Spatiotemporal Neuromorphic Operators for the Detection of Motion-in-depth

Silvio P. Sabatini, Fabio Solari, 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

A cortical-like model for the binocular perception of motion-in-depth, based on phase-based measurements of dynamic disparity, is presented. The analysis evidenced 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. The resulting architecture resorts to spatiotemporal operators organized as motion energy detectors that can be directly implemented in analog VLSI lattice networks. Experimental simulations on stereo sequences validated the approach.

1 Introduction

In many real-world visual application domains it is important to "measure" dynamic 3-D visual information from 2-D images impinging the retinas. In particular, the perception of motion-in-depth (MID), i.e. 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 upon a global analysis of the optic flow or on token matching techniques which combine stereo correspondence and visual tracking. In this paper, we demonstrate that reliable MID estimations can be obtained from local measurements of binocular parameters derived from direct comparison of the results of monocular spatiotemporal filtering operations performed on stereo image pairs. The approach is particularly suitable for an "economic" hardware implementation, since 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 are recently proposed as efficient computational structures for embedded analog filtering operations in smart vision sensors [1].

2 Phase-based dynamic stereopsis

Stereoscopy is a technique used to extract depth information from a pair of images taken from slightly different view points. Depth perception derives from the differences in the positions of corresponding points in the stereo image pair projected on the retinas. In a first approximation, the positions of corresponding points are related by a 1-D horizontal shift, the *disparity*, along the direction of the epipolar lines. Formally, the left and right observed intensities from the two eyes, respectively $I^L(x)$ and $I^R(x)$, result related as:

$$I^L(x) = I^R[x + \delta(x)] \tag{1}$$

where $\delta(x)$ is the (horizontal) binocular disparity.

2.1 Disparity as phase difference

Following [2] and [3], disparity can be estimated in terms of phase differences in the spectral components of the stereo image pair. Since the two images are locally related by a shift, in the neighborhood of each image point the local k spectral components of $I^L(x)$ and $I^R(x)$ are related by a phase difference equal to $\Delta\phi(k) = \phi^L(k) - \phi^R(k) = k\delta$. Spatially-localized phase measures can be obtained by filtering operations with complex-valued quadrature pair bandpass kernels (e.g. Gabor filters), approximating a local Fourier analysis on the retinal images. Considering a complex Gabor filter with a peak frequency k_0:

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

the resulting convolutions with the left and right binocular signals can be expressed as

$$Q(x) = \rho(x)\mathrm{e}^{i\phi(x)} = C(x) + iS(x) \tag{3}$$

where $$\rho(x) = \sqrt{C^2(x) + S^2(x)} \tag{4}$$

and $$\phi(x) = \arctan\left[\frac{S(x)}{C(x)}\right] \quad (5)$$

denote their amplitude and phase components and $C(x)$ and $S(x)$ are the responses of the quadrature pair of filters. 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 [4]. This property of the phase signal yields good predictions of binocular disparity by

$$\delta(x) = \frac{\phi^L(x) - \phi^R(x)}{k(x)} \quad (6)$$

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

$$k(x) = \frac{\phi_x^L(x) + \phi_x^R(x)}{2}. \quad (7)$$

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 $|\phi_x - k_0| < k_0\tau$ where τ is a proper threshold [4].

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 [5]. Considering jointly the binocular spatiotemporal constraints posed by moving objects in the 3-D space, the resulting dynamic disparity is defined as $\delta(x,t) = \delta[x(t),t]$. 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 (see Fig. 1). 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

$$\delta[x(t),t] = \frac{\phi^L[x(t),t] - \phi^R[x(t),t]}{k_0} \quad (8)$$

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

$$Q(x,t) = C(x,t) + iS(x,t) \quad (9)$$

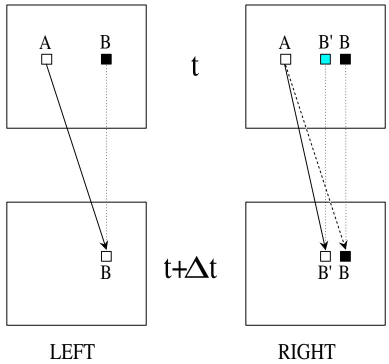

Figure 1: The stereo dynamic correspondence problem. A moving object in the 3-D space projects different trajectories onto the left and right images. The differences between the two trajectories carry information about motion-in-depth.

