A Model for the Estimate of Local Velocity

N. M. Grzywacz (C.B.I.P., M.I.T.) and A. L. Yuille (D.A.S. Harvard)

Abstract

Motion sensitive cells in the primary visual cortex are not selective to velocity, but rather are directionally selective and tuned to spatiotemporal frequencies. This paper describes physiologically plausible theories for computing velocity from the outputs of spatiotemporally oriented filters and proves several theorems showing how to combine the outputs of a class of frequency tuned filters to detect local image velocity. Furthermore, it can be shown (Grzywacz and Yuille 1990) that the filters' combination may simulate "Pattern" cells in the middle temporal area (MT), while each filter simulates primary visual cortex cells. This suggests that MT's role is not to solve the aperture problem, but to estimate velocities from primary cortex information. The spatial integration that accounts for motion coherence may be postponed to a later cortical stage.

1 Introduction

This paper gives a brief summary of a theory for motion estimation. We concentrate here on the mathematical aspects of the theory. The reader is referred to Grzywacz and Yuille (1990) for proofs of the theorems, comparisons of the theory to neurobiology and detailed references to the literature (these references alone take up over five pages).

Motivated by neuroscientific experiments the theory assumes that the motion is first filtered by spatiotemporally tuned filters and only later is velocity explicitly computed. It is, therefore, related to several existing models (Hassenstein and Reichardt 1956; Poggio & Reichardt 1976; van Santen and Sperling 1984; Watson & Ahumada 1985; Jasinschi 1988; Bulthoff, Little & Poggio 1989, Fleet & Jepson 1989) and, in particular, to theories involvinge spatiotemporally oriented motion energy filters (Adelson & Bergen 1985). Our model is closely related to the elegant model of Heeger (1987) that computes velocities through the spatiotemporal integration of the outputs of Gabor motion energy filters (Gabor 1946; Daugman 1985). Unfortunately his method of computing the velocities assumes that the image's power spectrum is flat, which is often incorrect. Our theoretical results show that this assumption is unnecessary.

The model has two stages: The first measures motion energies (the output of motion energy filters) and the second estimates velocity from these energies. Section 2 describes the first stage and Section 3 gives mathematical results used as a basis for the second stage in Section 4.

2 Model Description

Following previous work (Adelson & Bergen 1985, Heeger 1987) we use a (complex) spacetime Gabor filter (Gabor 1946; Daugman 1985) $F\left(\vec{x}, t : \Omega, \vec{n}, \Omega_t, \sigma, \sigma_t\right)$ where the σ's are the standard deviations and the Ω's are the frequencies of the filter.

We model the responses of directionally selective cells in primary cortex to an image, $I(\vec{x},t)$, as the nonlinear filter

$$N\left(\vec{x},t:\Omega,\vec{n},\Omega_t,\sigma,\sigma_t\right)=\left|F\left(\vec{x},t:\Omega,\vec{n},\Omega_t,\sigma,\sigma_t\right)*I\left(\vec{x},t\right)\right|^2, \tag{2.1}$$

where $*$ stands for convolution. This definition is similar to the one proposed by Adelson and Bergen (1985) who call it motion energy.

The Fourier tranform of s a Gabor is a Gaussian centered on $(\Omega\vec{n},\Omega_t)$. This shows that the nonfilter N is not directly tuned to velocity. However, although one filter cannot estimate the velocity, $\vec{v}$, as we will show in the next section, the set of filters responding most vigorously can.

A physiological interpretation for the model is developed in Grzywacz and Yuille (1990). A consequence of this interpretation is that the bandwidth of the temporal frequency tuning curves is relatively wide compared to the spatial bandwidth. Precisely, the assumption states that for all velocities, $\vec{v}$, to which the cells respond the following relationship holds: $(|\vec{v}|\sigma_t)^2 \ll \sigma^2$. Informally, it was verified by literature inspection that typically $3 \leq (\sigma/(|\vec{v}|\sigma_t))^2 \leq 60$.

3 Velocity Estimation

The spatiotemporal power spectrum of a translating image lies on the plane $\vec{\omega}\cdot\vec{v}+\omega_t=0$ in the frequency domain (Watson & Ahumada 1985; Heeger 1987; Daugman 1988). This suggests using the combination of the outputs of cells tuned to specific spatiotemporal frequencies to detect this plane. Our results show how to combine these cells' responses in a computationally sensible way.

