Phase-based Binocular Perception of Motion-in-depth: Cortical-like Operators and aVLSI Architectures

Silvio P. Sabatini, Fabio Solari, P. Cavalleri and Giacomo M. Bisio

Department of Biophysical and Electronic Engineering, University of Genoa

Via Opera Pia 11a - 16145 Genova - ITALY   silvio@dibe.unige.it

Abstract

In this paper we present a cortical-like strategy to obtain reliable estimates of the motions of objects in a scene toward-to/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 feed-forward 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 3-D structure of the scene, directly from sensorial data, without resorting to explicit scene reconstruction.

Keywords: cortical architectures, phase-based dynamic stereoscopy, motion processing, Gabor filters, lattice networks.
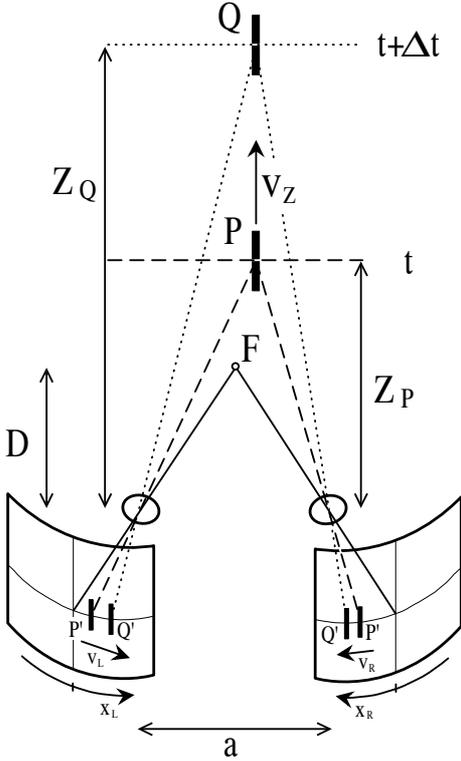
1   Introduction

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

To better introduce the subject, let us briefly consider the dynamic correspondence problem in the stereo image pairs acquired by a binocular vision system. Fig. 1 shows the relationships between an object moving in 3-D space and the geometrical projection of the image in the right and left retinas. 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 Fig. 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 1-D horizontal shift, the binocular disparity $\delta(x)$. Formally, the left and right observed intensities from the two eyes, respectively $I^L(x)$ and $I^R(x)$, result related as $I^L(x) = I^R[x + \delta(x)]$. 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's 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 Fig. 1. The predominance of one measure on the other one 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 comparison of the results of monocular spatiotemporal filtering operations performed on

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

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

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

$$\frac{\Delta \delta}{\Delta t} = \frac{\delta(t + \Delta t) - \delta(t)}{\Delta t} =$$

$$= \frac{(x_L^Q - x_L^P) - (x_R^Q - x_R^P)}{\Delta t} \approx v_L - v_R$$

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

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

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 feed-forward 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 3-D 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 δ 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:

$$h(x, k_0) = e^{-x^2/\sigma^2} e^{ik_0 x}, \tag{1}$$

we indicate convolutions with the left and right binocular signals as

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

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 $C(x), S(x)$ are the responses of the quadrature filter pair. Local phase measurements result stable and with a quasi-linear behaviour 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

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

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

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

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 μ 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 3-D 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 3-D 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 3-D 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 ($\phi \simeq k_0 x + \omega_0 t$), the phase differences in Eq. (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 motion in depth 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 Fig. 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 Eq. (5) with respect to time, the total rate of variation of δ is:

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

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]:

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

Combining Eq. (8) with Eq. (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:

