Emergence of Motion-in-depth Selectivity in the Visual Cortex through Linear Combination of Binocular Energy Complex Cells with Different Ocular Dominance

Silvio P. Sabatini and Fabio Solari

Department of Biophysical and Electronic Engineering
University of Genoa, I-16145 Genova, ITALY - *silvio@dibe.unige.it*

Abstract

An architectural hypothesis for the origin of motion-in-depth selectivity in the visual cortex is proposed. On the basis of a time extension of the phase-based techniques for disparity estimation, we consider the computation of the total temporal derivative of the time-varying disparity through the combination of the responses of disparity energy units. The emergence of motion-in-depth tuning is pointed out in relation to the unbalanced ocular dominance index of afferent binocular complex cells. The resulting cortical units of the model exhibit properties that can be directly compared with that reported in the literature for real cortical cells. The final complex cell is binocular, tuned to motion-in-depth, but not so sensitive ti disparity.

1 Introduction

There are at least two binocular cues that can be used to discriminate the 3-D component of the object's motion (i.e., its motion-in-depth (MID)) [1]: (1) the rate of change of binocular disparity, and (2) the inter-ocular velocity differences. The question of which mechanism is predominant on the other implies different hypotheses on the architectural solutions adopted by visual cortical cells to encode dynamic 3-D visual information. In particular, it has an specific bearing on the degree to which the brain mechanism for detecting motion-in-depth is independent of the mechanism for detecting static disparities. Recently, numerous experimental and computational studies (see e.g., [2] [3] [4] [5] [6]) addressed this issue, by analyzing the binocular spatio-temporal properties of simple and complex cells. The fact that the resulting disparity tuning does not vary with time, and that most of the cells in the primary visual cortex have the same motion preference for the two eyes, led to the conclusion that these cells are not tuned to motion-in-depth. In this paper, we demonstrate that, within a phase-based disparity encoding scheme, such cells relay phase temporal derivative components that can be combined, at a higher level, to yield a specific motion-in-depth selectivity. The rationale of this statement relies upon analytical considerations on phase-based dynamic stereopsis, as a time extension of the well-known phase-based techniques for disparity estimation [7]. The resulting model is based on the computation of the total temporal derivative of the disparity through the combination of the outputs of binocular disparity energy units [4] [5] characterized by different ocular dominance indices. Since each energy unit is just a binocular Adelson and Bergen's motion detector, this establishes a link between the information contained in the total rate of change of the binocular disparity and that held by the interocular velocity differences.

2 Phase-based dynamic stereopsis

Binocular 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 $I^L(x) = I^R[x+\delta(x)]$, where $\delta(x)$ is the horizontal binocular disparity. In the last decades, a computational approach for stereopsis, that rely on the phase information contained in the spectral components of the stereo image pair, has been proposed [7]. Spatially-localized phase measures on the left and right images can be obtained by filtering operations with a 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 $Q(x) = \rho(x)\mathrm{e}^{i\phi(x)} = C(x) + iS(x)$ where $\rho(x) = \sqrt{C^2(x) + S^2(x)}$ and $\phi(x) = \arctan\left(S(x)/C(x)\right)$ denote their amplitude and phase components, respectively, 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 $k(x) = [\phi^L_x(x) + \phi^R_x(x)]/2$, with ϕ_x spatial derivative of phase ϕ, is the average instantaneous frequency of the bandpass signal, that, under a linear phase model, can be approximated by the peak frequency of the Gabor filter k_0. Extending to time domain, the disparity of a point moving with the motion field can be estimated by:

$$\delta[x(t), t] = \frac{\phi^L[x(t), t] - \phi^R[x(t), t]}{k_0} \tag{1}$$

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 a 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 Eq. (1) provide only spatial information, useful for reliable disparity estimates.
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

$$\frac{d\delta}{dt} = \frac{\partial\delta}{\partial t} + \frac{v_L}{k_0}\left(\phi^L_x - \phi^R_x\right) \tag{2}$$

where v_L is the horizontal component of the velocity signal on the left retina. Considering the conservation property of local phase measurements [8], image velocities can be computed from the temporal evolution of constant phase contours, and thus:

$$\phi^L_x = -\frac{\phi^L_t}{v_L} \quad \text{and} \quad \phi^R_x = -\frac{\phi^R_t}{v_R} \tag{3}$$

with $\phi_t = \frac{\partial\phi}{\partial t}$. Combining Eq. (3) with Eq. (2) 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:

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi^L_t - \phi^R_t}{k_0} \simeq v_R - v_L \tag{4}$$

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):

$$\frac{\partial\delta}{\partial t} = \left[\frac{S^L_tC^L - S^LC^L_t}{(S^L)^2 + (C^L)^2} - \frac{S^R_tC^R - S^RC^R_t}{(S^R)^2 + (C^R)^2}\right]\frac{1}{k_0} \tag{5}$$

thus avoiding explicit calculation and differentiation of phase, and the attendant problem of phase unwrapping. Considering that, at first approximation $(S^L)^2 + (C^L)^2 \simeq (S^R)^2 + (C^R)^2$ and that these terms are scantly discriminant for motion-in-depth, we can formulate the cortical model taking into account the numerator terms only.

3 The cortical model

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: $S'(x,t) \simeq g' * S(x,t)$ and $C'(x,t) \simeq g' * C(x,t)$. S' and C' approximate S_t and C_t, respectively, if g and g' are a quadrature pair of temporal filters, e.g.: $g(t) = (kt)^5\ exp(-kt)(1/5! - (kt)^2/(7!)$ and $g'(t) = (kt)^3\ exp(-kt)(1/3! - (kt)^2/(5)!)$. By algebraic manipulation of the terms of the numerators in (5): 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 (RFs) of the primary visual cortex. Though, it is worthy to note that a direct interpretation of the computational model is not biologically plausible. Indeed, in the computational scheme (see Eq. (5)), the temporal variations of phases are obtained by processing monocular images separately and then the resulting signals are binocularly combined to give at an estimate of motion-in-depth in each spatial location.
To employ binocular RFs from the beginning, as they exist for most of the cells in the visual cortex, we manipulated the numerators of Eq. 5 by rewriting them as the combination of terms characterized by a dominant contributions from one eye with respect to the other. These contributions are referable to binocular disparity energy units [5] built from two pairs of binocular direction selective simple cells with left and right RFs weighted by an ocular dominance index $\alpha \in [0,1]$. The "tilted" spatio-temporal RFs of simple cells of the model are obtained by combining separable RFs according to an Adelson and Bergen's scheme [9]. It can be demonstrated that the information about motion-in-depth can be obtained with a minimum number of eight binocular simple cells, four with a left and four with a right ocular dominance, respectively (see Fig. 1):

$$s_1 = (1-\alpha)(C_t^L + S^L) - \alpha(C^R - S_t^R) \ ; \ s_2 = (1-\alpha)(C^L - S_t^L) + \alpha(C_t^R + S^R)$$

$$s_3 = (1-\alpha)(C_t^L - S^L) - \alpha(C^R + S_t^R) \ ; \ s_4 = (1-\alpha)(C^L + S_t^L) + \alpha(C_t^R - S^R)$$

$$s_5 = \alpha(C_t^L + S^L) - (1-\alpha)(C^R - S_t^R) \ ; \ s_6 = \alpha(C^L - S_t^L) + (1-\alpha)(C_t^R + S^R)$$

$$s_7 = \alpha(C_t^L - S^L) - (1-\alpha)(C^R + S_t^R) \ ; \ s_8 = \alpha(C^L + S_t^L) + (1-\alpha)(C_t^R - S^R)$$

$$c_{11} = s_1^2 + s_2^2 \ ; \ c_{12} = s_3^2 + s_4^2 \ ; \ c_{13} = s_5^2 + s_6^2 \ ; \ c_{14} = s_7^2 + s_8^2$$

$$c_{21} = c_{12} - c_{11} \ ; \ c_{22} = c_{13} - c_{14}$$

$$c_3 = (1-2\alpha)(S_t^L C^L - S^L C_t^L - S_t^R C^R + S^R C_t^R)\ .$$

The output of the higher cell in the hierarchy (c_3) truly encodes motion-in-depth information.