The following theorem shows that, if the filters are Gabor with constant standard deviation, then the maximal response lies on a plane in the space of cell parameters. Knowledge of this plane determines the velocity. This result does *not* follow trivialy from the knowledge that the spatiotemporal power spectrum of a translating image lies on a plane. Indeed, one can show that filters other than Gabor filters do not have the same property (this is related to the scale–space theorems; Yuille & Poggio 1986). The theorem is strictly correct only when the receptive field sizes and temporal windows are constant for all cells. However, in Theorem 3 and its corollary, we show that this constancy requirement can be relaxed under physiological conditions.

Theorem 1. *If σ and σ_t are constants, then the local maxima of $N(\vec{x},t:\Omega,\vec{n},\Omega_t,\sigma,\sigma_t)$ as a function of $(\Omega,\vec{n},\Omega_t)$ lie on the plane $\Omega\vec{n}\cdot\vec{v}+\Omega_t=0$ for all images that move with a constant velocity $\vec{v}$.*

This results follows from a corollary of a stronger result: Theorem 2.

Theorem 2 will provide the response distribution in the three–dimensional space defined by the cells' optimal spatial and temporal frequencies.

Theorem 2. *The response $N(\vec{x},t:\vec{\Omega},\Omega_t,\sigma,\sigma_t)$ is weakly separable as follows: A function p exists such that $N(\vec{x},t:\vec{\Omega},\Omega_t,\sigma,\sigma_t)=p(\vec{x},t:\sigma^2\vec{\Omega}-\sigma_t^2\Omega_t\vec{v},\sigma,\sigma_t)\exp(-(\sigma_t^2\sigma^2)\left(\Omega_t+(\vec{\Omega}\cdot\vec{v})\right)^2/2(\sigma^2+\sigma_t^2v^2))$. Hence the only dependence of N on the spatial characteristics of the stimuli occurs within the function p.*

Proof: See Grzywacz and Yuille (1990).

The following corollary shows that if the receptive field sizes and temporal windows are constant, then the responses follow a known Gaussian distribution centered on the plane of Theorem 1. Thus, the claim in Theorem 1 follows from this corollary.

Corollary 1. *If σ and σ_t are constants, then the variation of $N(\vec{x},t:\vec{\Omega},\Omega_t,\sigma,\sigma_t)$ in the $(\vec{\Omega},\Omega_t)$ space in the direction $((\vec{v}\sigma_t)/\sigma^2,1/\sigma_t)$ is a Gaussian function centered on the plane $\Omega\vec{n}\cdot\vec{v}+\Omega_t=0$, and dependent only on $\vec{v}$, σ, and σ_t.*

Proof: The arguments $(\sigma^2\vec{\Omega}-\sigma_t^2\Omega_t\vec{v})$ of the function p do not vary in this direction. The only variation is therefore due to the Gaussian function $\exp(-(\sigma_t^2\sigma^2)\left(\Omega_t+(\vec{\Omega}\cdot\vec{v})\right)^2/2(\sigma^2+\sigma_t^2v^2))$. This Gaussian is centered on the plane $\Omega\vec{n}\cdot\vec{v}+\Omega_t=0$.

Theorem 3 shows that the response distribution in the optimal-frequency space is of a simple form. This result is important, because the receptive field sizes and temporal windows may depend on the cells' optimal frequencies (Section 2). We denote these dependencies by $\sigma(\Omega) = K(|\vec{\Omega}|)/|\vec{\Omega}|$ and $\sigma_t(\Omega_t) = K_t(\Omega_t)/|\Omega_t|$, where K and $K_t(\sigma_t)$ are functions that are mildly dependent , or perhaps independent, of σ and σ_t respectively. More precisely, Theorem 3 assumes that for all velocities, $\vec{v}$, to which the cells respond, $(|\vec{v}|\sigma_t)^2 \ll \sigma^2$. An informal literature study seems to justify this assumption.

