

# Phase-Based Binocular Perception of Motion in Depth: Cortical-Like Operators and Analog VLSI Architectures

**Silvio P. Sabatini**

*Department of Biophysical and Electronic Engineering, University of Genoa, Via All' Opera Pia 11a, 16145 Genova, Italy  
Email: [silvio@dibe.unige.it](mailto:silvio@dibe.unige.it)*

**Fabio Solari**

*Department of Biophysical and Electronic Engineering, University of Genoa, Via All' Opera Pia 11a, 16145 Genova, Italy  
Email: [fabio@dibe.unige.it](mailto:fabio@dibe.unige.it)*

**Paolo Cavalleri**

*Department of Biophysical and Electronic Engineering, University of Genoa, Via All' Opera Pia 11a, 16145 Genova, Italy  
Email: [paolo.cavalleri@dibe.unige.it](mailto:paolo.cavalleri@dibe.unige.it)*

**Giacomo Mario Bisio**

*Department of Biophysical and Electronic Engineering, University of Genoa, Via All' Opera Pia 11a, 16145 Genova, Italy  
Email: [bisio@dibe.unige.it](mailto:bisio@dibe.unige.it)*

*Received 30 April 2002 and in revised form 7 January 2003*

We present a cortical-like strategy to obtain reliable estimates of the motions of objects in a scene toward/away from the observer (motion in depth), from local measurements of binocular parameters derived from direct comparison of the results of monocular spatiotemporal filtering operations performed on stereo image pairs. This approach is suitable for a hardware implementation, in which such parameters can be gained via a feedforward computation (i.e., collection, comparison, and punctual operations) on the outputs of the nodes of recurrent VLSI lattice networks, performing local computations. These networks act as efficient computational structures for embedded analog filtering operations in smart vision sensors. Extensive simulations on both synthetic and real-world image sequences prove the validity of the approach that allows to gain high-level information about the 3D structure of the scene, directly from sensorial data, without resorting to explicit scene reconstruction.

**Keywords and phrases:** cortical architectures, phase-based dynamic stereoscopy, motion processing, Gabor filters, lattice networks.

## 1. INTRODUCTION

In many real-world visual application domains it is important to extract dynamic 3D visual information from 2D images impinging the retinas. One of this kind of problems concerns the perception of motion in depth (MID), that is, the capability of discriminating between forward and backward movements of objects from an observer has important implications for autonomous robot navigation and surveillance in dynamic environments. In general, the solutions to these problems rely on a global analysis of the optic flow or on token matching techniques which combine stereo correspondence and visual tracking. Interpreting 3D motion estimation as a reconstruction problem [1], the goal of these approaches is to obtain from a monocular/binocular image se-

quence the relative 3D motion to every scene component as well as a relative depth map of the environment. These solutions suffer under instability and require a very large computational effort which precludes a real-time reactive behaviour unless one uses data parallel computers to deal with the large amount of symbolic information present in the video image stream [2]. Alternatively, in the light of behaviour-based perception systems, a more direct estimation of MID can be gained through the local analysis of the spatiotemporal properties of stereo image signals.

To better introduce the subject, we briefly consider the dynamic correspondence problem in the stereo image pairs acquired by a binocular vision system. Figure 1 shows the relationships between an object moving in 3D space and the geometrical projection of the image in the right and left retinas.

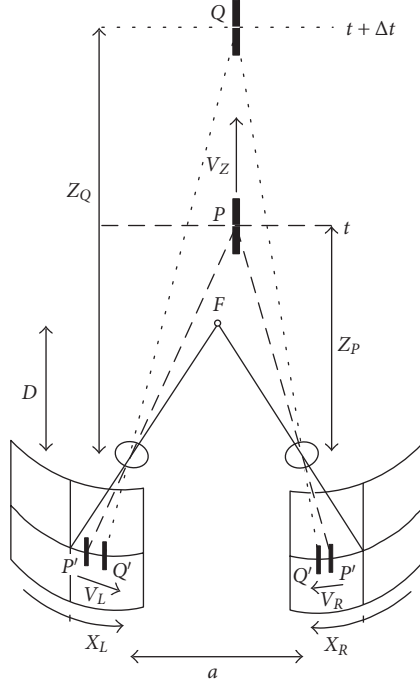


FIGURE 1: The stereo dynamic correspondence problem. A moving object in the 3D space projects different trajectories onto the left and right images. The differences between the two trajectories carry information about MID.

If an observer fixates at a distance  $D$ , the perception of depth of an object positioned at a distance  $Z_P$  can be related to the differences in the positions of the corresponding points in the stereo image pair projected on the retinas, provided that  $Z_P$  and  $D$  are large enough ( $D, Z_P \gg a$  in Figure 1, where  $a$  is the interpupillary distance and  $f$  is the focal length). In a first approximation, the positions of corresponding points are related by a 1D horizontal shift, the binocular disparity  $\delta(x)$ . The relation between the intensities observed by the left and right eye, respectively,  $I^L(x)$  and  $I^R(x)$ , can be formulated as follows:  $I^L(x) = I^R[x + \delta(x)]$ . If an object moves from  $P$  to  $Q$ , its disparity changes and projects different velocities on the retinas ( $v_L, v_R$ ). Thus, the  $Z$  component of the object motion (i.e., its motion in depth)  $V_Z$  can be approximated in two ways [3]: (1) by the rate of change of disparity and (2) by the difference between retinal velocities, as it is evidenced in the box in Figure 1. The predominance of one measure on the other corresponds to different hypotheses on the architectural solutions adopted by visual cortical cells in mammals. There are, indeed, several experimental evidences that cortical neurons with a specific sensitivity to retinal disparities play a key role in the perception of stereoscopic depth [4, 5]. Though, to date, it is not completely known the way in which cortical neurons measure stereo disparity and motion information. Recently, we showed [6] that the two measures can be placed into a common framework considering a phase-based disparity encoding scheme.

In this paper, we present a cortical-like (neuromorphic) strategy to obtain reliable MID estimations from local measurements of binocular parameters derived from direct com-

$$\delta(t) = (x_L^P - x_R^P) \approx a(D - Z_P)f/D^2$$

$$\delta(t + \Delta t) = (x_L^Q - x_R^Q) \approx a(D - Z_Q)f/D^2$$

$$V_Z \approx \frac{\Delta\delta}{\Delta t} D^2/af$$

$$\begin{aligned} \frac{\Delta\delta}{\Delta t} &= \frac{\delta(t + \Delta t) - \delta(t)}{\Delta t} \\ &= \frac{(x_L^Q - x_L^P) - (x_R^Q - x_R^P)}{\Delta t} \approx v_L - v_R \end{aligned}$$

$$V_Z \approx (v_L - v_R)D^2/af$$

parison of the results of monocular spatiotemporal filtering operations performed on stereo image pairs (see Section 2). This approach is suitable for a hardware implementation (see Section 3), in which such parameters can be gained via a feedforward computation (i.e., collection, comparison, and punctual operations) on the outputs of the nodes of recurrent VLSI lattice networks which have been proposed [7, 8, 9, 10] as efficient computational structures for embedded analog filtering operations in smart vision sensors. Extensive simulations on both synthetic and real-world image sequences prove the validity of the approach (see Section 4) that allows to gain high-level information about the 3D structure of the scene, directly from sensorial data, without resorting to explicit scene reconstruction (see Section 5).