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

It is worthy to note 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
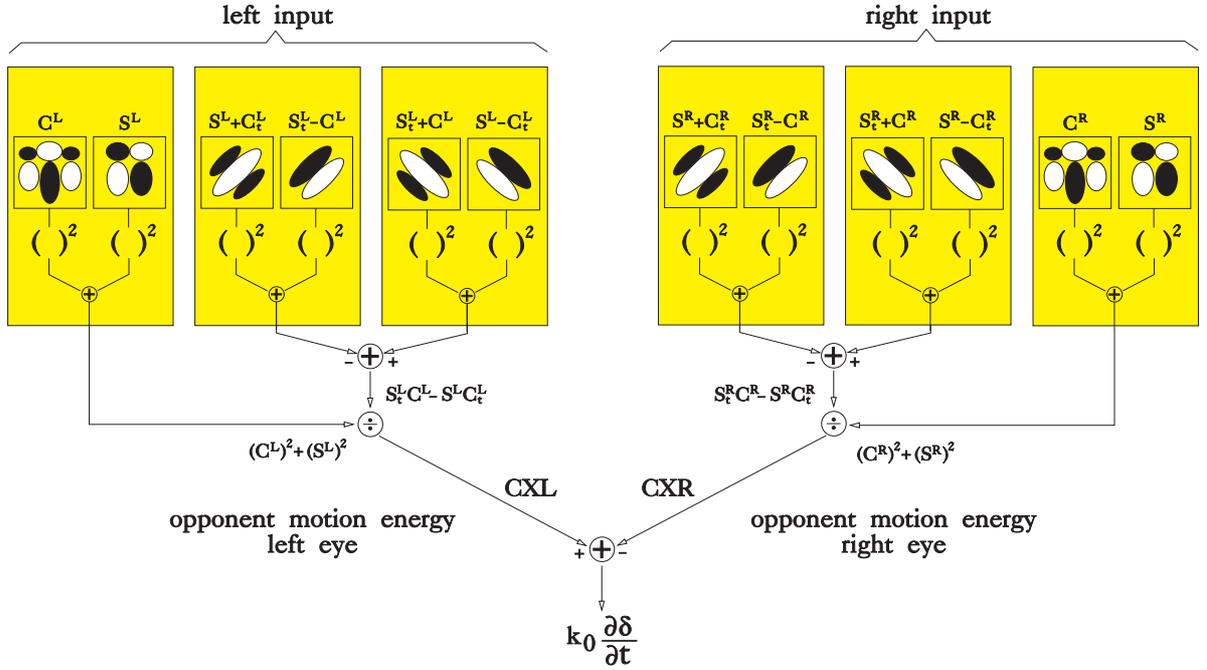
Figure 2: Cortical architecture of a motion-in-depth detector. The rate of variation of disparity can be obtained by a direct comparison of the responses of two monocular units labelled CXL and CXR. Each monocular unit receives contributions from a pair of directionally tuned "energy" complex cells that compute phase temporal derivative $(S_t C - S C_t)$ and a non-directional 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 and Bergen's motion detector, thus establishing a link between phase-based approaches and motion energy models (see text).

to compute correct and stable numerical derivates. 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 sub-band,

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

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

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

S' and C' approximate S_t and C_t, respectively, if f_1 and f_2 approximate a quadrature pair of temporal filters, e.g.:

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

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

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

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

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 sides of Eqs.(15)

