Stereoscopic depth analysis in video-real time based on visual cortical cell behavior - an FPGA solution

F. Wörgötter
Dept. of Psychology, University of Stirling, Stirling FK9 4LA, UK. Email: worgott@neurop.ruhr-uni-bochum.de

The goal of neuromorphic engineering is threefold: 1) Try to gain insight into neuronal behavior through electronically realized models (viz. chips) 2) Try to make advancements in the field of complex parallel electronic micro-circuitry design and 3) try to arrive at an industrially applicable product which could be relevant for high-tech domains like robotics or computer vision. In this article I will present a set of solutions addressing the problem of real-time depth analysis from stereoscopic images which have taken us in this field of problems - to my believe - a little closer to the first and third goal defined above. In a stereoscopic system both eyes or cameras have a slightly different view. As a consequence small variations between the projected images exist ("disparities") which are *spatially* evaluated in order to retrieve depth information [7, 25]. I will show that two related algorithmic versions can be designed which recover disparity. Both approaches are based on the comparison of filter outputs from filtering the left and the right image. The difference of the phase components between left and right filter responses encodes the disparity. The first approach, which will be described, very strongly relates to the behavior of visual cortical simple and complex cells. The second approach uses the apparently paradoxical similarity between the analysis of visual disparities and the determination of the azimuth of a sound source [27]. Animals determine the direction of the sound from the *temporal* delay between the left and right ear signals [12]. Similarly, in the second approach [22] I transpose the spatially defined problem of disparity analysis into the temporal domain and utilize two resonators implemented in the form of causal (electronic) filters to determine the disparity as local temporal phase differences between the left and right filter responses. This approach permits video real-time analysis of stereo image sequences (see movies at http://www.neurop.ruhr-uni-bochum.de/Real-Time-Stereo) and a FPGA-based PC-board has been developed which performs stereo-analysis at full PAL resolution in video real-time. The software version is already used in industrial applications.

1 Introduction

When talking about neuromorphic engeneering usually those approaches are discussed where cell- or membrane characteristics are modeled with sub-threshold transistor technology. In this article I will follow a different strategy and describe how electronic filter circuits can be used to mimic neuronal behavior while at the same time these circuits are extremely well suited to implement them in VLSI hardware. The target application is: *stereoscopic depth analysis.*

In general there are several strategies of how to retrieve depth information from a sequence of images, like depth from motion (flow-field analysis), depth from shading and depth from stereopsis, on which I concentrate in this article. In a stereoscopic approach usually two cameras are mounted with a horizontal distance between them. As a consequence objects displaced in depth from the fixation point are projected onto image regions which are horizontally shifted with respect to the image center. This shift is called *disparity* and it can be used to determine the depth of the object. Due to the geometry of the optic system it is thereby sufficient to restrict disparity analysis to the projection of corresponding linear segments (lines) in the left and right eye (epipolar line constraint). It is therefore not necessary to extend the problem to two dimensions, which raises computational complexity. This can, however, improve the results.

In the most straightforward approaches that address the problem of depth from stereo, the disparity is computed by searching the maximum of the cross-correlation between image windows along the epipolar lines of the left and right image [10]. This algorithmic solution, however, bears little realism in comparison to the behavior of visual cortical neurons. Thus, a different group of algorithms for disparity analysis has been designed more recently based on spatially localized band-pass filters. This method computes the convolution between Gabor kernels (Eq. 1) and the left and right image parts. It is by now largely accepted that the shape of the

receptive fields of visual cortical simple cells resembles such Gabor filters. Thus, in this approach cell responses are linearly approximated by the filter convolutions.

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{\frac{-(x-x_0)^2}{2\sigma^2}} \mathrm{e}^{\mathrm{i}(kx-\phi)} \qquad (1)$$

Each convolution result consists of an amplitude and a phase value. The disparity is computed from the difference of the two phase values obtained from the left and right images divided by the filter tuning frequency k. The amplitude of the filter response can be used to estimate the reliability of the obtained result: the bigger, the more reliable is the phase difference. If the amplitude is zero, obviously, the phase is ill-defined. Since this idea was introduced by Sanger in 1988 [25] a large body of literature has been devoted to these approaches [2, 6, 7, 8, 23, 24], which are commonly called *phase-based stereo algorithms.* These studies are concerned with the theory of cortical disparity processing and, as a consequence, fail to accommodate industrial requirements concerning processing speed and accuracy[2].