## 2. PHASE-BASED DYNAMIC STEREOPSIS

### 2.1. Disparity as phase difference

According to the Fourier shift theorem, a spatial shift of  $\delta$  in the image domain effects a phase shift of  $k\delta$  in the Fourier domain. On the basis of this property, several researchers [11, 12] proposed phase-based techniques in which disparity is estimated in terms of phase differences in the spectral components of the stereo image pair. Spatially localized phase measures can be obtained by filtering operations with complex-valued quadrature pair bandpass kernels (e.g., Gabor filters [13, 14]), approximating a local Fourier analysis on the retinal images. Considering a complex Gabor filter with

a peak frequency  $k_0$ :

$$h(x, k_0) = e^{-x^2/\sigma^2} e^{ik_0x}, \quad (1)$$

we indicate convolutions with the left and right binocular signals as

$$Q(x) = \rho(x)e^{i\phi(x)} = C(x) + iS(x), \quad (2)$$

where  $\rho(x) = \sqrt{C^2(x) + S^2(x)}$  and  $\phi(x) = \arctan[S(x)/C(x)]$  denote their amplitude and phase components, and  $C(x)$  and  $S(x)$  are the responses of the quadrature filter pair. In general, this type of local measurement of the phase results stable and a quasilinear behaviour of the phase vs. space is observed over relatively large spatial extents, except around singular points where the amplitude  $\rho(x)$  vanishes and the phase becomes unreliable [15]. This property of the phase signal yields good predictions of binocular disparity by

$$\delta(x) = \frac{\phi^L(x) - \phi^R(x)}{k(x)}, \quad (3)$$

where  $k(x)$  is the average *instantaneous frequency* of the bandpass signal, measured by using the phase derivative from the left and right filter outputs:

$$k(x) = \frac{\phi_x^L(x) + \phi_x^R(x)}{2}. \quad (4)$$

As a consequence of the linear phase model, the instantaneous frequency is generally constant and close to the tuning frequency of the filter ( $\phi_x \simeq k_0$ ), except near singularities where abrupt frequency changes occur as a function of spatial position. Therefore, a disparity estimate at a point  $x$  is accepted only if  $|\phi_x - k_0| < k_0\mu$ , where  $\mu$  is a proper threshold [15].

## 2.2. Dynamics of binocular disparity

When the stereopsis problem is extended to include time-varying images, one has to deal with the problem of tracking the monocular point descriptions or the 3D descriptions which they represent through time. Therefore, in general, dynamic stereopsis is the integration of two problems: static stereopsis and temporal correspondence [16]. Considering jointly the binocular spatiotemporal constraints posed by moving objects in the 3D space, the resulting dynamic disparity is defined as  $\delta(x, t) = \delta[x(t), t]$ , where  $x(t)$  is the trajectory of a point in the image plane. The disparity assigned to a point as a function of time is related to the trajectories  $x^R(t)$  and  $x^L(t)$  in the right and left monocular images of the corresponding point in the 3D scene. Therefore, dynamic stereopsis implies the knowledge of the position of objects in the scene as a function of time.

Extending to time domain the phase-based approach, the disparity of a point moving with the motion field can be estimated by

$$\delta[x(t), t] = \frac{\phi^L[x(t), t] - \phi^R[x(t), t]}{k_0}, \quad (5)$$

where phase components are computed from the spatiotemporal convolutions of the stereo image pair

$$Q(x, t) = C(x, t) + iS(x, t) \quad (6)$$

with directionally tuned Gabor filters with central frequency  $\mathbf{p} = (k_0, \omega_0)$ . For spatiotemporal locations where linear phase approximation still holds ( $\phi \simeq k_0x + \omega_0t$ ), the phase differences in (5) provide only spatial information, useful for reliable disparity estimates. Otherwise, in the proximity of singularities, an error occurs that is also related to the temporal frequency of the filter responses. In general, a more reliable disparity computation should be based on a combination of confidence measures obtained by a set of Gabor filters tuned to different velocities. Though, due to the robustness of phase information, good approximations of time-varying disparity measurements can be gained by a quadrature pair of Gabor filters tuned to null velocities ( $\mathbf{p} = (k_0, 0)$ ). A detailed analysis of the phase behaviour in the joint space-time domain, and of its confidence, in relation to the directional tuning of the Gabor filters, evades the scope of the present paper and it will be presented elsewhere.

## 2.3. Motion in depth

Perspective projections of a MID leads to different motion fields on the two retinas, that is a temporal variation of the disparity of a point moving with the flow observed by the left and right views (see Figure 1). The rate of change of such disparity provides information about the direction of MID and an estimate of its velocity. Disparity has been defined in Section 1 as  $I^L(x) = I^R[x + \delta(x)]$  with respect to the spatial coordinate  $x^L$ . Therefore, when differentiating (5) with respect to time, the total rate of variation of  $\delta$  is

$$\frac{d\delta}{dt} = \frac{\partial\delta}{\partial t} + \frac{v^L}{k_0}(\phi_x^L - \phi_x^R), \quad (7)$$

where  $v^L$  is the horizontal component of the velocity signal on the left retina. Considering the conservation property of local phase measurements, image velocities can be computed from the temporal evolution of constant phase contours [17]:

$$\phi_x^L = -\frac{\phi_t^L}{v^L}, \quad \phi_x^R = -\frac{\phi_t^R}{v^R}. \quad (8)$$

Combining (8) with (7), we obtain

$$\frac{d\delta}{dt} = \frac{\phi_x^R}{k_0}(v^R - v^L), \quad (9)$$

where  $(v^R - v^L)$  is the phase-based interocular velocity difference. When the spatial tuning frequency of the Gabor filter  $k_0$  approaches the instantaneous spatial frequency of the left and right convolution signals, one can derive the following approximated expressions:

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi_t^L - \phi_t^R}{k_0} \simeq v^R - v^L. \quad (10)$$

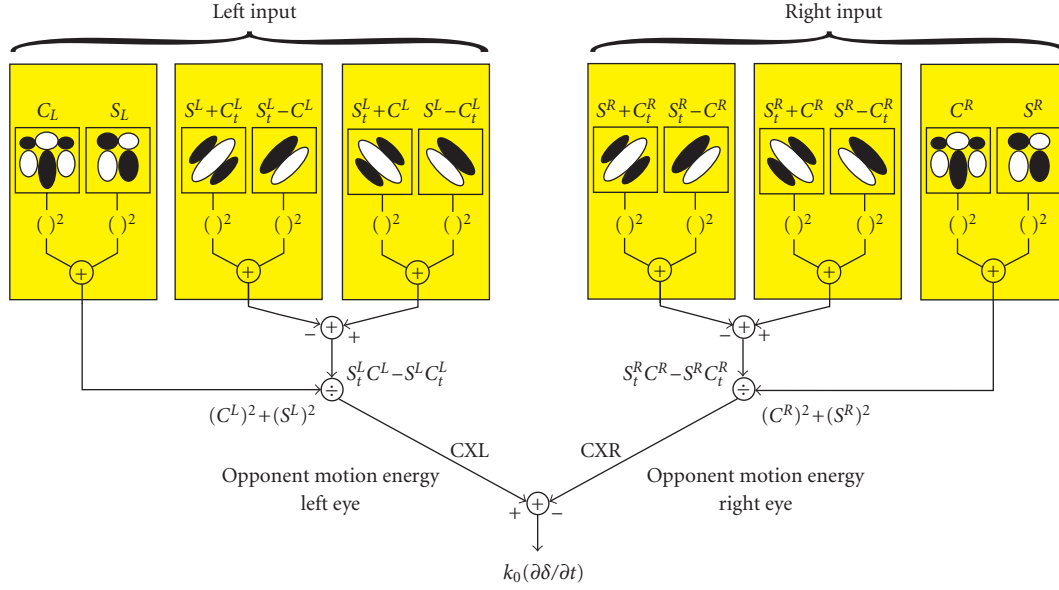


FIGURE 2: Cortical architecture of a MID detector. The rate of variation of disparity can be obtained by a direct comparison of the responses of two monocular units labelled CXL and CXR. Each monocular unit receives contributions from a pair of directionally tuned “energy” complex cells that compute phase temporal derivative ( $S_t C - S C_t$ ) and a nondirectional complex cell that supplies the “static” energy of the stimulus ( $C^2 + S^2$ ). Each monocular branch of the cortical architecture can be directly compared to the Adelson-Bergen motion detector, thus establishing a link between phase-based approaches and motion energy models.

It is worth noting that the approximations depend on the robustness of phase information, and the error made is the same as the one which affects the measurement of phase components around singularities [15, 17]. Hence, on a local basis, valuable predictions about MID can be made, without tracking, through phase-based operators which need not to know the direction of motion on the image plane  $x(t)$ .