is a component of a directionally tuned *energy detector* [19]. The overall MID cortical detector can be built as shown in Fig. 2. Each branch represents a monocular opponent motion energy unit of Adelson and Bergen's type where divisions by the responses of separable spatiotemporal filters (cf. the denominators of Eq.(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 worthy to note 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 Fig. 3): (1) spatiotemporal convolutions with 1-D Gabor ker-
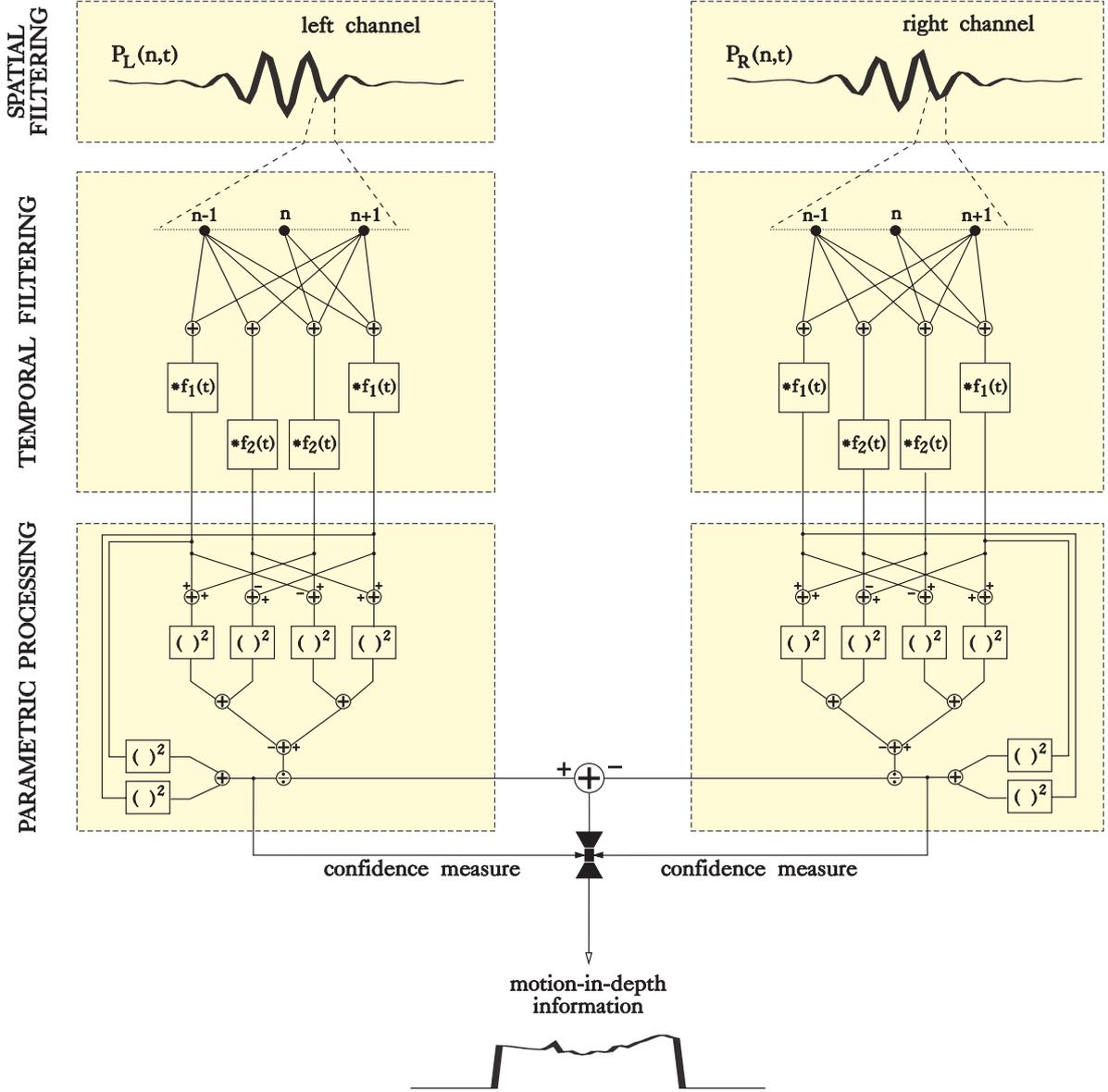
Figure 3: Architectural scheme of the neuromorphic motion-in-depth detector.