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. (8) 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. If disparity is defined as in Eq. (1), with respect to the spatial coordinate x^L, by differentiating Eq. (8) with respect to time, the total rate of variation of δ can be written as

$$\frac{d\delta}{dt} = \frac{\partial\delta}{\partial t} + \frac{v^L}{k_0}\left(\phi_x^L - \phi_x^R\right) \quad (10)$$

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 [6]:

$$\phi_x^L = -\frac{\phi_t^L}{v^L} \quad \text{and} \quad \phi_x^R = -\frac{\phi_t^R}{v^R}. \tag{11}$$

Combining Eq. (11) with Eq. (10) we obtain

$$\frac{d\delta}{dt} = \frac{\phi_x^R}{k_0}(v^R - v^L) \tag{12}$$

where $(v^R - v^L)$ is the phase-based interocular velocity difference along the epipolar lines. 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:

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi_t^L - \phi_t^R}{k_0} \simeq v^R - v^L \tag{13}$$

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 [4] [6]. 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):

$$\frac{\partial\delta}{\partial t} = \left[\frac{S_t^L C^L - S^L C_t^L}{(S^L)^2 + (C^L)^2} - \frac{S_t^R C^R - S^R C_t^R}{(S^R)^2 + (C^R)^2}\right]\frac{1}{k_0} \tag{14}$$

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" [3], 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 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,

$$S(x,t) \simeq f_1 * S(x,t) \quad ; \quad C(x,t) \simeq f_1 * C(x,t) \tag{15}$$

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

$$S'(x,t) \simeq f_2 * S(x,t) \quad ; \quad C'(x,t) \simeq f_2 * C(x,t) \tag{16}$$

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

$$f_1(t) = \mathrm{e}^{-t/\tau}\sin\omega_0 t \quad ; \quad f_2(t) = \mathrm{e}^{-t/\tau}\cos\omega_0 t. \tag{17}$$

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

By rewriting the terms of the numerators in (14):

$$\begin{aligned} 4S_tC &= (S_t + C)^2 - (S_t - C)^2 \quad \text{and} \\ 4SC_t &= (S + C_t)^2 - (S - C_t)^2, \end{aligned} \tag{18}$$

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 [7]. Specifically, each square term on the right sides of Eqs.(18) is a directionally tuned *energy detector* [8]. 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 stationary filters (cf. the denominators of Eq.(14)), yields to 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, as a side issue evidenced by our formulation, formal relationships exist between energy and phase-based approaches to motion modeling.

3 The neuromorphic motion-in-depth detector

In general, early vision processing relates to a parametric representation of perceptual "components" on the basis of measures of salient information obtained by localized filtering procedures. In the neuromorphic scheme proposed above, we can evidence two different processing stages: (1) spatiotemporal convolutions with 1-D Gabor kernels that extract amplitude and phase spectral components of the image

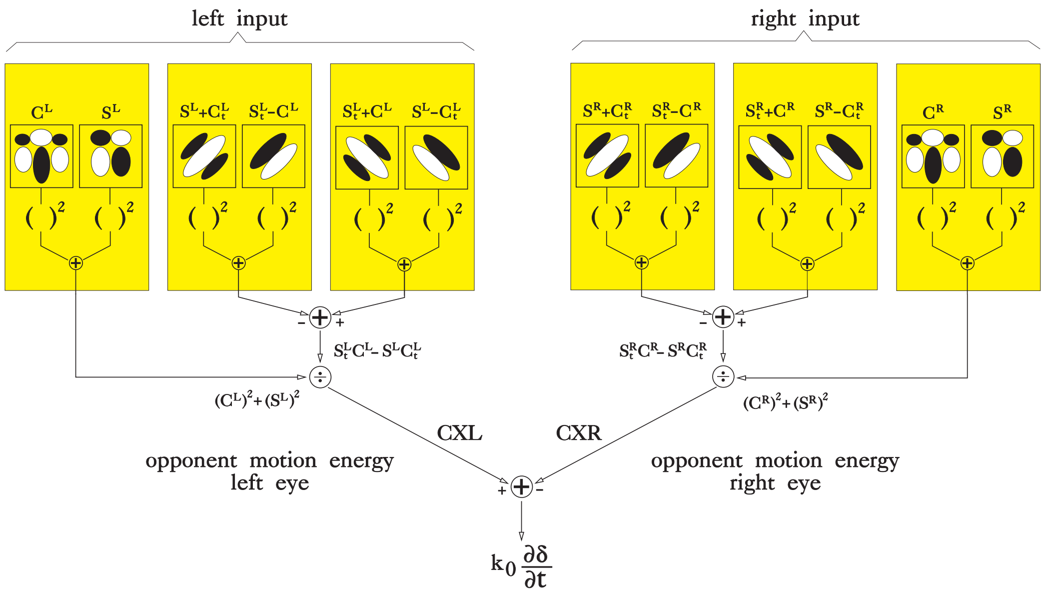

Figure 2: Cortical architecture of a motion-in-depth detector. The rate of variation of disparity can be obtained by a direct comparison of the responses of two monocular units labelled CXL and CXR. Each monocular unit receives contributions from a pair of directionally tuned "energy" complex cells that compute phase temporal derivative $(S_tC - SC_t)$ 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).