The partial derivative of the disparity can be directly computed by convolutions ( $S, C$ ) of stereo image pairs and by their temporal derivatives ( $S_t, C_t$ ):

$$\frac{\partial \delta}{\partial t} = \left[ \frac{S_t^L C^L - S^L C_t^L}{(S^L)^2 + (C^L)^2} - \frac{S_t^R C^R - S^R C_t^R}{(S^R)^2 + (C^R)^2} \right] \frac{1}{k_0}, \quad (11)$$

thus avoiding explicit calculation and differentiation of phase, and the attendant problem of phase unwrapping. Moreover, the direct determination of temporal variations of the disparity, through filtering operations, better tolerates the problem of the limit on maximum disparities due to “wrap-around” [11], yielding correct estimates even for disparities greater than one half the wavelength of the central frequency of the Gabor filter.

## 2.4. Spatiotemporal operators

Since numerical differentiation is very sensitive to noise, proper regularized solutions have to be adopted to compute correct and stable numerical derivatives. As a simple way to avoid the undesired effects of noise, band-limited filters can be used to filter out high frequencies that are amplified by differentiation. Specifically, if one prefilters the image signal

to extract some temporal frequency subband

$$S(x, t) \simeq f_1 * S(x, t), \quad C(x, t) \simeq f_1 * C(x, t) \quad (12)$$

and evaluates the temporal changes in that subband, time differentiation can be attained by convolutions on the data with appropriate bandpass temporal filters:

$$S'(x, t) \simeq f_2 * S(x, t), \quad C'(x, t) \simeq f_2 * C(x, t), \quad (13)$$

where  $S'$  and  $C'$  approximate  $S_t$  and  $C_t$ , respectively, if  $f_1$  and  $f_2$  approximate a quadrature pair of temporal filters, for example,

$$f_1(t) = e^{-t/\tau} \sin \omega_0 t, \quad f_2(t) = e^{-t/\tau} \cos \omega_0 t. \quad (14)$$

This formulation allows a certain degree of robustness of our MID estimates.

By rewriting the terms of the numerators in (11):

$$\begin{aligned} 4S_t C &= (S_t + C)^2 - (S_t - C)^2, \\ 4S C_t &= (S + C_t)^2 - (S - C_t)^2, \end{aligned} \quad (15)$$

one can express the computation of  $\partial \delta / \partial t$  in terms of convolutions with a set of oriented spatiotemporal filters whose shapes resemble simple cell receptive fields of the primary visual cortex [18]. Specifically, each square term on the right-hand sides of (15) is a component of a directionally tuned *energy detector* [19]. The overall MID cortical detector can be built as shown in Figure 2. Each branch represents a monocular opponent motion energy unit of Adelson-Bergen type where divisions by the responses of separable spatiotemporal

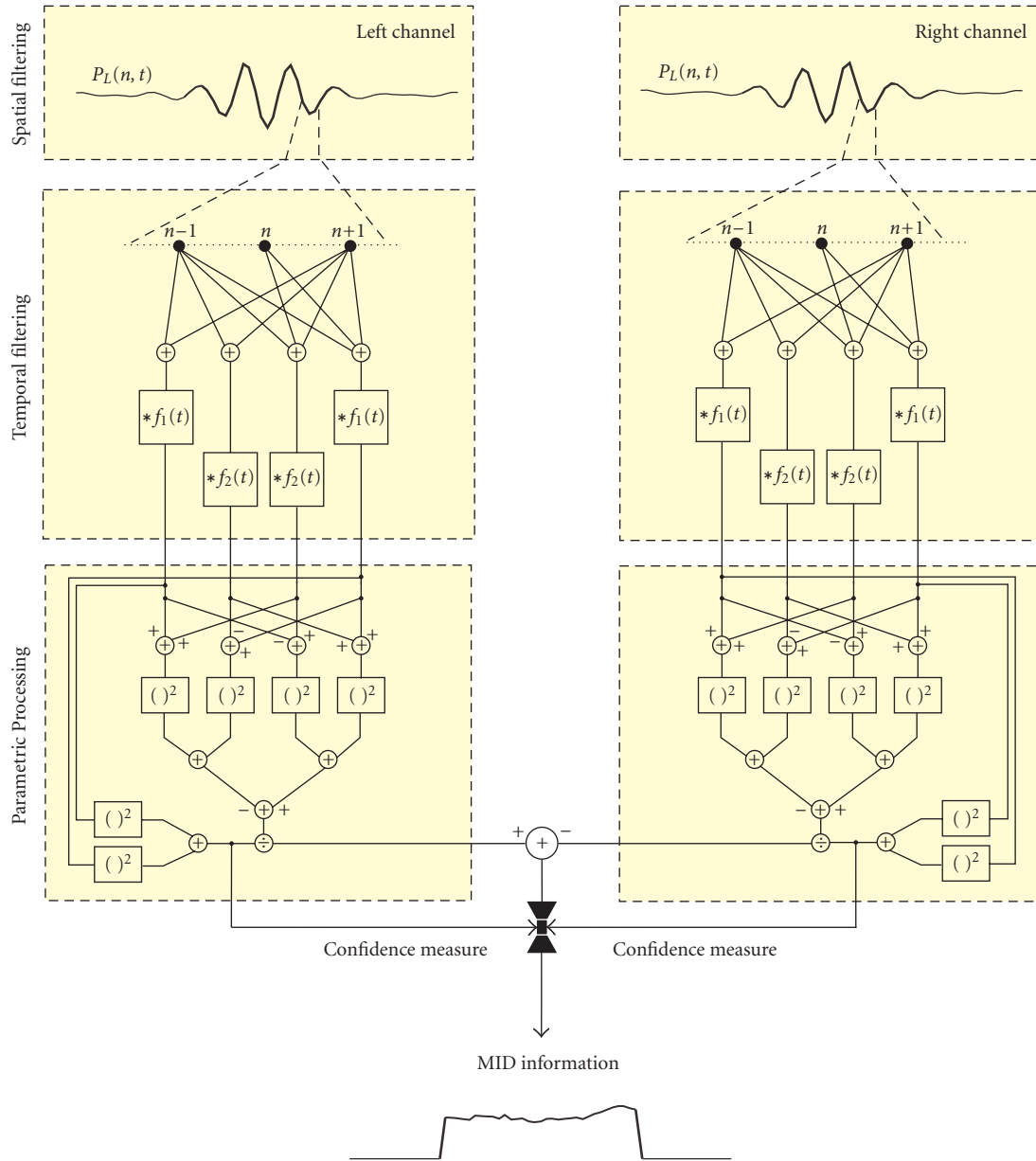


FIGURE 3: Architectural scheme of the neuromorphic MID detector.

filters (see the denominators of (11)) approximate measures of velocity that are invariant with contrast. We can extract a measure of the rate of variation of local phase information by taking the arithmetic difference between the left and right channel responses. Further division by the tuning frequency of the Gabor filter yields a quantitative measure of MID. It is worth noting that phase-independent motion detectors of Adelson and Bergen can be used to compute temporal variations of phase. This result is consistent with the assumption we made of the linearity of the phase model. Therefore, our model evidences a novel aspect of the relationships existing between energy and phase-based approaches to motion modeling to be added to those already presented in the literature [17, 20].

### 3. TOWARDS AN ANALOG VLSI IMPLEMENTATION

In the neuromorphic scheme proposed above, we can evidence two different processing stages (see Figure 3): (1) spatiotemporal convolutions with 1D Gabor kernels that extract amplitude and phase spectral components of the image signals, and (2) punctual operations such as sums, squarings, and divisions that yield the resulting percept. These computations can be supported by neuromorphic architectural resources organized as arrays of interacting nodes. In the following, we will present a circuit hardware implementation of our MID detector based on analog perceptual microsystems. Following the Adelson-Bergen model [19] for motion-sensitive cortical cell receptive fields, spatiotemporal oriented



filters can be constructed by pairs of separable (i.e., not oriented) filters. In this way, filters tuned to a specific direction can be obtained through a proper cascading combination of spatial and temporal filters (see Figure 3), thus decoupling the design of the spatial and temporal components of the motion filter [21, 22].

