Smart Materials for Visual Perception

S.P. Sabatini[a], L. Raffo[b], F. Solari[a], A. Bonfiglio[b] and G.M. Bisio[a]

[a]PSPC Research Group-DIBE University of Genoa, Italy
[b]DIEE University of Cagliari, Italy

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.

The paper presents a system solution to the generation of stereo depth maps based on a 1D 17-node lattice network implemented in CMOS technology. The essence of the computation performed by the recurrent lattice network, can be described most conveniently by introducing a basic cooperative kernel (the "perceptual engine") that can be used as a building block for a variety of early vision modules.

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) [1] [2] [3]. 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 [4] [5] [6] [7] [2].

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 [8]. 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 in Section 2 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. The use of this basic cooperative block in relevant perceptual problems such as stereo disparity estimation, or motion analysis is discussed (see Section 3).

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 [9] [10] [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:

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

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 $h(\cdot,\cdot)$:

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

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:

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

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

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

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:

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

one derives the Z-domain transfer function and obtains the spatial impulse response [12]:

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

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:

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

one derives the Z-domain transfer function and obtains the spatial impulse response [12]:

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

where $b_0 = C^{(2)}[(1-\mathrm{e}^{-4\lambda})(1+\mathrm{e}^{-2\lambda}) \sin(\omega_0) - 4\mathrm{e}^{-2\lambda}\ (1-\mathrm{e}^{-2\lambda}) \sin(\omega_0) \cos^2(\omega_0)](1-\mathrm{e}^{-4\lambda})$, $a_{-2} = a_2 = \mathrm{e}^{-2\lambda}$, $a_{-1} = a_1 = -2\mathrm{e}^{-\lambda}(1+\mathrm{e}^{-2\lambda}) \cos\omega_0$, $a_0 = 1 + 4\mathrm{e}^{-2\lambda} \cos^2\omega_0 + \mathrm{e}^{-4\lambda}$. 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) [14], Gabor functions [15] can be implemented by a 2nd order network [12]. 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:

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

where $\alpha = C\mathrm{e}^{-\lambda}(\mathrm{e}^{-2\lambda} - 1)\cos(\omega_0 - \varphi)$ $\beta = C(1 - \mathrm{e}^{-4\lambda})\cos(\varphi)$ $\gamma = C\mathrm{e}^{-\lambda}(\mathrm{e}^{-2\lambda} - 1)\cos(\omega_0 + \varphi)$. 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 [6] [16] [1] [17], or by current generators driven by the voltage at other nodes (VCCS) [18]. 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 architectureof 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 [19], 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

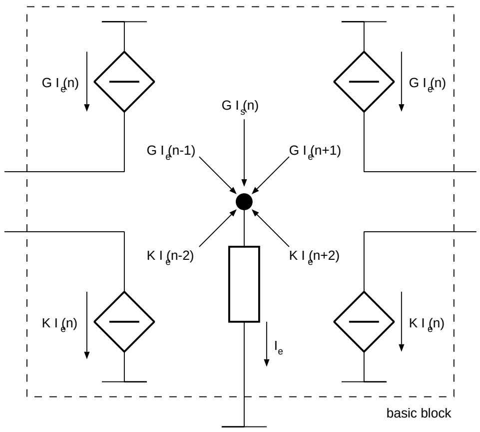


Figure 1: The node basic block and its internal circuital structure: the black dot represents the node which is stimulated by the photocurrent $I_s(n)$ and receives inhibitory (K) and excitatory (G) contributions from neighboring nodes. The sum of the current contributions from other nodes is performed 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.

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 (see Fig. 1(b)).

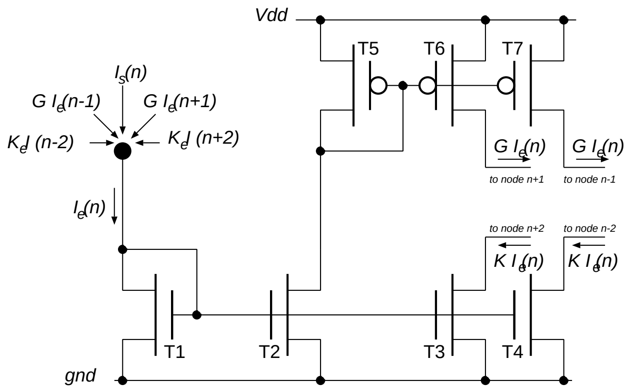


Figure 2: Description of the node basic block at transistor level. $I_e(n)$ is copied (weighted according to G and K values) to provide excitation or inhibition to the neighboring nodes.