Correlation techniques and phase based stereo algorithms are acausal in the sense that data acquisition of at least parts of the image needs to be completed before the computation of the disparity can start. It is believed that these techniques could play a major role in the process of disparity analysis in the mammalian brain because visual cortical cells have receptive fields with a Gabor filter profile[4, 11]. The vast number of cortical cells allows for an efficient parallel processing and thereby animals and humans can react to a changing depth structure in their environment in "real-time". When dealing with an artificial computer vision system, one has to realize that phase-bases disparity analysis is indeed a computationally rather slow process, because many convolutions have to be calculated.

In this study I describe two different approaches: In the first part I will briefly introduce a very compact formalism for the conventional spatial phase-based stereo algorithm. In the second part I describe a novel, causal, real-time phase-based algorithm to determine the disparities in two stereo images[22]. The central idea behind this approach is to transpose the spatially-defined problem of disparity estimation into the temporal domain and compute the disparity simultaneously with the incoming data flow.

2 The acausal, spatial filtering, neuronal approach

Simple cell responses in the visual cortex can be described by Gabor filters [4, 11]. A Gabor function (Eq. 1) is a sine-wave multiplied and, thus, damped by a Gaussian envelope [9]. Thus, these cells represent localized spatial band-pass filters which are tuned to the resonance frequency k of the sine-wave and located at x_0 in the visual field where the Gaussian envelope has its center. In Eq. 1 σ is related to the width of the receptive field. The phase parameter ϕ represents the fact that most cells in the visual cortex have a receptive field which is mixed from a pure cosine- and a pure sine-type. I will set x_0 and ϕ to zero because they do not affect our results except by adding unnecessary mathematical complexity. Thus, in Eq. 1 the real component represents a cosine- and the imaginary component a sine-shaped receptive field. These are the archetypes of receptive fields that exist in monocularly driven simple cells. The linear part of the response of such a cell is given by the convolution of the receptive field with the stimulus f(x):

$$M_{l,r}(x) = G(x) * f_{l,r}(x) \qquad (2)$$
$$= \int_{-\infty}^{+\infty} G(x-x') f_{l,r}(x') dx'$$

It can be shown that these monocular simple cell responses can be combined to construct complex cells responses by means of designing quadrature pairs [13, 19, 20, 21] utilizing a push-pull arrangement [5, 18, 20, 26, 28] between corresponding simple cells. Details of this algorithms cannot be publicly laid open because of an IPR protection agreement with our industrial partner (I-to-I, Hamburg, http://www.I-to-I.de).

The theory outlined in the first approach towards stereo disparity analysis suggests that neuronal operations in simple and complex cells - many of which have already been observed experimentally [3, 14, 15, 16, 17, 23] - can in a very direct way lead to disparity estimates of the objects in a visual scene. Thus, it seems that the computation of visual disparities, which is a central component for the perception of depth, is already to a large degree solved by the cells in the primary visual cortex. It should be noted that this approach is very well suited for a parallel system like the cortical neural network. There, parallel spatial filtering operations will permit real-time stereo vision in animals and humans. Conventional computer vision system, however, do not operate in parallel. Here, spatial convolutions are very time-consuming operations and real-time performance is prevented this way.