nels 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 shall present a circuit hardware implementation of our MID detector based on analog perceptual microsystems. Following the Adelson and Bergen's 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 Fig. 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 1-D Gabor-like kernels can be made isomorphic to the behaviour of a 2nd-order lattice network with diffusive excitatory nearest couplings and next nearest neighbors inhibitory reactions among nodes. Fig. 4a shows a block reprentation 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 circuital level, each node is fed by a current generator whose value is proportional to the incident light intensity at that point, 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
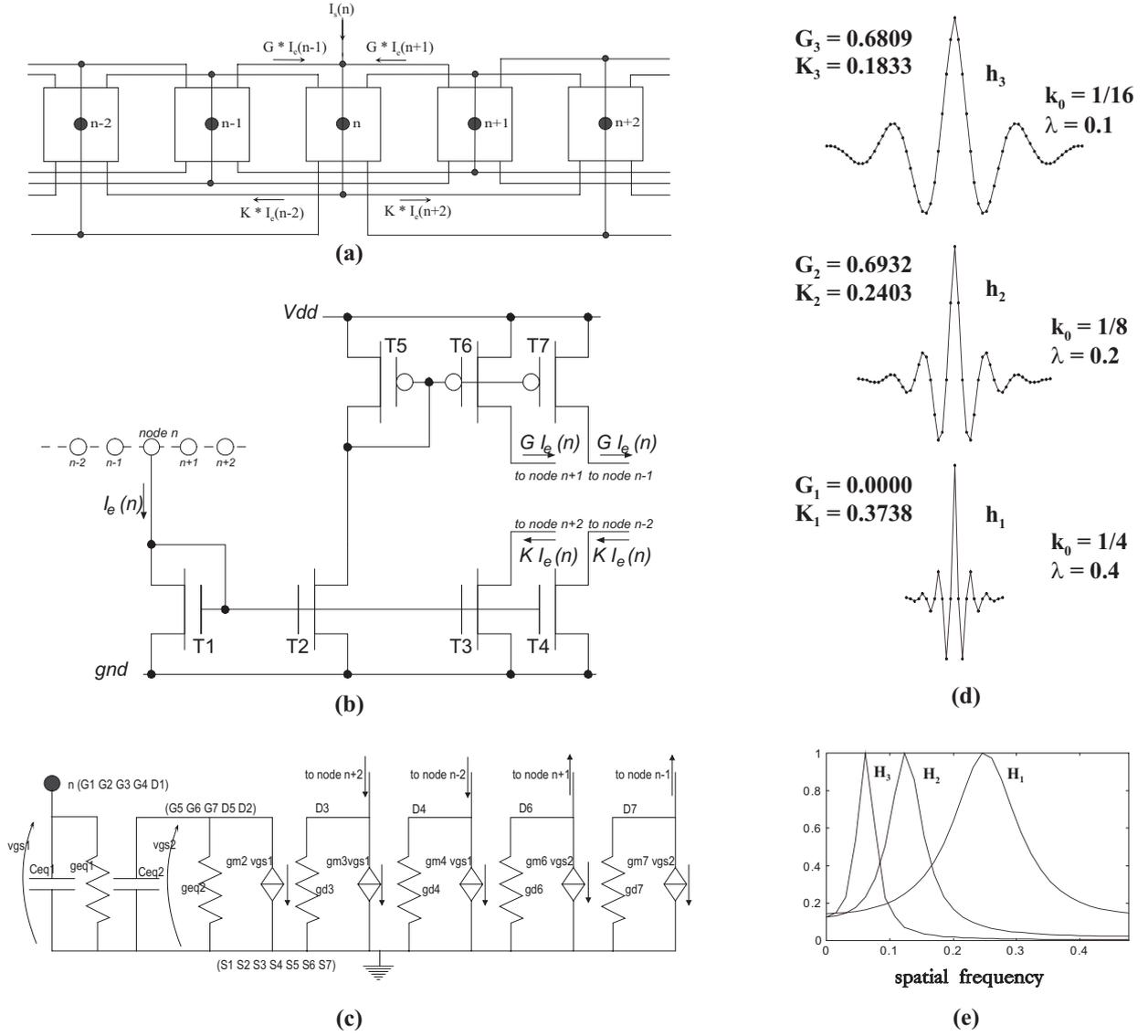
5

Figure 4: Spatial filtering. (a) 2nd-order lattice network represented as an array of cells interacting through currents. (b) Transistor level representation of a single computational cell. (c) Small-signal circuital representation of a single cell. (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.

computational node has two output currents $GI_e(n)$ toward the 1st nearest nodes, two (negative) output currents $KI_e(n)$ toward the 2nd 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 Fig. 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 φ:

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

where λ is the decay rate and k_0 is the oscillating frequency of the impulse response. The values of λ and k_0 depend on the interaction coefficients G and K. The phase φ depends on α, β, γ, given the values of λ and k_0; D is a normalization constant. The decay rate and frequency, though hard-wired in the underlying perceptual engine, can be controlled by adjustable circuit parameters [23].