Transistor level As shown in Fig. 2, the CCCSs are implemented by using standard current mirrors which provide, when the transistors are operating in the saturation region, a weighted copy of their input currents according to the W/L ratios of the two transistors. To improve the matching between devices, a current gain p/q, different from one, is best obtained by using two sets of identical transistors connected in parallel (i.e. q for T1 and a proper p_g for T2 and p_k for T3 and T4). The two CCCSs with gain K are obtained through the current mirror transistors T1-T3(T4); the two CCCSs with gain G are obtained through the current mirror transistors T1-T2-T5-T6(T7). It is worth noting that in this way the values of the 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; these combinations, selectable by proper switches, are characterized by the same level of precision.

If one considers that a current mirror is a cell in which the input and output currents are monodirectional, whereas the signals of the filter (mapped on currents) can be both positive and negative, one concludes that the circuit implementation of each generic node should be composed of two complementary parts, one for positive input currents (i.e., the current is injected into the node), and one for negative input currents (when the current is sourced from the node). The circuit solution for the negative input current is the same of the one shown in Fig. 2, replacing the NMOS with PMOS transistors and vice versa. In the complete schematic of a node two transistors are introduced to cancel the current flowing through VDD and GND 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 or sourced to the cell.

The present implementation of the chip (see Fig. 3) is characterized by $q = 7$, $p_g = 1$, and $p_k = 2$. In this way, the gain of the CCCSs are $G = 1/7$ and $K = 2/7$ which correspond to have $\lambda = 0.5713$ and $\omega_0 = 1.46354$.

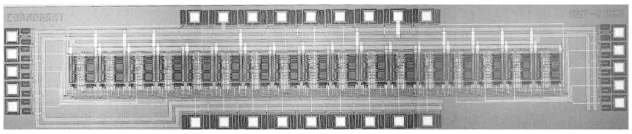


Figure 3: Microphotograph of a 1-D 2nd order network chip

As discussed, by combining the response of neighboring nodes it is possible to obtain a Gabor-like function of proper phase (decay rate and frequency are hard-wired in the underlying perceptual engine). Figure 4 shows the measured impulse response for different values of the injected input current in the central node.

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 [15] [20]. 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 in detail its potential application in stereoscopic depth perception and , in perspective, in motion analysis.

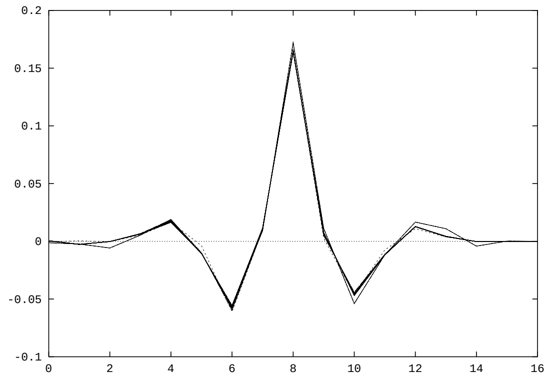


Figure 4: Measured impulse response for different values of the injected input current in the central node of the perceptual engine.

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. Hence, a joint representation of the information flows coming from the two eyes is a definitive need for stereoscopic depth perception. Such a joint representation can be related to the spatial properties of binocular filters tuned to specific disparity values. This problem has an elegant formulation in the frequency domain, in relation to the so-called phase-based matching methods [21]. 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. In the simplest configuration, 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 casted as a 1D problem. Hence, we can use a rank of 1D perceptual engines acting separately on different rows of the input image (see Fig. 5). The local phase is computed following Sanger's algorithm [21], from the output of a pair

