A Fast Joint Bioinspired Algorithm for Optic Flow and Two-Dimensional Disparity Estimation

Manuela Chessa, Silvio P. Sabatini, and Fabio Solari

Department of Biophysical and Electronic Engineering, University of Genoa
Via all'Opera Pia 11a - 16145 Genova - Italy
{manuela.chessa,silvio.sabatini,fabio.solari}@unige.it

Abstract. The faithful detection of the motion and of the distance of the objects in the visual scene is a desirable feature of any artificial vision system designed to operate in unknown environments characterized by conditions variable in time in an often unpredictable way. Here, we propose a distributed neuromorphic architecture, that, by sharing the computational resources to solve the stereo and the motion problems, produces fast and reliable estimates of optic flow and 2D disparity. The specific joint design approach allows us to obtain high performance at an affordable computational cost. The approach is validated with respect to the state-of-the-art algorithms and in real-world situations.

1 Introduction

The extraction of stereo and motion information, which can be directly used by a robot to interact with the environment, requires the processing of the visual information in a fast and reliable way. In the literature, several algorithms for the computation either of the optic flow [1] or the binocular disparity [2] have been proposed, which are characterized by high performances either in term of accuracy of the estimates or in term of the execution time [3]. Though, real-time system solutions for both motion and depth, which partially share the computational resources are very seldom. On the contrary, in the visual system of mammals the analysis of the dynamics and of the 3D spatial relationships occur jointly in the *dorsal* cortical pathway, which projects visual information from the primary visual cortex (V1) to the occipital areas V3 and MT (V5), up to the parietal lobe, by specializing as an "action stream" for guiding in real-time the actions that we direct at objects in the world [4]. Bioinspired computer vision techniques, although conceptually valuable, have always been characterized by a high computational cost, as the price to pay for their higher flexibility. Usually, these bioinspired solutions were indeed developed as models of how the visual cortex works, and their functionality were demonstrated on simple synthetic images, but were not designed to form a systematic alternative to computer vision, working on raw images and real video sequences [5] [6] [7] [8], but see [9]. It is also worth mentioning, that several neuromorphic solutions, at an affordable cost, have been proposed [10][11][12] with the aim of providing efficient sensor modules for machine vision. Yet, even though such solutions have the merit of

M. Fritz, B. Schiele, and J.H. Piater (Eds.): ICVS 2009, LNCS 5815, pp. 184–193, 2009.

solving excellently the specific sensorial problem, they do not have the necessary generality to guarantee a rich and flexible perceptual representation of the visual signal to be used for a more advanced visual analysis.

In this paper, we propose a joint neural architecture for the computation of both 2D (horizontal and vertical) disparity and optic flow, whose performances are comparable to the state-of-the-art algorithms (also not bioinspired) in terms of accuracy (see Section 4) and execution time. Specific emphasis is given to the sharing of the resources for the computation of both the features and to a proper combination of the information extracted by a set of spatial orientation channels, that allows us to tackle the aperture problem that arises in both optic flow and 2D disparity estimation.

2 Neuromorphic Computational Paradigms for Visual Processing

Disparity and optic flow features can be extracted from a sequence of stereo image pairs, using a distributed bioinspired architecture that resorts to a population of tuned cells. In the literature, many authors analyze different approaches to design those populations of neurons, and to properly combine their responses in order to obtain reliable information from the visual signal [13]. In distributed representations, or population codes, the information is encoded by the activity pattern of a network of *simple* and *complex* neurons, that are selective for elemental vision attributes: oriented edges, direction of motion, color, texture, and binocular disparity [14]. In particular, we consider neurons tuned to specific values of disparity and velocity, whose receptive fields (RFs) overlap with the ones of the other neurons of the population. To decode information at a specific image location the whole activity of the population within a neighborhood of this location is considered.