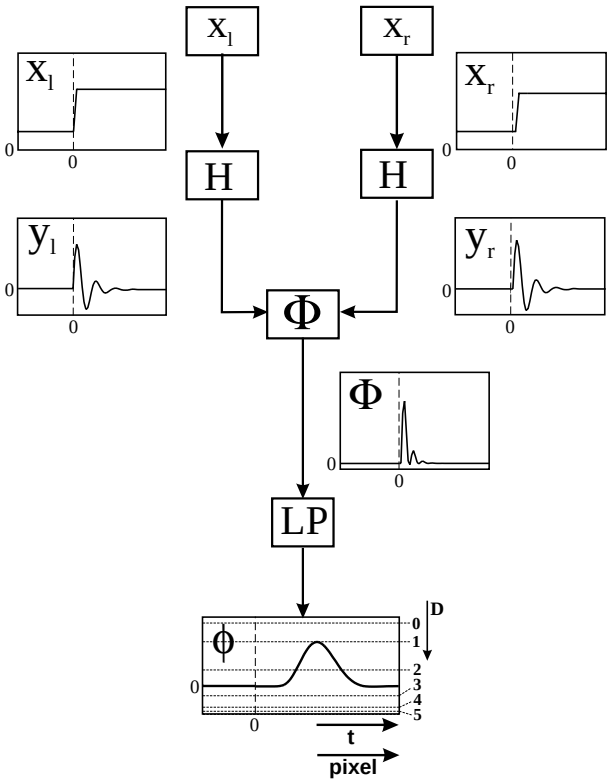

Figure 1. *Block diagram of the computational process and results of a disparity estimation from the two input step functions x_r and x_l.*

3 The causal, temporal filtering, computer vision approach

In order to deal with the restrictions of a conventional computer vision system we realize that camera signals are temporal signals [22] [patent pending]. Thus, one can now take the luminance signal of the image scan-lines from the left and the right image and pipe it through a left and a right temporal band-pass filter (a resonator). This filter function looks like the right half of a sine-wave Gabor filter. This way two signals are generated which are quasi-oscillatory at the resonance frequency. As before it is the (local) phase difference between these two oscillations which is directly equivalent to the disparity. Thus, subsequently our system measures this phase difference by two more simple electronic operations as shown in Fig. 1 and explained below. In order to be allowed to do this I assume a fronto-parallel camera arrangement which leads to horizontal epipolar lines.

Let $x_l(t)$, $x_r(t)$ be the two corresponding pixel lines of a stereo image pair in which a single contrast step exists at different disparities (viz. different times t_l and t_r).

The two step functions $x_l(t) \leftrightarrow X_l(s)$ and $x_r(t) \leftrightarrow X_r(s)$ are defined in the Laplace domain by (Fig.1):

$$X_l(s) := \frac{1}{s}\mathrm{e}^{-t_l s}, \quad and \quad X_r(s) := \frac{1}{s}\mathrm{e}^{-t_r s}, \tag{3}$$

and the transfer function of the resonator is given as:

$$H(s) = \frac{s}{(s - s_\infty)(s - s_\infty^*)} \tag{4}$$

where s_∞ is a filter pole and specifies the filter characteristic defined by f_0 and the filter quality Q, which determines the damping; the "$*$" denotes the complex conjugate.

$$\mathrm{Re}(s_\infty) = -2\pi f_0/2Q \tag{5}$$

$$\mathrm{Im}(s_\infty) = \sqrt{(2\pi f_0)^2 - (\mathrm{Re}(s_\infty))^2} \tag{6}$$

Convolution of signal and filter yields for the right image:

$$Y_r(s) = X_r(s)H(s) = \frac{s}{(s - s_\infty)(s - s_\infty^*)}\frac{1}{s}\mathrm{e}^{-t_r s} \tag{7}$$

A similar convolution is performed for the left image. I define $a := (s_\infty - s_\infty^*)^{-1}$, then the inverse Laplace transformation of $Y_r(s)$ yields:

$$y_r(t) = \begin{cases} a\mathrm{e}^{s_\infty(t-t_r)} + a^*\mathrm{e}^{s_\infty^*(t-t_r)} & \text{if } t \geq t_r \\ 0 & \text{if } t < t_r \end{cases} \tag{8}$$

The temporal resonator signal $y(t)$ reflects a damped sine-wave with frequency f_0 (Fig. 1, y_l, y_r). The number of full cycles until the signal fades is roughly equivalent to the value of Q. Note that any DC component present in the input signal is removed by the resonator. This is an advantage of the new method because the DC usually poses a severe problem in all spatial filter approaches [2, 7, 25].