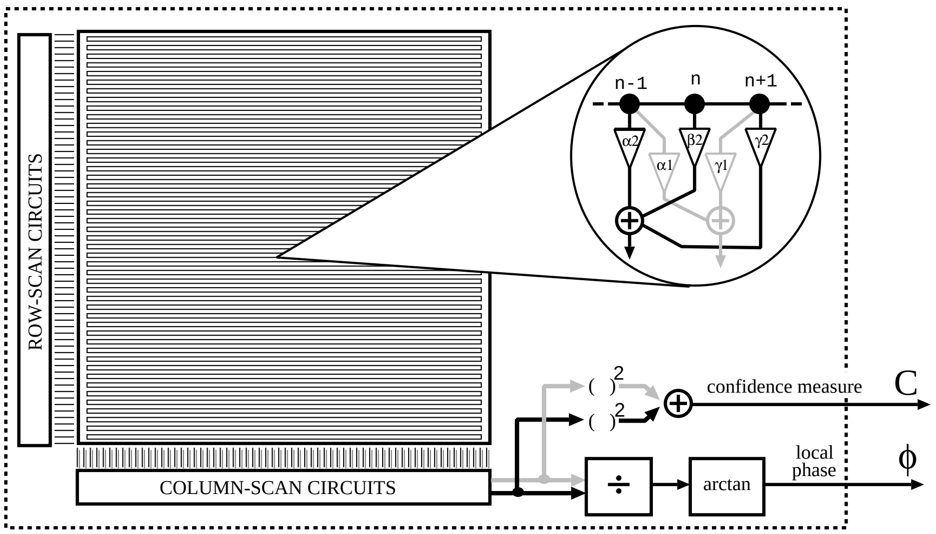


Figure 5: Overall architecture of the module for stereo depth perception system. The column-scan circuits perform the local weighted combination of the output currents coming from the chip core to obtain the pair of quadrature Gabor filters. The ouput of the overall system are the local phase (ϕ) and its confidence value (C).

of Gabor filters in quadrature (see Fig. 5). The functional organization of the stereoscopic perception system is sketched in Fig. 6: the differential measure of the local phase, obtained by the left and right retinas, is validated by a confidence measure, thus deriving a reliable disparity map. In Fig. 7 the image reproduces a simple toy-car reconstruction of a traffic/vehicle scene. The disparity map, obtained through a simulation of the system depicted in Fig. 6, evidences in light gray the car closer to the viewer, and in darker gray the approaching car in the background.

Motion analysis Several approaches to motion analysis rely upon correspondence measures in the spatio-temporal domains of a visual scene. Among the others, the so-called "energy" or "spatio-temporal" models [22] are based on a

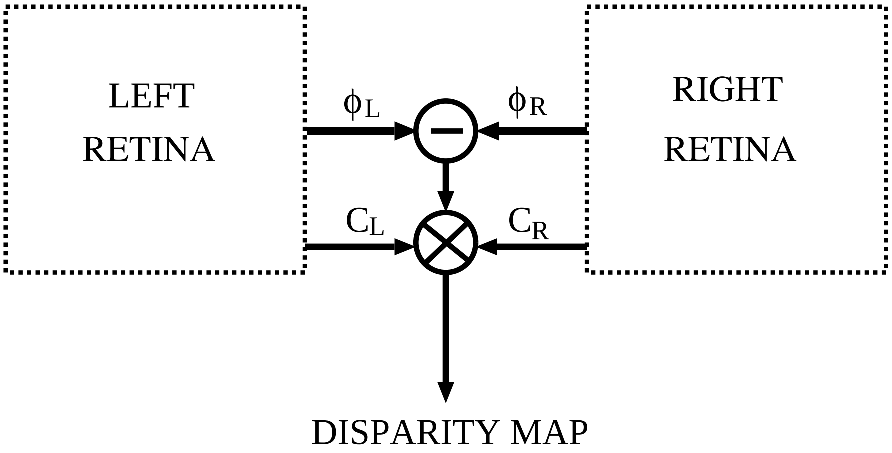


Figure 6: Schematic representation of the architecture of the stereo depth perception system.

linear filtering operation with kernels with a specific orientation in the spatio-temporal plane which gives them a specific direction sensitivity. A set of such filters are often used as a first stage of motion algorithms in order to extract information about local velocity [23].

