A Cortical Architecture for the Binocular Perception of Motion-in-depth

Silvio P. Sabatini, Fabio Solari and Giacomo M. Bisio
DIBE, PSPC-Group - University of Genoa [pspc@dibe.unige.it]

Abstract

A model for the generation of cortical cells selective to motion-in-depth is presented. The model relies upon the computation of the total rate of change of the disparity through the combination of the outputs of monocular cortical units characterized by spatiotemporal receptive fields extracting temporal variations of phase information on the left and right retinal images. Each monocular unit of the cortical architecture can be directly compared to the Adelson and Bergen's motion detector, thus establishing a link between the information contained in the total derivative of the binocular disparity and those hold in the interocular velocity differences. Experimental simulations on stereo sequences evidenced that the model can quantitatively predict motion-in-depth information.

1 Introduction

The analysis of a dynamic scene implies the estimates of motion parameters to infer spatio-temporal information about the visual world. Among them, 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, a reliable estimate of motion-in-depth can be helped by considering the dynamic stereo correspondence problem in the stereo image signals 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). 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 $I^L(x) = I^R[x + \delta(x)]$, where $\delta(x)$ is the (horizontal) binocular disparity. 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 [1]: (1) by the rate of change of disparity, and (2)

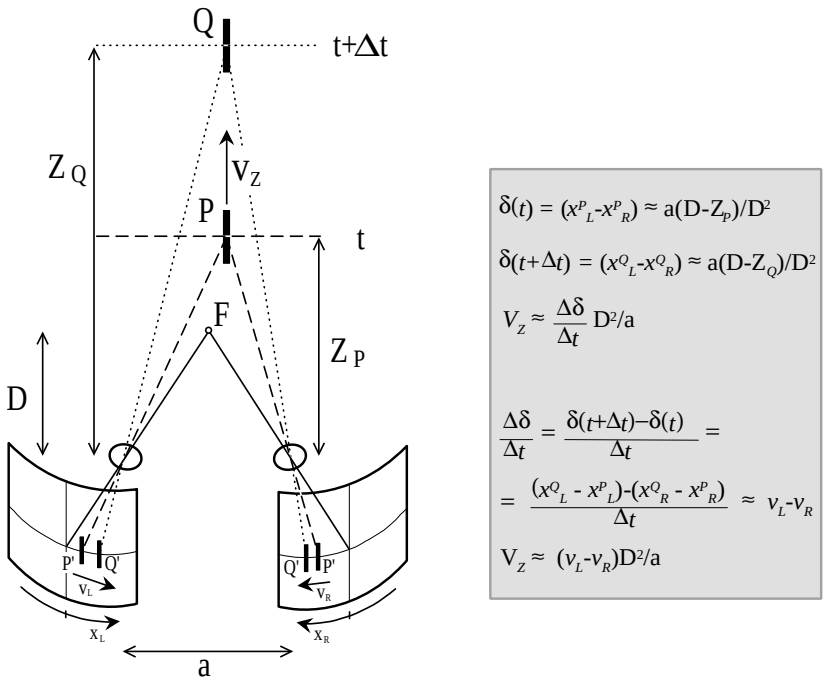

Figure 1: The stereo dynamic correspondence problem. A moving object in the 3-D space projects different trajectories onto the left and right images. The differences between the two trajectories carry information about motion-in-depth.

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 [2][3]. Though, to date, it is not completely known the way in which cortical neurons measure stereo disparity and motion information. In this paper, we show that the two measures can be placed into a common framework considering a phase-based disparity encoding scheme.

2 Phase-based measurements of local disparity

According to the *Fourier Shift Theorem*, the spatial shift δ in an image domain effects a phase shift $k\delta$ in the Fourier domain. On the basis of this property, several researchers (e.g., [4]) 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 of Gabor filters $h(x, k_0) = \mathrm{e}^{-x^2/\sigma^2}\mathrm{e}^{ik_0x}$, where k_0 is the peak frequency of the filter and σ

relates to its spatial extension. The resulting convolutions with the left and right binocular signals can be expressed as $Q(x) = \rho(x)e^{i\phi(x)} = C(x) + iS(x)$ 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 pair of filters. Hence, binocular disparity can be predicted by $\delta(x) = [\phi^L(x) - \phi^R(x)]/k(x)$ where $k(x) = [\phi^L_x(x) + \phi^R_x(x)]/2$ is the average *instantaneous frequency* of the bandpass signal, and can be approximated by the peak of the Gabor filter k_0. Extending to time domain, the disparity of a point moving with the motion field can be estimated by $\delta[x(t), t] = (\phi^L[x(t), t] - \phi^R[x(t), t])/k_0$, where phase components are computed from the spatiotemporal convolutions of the stereo image pair $Q(x, t) = C(x, t) + iS(x, t)$ with directionally tuned Gabor filters with central frequency $\mathbf{p} = (k_0, \omega_0)$.