2.1 Population Coding

Considering that the image motion is described as an orientation in the space-time domain [15], numerous models for the detection of velocity are based on spatio-temporal filtering of the image sequence with 3D Gabor filters (e.g., [5][6]). Similarly, the tuning to binocular disparity is obtained from physiological and modeling studies [16][8], that demonstrated that a difference in the phase of the left and right spatial receptive field profiles of a binocular simple cell of area V1 can encode the disparity information (*phase-shift-model*)[1] [8][7]. Given the distributed character of the representation, proper decoding mechanisms have to be considered to estimate the value of a given feature from the whole activity of the population. Common and simple approaches for decoding the population codes are the Winner Takes All (WTA) method and the weighted sum [13].

[1] Yet, models based on a difference in the position of the left and right RFs (*position-shift-model*) or hybrid approaches have been proposed.

2.2 Multichannel Representation and Cooperation

To come up with more complex visual descriptors and to solve the problem of visual motion and depth perception, the outputs from area V1 have to be combined in higher cortical levels, through feed-forward convergence, recursive interactions and selection processes. The combination of the information extracted by single cells with respect to different orientation channels represents a widely used computational paradigm in the visual cortex. The different orientation channels could be used to remove the inherent ambiguities of local motion and stereo estimates, consequence of the well known aperture problem. In optic flow computation, motion along an edge cannot be discriminated when the edge is larger than the population RFs, used to estimate *image velocity* $\mathbf{v}$, since only the component of the feature orthogonal to an edge can be computed. A similar problem arises for a stereo active vision system with convergent axes [17], where 2D (horizontal and vertical) disparities are present and it is possible to define the *vector disparity* $\boldsymbol{\delta}$ as the vector difference in positions of identified corresponding points in the left and right eyes, each measured with respect to the point of fixation as origin [18]. From this perspective, a multidimensional representation of the visual signal over several spatial orientation channels is instrumental to provide a structural reference with respect to which to evaluate motion and stereo information.

3 Neural Architecture

The proposed population approaches for the computation of horizontal and vertical disparities and optic flow share a common algorithmic structure (see Fig. 1): (i) the distributed coding of the features across different orientation channels, through a filtering stage (that resembles the filtering process of area V1), (ii) the decoding stage for each channel, (iii) the estimation of the features through channel interactions, and (iv) the coarse-to-fine refinement.

3.1 Feature Coding Strategy

A population of spatio-temporal units, characterized by a spatial orientation θ and tuned to different velocities v_i^θ and to different disparities δ_i^θ, is used to represent the feature values. Each unit (a quadrature pair of simple cells) is described by a 3D Gabor filter, in order to maintain the sensitivity to the 3D orientation in the spatio-temporal domain.

A set of Gabor filters [19], with the form $h(\mathbf{x}, t) = g(\mathbf{x})f(t)$, that uniformly cover the orientation space and optimally sample the spatio-temporal domain, is chosen. The spatial component of the filters, $g(\mathbf{x})$, is built by exploiting its separability, to keep low the computational cost [20]. A Gabor filter rotated by an angle θ with respect to the horizontal axis is defined by:

$$g(x, y; \psi, \theta) = e^{\left(-\frac{x_\theta^2}{2\sigma_x^2} - \frac{y_\theta^2}{2\sigma_y^2}\right)} e^{j(\omega_0 x_\theta + \psi)} \tag{1}$$