Finally, disparity is determined from the phase difference between the resonator signals from both images. Phase comparison is achieved by multiplication of the two signals in the time domain and subsequent low-pass filtering (Fig. 1, Φ, LP). Multiplication yields (Fig. 1, Φ):

$$\Phi(t) = y_l(t)y_r(t) = \begin{cases} g_{2f_0}(t) + \phi(t) & \text{if } t \geq t_r \\ 0 & \text{if } t < t_r \end{cases} \tag{9}$$

with:

$$g_{2f_0}(t) = \underbrace{a^2\mathrm{e}^{s_\infty(2t-t_l-t_r)} + a^{*2}\mathrm{e}^{s_\infty^*(2t-t_l-t_r)}}_{\text{double frequency term}} \tag{10}$$

and

$$\phi(t) = \underbrace{2|a|^2\cos\left[(t_r - t_l)\mathrm{Im}(s_\infty)\right]}_{K}\mathrm{e}^{\mathrm{Re}(s_\infty)(2t-t_r-t_l)} \tag{11}$$

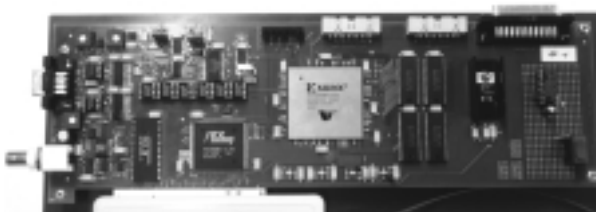

Figure 2. The PC-compatible board for real-time stereo analysis.

The term $g_{2f_0}(t)$ reflects an oscillation with $2f_0$. In an implementation it will be eliminated by low-pass filtering with low cut-off (Fig. 1 LP). The second part represents the phase $\phi(t)$ between the two signals and contains an exponential relaxation term and a constant term K, which encodes the true disparity.

$$K = \frac{Q^2}{2\pi^2 f_0^2 (4Q^2 - 1)} \cos\left[(t_r - t_l)\mathrm{Im}(s_\infty)\right] \quad (12)$$

The disparity which is the spatial equivalent of $t_r - t_l$ can be computed by inverting Eq. 12 and is obtained immediately at the second contrast step (i.e., for $t = t_r$), after which the signal relaxes to zero. This relaxation behavior which originates from the characteristic of the resonator assures temporal (viz. spatial) locality. Otherwise only the average phase (viz. disparity) of each image line could be computed. In order to make this algorithm applicable the output signal needs to be normalized to be independent of overall luminance variations.

4 The board

The block diagram in Fig. 1 shows that this system can be easily implemented in hardware. We used an XILINX FPGA processor to design an prototype PC-compatible board for real-time stereo image analysis (Fig. 2). This was done in cooperation with the "Mikroelektronik Anwenderzentrum (MAZ) Hamburg-Harburg". In such a system the disparity is determined continuously from the incoming data and thereby real-time performance is achieved. The board currently operates at 25 Hz and full PAL resolution. Thereby it is a factor of more than 100 faster than any stereo-system based on a conventional Pentium-PC. Real-time MPEG encoded movies which demonstrate the performance of the board can be viewed on our internet page (http://www.neurop.ruhr-uni-bochum.de/Real-Time-Stereo).