3 The cortical model

If disparity is defined with respect to the spatial coordinate x^L, by differentiating with respect to time, its total rate of variation can be written as

$$\frac{d\delta}{dt} = \frac{\partial\delta}{\partial t} + \frac{v^L}{k_0}\left(\phi^L_x - \phi^R_x\right) \tag{1}$$

where v^L is the horizontal component of the velocity signal on the left retina. Considering the conservation property of local phase measurements [5], image velocities can be computed from the temporal evolution of constant phase contours, and thus:

$$\phi^L_x = -\frac{\phi^L_t}{v^L} \quad \text{and} \quad \phi^R_x = -\frac{\phi^R_t}{v^R}. \tag{2}$$

Combining Eq. (2) with Eq. (1) we obtain $d\delta/dt = (v^R - v^L)\phi^R_x/k_0$, 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:

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi^L_t - \phi^R_t}{k_0} \simeq v^R - v^L \tag{3}$$

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):

$$\frac{\partial\delta}{\partial t} = \left[\frac{S^L_t C^L - S^L C^L_t}{(S^L)^2 + (C^L)^2} - \frac{S^R_t C^R - S^R C^R_t}{(S^R)^2 + (C^R)^2}\right]\frac{1}{k_0} \tag{4}$$

thus avoiding explicit calculation and differentiation of phase, and the attendant problem of phase unwrapping.

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, $S(x,t) \simeq g * S(x,t)$ and $C(x,t) \simeq g * C(x,t)$, and evaluates the temporal changes in that sub-band, differentiation can be attained by convolutions on the data with appropriate bandpass temporal filters:

$$S'(x,t) \simeq g' * S(x,t) \quad ; \quad C'(x,t) \simeq g' * C(x,t). \tag{5}$$

S' and C' approximate S_t and C_t, respectively, if g and g' are a quadrature pair of temporal filters, e.g.: $g(t) = \mathrm{e}^{-t/\tau} \sin \omega_0 t$ and $g'(t) = \mathrm{e}^{-t/\tau} \cos \omega_0 t$. By rewriting the terms of the numerators in (4):

$$4S_tC = (S_t + C)^2 - (S_t - C)^2 \quad \text{and} \quad 4SC_t = (S + C_t)^2 - (S - C_t)^2, \tag{6}$$

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 [6]. Specifically, each square term on the