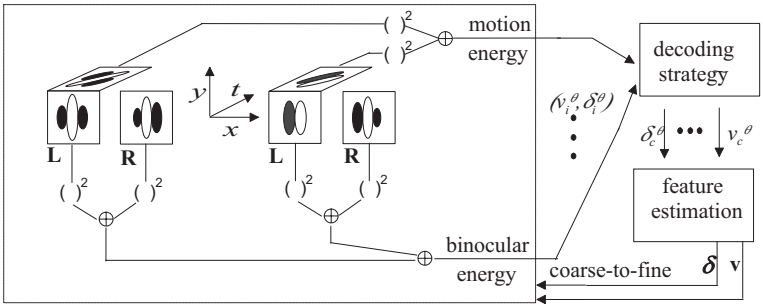

Fig. 1. The joint architecture for the computation of 2D disparity and optic flow. The box on the left represents a complex unit of the distributed population. Each unit, characterized by a spatial orientation θ, codes for its preferred velocity v_i^θ and for its preferred disparity δ_i^θ. The motion energy and the binocular energy responses, obtained by the $N \times M \times K$ cells, are decoded (upper-right box), thus obtaining an estimate of the component velocity v_c^θ and disparity δ_c^θ for each spatial orientation. These estimates are then combined to obtain the full velocity $\mathbf{v}$ and the full disparity $\boldsymbol{\delta}$ (lower-right box). The obtained features are used to perform a coarse-to-fine refinement.

where ω_0 is the spatial radial peak frequency, σ_x and σ_y determine the spatial supports of the filter, ψ is the phase of the sinusoidal modulation and (x_θ, y_θ) are the rotated spatial coordinates. In the spatial domain, the orientation space is uniformly sampled using N filters oriented from 0 to 2π and having the same radial peak frequency. It is worthy to note that, to avoid the introduction of a loss of balance between the convolutions with the even and odd Gabor filters, the contribution of the DC component is removed. In the following, we describe how the specific tuning to velocity and disparity values is obtained.

Optic flow. The temporal component of the 3D filter is defined by:

$$f(t;\omega_t) = e^{\left(-\frac{t^2}{2\sigma_t^2}\right)} e^{j\omega_t t} 1(t) \tag{2}$$

where σ_t determines the filter support in time domain, ω_t is the temporal peak frequency and $1(t)$ denotes the unit step function. Each cell is tuned to the velocity with magnitude v^θ and direction orthogonal to the preferred spatial orientation θ of the filter. The spatial frequency ω_0 is kept constant while the temporal peak frequency is varied by the rule $\omega_t = v^\theta \omega_0$. For each spatial orientation, a set of M tuning velocities are chosen (accordingly to the limit imposed by the temporal support of the filter and by the Nyquist theorem).

The described quadrature pair of spatio-temporal receptive fields $h(\mathbf{x}, t)$ are applied to the sequence of images in input $I(\mathbf{x}, t)$, thus obtaining a complex response:

$$Q(\mathbf{x_0}, t; v^\theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\mathbf{x_0} - \mathbf{x}, t - \tau) I(\mathbf{x}, \tau) d\mathbf{x} d\tau \tag{3}$$

Following [15] it is possible to compute the *motion energy*, as the squared modulus of the complex response:

$$E(\mathbf{x_0}, t; v^\theta) = |Q(\mathbf{x_0}, t; v^\theta)|^2 = \left| e^{j\psi(t)} \int_0^t Q(\mathbf{x_0}, \tau; v^\theta) e^{-j\omega_t \tau} d\tau \right|^2 \tag{4}$$

where $\psi(t) = \psi + \omega_t t = \psi + \omega_0 v^\theta t$.
The response of the motion energy unit has its maximum when the tuning velocity is equal to the velocity present in the stimulus.

Disparity. Following the *phase-shift model*, to obtain the tuning to a specific disparity the left and right RFs, $g^L(\mathbf{x})$ and $g^R(\mathbf{x})$ respectively, are centered at the same position in the left and in the right images, but have a proper binocular phase difference $\Delta\psi = \psi^L - \psi^R$. For each spatial orientation, a set of K binocular phase differences, uniformly distributed between $-\pi$ and π, are chosen to obtain the tuning to different disparities $\delta^\theta = \Delta\psi/\omega_o$, oriented along the direction orthogonal to the orientation of the RF. Then, the left and right RFs are applied to the binocular image pair in input $I^L(\mathbf{x})$ and $I^R(\mathbf{x})$, thus obtaining a complex response:

$$Q(\mathbf{x_0}; \delta^\theta) = \int_{-\infty}^{\infty} g^L(\mathbf{x_0} - \mathbf{x}) I^L(\mathbf{x}) d\mathbf{x} + \int_{-\infty}^{\infty} g^R(\mathbf{x_0} - \mathbf{x}) I^R(\mathbf{x}) d\mathbf{x} \tag{5}$$