signals, and (2) simple punctual operations such as sums, squarings and divisions that yield the resulting percept (see Fig. 4). These computations can be supported by neuromorphic architectural resources organized as networks of interacting nodes. In the following, we shall present a possible hardware implementation of our MID detector based on analog perceptual microsystems. These circuits extract visual information analysing the input signal in space and time on a local basis. A lattice structure is employed for the spatial analysis. The processing nodes have a suitable dynamics for performing temporal analysis.

Spatial filtering: the perceptual engine It has been demonstrated [1] that image convolutions with 1-D Gabor-like kernels can be made isomorphic to the behaviour of a 2nd-order lattice network with diffusive nearest couplings and next nearest neighbors inhibitory reactions among nodes. If one uses a current-mode technique, all signals -stimuli and responses- are encoded by currents, and the behaviour of a 2nd-order lattice network can be described by the following equation:

$$\begin{aligned} I_e(n) &= I_s(n) + GI_e(n-1) + GI_e(n+1) \\ &- KI_e(n-2) - KI_e(n+2) \end{aligned} \qquad (19)$$

where $I_s(n)$ 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. This equation can be interpreted as an array of circuital computational nodes, with local interconnections both at the 1st nearest and the 2nd nearest neighbors (see Fig. 3b). 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 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 $I_s(n)$ (see Fig. 3c). The circuit representation of a node is based on the use of CCCSs with the desired current gains G and K (see Fig. 3d). The spatial impulse response of the network, $h(n)$ has been interpreted as the *perceptual engine* of our 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:

$$\begin{aligned} g(n) &= \alpha h(n-1) + \beta h(n) + \gamma h(n+1) \\ &= De^{-\lambda|n|}\cos(2\pi k_0 n + \varphi) \end{aligned} \qquad (20)$$