4 Results

To assess model performances, for each layer, the tuning characteristics of the cells are analyzed as sensitivity maps in the $(x_L - x_R)$ and $(v_L - v_R)$ domains for the static and dynamic properties, respectively. The $(x_L - x_R)$ represents the binocular RF [5] of a cell, evidencing its disparity tuning. The $(v_L - v_R)$ response represents the binocular tuning curve of the velocities along the epipolar lines. Fig. 2 shows, as contour plots, the binocular responses in the $(x_L - x_R)$ domain of three complex cells at increasing position in the hierarchy, for two different values of the ocular dominance index ($\alpha = 0.8$, $\alpha = 0.5$). The cells of the cortical model exhibit properties and typical

profiles similar to those observed in the visual cortex [5]. Specifically, for $\alpha = 0.5$ we fall exactly in the Ohzawa et al. energy model. It is worthy to note the lack of disparity tuning in the cell of the last layer. To investigate motion-in-depth sensitivity, we derived cells' responses to drifting

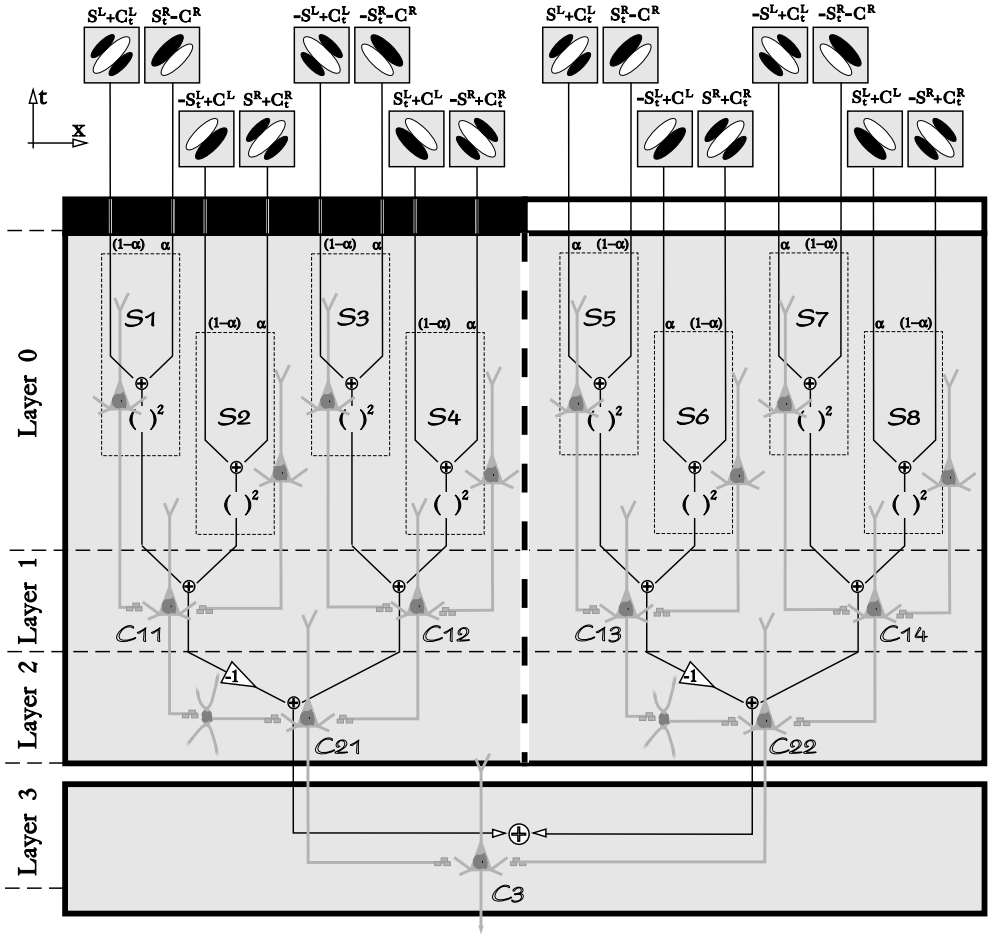

Figure 1: Functional representation of the proposed cortical architecture. Each branch groups cells belonging to an ocular dominance column (black and white slabs in the figure). The afferent signals from left and right ocular dominance columns are combined in layer 3. The basic units are binocular simple cells tuned to motion directions $(s_1, \ldots, s_8)$. The responses of the complex cells in layers 1, 2 and 3 are obtained by linear combinations of the outputs of those basic units.