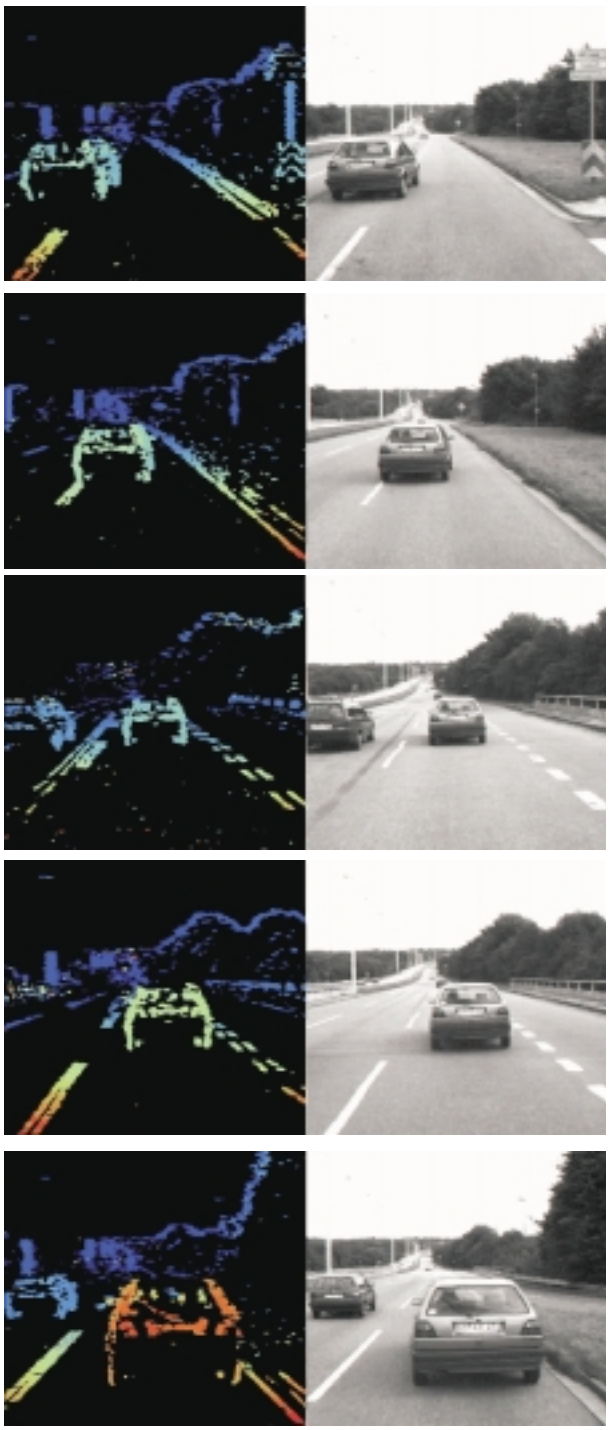

Figure 3. Original images and disparity maps obtained from the causal algorithm and taken from an outdoor driving scene (http://www.neurop.ruhr-uni-bochum.de/Real-Time-Stereo/motions/dra1_s.mpg)

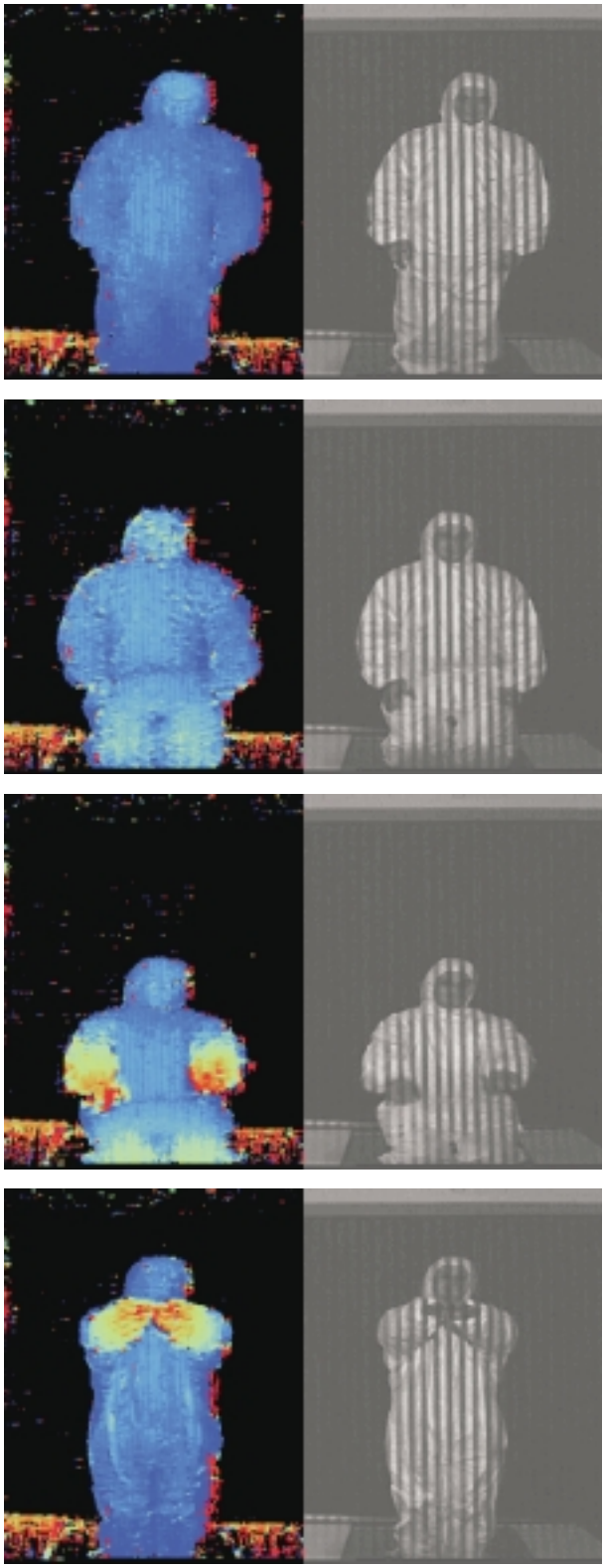

Figure 4. Original images and disparity maps obtained from the causal algorithm and taken from an indoor scene of a moving person illuminated with a grating pattern to increase texture (http://www.neurop.ruhr-uni-bochum.de/Real-Time-Stereo/motions/hanjo_s.mpg)