Temporal filtering The signal processing requirements specified by Eq. 14 in the time domain provide the functional characterization of the filter blocks f_1 and f_2 shown in Fig. 3. The Laplace transforms of
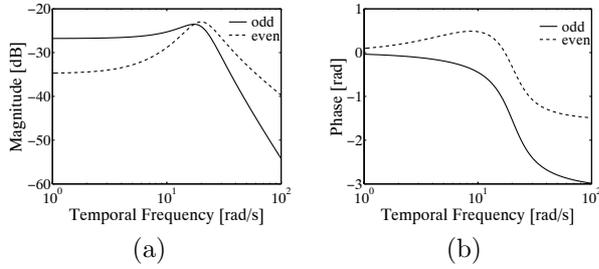
6

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

the impulse responses determine the desired transfer functions:

$$\mathcal{L}\left\{e^{-t/\tau}\sin\omega_0 t\right\} = \frac{\omega_0}{(s+1/\tau)^2+\omega_0^2}$$

$$\mathcal{L}\left\{e^{-t/\tau}\cos\omega_0 t\right\} = \frac{(s+1/\tau)}{(s+1/\tau)^2+\omega_0^2}.$$

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 Fig. 5a,b: 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/s, $\tau = 0.13$ s.

The circuital implementation of these filters can be based on continuous-time current-mode integrators [25]. The same two-integrator-loop circuital 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 2nd-order lattice network $P_{L/R}(n,t) \overset{def}{=} \int I^{L/R}(n',t) * h(n-n')dn'$, it is possible to compute convolutions with cortical-like spatiotemporal operators:

$$S(n,t) = [\alpha_1 P(n-1,t) + \beta_1 P(n,t) + \gamma_1 P(n+1,t)] * f_1(t)$$
$$C(n,t) = [\alpha_2 P(n-1,t) + \beta_2 P(n,t) + \gamma_2 P(n+1,t)] * f_1(t)$$
$$S_t(n,t) = [\alpha_1 P(n-1,t) + \beta_1 P(n,t) + \gamma_1 P(n+1,t)] * f_2(t)$$
$$C_t(n,t) = [\alpha_2 P(n-1,t) + \beta_2 P(n,t) + \gamma_2 P(n+1,t)] * f_2(t)$$

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

Parametric processing The high information content of the parameters provided by the spatiotemporal filtering units, makes it possible their direct (i.e., feed-forward) use *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 are 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, i.e., according to standard computational schemes (sequential, parallel, pipeline), or still resorting to analog computing circuits. In this way, one can take 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 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 ω. In relation to the architectural solution adopted for motion estimation, we will require that the spatial filter would still behave as a band-pass spatial filter for temporal frequencies up to and beyond ω_0 (see eq. 14, and Fig. 5). To perform this check, let us consider the small-signal low-frequency representation of the MOS transistor, governed by the gate-source capacitance. Our circuital implementation of the array will be characterized by two C/g_m time constants (Fig 4c). Other implementations in the literature, e.g. [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 have then the following form:

$$H(k,\omega_n) = \frac{L(\omega_n)}{M(k,\omega_n)+jN(k,\omega_n)} \qquad (17)$$

with

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

where $\omega_n = \omega\tau_1$ is the normalized temporal frequency, $\rho = \tau_2/\tau_1$, $\tau_1 = C_{eq1}/g_{eq1}$ and $\tau_2 = C_{eq2}/g_{eq2}$.

Fig. 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 band-pass 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 Fig. 7. For each spatial filter, characterized by the behavioural parameters (k_0, λ), or, in a 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 band-pass processing step, taking note of the value of the effective relative bandwidth, at

7

-3 dB points. Fig. 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, τ_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 τ_2/τ_1 on the overall performance. Let us compare for this purpose, solid and dashed curves. The solid ones are traced with $\tau_1 = \tau_2$, 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 τ_2/τ_1 is not influent on the transfer function.

If we consider the typical temporal bandwidth of perceptual tasks [28], and assume the value of τ_1 in the range of $10^{-7}s$, 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 65×65 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 Gabor-like filters considered ($k_0 = 1/4$, $k_0 = 1/8$, $k_0 = 1/16$). We use random dot stereogram sequences where a central square moves forward and backward on a static background with the same pattern. The 3-D motion of the square results in opposite horizontal motions of its projections on the left and right retinas, as evidenced in Fig. 8a. The resulting estimates of V_Z (see Figs. 8b,c,d) 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 confidence is below a given threshold (in our case the 10% of the energy peak), the estimates of V_Z are considered unreliable and therefore are discarded (cf. grayed regions in Figs. 8b-d). We observe that estimates of V_Z with high confidence values are always correct.