The *binocular energy* [16][8] is obtained as the squared responses of a quadrature pair of binocular units:

$$E(\mathbf{x_0}; \delta^\theta) = |Q(\mathbf{x_0}; \delta^\theta)|^2 = |Q^L(\mathbf{x_0}; \delta^\theta) + e^{-j\Delta\psi} Q^R(\mathbf{x_0}; \delta^\theta)|^2 \tag{6}$$

and it has its maximum when the product of the stimulus disparity and the spatial peak frequency equals the binocular phase difference.

3.2 Feature Decoding and Estimation

Once the features along each spatial orientation have been coded by the population activity, it is necessary to read out this information, to obtain a reliable estimate. The decoding strategy, the number of the cells in the population and their distribution are jointly related. To decode the population by a WTA strategy a large number of cells along each spatial orientation would be necessary, thus increasing the computational cost and the memory occupancy of the approach. To obtain precise feature estimation, while keeping the number of cells as low as possible, thus an affordable computational cost, a *weighted sum* (i.e. a center of gravity) of the responses for each orientation is calculated. The *component velocity* v_c^θ is obtained by:

$$v_c^\theta(\mathbf{x_0}, t) = \frac{\sum_{i=1}^{M} v_i^\theta E(\mathbf{x_0}, t; v_i^\theta)}{\sum_{i=1}^{M} E(\mathbf{x_0}, t; v_i^\theta)} \tag{7}$$

where v_i^θ are the M tuning velocities and $E(\mathbf{x_0}, t; v_i^\theta)$ are the motion energies for each spatial orientation. Similarly, we decode the *component disparity* δ_c^θ. Other decoding methods [13], such as the *maximum likelihood* estimator, have been considered, but the center of gravity of the population activity is the best compromise between simplicity, low computational cost and reliability of the estimates.

As we have described in Section 2.2, a single oriented filter cannot estimate the feature, but only the component orthogonal to the filter orientation i.e. the velocity v_c^θ or the disparity δ_c^θ. Following [21][22] the aperture problem is tackled by combining the estimates of v_c^θ and δ_c^θ for each spatial orientation, in order to obtain a robust estimate of the full velocity $\mathbf{v}$ and of the full disparity $\boldsymbol{\delta}$.

3.3 Multiscale Processing and Coarse-to-Fine Refinement

In the frequency domain, the Gabor filters, used in the architecture, act as band pass filters centered in a single spatial radial peak frequency. However, experimental studies have shown the presence of different spatial frequency channels, and, since information in natural images is spread over a wide range of frequencies, it is necessary to use a technique that allows us to get information from the whole range. Here, we have adopted a pyramidal decomposition, in order to keep as low as possible the computational cost. Moreover, by exploiting the pyramidal decomposition, a coarse-to-fine refinement is implemented. The features, obtained at a coarser level of the pyramid, are expanded and used to warp the sequence of the spatially convolved images, then the residual optic flow and disparity are computed. In this way, the "distance" between corresponding points is reduced, thus yielding to a more precise estimate, since the remaining feature values lie in the filters' range.

4 Comparisons and Results

The proposed algorithm analyzes the visual information by using a biologically plausible strategy. However, in order to use effectively this model in a robotic system, it is important to compare the obtained results with the ones of the well-established algorithms from the literature. It is worth noting that both the accuracy and the computational requirements (e.g. execution time) should be taken into account for the extraction of features, especially when aiming to embed the algorithm in a robotic system. For what concerns the accuracy, the quantitative evaluation of optical flow and disparity algorithm is performed by comparing the results for selected test sequences, for which the ground truth data are available [1][2] (Fig. 2(a-h)). It is worth noting that these test beds contain horizontal disparities, only. Thus, to benchmark the validity of our approach in active vision systems, and thus with images that contain 2D disparities, we have used the dataset described in [23] (Fig. 2(i-n)). The presented results have been obtained by using $N = 16$ oriented filters, each tuned to $M = 3$

Table 1. Comparison between the proposed distributed population code and some state-of-the-art algorithms for optic flow estimation. The reliability has been computed by using the average angular error (AAE) proposed by Barron [24]. (*)Results from `http://vision.middlebury.edu/flow/`.