Theorem 3. *Given the approximation $|\vec{v}|^2 \ll (\sigma/\sigma_t)^2$, and remembering that $\sigma_t = \sigma_t(\Omega_t)$ and $\sigma = \sigma(\Omega)$, we find that the response $N(\vec{x},t : \vec{\Omega}, \Omega_t, \sigma, \sigma_t)$ is weakly separable in the sense that there exists a function r, independent of Ω_t and σ_t, such that $N(\vec{x},t : \vec{\Omega}, \Omega_t, \sigma, \sigma_t) \approx r(\vec{x},t : \vec{\Omega}, \sigma)$ $\times \exp\{-\sigma_t^2(\Omega_t + (\vec{\Omega} \cdot \vec{v}))^2/2\}$.* *Proof*: See Grzywacz and Yuille (1990).

Corollary 2 shows that in the three-dimensional space of optimal frequencies, the response distributions as function of temporal frequency have maxima on the plane defined in Theorem 1. This means that the overall distribution has a maximal ridge on the plane. Under the approximation $(|\vec{v}|\sigma_t)^2 \ll \sigma^2$, the distribution of motion energies is oriented parallel to temporal frequency axis.

Corollary 2. *With the same assumptions and approximations as Theorem 3, along a one-dimensional line parallel to Ω_t axis, the maximum of $N(\vec{x},t : \vec{\Omega}, \Omega_t, \sigma, \sigma_t)$ lies on the plane $\vec{\Omega} \cdot \vec{v} + \Omega_t = 0$.*

Proof: Consider the set of lines parallel to the Ω_t axis. The only variation of $N(\vec{x},t : \vec{\Omega}, \Omega_t, \sigma, \sigma_t)$ is due to the $\exp\{-\sigma_t^2(\Omega_t + (\vec{\Omega} \cdot \vec{v}))^2/2\}$ term, which is unimodal with maximum at $\vec{\Omega} \cdot \vec{v} + \Omega_t = 0$.

Strictly speaking these theorems assume the velocity is constant over the whole image. Since, however, the filters have limited spatiotemporal range (determined by σ and σ_t) the velocity need only be approximately constant over this range.

4 Strategies and Neural Implementations for Velocity Estimation

We now describe three related methods for finding the velocity of the stimulus using the mathematical results of the previous section, and discuss possible neural implementations. The computational, psychophysical, and implementational aspects of this problem are described in Grzywacz and Yuille (1990).

The Ridge Strategy: This strategy uses Corollary 2 as a starting point and proposes to make excitatory connections from each motion-energy cell to the velocity selective cells most consistent with it. These connections should weakly prefer velocities with small components perpendicular to the preferred direction, so as to give a unique answer for the aperture problem in the large. Suppose we have a set of M motion-energy cells $(\vec{\Omega}^\mu, \Omega_t^\mu, \sigma^\mu, \sigma_t^\mu)$ with $\mu = 1, ..., M$. A possible implementation is to define the response, $R(\vec{x}, t : \vec{v})$, at time t of a velocity selective cell tuned to velocity $\vec{v}$, and whose receptive field is centered at position $\vec{x}$, by

$$R(\vec{x}, t : \vec{v}) = A \sum_\mu N(\vec{x}, t : \vec{\Omega}^\mu, \Omega_t^\mu, \sigma^\mu, \sigma_t^\mu) e^{-(\sigma_t^\mu)^2(\Omega_t^\mu + (\vec{\Omega}^\mu \cdot \vec{v}))^2/2} e^{-(\vec{v} \cdot \vec{\Omega}^{\mu *}/k)^2}, \tag{4.1}$$

where $\vec{\Omega}^*$ is orthogonal to $\vec{\Omega}$, and A and k are constant parameters.

This equation suggests that the strength of the connection between cell $(\vec{\Omega}^\mu, \Omega_t^\mu, \sigma^\mu, \sigma_t^\mu)$ and the velocity selective cell tuned to the velocity $\vec{v}$ should be $\exp\{-(\sigma_t^\mu)^2(\Omega_t^\mu + (\vec{\Omega}^\mu \cdot \vec{v}))^2/2\} \exp\{-(\vec{v} \cdot \vec{\Omega}^{\mu *}/k)^2\}$.

