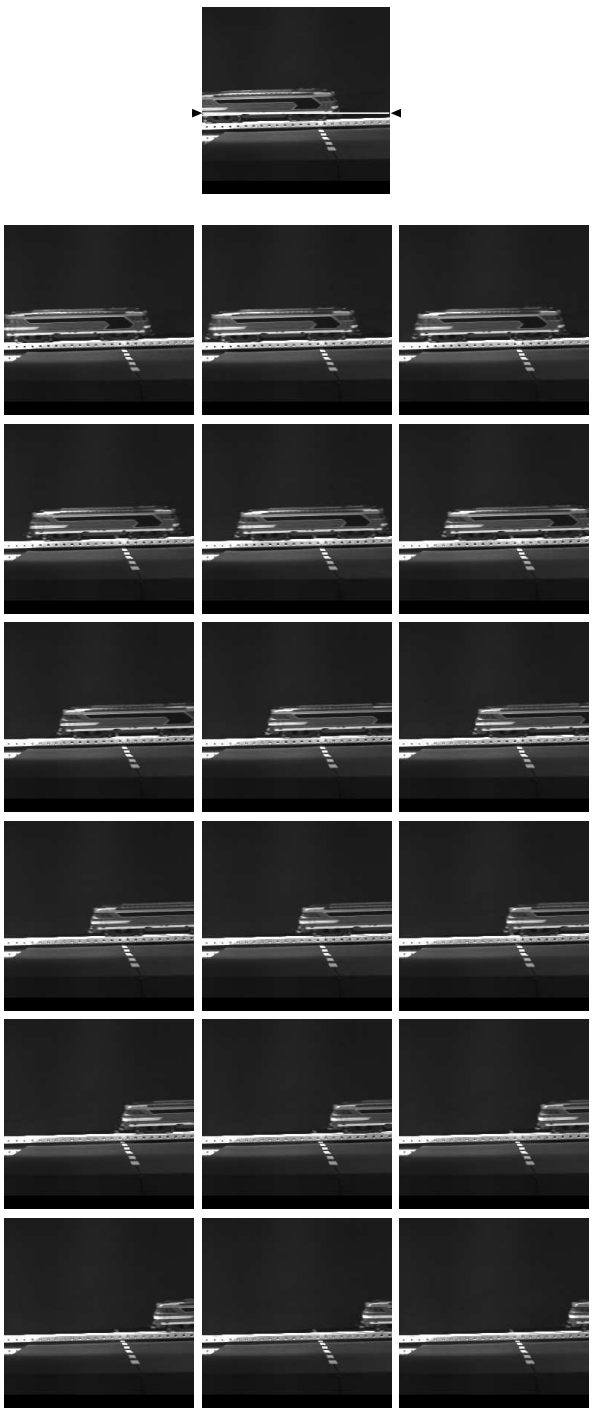
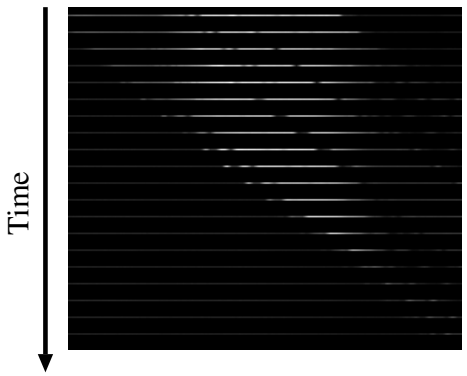
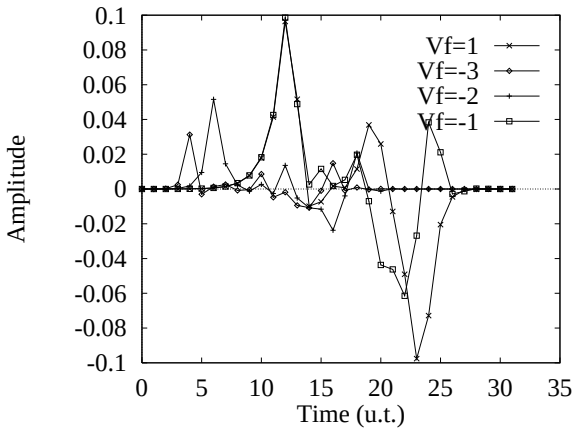


Figure 7: Sequence of a moving toy-train: in the first frame the row extracted to generate a 1D input sequence is evidenced.



(a)



(b)

Figure 8: (a) Spatio-temporal input provided to the system. (b) Results of the simulations for different values of v_f. Each curve represents the output of the matched filters.