Algorithm	Yosemite	Rubberwhale	Hydrangea
2D-CLG [Bruhn & Weickert, IJCV 61(3), 2005]	1.76	16.75	
SPSA-learn [Li & Huttenlocher, ECCV 2008]	2.56	5.22	2.43
Black-Anandan 2 modified by Simon Baker (*)	2.61	8.14	
Black-Anandan modified by Deqing Sun (*)	3.1		
Distributed population code	**3.19**	**8.01**	**5.79**
Dynamic MRF [Glocker et al., CVPR 2008]	3.63		
Fusion [Lempitsky et al., CVPR 2008]	4.55	3.68	
Pyramidal LK modified by Bouguet (*)	6.41	18.69	15.86
CBF [Trobin et al., ECCV 2008]	6.57		
Group Flow [Ren, CVPR 2008]		5.32	
FMT [Ho & Goecke, CVPR 2008]		10.07	
Proesman [Ho & Goecke, CVPR 2008]		17.43	

Table 2. Comparison between the proposed distributed population code and some state-of-the-art algorithms for disparity estimation. The reliability has been computed in terms of percentage of bad pixels for non-occluded regions (see `http://vision.middlebury.edu/stereo/`).

Algorithm	Venus	Teddy	Cones
Reliability DP [Gong & Yang, CVPR 2005]	2.35	9.82	12.9
Go-Light [Su & Khoshgoftaar, ICIP 2007]	2.47	14.5	9.78
SSD + MF [Scharstein & Szeliski, IJCV 47, 2002]	3.74	16.5	10.6
Infection [Olague et al., Artificial Life 2006]	4.41	17.7	14.3
Distributed population code	**4.5**	**11.7**	**6.4**
Adaptive weight[Yoon & Kweon, PAMI 28, 2006]	4.61	12.70	5.50
Phase-based [El-Etriby et al., ISIE 2007]	6.71	14.5	10.8
Phase-diff [El-Etriby et al., ICCVG 2006]	8.34	20.0	19.8
SO [Scharstein & Szeliski, IJCV 47, 2002]	9.44	19.9	13.0
DP [Scharstein & Szeliski, IJCV 47, 2002]	10.1	14.0	10.5

different velocities and to $K = 9$ binocular phase differences. The used Gabor filters have a spatio-temporal support of $(11 \times 11) \times 7$ pixels $\times$ frames and are characterized by a bandwidth of 0.833 octave and spatial frequency $\omega_0 = 0.5\pi$. Tables 1 and 2 show the comparison with some state-of-the art algorithms. Even if there are some algorithms that perform better than the proposed approach, the feature maps we obtain are reliable and accurate. The approach is also applied to real-world scenes, acquired by two stereo cameras, moving in a non-static environment. Figure 3 shows some of the obtained optic flow fields and disparity maps.