It is worthy to note that in those circumstances where it is not important to perform a quantitative
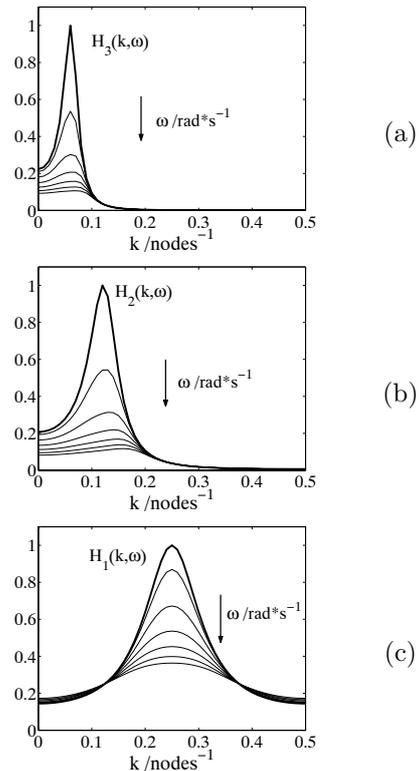

Figure 6: The intrinsic spatiotemporal transfer function of the analog lattice networks implementing Gabor-like spatial filters, designed for band-pass spatial operation; the three types of filters considered are those introduced in Fig. 4d,e. The different curves, parametrized respect to the temporal frequency ω, describe how the spatial filtering is modified when the input stimulus varies with time (see text).
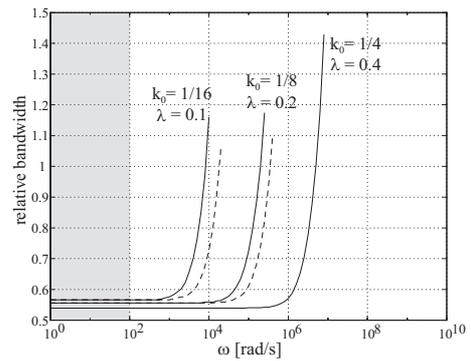

Figure 7: The overall equivalent lattice network relative spatial bandwith as a function of the input stimulus temporal frequency, for the time constant characteristic of the interaction among cells $\tau_1 = 10^{-7}s$. Solid and dashed curves describe the effect of the ratio of the two time constants (see text). The shaded region evidences the temporal bandwidth of perceptual tasks.
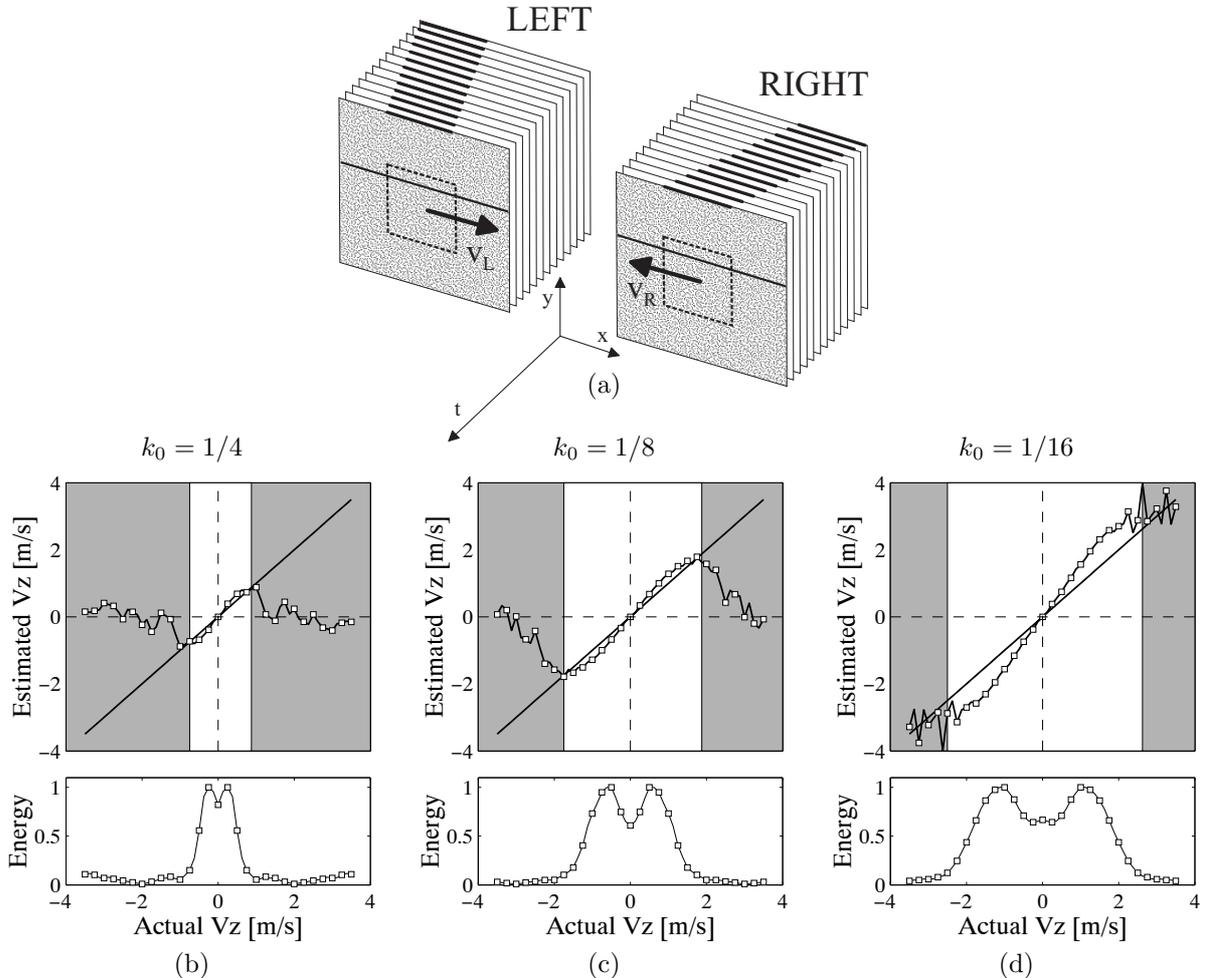
8