### Spatial filtering: the perceptual engine

It has been demonstrated [8, 9, 10] that image convolutions with 1D Gabor-like kernels can be made isomorphic to the behaviour of a second-order lattice network with diffusive excitatory nearest couplings and next nearest neighbors inhibitory reactions among nodes. Figure 4a shows a block representation of such network when one encodes all signals—stimuli and responses—by currents:  $I_s(n)$  is the input current (i.e., stimulus),  $I_e(n)$  is the output current (i.e., response), and the coefficients  $G$  and  $K$  represent the excitatory and inhibitory couplings among nodes, respectively. At circuitual level, each node is fed by a current generator whose value is proportional to the incident light intensity at that point and the interaction among nodes is implemented by current-controlled current sources (CCCSs) that feed or sink currents according to the actual current response at neighboring nodes. Each computational node has two output currents  $GI_e(n)$  toward the first nearest nodes and two (negative) output currents  $KI_e(n)$  toward the second nearest nodes, and receives the corresponding contributions from its neighbors, besides its input  $I_s(n)$ . The circuit representation of a node is based on the use of CCCSs with the desired current gains  $G$  and  $K$ . A CMOS transistor level implementation of a cell is illustrated in Figure 4b. The spatial impulse response of the network  $g(n)$  can be interpreted as the *perceptual engine* of the system since it provides a computational primitive that can be composed to obtain more powerful image descriptors. Specifically, by combining the responses of neighboring nodes, it is possible to obtain Gabor-like functions of any phase  $\varphi$ :

$$\begin{aligned} h(n) &= \alpha g(n-1) + \beta g(n) + \gamma g(n+1) \\ &= De^{-\lambda|n|} \cos(2\pi k_0 n + \varphi), \end{aligned} \quad (16)$$

where  $D$  is a normalization constant,  $\lambda$  is the decay rate, and  $k_0$  is the oscillating frequency of the impulse response. The values of  $\lambda$  and  $k_0$  depend on the interaction coefficients  $G$  and  $K$ . The phase  $\varphi$  depends on  $\alpha$ ,  $\beta$ , and  $\gamma$ , given the values of  $\lambda$  and  $k_0$ . The decay rate and frequency, though hardwired in the underlying perceptual engine, can be controlled by adjustable circuit parameters [23].

### Temporal filtering

The signal processing requirements specified by (14) in the time domain provide the functional characterization of the filter blocks  $f_1$  and  $f_2$  shown in Figure 3. The Laplace transforms of the impulse responses determine the desired trans-

fer functions:

$$\begin{aligned} \mathcal{L}\{e^{-t/\tau} \sin \omega_0 t\} &= \frac{\omega_0}{(s + 1/\tau)^2 + \omega_0^2}, \\ \mathcal{L}\{e^{-t/\tau} \cos \omega_0 t\} &= \frac{(s + 1/\tau)}{(s + 1/\tau)^2 + \omega_0^2}. \end{aligned} \quad (17)$$

They are (temporal) filters of the second order with the same characteristic equation. The pole locations determine the frequency peak and the bandwidth. The magnitude and phase responses of these filters are shown in Figure 5: they have nearly identical magnitude responses and a phase difference of  $\pi/2$ . The choice of the filter parameters is performed on the basis of typical psychophysical perceptual thresholds [24]:  $\omega_0 = 6\pi$  rad/seconds and  $\tau = 0.13$  second.

The circuitual implementation of these filters can be based on continuous-time current-mode integrators [25]. The same two-integrator-loop circuitual structure can be shared for realizing the two filters [26].

### Spatiotemporal processing

By taking appropriate sums and differences of the temporally convoluted outputs of a second-order lattice network  $P_{L/R}(n, t) \stackrel{\text{def}}{=} \int I^{L/R}(n', t) h(n-n') dn'$ , it is possible to compute convolutions with cortical-like spatiotemporal operators:

$$\begin{aligned} S(n, t) &= [\alpha_1 P(n-1, t) + \beta_1 P(n, t) + \gamma_1 P(n+1, t)] * f_1(t), \\ C(n, t) &= [\alpha_2 P(n-1, t) + \beta_2 P(n, t) + \gamma_2 P(n+1, t)] * f_1(t), \\ S_t(n, t) &= [\alpha_1 P(n-1, t) + \beta_1 P(n, t) + \gamma_1 P(n+1, t)] * f_2(t), \\ C_t(n, t) &= [\alpha_2 P(n-1, t) + \beta_2 P(n, t) + \gamma_2 P(n+1, t)] * f_2(t), \end{aligned} \quad (18)$$

where  $\alpha_1 = -\gamma_1 = De^{-\lambda}(e^{-2\lambda} - 1) \cos 2\pi k_0$ ,  $\beta_1 = 0$ ,  $\alpha_2 = \gamma_2 = De^{-\lambda}(e^{-2\lambda} - 1) \cos 2\pi k_0$ , and  $\beta_2 = D(1 - e^{-4\lambda})$ .

### Parametric processing

The high information content of the parameters provided by the spatiotemporal filtering units makes it possible to use them directly (i.e., feedforward) via a feedforward computation (i.e., collection, comparison, and punctual operations). The distinction between local and punctual data is particularly relevant when one considers the medium used for their representation with respect to the processing steps to be performed. In the approach followed in this work, local data is the result of a distributed processing on lattice networks whose interconnections have a local extension. Conversely, the output data from these processing stages can be treated in a punctual way, that is, according to standard computational schemes (sequential, parallel, pipeline), or still resorting to analog computing circuits. In this way, one can take full advantage of the potentialities of analog processing together with the flexibility provided by digital hardware.

### 3.1. The intrinsic dynamics of spatial filtering

In this Section, we discuss the temporal properties of the spatial array and analyze how its intrinsic temporal behaviour