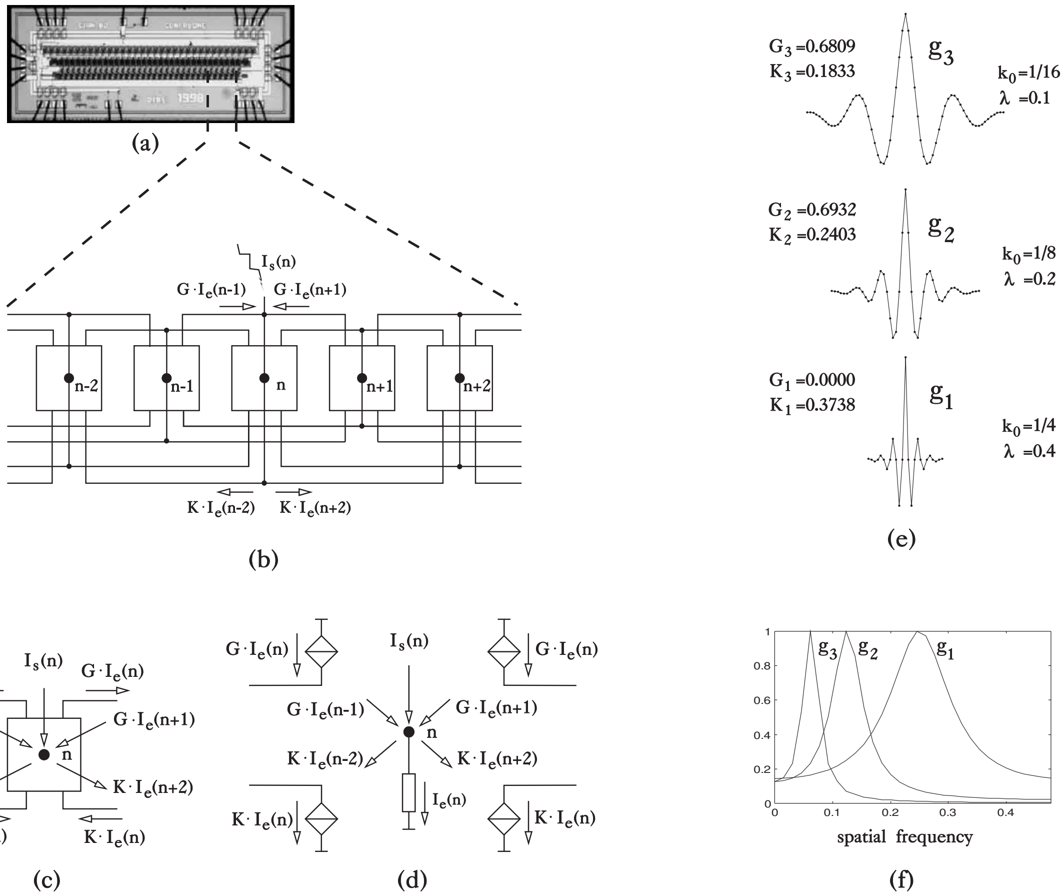

Figure 3: (a) Microphotograph of a 1-D 2nd-order network chip. (b) Schematic representation of a circuit implementing 2nd-order lattice networks. (c) Schematic and circuital (d) representation of a single node. (e) The three Gabor-like filters considered together with the corresponding values of G and K. The filters have been chosen to have a bandwidth of 1 octaves as shown in the frequency domain (f).

where λ is the decay rate, k_0 the oscillating frequency of the impulse response, and α, β, γ are coefficients depending on φ for any given set (λ, k_0).

Prototype chips have been fabricated and tested (see Fig. 3a). The decay rate and frequency, though hard-wired in the underlying perceptual engine, can be controlled by adjustable parameters [9].

Temporal filtering The signal processing *time-domain* requirements specified by Eq. 17 provide the functional characterization of the filter blocks f_1 and f_2. The Laplace transforms of the impulse responses determine the desired transfer functions:

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

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

They are (temporal) filters of the second order, often called biquad circuits, that can be implemented using continuous-time current-mode integrators [10].

By taking appropriate sums and differences of the temporally modulated outputs of a 2nd-order lattice network $P(n,t)$, and through squaring operations, it is possible to compute convolutions with cortical-like spatiotemporal operators:

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

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

Parametric processing The high information content of the parameters coming from the spatio-temporal 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 consid-