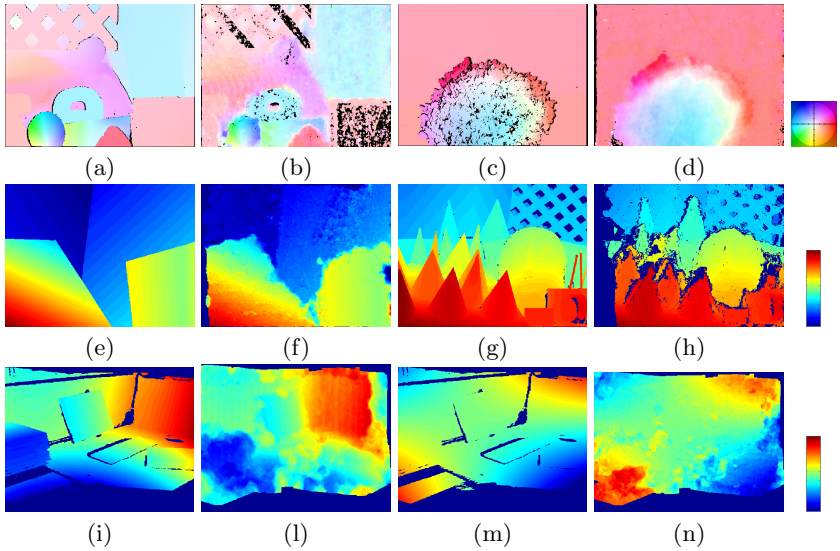

Fig. 2. The ground truth (a)(c) and the estimated (b)(d) optic flows for the Rubberwhale and the Hydrangea sequences. The different colors represent velocity direction. The ground truth (e)(g) and the estimated (f)(h) disparities for the Venus and the Cones image pairs. Disparity values are coded from red (near objects) to blue (far objects). The ground truth (i)(m) and the estimated (l)(n) horizontal and vertical disparities for a stereo pair acquired by two virtual cameras with convergent axes.

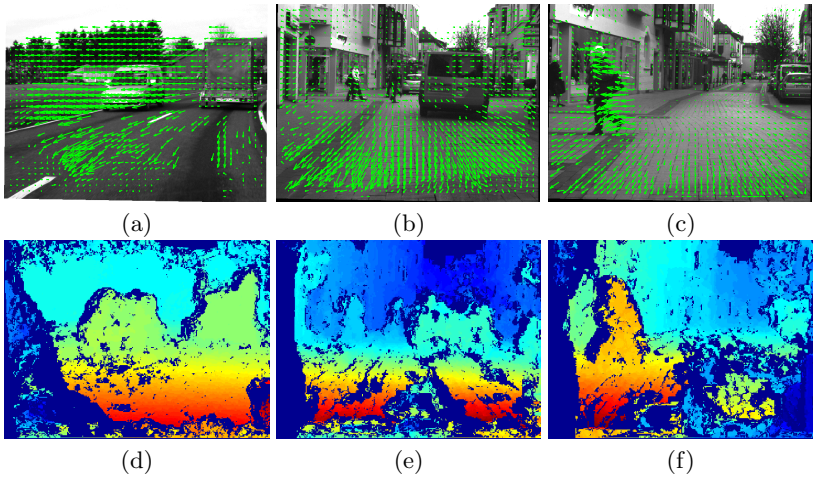

Fig. 3. Optic flow (a-c) and disparity maps (d-f) computed in different real-world situations. The images are acquired by moving stereo cameras, thus both ego-motion and independent motion of other objects in the scene are present. Disparity values are coded from red (near objects) to blue (far objects).

5 Conclusions and Future Work

The major contributions of this paper are: (1) the development of a distributed neuromorphic architecture for the estimation of motion and 2D disparity fields in a sequence of binocular stereo pairs, by mimicking the sharing of computational resources evidenced in cortical areas; (2) the handling both horizontal and vertical disparities; (3) the application of such bioinspired approach in real-world situations; (4) a good compromise between reliability of the estimates and execution time. The proposed solution, based on a distributed code, is able to keep the computational cost of the algorithm as low as possible, without loosing in accuracy and reliability. For what concerns the execution time, we reach near real-time performances (7 frames/second for images of 256×256 pixels on a QuadCore processor), by using Intel IPP libraries. To obtain real-time performances on standard VGA image we are implementing the proposed algorithm on architectures based on GPU.

From a broader perspective, we observe that the design approach followed for the cortical architecture leads to computational advantages, when the solution of a perceptual task of higher complexity requires the integration of several features, such as orientation, binocular disparity, and motion. By example, in typical real-world situations, as it occurs during stereo camera movements, stereopsis extends to time-varying images, thus requiring the integration of static (binocular) and temporal correspondence [25]. In these situations, by generalizing the local visual operators to binocular motion energy units [26], the architecture will be able to provide an early rich description of 3D motion events. In this way, coherent stereo-motion correspondence constraints will be directly embedded in the structure of such visual operators, rather than being considered at higher semantic level of data fusion.