5 Results

Figures 3 and 4 show a few results obtained with the causal algorithm applying it to an outdoor scene with natural light or alternatively indoors using artificial stripe-illumination. The outdoor scene was obtained mounting the stereo-camera setup on a regular car ski-rack and driving on a main motorway with about 100 km/h. Camera frame rate was 25 Hz. The images were recorded while a car was overtaking our vehicle after which we accelerated in order to reduce the distance. The accuracy of the algorithm is on average 0.2 pixels disparity which with this setup amounts to about 50cm at a distance of 50m. In general, disparity and distance scale hyperbolically such that the accuracy increases with decreasing distance.

The indoor scene mimics a situation often encountered in industrial environments: objects with little texture (structure). All triangulations methods (to which these two algorithms also belong), however, require texture boundaries which can be compared between the left and the right image. Without texture depth cannot be measured. Commonly active illumination is used in order to cope with this problem. However, most conventional algorithms infer depth directly from the perspective distortions which occur as a consequence of the depth-structure of the illuminated object. They are, therefore, extremely sensitive to optically induced distortions in the illumination pattern itself. Both algorithms presented here need texture only in order to "excite" the filter circuits and they do not care about the special structure of the illumination pattern (also white-noise would have done the job). As a consequence, the depth structure of the scene can be retrieved much more reliably.

6 Conclusions

The first goal of this study was to demonstrate that the stereoscopic depth analysis problem can be successfully tackled within the well-known framework of phase-based stereo algorithms reaching a degree of performance that permits industrial application. To this end I have shown that it is possible to design two versions of phase-based stereo algorithms: One which is compatible with neuronal operations in the visual cortex [14, 15, 16, 17, 23] and another one which operates causally and is therefore much faster in a conventional computer vision environment. As a final goal we were able to design an FPGA based version of the algorithm which operates in video real time. Even in a time of ever increasing conventional computer power, video real-time image processing is

still a massive challenge.

The second goal was to devise an alternative strategy in neuromorphic engeneering: The use of neural filter algorithms and their embedding into VLSI hardware structures. Filter algorithms have long been discussed as a valid level of abstraction for the description of neuronal behavior, like the receptive field structure of cortical simple cells. Strictly speaking, these filters cover only the linear part of the cell responses and certain "tricks" have to be applied as laid out in the cited literature in order to create a reasonable match between cell and filter behavior. For example, the most obvious and most easily taken care of deviation from linearity is the half-wave rectification behavior of neuronal impulse rates. However, despite the fact that concept receptive field filters has been so successful when trying to describe cell behavior, so far these filters have not played any great role in electronic circuit design. This is probably due to the fact that almost all visual problems rest on the analysis of 2-dimensional image "patches" and therefore require 2-dimensional receptive field filters. The 2-d convolutions involved are computationally expensive and prevent efficient on-chip implementation. Stereoscopic depth analysis is intrinsically a 1-dimensional problem. Even if we give up the parallel camera setup and allow for vergent viewing angles matching stereo points will still fall onto lines. In this case the epipolar lines are not anymore horizontal but instead tilted. Electronically the problem of tilted epipolar lines can be easily solved by an image rectification step, which can in principle also be performed in hardware as a pixel index warping. Thus, the stereoscopic depth analysis problem is ideally suited to apply a 1-dimensional neuronal filter algorithm which in itself is very well suited for VLSI implementation. Therefore, in this case the transition from neuronal behavior to hardware implementation can be performed rather smoothly.

7 Acknowledgements

The author acknowledges the support of the Transfer Programm within SFB 509 of the Deutsche Forschungsgemeinschaft.

References

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A. 2:284-299, 1985.

[2] A. Cozzi, B. Crespi, F. Valentinotti and F. Wörgötter. Performance of phase-based algorithms for disparity estimation. Machine Vis. Appl. 9:334-340, 1997.