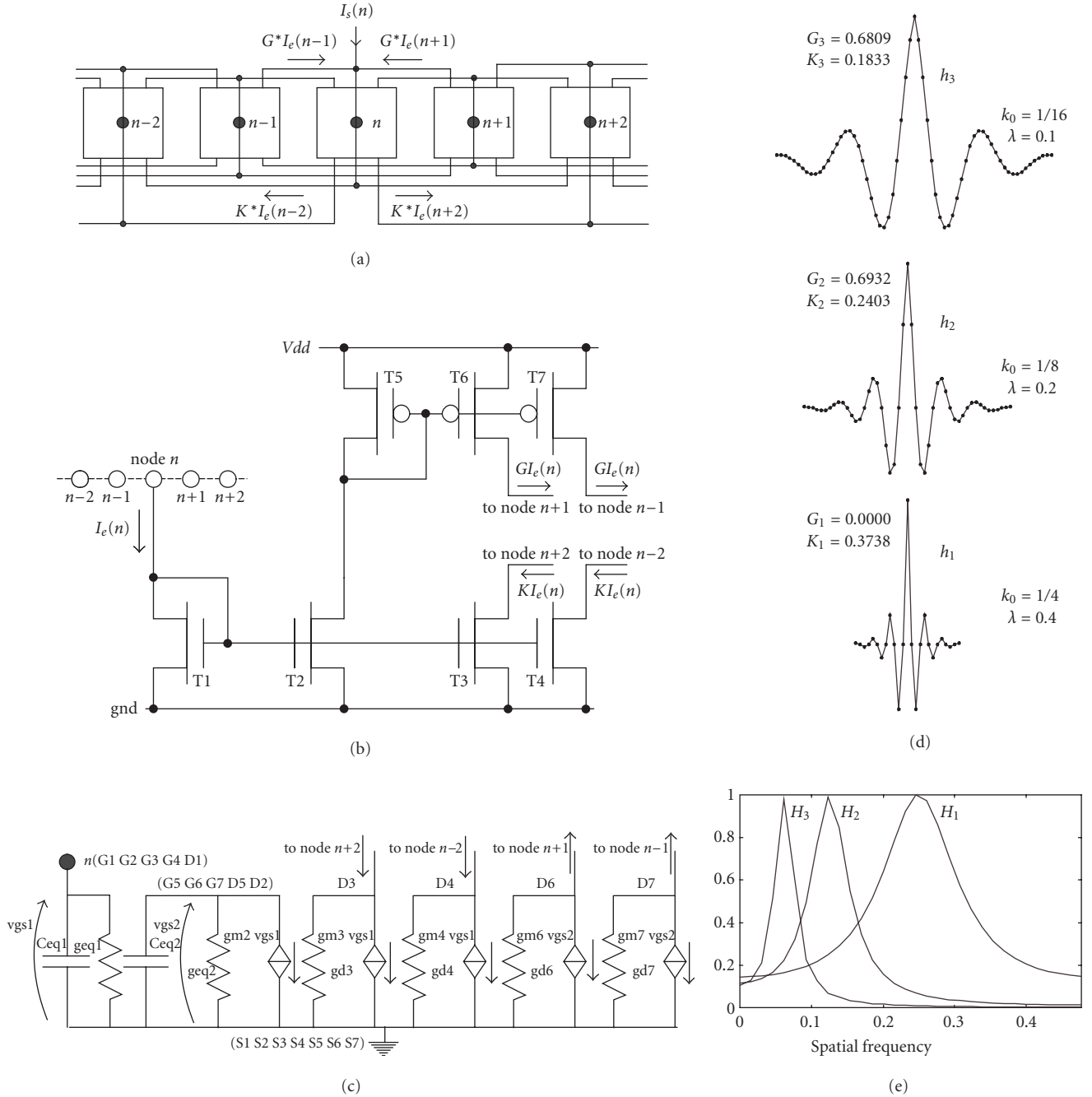


FIGURE 4: Spatial filtering. (a) Second-order lattice network represented as an array of cells interacting through currents. (b) Transistor-level representation of a single computational cell; (c) its small-signal circuitual representation. (d-e) Spatial and spatial-frequency plots of the three Gabor-like filters considered; the filters have been chosen to have in the frequency-domain constant octave bandwidth.

could affect the spatial processing. More specifically, we focus our analysis on how the array of interacting nodes modifies its spatial filtering characteristics, when the stimuli signals vary in time at a given frequency  $\omega$ . In relation to the architectural solution adopted for motion estimation, we will require that the spatial filter would still behave as a bandpass spatial filter for temporal frequencies up to and beyond  $\omega_0$  (see (14) and Figure 5). To perform this check, we consider

the small-signal low-frequency representation of the MOS transistor, governed by the gate-source capacitance. Our circuitual implementation of the array will be characterized by two  $C/g_m$  time constants (Figure 4c). Other implementations in the literature, for example, [27], are adequately modeled with a single time constant; as shown below the present analysis will cover both types of implementations. The intrinsic spatiotemporal transfer function of the array will then have

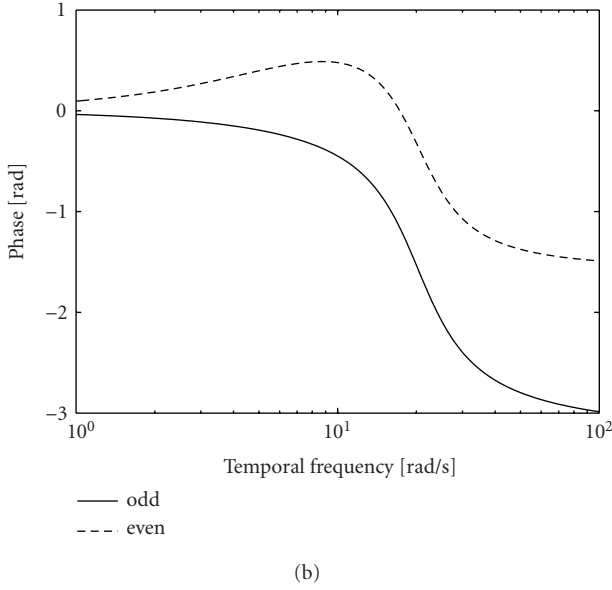
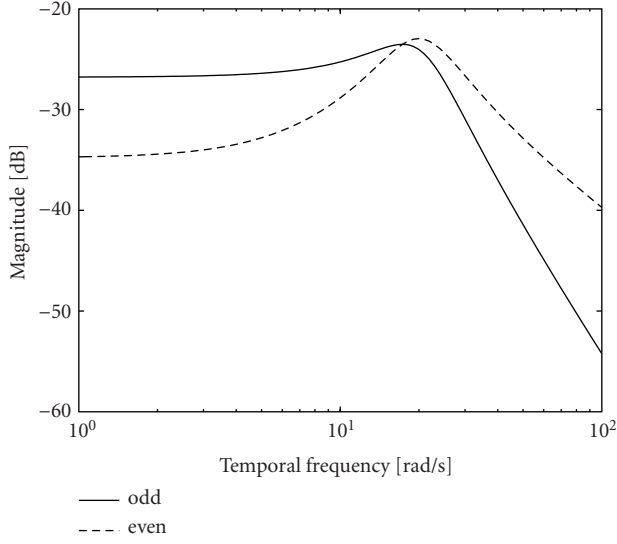


FIGURE 5: (a) The magnitude and (b) phase plots for the even and odd temporal filters used ( $\omega_0 = 6\pi$  rad/s and  $\tau = 0.13$  s).

the following form:

$$H(k, \omega_n) = \frac{L(\omega_n)}{M(k, \omega_n) + jN(k, \omega_n)} \quad (19)$$

with

$$\begin{aligned} L(\omega_n) &= 1 - \omega_n^2 \rho + j\omega_n(1 + \rho), \\ M(k, \omega_n) &= 1 - 2G \cos(2\pi k) - \omega_n^2 \rho + 2K \cos(4\pi k), \\ N(k, \omega_n) &= \omega_n[1 + \rho + 2\rho K \cos(4\pi k)], \end{aligned} \quad (20)$$

where  $\omega_n = \omega \tau_1$  is the normalized temporal frequency,  $\rho = \tau_2/\tau_1$ .

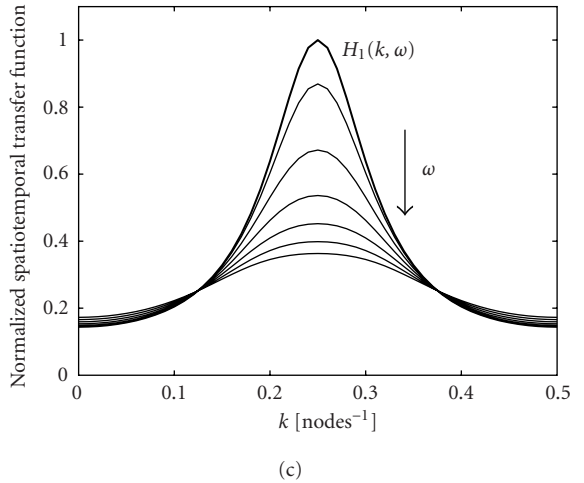
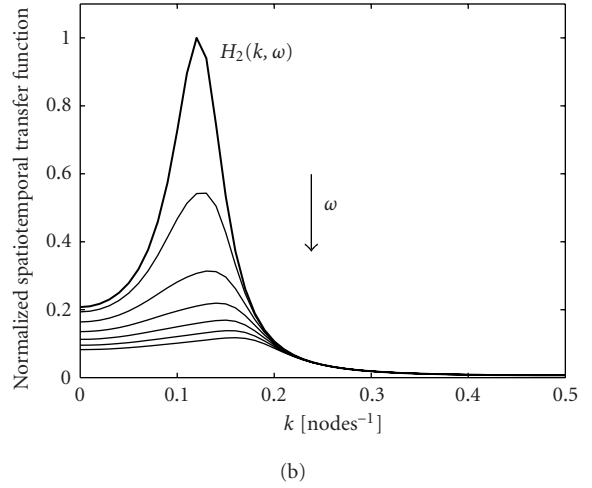
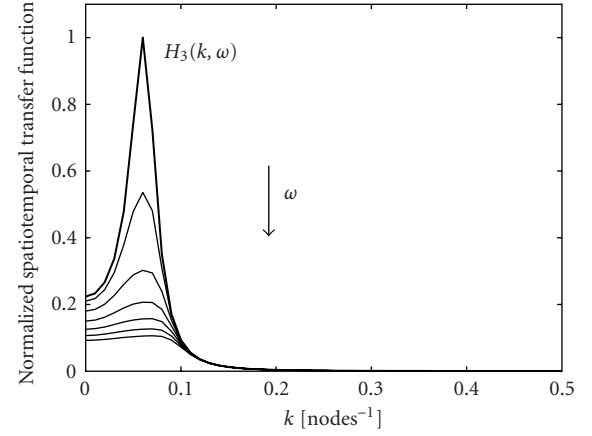


FIGURE 6: The intrinsic spatiotemporal transfer function of the analog lattice networks implementing Gabor-like spatial filters, designed for bandpass spatial operation; the three considered types of filters are those introduced in Figures 4d and 4e. The curves, normalized to the peak value of the static transfer function and parametrized with respect to the temporal frequency  $\omega$ , describe how the spatial filtering is modified when the input stimulus varies with time.