sinusoidal gratings with different speeds in the left and right eye. The spatial frequency of the gratings has been chosen as central to the RF's bandwidth. The responses of the model cells with respect to the interocular velocities ratio for twelve different motion trajectories in depth (labeled 1 to 12) are represented in Fig. 3 as polar plots (cf. [10]). The 3 and 9 paths represent the rightward and leftward motions, respectively, in a frontoparallel plane; the 12 and 6 represent motions straight away from and toward the observer, respectively. The remaining eight trajectories represent intermediate oblique paths in depth. As one ascends in the hierarchy of the cortical network, model cells combine their outputs in such a way to gain a still more distinct tuning to motion-in-depth. A tuning that is clearly revealed in layer 3 cell, which is specifically tuned to motion toward the observer. By combining subunits with opposite ocular dominance indices,

opposite direction tuning can be straightforwardly obtained. It is worthy to note that the ocular dominance plays a key role for the origin of motion-in-depth tuning: if $\alpha = 0.5$ (i.e., balanced contributions from the two eyes) there is, indeed, no direction-in-depth selectivity (see Fig. 3). Summarizing, the complex cells belonging to the middle two layers exhibit a strong selectivity to static disparity, but no specific tuning to motion-in-depth. On the contrary, the output cell c_3 shows a narrow tuning to the Z direction of the object's motion, while lacking disparity tuning.

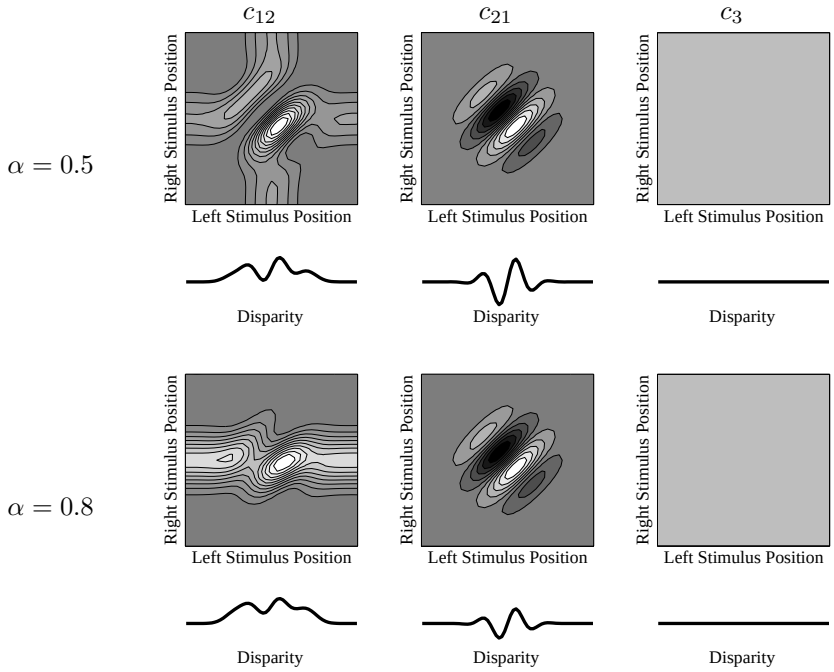

Figure 2: Binocular receptive fields of three complex cells of the cortical network, for two different values of ocular dominance index. The ligher the shading the stronger the response. The solid curves below each $(x_L - x_R)$ plot, represent the corresponding disparity tuning curves, obtained by integrating the 2-D $(x_L - x_R)$ profile along constant disparity lines parallel to the 45^o diagonal (cf. [5]). Responses are normalized with respect to each other, in order to have a common reference.