Adelson and Bergen [22] demonstrated that spatio-temporal oriented filters can be constructed from pairs of separable (i.e., not oriented) filters. According to the method of Adelson and Bergen, filters tuned to specific directions can be obtained through a temporally modulated combination of perceptual engines. Specifically, by taking appropriate sums and differences of the outputs of a 2nd-order lattice network, it is possible to construct pairs of spatio-temporal oriented filters to be used to measure and discriminate the energy of motion. For example, an even filter sensitive to leftward motion can be obtained with the following combination of the perceptual engine responses:

$$g_{LE} = \alpha_{odd} f_3(t)h(n-1) + \gamma_{odd} f_3(t)h(n+1) + \alpha_{even} f_5(t)h(n-1) + \beta_{even} f_5(t)h(n) + \gamma_{even} f_5(t)h(n+1) \quad (11)$$

where f_3 and f_5 are modulating functions temporally in quadrature.

4 Conclusions

The approach followed in this paper can be related to the research path followed by Carver Mead and his colleagues [6] towards efficient silicon implementations of visual tasks. Simple cooperative networks are seen both as a medium whose structure is in many ways similar to neural structures [24] [25], and as the circuit analogue of the discretized forms of machine vision algorithms [26]. The main innovative feature of our approach concerns (i) the architectural specification of an analog perceptual agent –the *perceptual engine*– to be used as a powerful building block for early vision perceptual tasks; (ii) the VLSI implementation of such recurrent network as an analog VLSI CMOS circuit; (iii) the overall specification of a microsystem combining sensorial and computational capabilities based on the perceptual engine.

The perceptual applications presented focus on low-level vision tasks such as extraction of stereo disparity, and analysis of motion. In perspective, information processing capabilities at intermediate and high levels could be obtained by our microsystem by adding further processing layers of similar cells.

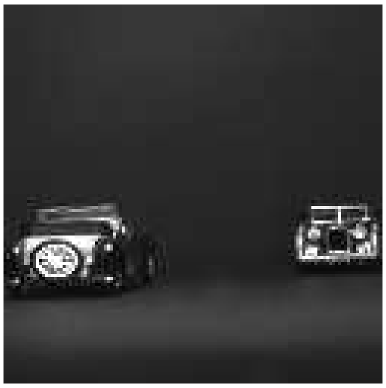
Acknowledgments The authors wish to thank Prof. G. Soncini and Ing. G.F. Dalla Betta for the support in the fabrication of the chip. This work was partially supported by the *CEC- Esprit BR Project CORMORANT-8503*, and by the *University of Genoa Project "Strutture Fisiche per la Percezione e l'Elaborazione", and by* ASI Project "Microsistemi Ottici per la Robotica Spaziale".

References

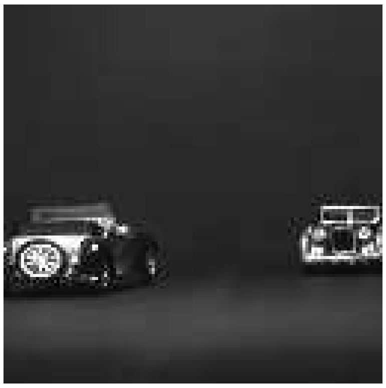
[1] X. Arreguit and E.A. Vittoz. Perception systems implemented in analog VLSI for real-time applications. In *Proc. of the PerAc'94 Conference: From Perception to Action*, Lausanne, Sept.5-9, 1994, 1994.

[2] 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.

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



left image



right image

disparity map

Figure 7: Example of depth estimation from a stereo image pair. The disparity map, obtained through a simulation of the system depicted in Figs. 5-6, evidences in light gray the car closer to the viewer, and in darker gray the approaching car in the background.

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

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

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

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

[8] 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.

[9] 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.

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

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

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

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

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

[15] 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.

[16] 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.

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

[18] 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.

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

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

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

[22] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Amer.*, 2:284–321, 1985.

[23] S.J. Nowlan and T.J. Sejnowski. Filter selection model for motion segmentation and velocity integration. *J. Opt. Soc. Amer.*, 11:3177–3200, 1994.

[24] M.A. Sivilotti, M.A. Mahowald, and C.A. Mead. Real-time visual computations using analog CMOS processing arrays. In *Advanced Research in VLSI:Proc. Stanford Conference 1987*, pages 295–312, 1987.

[25] C.A. Mead and M.A. Mahowald. A silicon model of early visual processing. *Neural Networks*, 1:91–97, 1988.

[26] C. Koch. Seeing chips: Analog VLSI circuits for computer vision. *Neural Comp.*, 1:184–200, 1989.