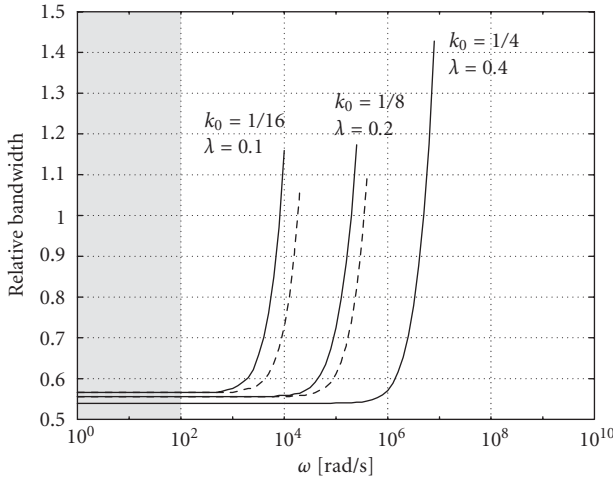


FIGURE 7: The overall equivalent lattice network relative spatial bandwidth as a function of the input stimulus temporal frequency, for the time constant characteristic of the interaction among cells  $\tau_1 = 10^{-7}$  second. Solid and dashed curves describe the effect of the ratio of the two time constants. The shaded region evidences the temporal bandwidth of perceptual tasks.

Figure 6 shows the spatial frequency behaviour of the array for three values of their central frequency, spanning a two-octave range:  $k_0 = 1/16, 1/8, 1/4$ . In all three cases, when the temporal frequency increases, the array tends to maintain its bandpass character up to a limit frequency, beyond which it assumes a low-pass behaviour. A more accurate description of the modifications that occur is presented in Figure 7. For each spatial filter, characterized by the behavioural parameters  $(k_0, \lambda)$ , or, in an equivalent manner, by the structural parameters  $(G, K)$ , we consider its spatial performance when the stimulus signal varies in time. At any temporal frequency we can characterize the spatial filtering as a bandpass processing step, taking note of the value of the effective relative bandwidth, at  $-3$  dB points. Figure 7 reports the result of such analysis for the three filters considered. We can observe that the array maintains the spatial frequency character it has for static stimuli, up to a frequency that basically depends on the time constant,  $\tau_1$ , of its interaction couplings, and in a more complex way on the strength  $G$  and  $K$  of these couplings. We can note that the higher is the static gain at the central frequency of the spatial filter, the higher is the overall equivalent time constant of the array. This effect has to be related to the fact that high gains in the spatial filter are the result of many-loop recurrent processing.

We can also evidence the effect of the ratio  $\tau_2/\tau_1$  on the overall performance. We compare for this purpose solid and dashed curves. The solid ones are traced with  $\tau_1 = \tau_2$  and the dashed ones with  $\tau_2 = 0$ . It is worth noting that when  $k_0 = 1/4$  the interaction coefficient  $G$  is null and the ratio  $\tau_2/\tau_1$  is not influent on the transfer function.

If we consider the typical temporal bandwidth of perceptual tasks [28] and assume the value of  $\tau_1$  in the range of  $10^{-7}$  second, we can conclude that the neuromorphic lattice

network adopted for spatial filtering has an intrinsic temporal dynamics more than adequate for performing visual tasks on motion estimation.

#### 4. RESULTS

We consider a  $65 \times 65$ -pixel target implementation of our neuromorphic architecture—compatible with current hardware constraints—and we test its performance at system level through extensive simulations on both synthetic and real-world image sequences.

The output of the MID detector provides a measure of  $\partial\delta/\partial t$  (i.e.,  $V_Z$ ), except for the proportionality constant  $k_0$ . We evaluate the correctness of the estimation of  $V_Z$  for the three considered Gabor-like filters ( $k_0 = 1/4, k_0 = 1/8$ , and  $k_0 = 1/16$ ). We use random dot stereogram sequences where a central square moves forward and backward on a static background with the same pattern. The 3D motion of the square results in opposite horizontal motions of its projections on the left and right retinas, as evidenced in Figure 8a. The resulting estimates of  $V_Z$  (see Figures 8b, 8c, and 8d) are derived from the measurements of the interocular velocity differences ( $v^L - v^R$ ) obtained by our architecture, taking into account the geometrical parameters of the optic system: fixation distance  $D = 1$  m, focal length  $f = 0.025$  m, and interpupillary distance  $a = 0.13$  m. The estimation of the velocity in depth  $V_Z$  should be always considered jointly with a confidence measure related to the binocular average energy value of the filtering operations [ $\rho = (\rho^L + \rho^R)/2$ ]. When the below confidence is a given threshold (in our case the 10% of the energy peak), the estimates of  $V_Z$  are considered unreliable and therefore are discarded (see grayed regions in Figures 8b, 8c, and 8d). We observe that estimates of  $V_Z$  with high confidence values are always correct.

It is worth noting that in those circumstances, where it is not important to perform a quantitative measure on  $V_Z$  but it is sufficient to discriminate its sign, all the necessary information is “mostly” contained in the numerators of (11) since the denominators are of the same order when the confidence values are high. In this case, the architecture of the MID detector can be simplified by removing the two normalization stages on each monocular branch, thus saving two divisions and four squaring operations for each pixel. The results on correct discrimination between forward and backward movements of objects from the observer are shown in Figure 9 for a real-world stereo sequence. Also in this case, points where phase information is unreliable are discarded, according to the confidence measure, and represented as static.

#### 5. CONCLUSION

The general context in which this research can be framed concerns the development of artificial systems with cognitive capabilities, that is, systems capable of collecting information from the environment, of analyzing and evaluating them to properly react. To tackle these issues, an approach that