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 respect to a static background with the same random pattern. (b,c,d) The upper plots show the estimated speed as a function of the actual speed V_Z for the three Gabor-like filters considered ($k_0 = 1/4$, $k_0 = 1/8$, $k_0 = 1/16$); 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.

measure on V_Z, but it is sufficient to discriminate its sign, all the necessary information is "mostly" contained in the numerators of Eq. 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 Fig. 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, i.e., systems capable of collecting information from the environment, of analyzing and evaluating them to properly react. To tackle these issues, an approach that finds increasing favour is the one which establishes a bi-directional relation with brain sciences, from one side, trasferring 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, are 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-

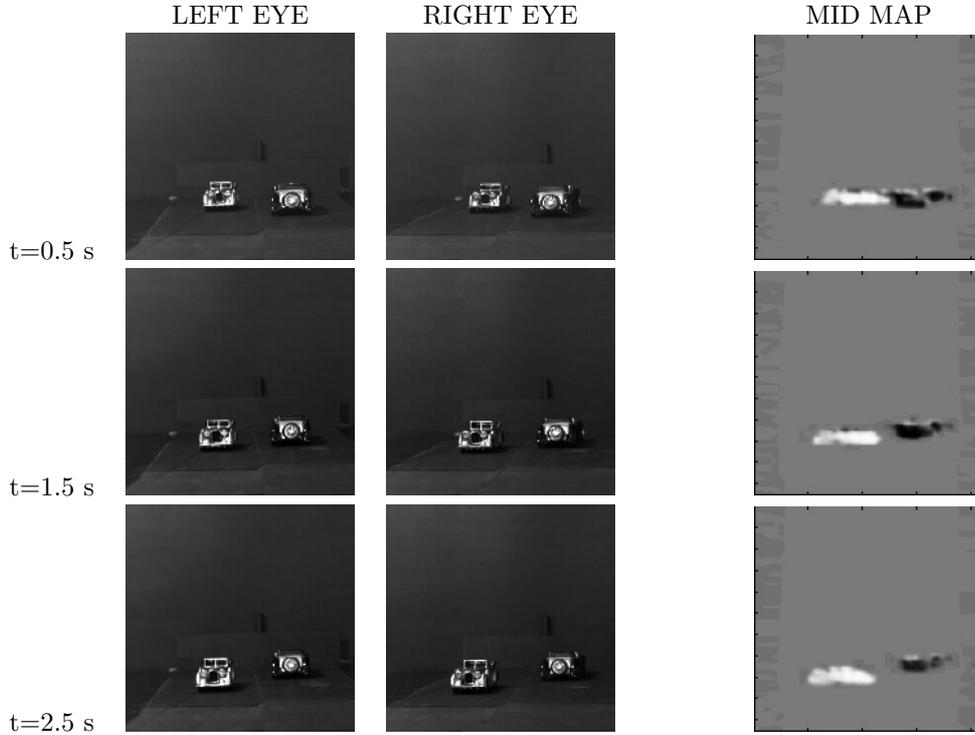|	LEFT EYE	RIGHT EYE	MID MAP
t=0.5 s			
t=1.5 s			
t=2.5 s			