[3] G.C. DeAngelis, I. Ohzawa and R.D. Freeman. Depth is encoded in the visual cortex by a specialized receptive field structure. Nature 352: 156-159, 1991.

[4] J.G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. Vision Res. 20:847-856, 1980.

[5] D. Ferster. Spatially opponent excitation and inhibition in simple cells of the cat visual cortex. J. Neurosci. 8:1172-1180, 1988.

[6] D.J. Fleet and A.D. Jepson. Stability of phase information. IEEE Trans. PAMI, 15(12):1253-1268, 1993. http://www.qucis.queensu.ca:1999/ fleet /stability.ps.

[7] D. Fleet, A. Jepson and M. Jenkin. Phase-based disparity measurement. Comp. Vis., Graph. Image Proc. 53:198-210, 1991.

[8] D.J. Fleet, H. Wagner and D.J. Heeger. Neural Encoding of Binocular Disparity: Energy Models, Position Shifts and Phase Shifts, Vision Res. 36:1839-1858, 1996.

[9] D. Gabor. Theory of communication. J. IEEE Lond. 93:429-457, 1946.

[10] R.M. Haralick and L.G. Shapiro. *Computer and Robot Vision*, Vol. 2. Addison Wesley, 1992.

[11] J.P. Jones and L.A. Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. J. Neurophysiol. 58:1187-1211, 1987.

[12] M. Konishi and W. Sullivan. Neural map of interaural phase difference in the owl's brainstem. Proc Natl Acad Sci USA, 83:8400-8404, 1986.

[13] Z. Liu, J.P. Gaska, L.D. Jacobson and D.A. Pollen. Interneuronal interaction between members of quadrature phase and anti-phase pairs in the cat's visual cortex. Vision Res. 32:1193-1198, 1992.

[14] M. Nomura, G. Matsumoto and S. Fujiwara. A binocular model for the simple cell. Biol. Cybern. 63:237-242, 1990.

[15] I. Ohzawa, G.C. DeAngelis and R.D. Freeman. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249:1037-1041, 1990.

[16] 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.

[17] A. Anzai, I. Ohzawa and R.D. Freeman. Neural mechanism for encoding binocular disparity: receptive field position versus phase. J. Neurophysiol. in press, 1999.

[18] L.A. Palmer, J.P. Jones and R.A. Stepnowski. Striate receptive fields as linear filters: characterization in two dimensions of space. In: Cronly-Dillon J, Leventhal AG (eds). *The neural basis of visual function.* (Vision and visual dysfunction, vol 4) Macmillan, London, 1991.

[19] D.A. Pollen and S.E. Ronner. Phase relationships between adjacent simple cells in the visual cortex. Science 212:1409-141, 1981.

[20] D.A. Pollen and S.E. Ronner. Spatial computation performed by simple and complex cells in the visual cortex of the cat. Vision Res. 22:101-118, 1982.

[21] D.A. Pollen and S.E. Ronner. Visual cortical neurons as localized spatial frequency filters. IEEE Trans. SMC 13:907-915, 1983.

[22] B. Porr, A. Cozzi and F. Wörgötter. How to "hear" visual disparities: real-time stereoscopic depth analysis using temporal resonance. Biol. Cybern. 78:329-336, 1998.

[23] N. Qian and Y. Zhu. Physiological Computation of Binocular Disparity, Vision Res. 37:1811-1827, 1997.

[24] N. Qian and S. Mikaelian. Relationship between Phase and Energy Methods for Disparity Computation, Neural Comp., in press, 1999.

[25] T.D. Sanger. Stereo disparity computation using gabor filters. Biol. Cybern. 59:405-418, 1988.

[26] D.J. Tolhurst and A.F. Dean. Spatial summation by simple cells in the striate cortex of cat. Exp. Brain Res. 66:607-620, 1987.

[27] H. Wagner and B. Frost. Disparity-sensitive cells in the owl have a characteristic disparity. Nature, 364:796-757, 1993.

[28] F. Wörgötter, E. Nelle, B. Li, L. Wang and Y. Diao. A possible basic cortical microcircuit called "cascaded inhibition" Results from cortical network models and recording experiments from striate simple cells. Exp. Brain Res. 122:318-332, 1998.