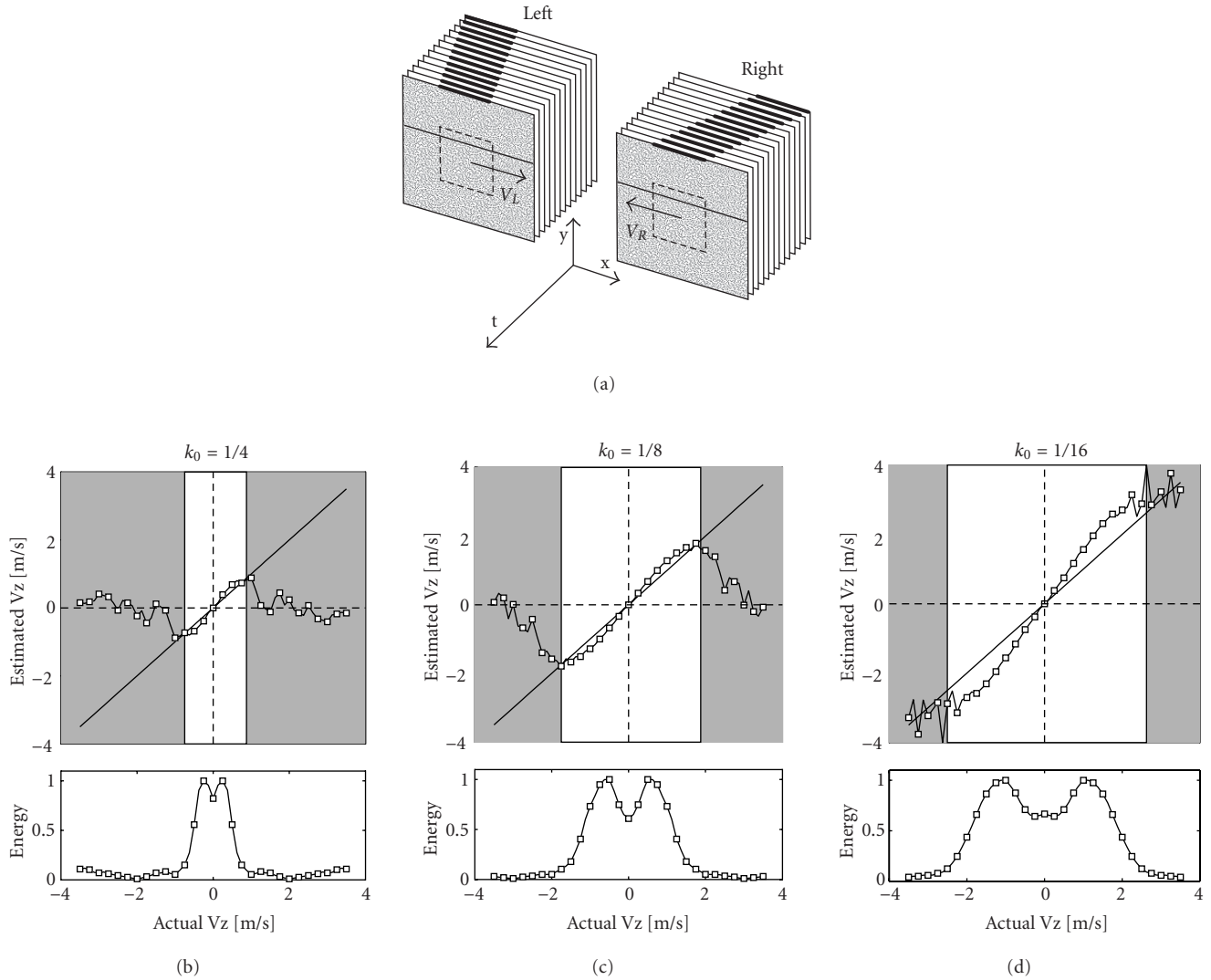


FIGURE 8: Results on synthetic images. (a) Schematic representation of the random dot stereogram sequences where a central square moves, with speed  $V_z$ , forward and backward with respect to a static background with the same random pattern. (b–d) The upper plots show the estimated speed as a function of the actual speed  $V_z$  for the three considered Gabor-like filters ( $k_0 = 1/4$ ,  $k_0 = 1/8$ , and  $k_0 = 1/16$ ); the lower plots show the binocular average energy taken as a confidence measure of the speed estimation. The ranges of  $V_z$  for which the confidence goes below 10% of the maximum are evidenced in the gray shading.

finds increasing favour is the one which establishes a bidirectional relation with brain sciences, from one side, transferring the knowledge from the studies on biological systems toward artificial ones (developing hardware, software, and wetware models that capture architectural and functional properties of biological systems) and, on the other side, using artificial systems as tools for investigating the neural system. Considering more specifically vision problems, this approach pays attention to the architectural scheme of visual cortex that, with respect to the more traditional computational schemes, is characterized by the simultaneous presence of different levels of abstraction in the representation and computation of signals, hierarchically/structurally organized and interacting in a recursive and adaptive way [29, 30]. In

this way, high-level vision processing can be rethought in structural terms by evidencing novel strategies to allow a more direct (i.e., structural) interaction between early vision and cognitive processes, possibly leading to a reduction of the gap between PDP and AI paradigms. These neuromorphic paradigms can be employed by new artificial vision systems, in which a “novel” integration of bottom-up (data-driven) and top-down approaches occurs. In this way, it is possible to perform perceptual/cognitive computations (such as those considered in this paper) by properly combining the outputs of receptive fields characterized by specific selectivities, without introducing explicitly a priori information. The specific vision problem tackled in this paper is the binocular perception of MID. The assets of the approach can be considered

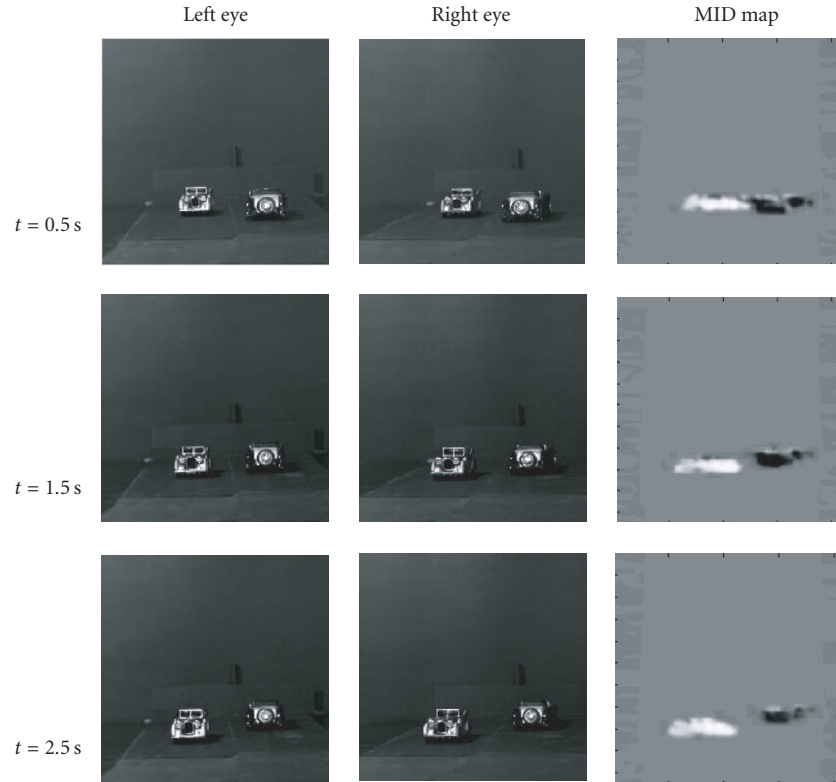


FIGURE 9: Experimental results on a natural scene. Two toy cars are moving in opposite directions with respect to the observer. Left and right frames at three different times are shown. The gray levels in the MID maps code the MID of the two cars: the lighter gray blob represents the car moving toward the observer, whereas the darker gray blob represents the car moving away. The background gray level represents points discarded according to the confidence measure. The few still present error points do not impair the interpretation of the MID map.

under different perspectives: modeling, computational, and implementation.

### Modeling

Psychophysical studies evidenced that perception of MID can be based on binocular cues such as interocular velocity differences or temporal variations of binocular disparity [3]. We analytically demonstrated that information hold in the interocular velocity difference is the same of that derived by the evaluation of the total derivative of the binocular disparity if a phase-based disparity encoding scheme is assumed.

### Computational

By exploiting the chain rule in the evaluation of the temporal derivative of phases, one can obtain information about MID directly from the convolutions of the two stereo images with complex spatiotemporal bandpass filters. This formulation eliminates the need for an explicit trigonometric function to compute the phase signal from  $Q(x, t)$ , thus avoiding also problems arising from phase unwrapping and discontinuities. Moreover, the approximation of temporal derivatives by temporal filtering operations yields to regularized solutions in which noise sensitivity is reduced.

### Implementation

The algorithmic approach followed allows a fully analog computation of MID through spatiotemporal filtering with quadrature pairs of Gabor kernels, that can be directly implemented in VLSI, as demonstrated by recent prototypes of our group [10]. Simulations have been performed to analyze the effects on system performance of constraints posed by analog and digital hardware implementation.

### ACKNOWLEDGMENTS

We wish to thank L. Raffo and G. M. Bo for their contribution to the accomplishment of this work. This work was partially supported by EU Project IST-2001-32114 *ECOVISION* "Artificial vision systems based on early cognitive cortical processing."

### REFERENCES

- [1] O. Faugeras, *Three Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, Cambridge, Mass, USA, 1993.
- [2] C. E. Thorpe, *Vision and Navigation: The Carnegie Mellon Navlab*, Kluwer Academic Publishers, Boston, Mass, USA, 1990.