This method is similar to correlation and template matching methods in computer vision. If we fix $\vec{\Omega}$ and let Ω_t vary, then from Theorem 3, we know that the form of the variation of the filtered response is $\exp\{-\sigma_t^2(\Omega_t + (\vec{\Omega} \cdot \vec{v}))^2/2\}$; this defines our template. The largest value of the correlation of this template with $N(\vec{x},t:\vec{\Omega}^{\mu},\Omega_t^{\mu})$, as we vary the value of $\vec{v}$ while fixing $\vec{\Omega}$, gives an estimate for the velocity. To combine the results as $\vec{\Omega}$ varies, we simply add the magnitude of the responses for each $\vec{\Omega}$. The factor $\exp -(\frac{\vec{v}\cdot\vec{\Omega}^{\mu *}}{k})^2$ is designed to prevent the aperture problem in the large (if the image motion is consistent with an infinite set of possible velocities, then the smallest velocity is perceived). The parameter k should be sufficiently large to maintain the validity of the results of Section 3. A number of velocity selective cells will be excited and the one with the largest response corresponds to the velocity estimate. A winner–take–all mechanism may then select the maximally responding cell.

The Estimation Strategy: This strategy attempts to estimate the image's spatial characteristics and compute the velocity simultaneously by minimizing a goodness–of–fit criterion. It is based on Theorem 3 which shows that the response N is the product of two functions, the first of which, r, is independent of Ω_t while the second depends only on the velocity of the image and the filter parameters. Thus several filters with the same Ω_t willput strong constraints on the possible velocity (since r will be constant for these filters).

A robust way of exploiting this idea is to minimize a goodness–of–fit criterion $E\left(\vec{v},r(\vec{\Omega})\right)$, both with respect to $\vec{v}$ and $r(\vec{\Omega})$, given a set of measurements $N(\vec{x},t:\vec{\Omega}^{\mu},\Omega_t^{\mu},\sigma^{\mu},\sigma_t^{\mu})$ for $\mu = 1,...,M$. We choose the standard least–squares fit criterion

$$E\left(\vec{v},r(\vec{\Omega}),\sigma,\sigma_t\right) = \sum_{\mu}\left(N(\vec{x},t:\vec{\Omega}^{\mu},\Omega_t^{\mu},\sigma,\sigma_t) - r(\vec{x},t:\vec{\Omega}^{\mu})e^{-\sigma_t^2(\Omega_t^{\mu}+(\vec{\Omega}^{\mu}\cdot\vec{v}))^2/2}\right)^2, \tag{4.2}$$

Suppose we have several lines of filters with constant Ω_t. Denote the values of $r(\vec{\Omega})$ on the lines as r^{ν} for $\nu = 1,...,L$. This gives $E\left(\vec{v},r^{\nu}\right) = \sum_{\mu}\left(N(\vec{x},t:\vec{\Omega}^{\mu},\Omega_t^{\mu},\sigma,\sigma_t) - r^{\nu}(\mu)e^{-\sigma_t^2(\Omega_t^{\mu}+(\vec{\Omega}^{\mu}\cdot\vec{v}))^2/2}\right)^2$. One of the ways to find the velocity $\vec{v}$ that minimizes this equation is as follows. Since the goodness–of–fit criterion, $E\left(\vec{v},r^{\nu}\right)$, is quadratic in r^{ν}, we can obtain by differentiation a system of L linear equations and L variables, whose solution gives the best r^{ν} as a function of $\vec{v}$. By substituting back for r^{ν} one obtains a cost function $\overline{E}(\vec{v})$. This function may be fed to velocity selective cells, that is, a cell selective to velocity $\vec{v}$ would receive input $\overline{E}(\vec{v})$. Among these cells, the one with the smallest response corresponds to the velocity estimate.

The Extra Information Strategy: This strategy uses the outputs of purely spatial frequency tuned cells to calculate the spatial characteristics of the image. This information can then be used to modify the Estimation Strategy by giving estimates for the form of $r(\vec{\Omega})$. We do not discuss this method in detail here.

5 Summary

The Gabor function is, strictly speaking, the only filter for which we can guarantee that the extrema of responses in the cells' optimal-frequency space lie on a ridge (unpublished calculations). This can be traced to the fact that the Gaussian is the only separable rotationally invariant function. If, however, the filters are similar, but not exactly, like Gabors, then we expect the results of Section 3 to be true most of the time. This expectation is confirmed by the velocity computation in real images with filters that were built by a self–organizing developmental model, and which resemble Gabor functions only roughly (Yuille & Cohen 1989).