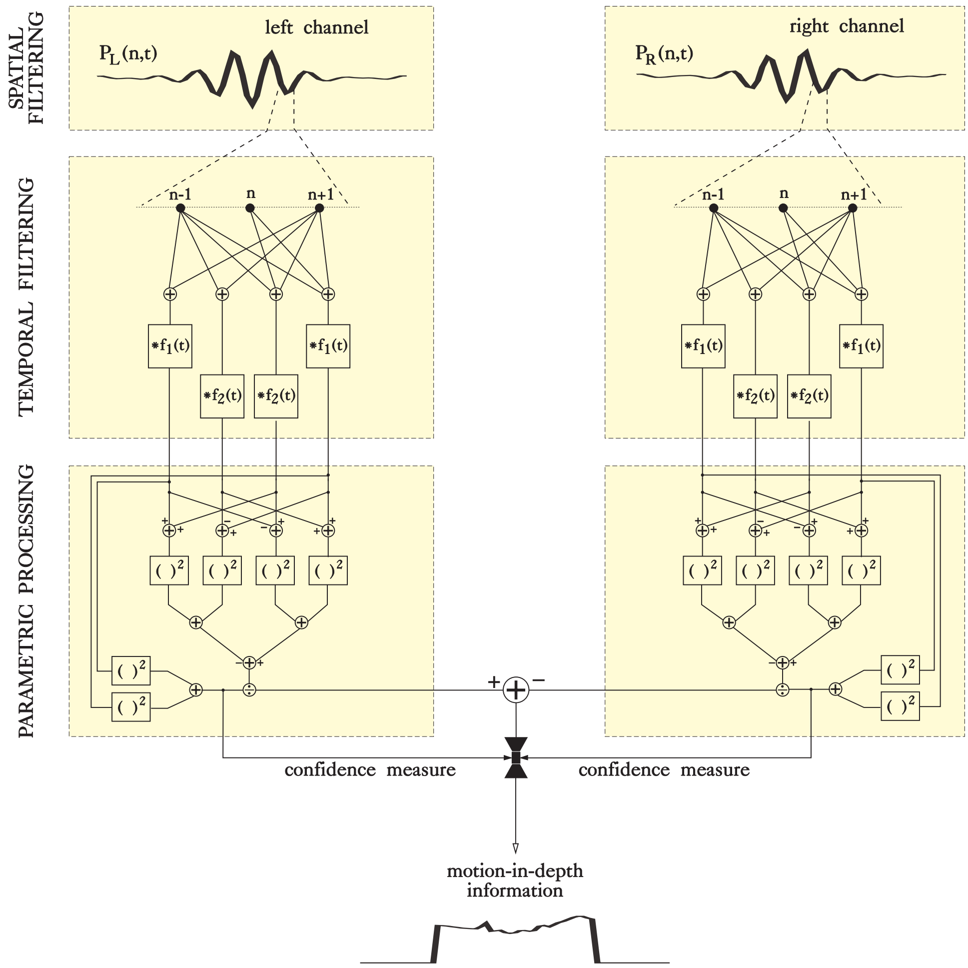

Figure 4: Architectural scheme of the neuromorphic motion-in-depth detector.

ers 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.

4 Experimental results

To test the performance of our neuromorphic architecture at system level we develop an emulation environment based on LabVIEW modules, which take into account the hardware constraints. In perspective, this solution will allow a direct control of data acquisition boards, thus being an ideal interface for the validation of MID, and other early vision modules, by direct use of the outputs of the perceptual microsystems mounted on boards.

Extensive simulations on both synthetic and real-world image sequences, yield to excellent performaces (see Fig. 5), resulting in correct discrimination between forward and backward movements of objects from the observer. Points where phase information are unreliable are discarded according to a confidence measure that is related to the local energy of the binocular filter output. To improve system performance, the different results obtained by the three Gabor filters used, can be weighted by the corresponding energies and combined in an average measure of MID.

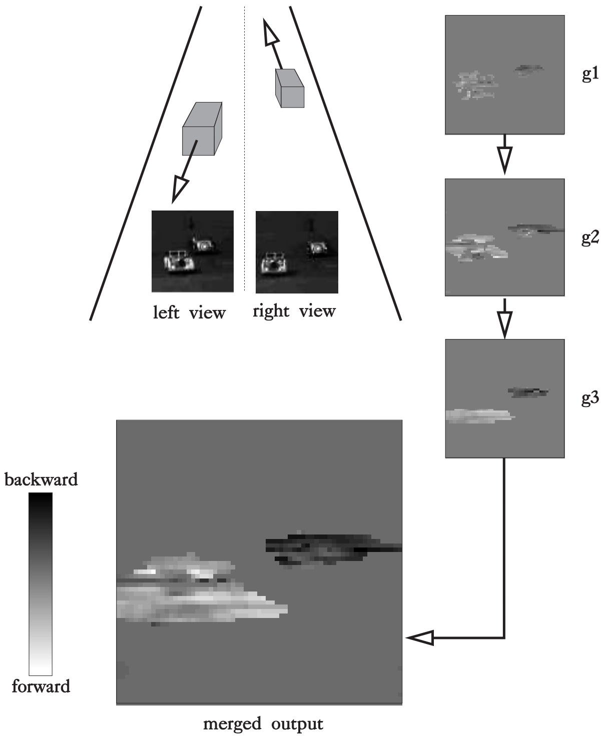

Figure 5: Experimental results on a natural scene for the three Gabor filters considered (g_1, g_2, g_3), and the resulting merged output.

5 Summary

In this paper the problem of binocular perception of MID has been tackled. 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 [11]. The predominance of one measure on the other one corresponds to different hypotheses on the architectural solutions which evolved, by nature, in the visual cortex. 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 spatio-temporal filtering with quadrature pairs of Gabor kernels, that can be directly implemented in VLSI, as demonstrated by recent prototypes of our group [1]. A simulation environment, based on LabVIEW modules, has been developed to analyze the effects on system performance of constraints posed by analog and digital hardware implementation.

Acknowledgments

This work was partially supported by the *MURST Project "Visual Perception Microsystems based on Structural Computational Paradigms"* and by the *ASI Project "Optical Microsystems for Robotic Applications in Space"*.

References

[1] 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.

[2] 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.

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

[4] D.J. Fleet, A.D. Jepson, and M. Jenkin. Phase-based disparity measurements. *CVGIP: Image Understanding*, 53:198–210, 1991.

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

[6] 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.

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

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

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

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

[11] J. Harris and S. N.J. Watamaniuk. Speed discrimination of Motion-in depth using binocular cues. *Vision Research*, 35(7):885–896, 1995.