Acknowledgements

This work has been partially supported by EU Projects FP7-ICT 217077 “EYESHOTS” and FP7-ICT 215866 “SEARISE” and by “Progetto di Ricerca di Ateneo 2007” (University of Genoa).

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. In: ICCV (2007)
2. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. of Computer Vision 47, 7–42 (2002)
3. Tombari, F., Mattoccia, S., Di Stefano, L., Addimanda, E.: Classification and evaluation of cost aggregation methods for stereo correspondence. In: CVPR (2008)
4. Milner, A.D., Goodale, M.: The visual brain in action. Oxford Univ. Press, Oxford (1995)
5. Heeger, D.: Model for the extraction of image flow. JOSA 4(8), 1455–1471 (1987)

6. Grzywacz, N., Yuille, A.: A model for the estimate of local image velocity by cells in the visual cortex. Proc. R. Soc. Lond. B 239, 129–161 (1990)
7. Chen, Y., Qian, N.: A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. Neural Computation 16, 1545–1577 (2004)
8. Fleet, D., Wagner, H., Heeger, D.: Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. Vision Res. 36(12), 1839–1857 (1996)
9. Bayerl, P., Neumann, H.: A fast biologically inspired algorithm for recurrent motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. 29(2), 246–260 (2007)
10. Shimonomura, K., Kushima, T., Yagi, T.: Binocular robot vision emulating disparity computation in the primary visual cortex. Neural Networks 21(2-3), 331–340 (2008)
11. Higgins, C., Shams, S.: A neuromorphic vision processor for spatial integration of optical flow. In: ICCNS 2001 (2001)
12. Dale, J., Johnston, A.: A real-time implementation of a neuromorphic optic-flow algorithm. Perception 31, 136 (2002)
13. Pouget, A., Dayan, P., Zemel, R.S.: Inference and computation with population codes. Ann. Rev Neurosci 26, 381–410 (2003)
14. Adelson, E., Bergen, J.: The plenoptic and the elements of early vision. In: Landy, M., Movshon, J. (eds.) Computational Models of Visual Processing, pp. 3–20. MIT Press, Cambridge (1991)
15. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. JOSA 2, 284–321 (1985)
16. Ohzawa, I., De Angelis, G., Freeman, R.: Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249, 1037–1041 (1990)
17. Morgan, M.J., Castet, E.: The aperture problem in stereopsis. Vision Res. 37(19), 2737–2744 (1997)
18. Serrano-Pedraza, I., Read, J.C.A.: Stereo vision requires an explicit encoding of vertical disparity. J.Vision 9(4), 1–12 (2009)
19. Daugman, J.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. JOSA A/2, 1160–1169 (1985)
20. Nestares, O., Navarro, R., Portilla, J., Tabernero, A.: Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. J. of Electronic Imaging 7(1), 166–173 (1998)
21. Pauwels, K., Hulle, M.V.: Optic flow from unstable sequences containing unconstrained scenes through local velocity constancy maximization. BMVC 1, 397–406 (2006)
22. Theimer, W., Mallot, H.: Phase-based binocular vergence control and depth reconstruction using active vision. CVGIP: Image Understanding 60(3), 343–358 (1994)
23. Chessa, M., Solari, F., Sabatini, S.: A virtual reality simulator for active stereo vision systems. In: VISAPP (2009)
24. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. Int. J. of Computer Vision 12, 43–77 (1994)
25. Jenkin, M., Tsotsos, J.: Applying temporal constraints to the dynamic stereo problem. In: CVGIP, vol. 33, pp. 16–32 (1986)
26. Sabatini, S., Solari, F., Cavalleri, P., Bisio, G.: Phase-based binocular perception of motion in depth: Cortical-like operators and analog VLSI architectures. EURASIP Journal on Applied Signal Processing 7, 690–702 (2003)