Figure 9: Experimental results on a natural scene. Two toy cars are moving in opposite directions respect to the observer. Left and right frames at three different times are shown. The gray levels in the MID maps code the motion-in-depth 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.

level vision processing can be re-thought 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 motion-in-depth. The assets of the approach can be considered 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 demonstrated analytically 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

References

[1] O Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint.* MIT Press, 1993.

[2] C. E. Thorpe. *Vision and Navigation: The Carnegie Mellon Navlab.* Kluwer Academic Publishers, Boston, Massachusetts, 1990.

[3] J. Harris and S. N.J. Watamaniuk. Speed discrimination of Motion-in depth using binocular cues. *Vision Research*, 35(7):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.*, 75:1779–1805, 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.*, 77:2879–2909, 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. NIPS'01*, Vancouver, CA, 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 Circuit and Systems: II. Analog and Digital Signal Processing*, 40:320–331, 1993.

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

[9] B. Shi. Gabor-type filtering in space and time with cellular neural networks. *IEEE Trans. Circuits and System I*, 45:121–132, 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 Trans. Neural Net.*, 9(6):1483–1494, 1998.

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

[12] A.D. Jenkin and M. Jenkin. The measurement of binocular disparity. In Z. Pylyshyn, editor, *Computational Processes in Human Vision*. Ablex Publ., New Jersey, 1988.

[13] D. Gabor. Theory of communication. *J. Inst. Elec. Eng.*, 93:429–459, 1946.

[14] 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.

[15] D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210, 1991.

[16] M. Jenkin and J.K. Tsotsos. Applying temporal constraints to the dynamic stereo problem. *CVGIP*, 33:16–32, 1986.

[17] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 1:77–104, 1990.

[18] G.C. DeAngelis, I. Ohzawa, and R.D. Freeman. Receptive-field dynamics in the central visual pathways. *Trends in Neurosci.*, 18:451–458, 1995.

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

[20] N. Qian and S. Mikaelian. Relationship between phase and energy methods for disparity computation. *Neural Comp.*, 12(2):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. MicroNeuro'97*, pages 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 System II*, 46:1121–1136, 1999.

[23] G.M. Bisio, L. Raffo, and S.P. Sabatini. Analog VLSI primitives for perceptual tasks in machine vision. *Neural Computing & Applications, special issue on Machine Vision*, 7:216–228, 1998.

[24] J.G. Robson. Spatial and temporal contrast sensitivity functions of the visual system. *J. Opt. Soc. Am.*, 56:1141–1142, 1966.

[25] M. Ismail and T. Fiez. *Analog VLSI signal and information processing.* Mc Graw-Hill International Editions, 1994.

[26] J. Silva-Martínez, M. Steyaert, and W. Sansen. *High-performance CMOS Continuous-time filters.* Kluwer Academic Publishers, 1993.

[27] B.E. Shi. A one-dimensional CMOS focal plane array for gabor-type image filtering. *IEEE Trans. on Circuit and Systems: I. Fundamental Theory and Applications*, 46:323–327, 1999.

[28] V. Bruce, P.R. Green, and M.A. Georgeson. *Visual Perception - Physiology, Psychology, and Ecology, 3rd ed.* Psychology Press, 2000.

[29] T.J. Sejnowski, C. Koch, and P.S. Churchland. Computational neuroscience. *Science*, 241:1299–1306, 1988.

[30] G. Deco. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vis. Res.*, 40:2845–2859, 2000.