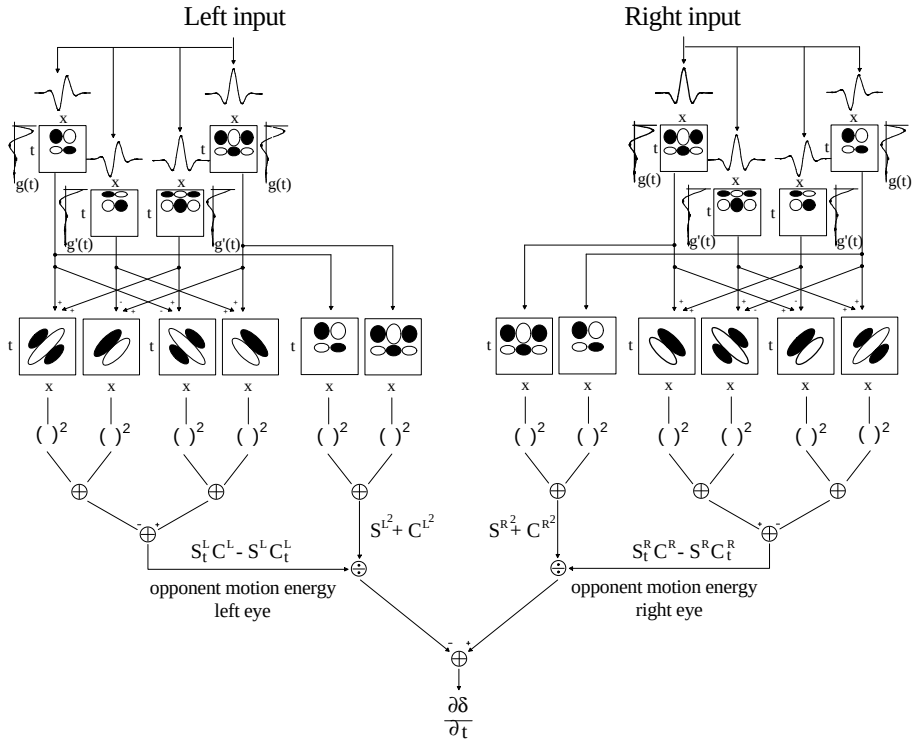

Figure 2: Cortical architecture of a motion-in-depth detector. The rate of variation of disparity can be obtained by a direct comparison of the responses of two monocular units labelled CXL and CXR. Each monocular unit receives contributions from a pair of directionally tuned "energy" complex cells that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

right sides of Eqs.(6) is a directionally tuned *energy detector* [7]. 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.(4)), 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, our formulation evidences that formal relationships exist between energy and phase-based approaches to motion modeling.

4 Experimental results

Extensive simulations on both synthetic and real-world image sequences, yield to excellent performaces (see Fig. 3), 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.

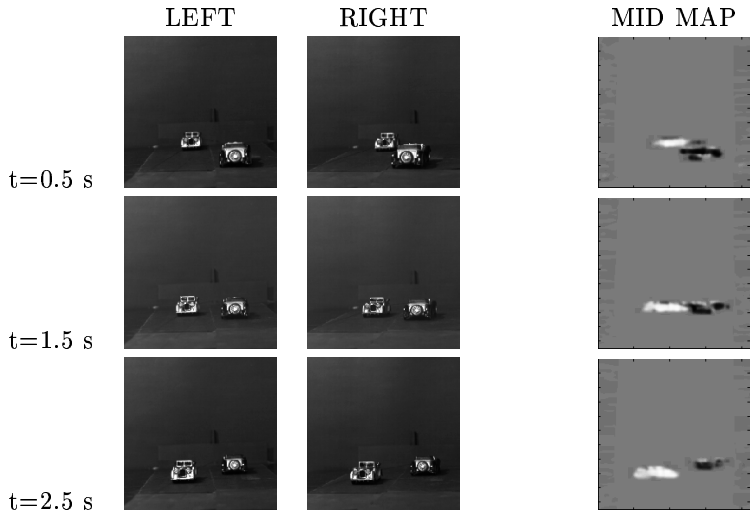

Figure 3: Experimental results on a natural scene. Two toy cars are moving in opposite directions respect to the observer. Left and right frames at three different times are shown. The gray levels in the MID maps code the motion-in-depth of the two cars: the lighter gray blob represents the car moving toward the observer, whereas the darker gray blob represents the car moving away. The background gray level codes all the static elements present in the scene.

5 Discussion and conclusions

There are at least two binocular cues that can be used to determine the motion of an object toward or away from an observer [1]: binocular combination of monocular velocity signals or the rate of change of retinal disparity. Assuming a phase-based disparity encoding scheme [4], we 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. The resulting computation relies upon spatiotemporal differentials of the left and right retinal phases that can be approximated by linear filtering operations with spatiotemporal receptive fields. Accordingly, we proposed a cortical model for the generation of binocular motion-in-depth selective cells as a hierarchical combination of monocular spatiotemporal subunits. Each monocular branch of the cortical architecture can be directly compared to the Adelson and Bergen's motion detector[7], thus establishing a link between phase-based approaches and motion energy models.

The algorithmic approach followed 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 a Gabor filtering stage that can be directly implemented in analog VLSI, as demonstrated by recent prototypes of our group [8]. Conversely, the feed-forward computations can be treated in a punctual way, i.e., according to standard computational schemes (sequential, parallel, pipeline). In this way, one can take take full advantage of the potentialities of analog processing together with the flexibility provided by digital hardware.

References

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

[2] 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.

[3] 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.

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

[5] 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.

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

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

[8] 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.