In Grzywacz and Yuille (1990) we show that our model is consistent with four experimental phenomena in the primary visual cortex and the middle temporal area. There are, however, three psysiological problems with Gabor models (discussed in Grzywacz and Yuille 1990), despite their nice mathematical properties. We hope that the substance of our mathematical analysis will remain when we replace Gabors with more realistic filters and make our theory satisfy psysiological constraints (Grzywacz & Poggio 1989).

Our theory provides local estimations of velocity and hence gives a partial solution to the aperture problem. It does not, however, globally integrate these estimates to give a coherent motion flow. We therefore suggest that these estimates should be input to a motion coherence theory (such as Yuille & Grzywacz 1988) which might be implemented in later cortical areas that perform spatial integration over large receptive fields.

Acknowledgements

Some of the motivations for this work were provided by discussions with Dave Heeger. N.M.G. was supported by grants BNS-8809528 IRI-8719394 from the NSF. A.L.Y. was supported by the Brown–Harvard–MIT Center for Intelligent Control Systems with the U.S. Army Research Office grant DAAL03-86-K-0171.

References

Adelson, E.H. and Bergen, J. 1985 Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.* A **2**, 284–299.

Bulthoff, H., Little, J. and Poggio, T. 1989 A parallel algorithm for real–time computation of optical flow *Nature* **337**, 549–553.

Daugman, J.G. 1985 Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two dimensional visual cortical filters *J. Opt. Soc. Am.* A **2**, 1160–1169.

Daugman, J.G. 1988 Pattern and motion vision without Laplacian zero–crossings *J. Opt. Soc. Am.* A **5**, 1142–1148.

Fleet, D.J. and Jepson, A.D. 1989 Computation of normal velocity from local phase information *Technical Reports on Research on Biological and Computational Vision* **RBCV–TR–89–27**, Department of Computer Science, University of Toronto, Toronto, Canada.

Gabor, D. 1946 Theory of communication *J. Inst. Electr. Eng.* **93**, 429–457.

Gaddum, J.H. 1945 Lognormal distributions *Nature* **156**, 463–466.

Grzywacz, N.M. and Poggio, T. 1989 Computation of motion by real neurons. In *An Introduction to Neural and Electronic Networks* (eds. S.F. Zornetzer, J.L. Davis and C. Lau) Orlando, FL, USA: Academic Press. In Press.

Grzywacz, N.M. and Yuille, A.L. 1990 A model for the estimate of local image velocity by cells in the visual cortex. *Proceedings of the Royal Society of London.* In press.

Hassenstein, B. and Reichardt, W.E. 1956 Systemtheoretische analyse der zeit-, reihenfolgen- und vorzeichenauswertung bei der bewegungsperzeption des russelkafers *chlorophanus*, *Z. Naturforsch.* **11b**, 513–524.

Heeger, D. 1987 A model for the extraction of image flow, *J. Opt. Soc. Am.* A **4**, 1455–1471.

Hildreth, E.C. 1984 *The Measurement of Visual Motion.* Cambridge, MA: MIT Press.

Jasinschi, R.S. 1988 Space–time sampling with motion uncertainty: Constraints on space–time filtering. In *Proc. Second Int. Conf. Computer Vision* Tampa, FL, USA, pp. 428–434. Washington, DC: IEEE Computer Society Press.

Poggio, T. and Reichardt, W.E. 1976 Visual control of orientation behaviour in the fly: Part II: Towards the underlying neural interactions *Q. Rev. Biophys.* **9**, 377–438.

van Santen, J.P.H. and Sperling, G., 1984 A temporal covariance model of motion perception J. Opt. Soc. Am. A **1**, 451–473.

Watson, A.B. and Ahumada, A.J. 1985 Model of human visual-motion sensing *J. Opt. Soc. Am.* A **2**, 322–341.

Yuille, A.L. and Cohen, D.S. 1989 The Development and training of motion and velocity sensitive cells. Harvard Robotics Laboratory Technical Report 89–9.

Yuille, A.L. and Grzywacz, N.M. 1988. The motion coherence theory. In *Proc. Second Int. Conf. Computer Vision*, Tampa, Florida, USA. pp. 344–353. Washington, DC: IEEE Computer Society Press.

Yuille, A.L. and Poggio, T. 1986 Scaling theorems for zero–crossings *PAMI–8* **1**, 15–25.