- [3] J. Harris and S. N. J. Watamaniuk, "Speed discrimination of motion-in-depth using binocular cues," *Vision Research*, vol. 35, no. 7, pp. 885–896, 1995.
- [4] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Encoding of binocular disparity by simple cells in the cat's visual cortex," *J. Neurophysiol.*, vol. 75, no. 5, pp. 1779–1805, 1996.
- [5] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, "Encoding of binocular disparity by complex cells in the cat's visual cortex," *J. Neurophysiol.*, vol. 77, no. 6, pp. 2879–2909, 1997.
- [6] S. P. Sabatini, F. Solari, G. Andreani, C. Bartolozzi, and G. M. Bisio, "A hierarchical model of complex cells in visual cortex for the binocular perception of motion-in-depth," in *Proc. Neural Information Processing Systems (NIPS '01)*, pp. 1271–1278, Vancouver, British Columbia, Canada, December 2001.
- [7] B. E. Shi, T. Roska, and L. O. Chua, "Design of linear cellular neural networks for motion sensitive filtering," *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 40, no. 5, pp. 320–331, 1993.
- [8] L. Raffo, "Resistive network implementing maps of Gabor functions of any phase," *Electronics Letters*, vol. 31, no. 22, pp. 1913–1914, 1995.
- [9] B. Shi, "Gabor-type filtering in space and time with cellular neural networks," *IEEE Trans. on Circuits and Systems I*, vol. 45, no. 2, pp. 121–132, 1998.
- [10] L. Raffo, S. P. Sabatini, G. M. Bo, and G. M. Bisio, "Analog VLSI circuits as physical structures for perception in early visual tasks," *IEEE Transaction on Neural Networks*, vol. 9, no. 6, pp. 1483–1494, 1998.
- [11] T. D. Sanger, "Stereo disparity computation using Gabor filters," *Biological Cybernetics*, vol. 59, pp. 405–418, 1988.
- [12] A. D. Jenkin and M. Jenkin, "The measurement of binocular disparity," in *Computational Processes in Human Vision*, Z. Pylyshyn, Ed., Ablex Publishing, Norwood, NJ, USA, 1988.
- [13] D. Gabor, "Theory of communication," *Journal of the Institute of Electrical Engineers*, vol. 93, no. 3, pp. 429–459, 1946.
- [14] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, no. 7, pp. 1160–1169, 1985.
- [15] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *Computer Vision, Graphics and Image Processing: Image Understanding*, vol. 53, no. 2, pp. 198–210, 1991.
- [16] M. Jenkin and J. K. Tsotsos, "Applying temporal constraints to the dynamic stereo problem," *Computer Vision, Graphics and Image Processing: Image Understanding*, vol. 33, no. 1, pp. 16–32, 1986.
- [17] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, no. 1, pp. 77–104, 1990.
- [18] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Receptive-field dynamics in the central visual pathways," *Trends in Neurosciences*, vol. 18, no. 10, pp. 451–458, 1995.
- [19] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Journal of the Optical Society of America*, vol. 2, no. 2, pp. 284–299, 1985.
- [20] N. Qian and S. Mikaelian, "Relationship between phase and energy methods for disparity computation," *Neural Computation*, vol. 12, no. 2, pp. 279–292, 2000.
- [21] G. M. Bisio, G. M. Bo, M. Confalone, L. Raffo, S. P. Sabatini, and M. P. Zizola, "A current-mode computational engine for stereo disparity and early vision tasks," in *Proc. International Conference on Microelectronics for Neural Networks and Fuzzy Systems (MicroNeuro '97)*, pp. 83–90, Dresden, Germany, September 1997.
- [22] R. Etienne-Cummings, J. Van der Spiegel, and P. Mueller, "Hardware implementation of a visual motion pixel using oriented spatiotemporal neural filters," *IEEE Trans. Circuits and Systems II*, vol. 46, no. 9, pp. 1121–1136, 1999.
- [23] G. M. Bisio, L. Raffo, and S. P. Sabatini, "Analog VLSI primitives for perceptual tasks in machine vision," *Neural Computing & Applications*, vol. 7, pp. 216–228, 1998, Special Issue on Machine Vision.
- [24] J. G. Robson, "Spatial and temporal contrast sensitivity functions of the visual system," *Journal of the Optical Society of America*, vol. 56, no. 8, pp. 1141–1142, 1966.
- [25] M. Ismail and T. Fiez, *Analog VLSI Signal and Information Processing*, McGraw-Hill, New York, NY, USA, 1994.
- [26] J. Silva-Martinez, M. Steyaert, and W. Sansen, *High-Performance CMOS Continuous-Time Filters*, Kluwer Academic Publishers, Boston, Mass, USA, 1993.
- [27] B. E. Shi, "A one-dimensional CMOS focal plane array for Gabor-type image filtering," *IEEE Trans. on Circuits and Systems I: Fundamental Theory and Applications*, vol. 46, no. 2, pp. 323–327, 1999.
- [28] V. Bruce, P. R. Green, and M. A. Georgeson, *Visual Perception: Physiology, Psychology, and Ecology*, Psychology Press, Hove, East Sussex, UK, 3rd edition, 1996.
- [29] T. J. Sejnowski, C. Koch, and P. S. Churchland, "Computational neuroscience," *Science*, vol. 241, pp. 1299–1306, 1988.
- [30] G. Deco and B. Schurmann, "A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition," *Vision Research*, vol. 40, pp. 2845–2859, 2000.

**Silvio P. Sabatini** graduated in electronic engineering from the University of Genoa, Italy (1992). He received his Ph.D. degree in electronic engineering and computer science from the Department of Biophysical and Electronic Engineering (DIBE) of Genoa University (1996). Since 1999, he is an Assistant Professor in computer science at the University of Genoa. In 1995, he promoted the creation of the "Physical Structure of Perception and Computation" (PSPC) Research Group at the DIBE to develop models that capture the "physicalist" nature of the information processing that takes place in the visual cortex, to understand the signal processing strategies adopted by the brain, and to build novel algorithms and hardware devices for artificial perception machines. His current research interests include biocybernetics of vision, theoretical neuroscience, neuromorphic engineering, and artificial vision. He is an author of more than 50 international papers in peer-reviewed journals and conferences.



**Fabio Solari** received the Laurea degree in electronic engineering from the University of Genoa, Italy, in 1995. In 1999, he obtained his Ph.D. degree in electronic engineering and computer science from the same University. He is currently a Postdoctoral Fellow at the Department of Biophysical and Electronic Engineering (DIBE), University of Genoa. His research activity concerns the study of the physical processes of biological vision to inspire the design of artificial perceptual machines based on neuromorphic computational paradigms. In particular, he is interested in cortical modelling, dynamic stereopsis, visual motion analysis, and probabilistic modelling.



**Paolo Cavalleri** was born in 1973. He received the M.S. degree in electronic engineering from the University of Genoa, Italy, in 1999. He is currently working toward the Ph.D. degree in electronic engineering and computer science at the Department of Biophysical and Electronic Engineering (DIBE), Genoa, Italy. His research activity concerns cortical modelling, visual motion analysis, and artificial vision.



**Giacomo Mario Bisio** is a Full Professor of microelectronics at the School of Engineering, University of Genoa, and a member of the Department of Biophysical and Electronic Engineering (DIBE). He teaches courses on Electronic Measurements and Models of Perceptual Systems, and contributes to the Graduate Program on electronic engineering and computer science. Born in Genoa, Italy, in 1940, he graduated in electronic engineering from the University of Genoa in 1965, and received the M.S. degree in electrical engineering from Stanford University, USA, in 1971. He was “Alessandro Volta” Research Fellow at the Microwave Laboratory of Stanford University (1969–1972) and a CNR Scientist at IROE Institute and ICE Institute, Genoa (1968–1983). He has been a Lecturer at the School of Engineering, University of Genoa, since 1972. Formerly, he was a Secretary of the CNR-CCTE National Group, Director of the DIBE, member of the program committee, and Chairperson of Eurochip Workshop on VLSI Design Training, Neuro-Nimes, NEURAP, and EUSIPCO. He is a member of the AEI and IEEE. He was awarded the AEI “E. Bottani” medal for contributions to the teaching of electronics. He coauthored more than 150 papers in refereed journals and conferences. His present researches concern microsystems considered as physical structures for perception and computation. The activities of his laboratory (PSPC-Lab) are described at [www.pspc.dibe.unige.it](http://www.pspc.dibe.unige.it).