5 Conclusions

Considering the dynamic stereo correspondence problem, the Z component of the object's motion (i.e., its motion-in-depth) can be approximated by binocular combination of monocular velocity signals or by the rate of change of retinal disparity [1]. Assuming a phase-based disparity encoding scheme [7], we demonstrated that information held 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 spatio-temporal differentials of the left and right retinal phases that can be approximated by linear filtering operations with spatio-temporal RFs. Accordingly, we proposed a cortical model for the generation of binocular motion-in-depth selective cells as a hierarchical combination of binocular energy complex cells. It is worth noting that the phase response and the associated characteristic disparity of simple and complex cells in layers from 0 to 2 do not change with time, but the amplitudes of their responses carry information on temporal phase derivatives, that can be related to both retinal velocities and temporal changes in disparity. Moreover, the model evidences the different roles of simple and complex cells. Simple cells provide

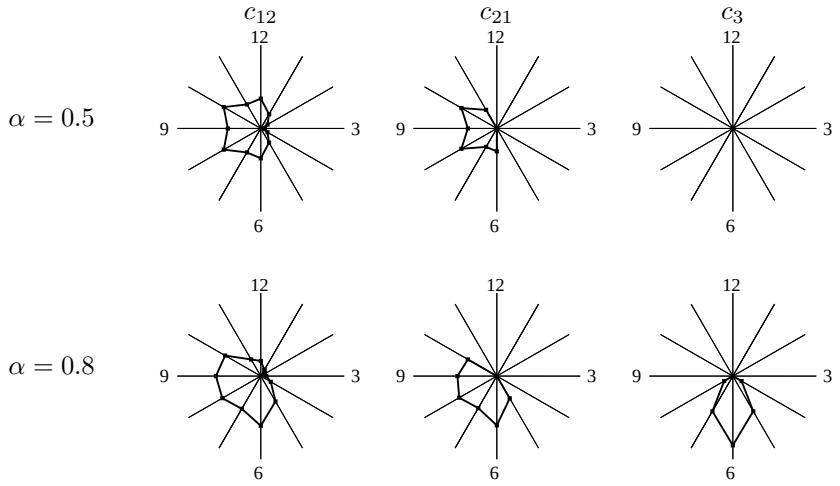

Figure 3: Motion-in-depth tuning curves of three model complex cells for two different values of ocular dominance index. The responses are normalized to the largest amplitude value.

a Gabor-like spatio-temporal transformation of the visual space, on which to base a variety of visual functions (perception of form, depth, motion). Complex cells, by proper combinations of the same signals provided by simple cells, actively eliminate sensitivity to a selected set of parameters, thus becoming specifically tuned to different features, such as disparity but not motion-in-depth (layer 1 and 2), motion-in-depth but not disparity (layer 3).

Acknowledgments

This work was partially supported by the *CEC IST-FET Project 2001-32114 "Ecovision"*.

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] N. Qian and S. Mikaelian. Relationship between phase and energy methods for disparity computation. *Neural Comp.*, 12(2):279–292, 2000.

[3] Y. Chen, Y. Wang, and N. Qian. Modelling V1 disparity tuning to time-varying stimuli. *J. Neurophysiol.*, pages 504–600, 2001.

[4] D. J. Fleet, H. Wagner, and D. J. Heeger. Neural encoding of binocular diparity: energy models, position shift and phase shift. *Vision Research*, 17:345–398, 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] A. Anzai, I. Ohzawa, and R.D. Freeman. Joint-encoding of motion and depth by visual cortical neurons: Neural basis of Pulfrich effect. *Nature Neurosci.*, 4:513–518, 2001.

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

[8] 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.

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

[10] W. Spileers, G.A. Orban, B. Gulyàs, and H. Maes. Selectivity of cat area 18 neurons for direction and speed in depth. *J. Neurophysiol.*, 63(4):936–954, 1990.