From *Proceedings from the Workshop on Motion: Representation and Analysis* (pp. 151-155), Charleston, SC, May 7-9 (1986).

The extraction of Spatio-temporal Energy in Human and Machine Vision

Edward H. Adelson and James R. Bergen

RCA Labs, David Sarnoff Research Center
Princeton, NJ 08540

Abstract:

Recent work in human motion perception has conceptualized motion detection in terms of filters selective for spatio-temporal (ST) energy. One class of such models are called "energy models." They code motion energy, not velocity as such, but we describe a velocity coding model based on the ST energy within given bands of spatial and temporal frequency.

The motion energy computations appear at first to be quite different from the gradient-based computations that have been used in machine vision. But gradient systems that make appropriate use of confidence can be considered to derive velocity from opponent energy measures.

We also note that in human vision, dynamic energy seems to be coded in a set of transformed axes, which are opponent motion energy (e.g. right - left) and counterphase flicker energy. The flicker axis is often neglected, we note its possible utility in visual processing.

Introduction:

In several recent models of human motion perception, the motion analyzers are considered to be built from linear filters that are specifically tuned to the "motion energy" that is physically present in the stimulus (Fahle and Poggio, 1981; van Santen and Sperling, l984,1985; Adelson and Bergen, 1985; Watson and Ahumada, 1985; Fleet and Jepson, 1981, Ross and Burr, 1985, Wilson, 1985). Our interest here is to explore the computational properties of such systems, and to relate them to some machine vision systems.

Spatio-temporal filtering is readily implemented in the fast, parallel architecture of visual cortex, as well as the proposed architectures for many machine vision systems. Spatio-temporal energy analysis is a useful tool for understanding and developing motion systems.

Motion as tilt in (x,y,t) space:

The fundamental motivation is shown in figure 1, which shows three depictions of the same dynamic scene: a black square moving to the right over time. In fig. 1(a), it is shown at one instant, with an arrow suggesting the rightward motion. In fig. l(b), an interval of time is shown as a three-dimensional volume, this is a complete record of the object's appearance over that time. In (x,y,t) space, the moving square becomes a sheared parallelopiped. Finally, fig. l(c) shows an (x,t) slice through this volume, viewed from above. For convenience, we will show motion with (x,t) diagrams rather than the full (x,y,t) volume. Thus we will consider the spatial stimuli to be one-dimensional, with time acting as the second dimension.

It is plain that, in (x,y,t) space, motion corresponds to a spatio-temporal tilt. The faster an object moves, the greater its tilt. The problem of motion analysis is to measure the energy corresponding to these locally oriented contours, and to use this energy in analyzing the dynamics of the scene.

In space, one can analyze orientation with a set of oriented filters, such as Gabor functions (Gabor, 1946). The same approach can be used in space-time. One convolves the input with a collection of linear filters, each of which is tuned to a specific region in ST frequency space. Each filter in this bank gives a new (x,y,t) image that has been selectively filtered to extract one type of spatio-temporal energy.

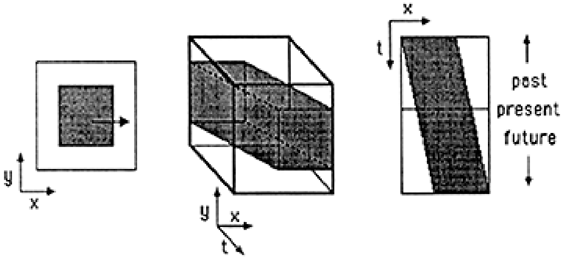

Figure 1

Some examples of such filters are suggested in figure 2.

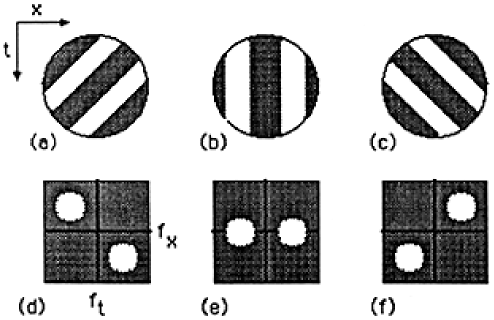

Figure 2

Figure 2(a) shows, in schematic fashion, a tilted spatio-temporal Gabor function, selective for leftward motion; fig. 2(b) shows a static one, and fig. 2(c) shows one selective for rightward motion. The functions have the form of moving sinusoids windowed by a Gaussian in space-time:

$$F(x,t) = \cos(ux + vt)\exp(-k_x x^2 - k_t t^2)$$

where u and v determine the SF and TP respectively, and k_x and k_t determine the window size.

Figures 2(d-f) show, schematically, the spatio-temporal power spectra of the Gabor functions. A Gabor function picks out a pair of Gaussian blobs in ST frequency space; the width of the blobs is inversely proportional to the width of the spatial window, and the locations of the blobs are given by the spatial and temporal frequencies of the sinusoid.

Gabor functions are not practical for actual ST energy analysis because they are non-causal. We use them here for their mathematical convenience; more realistic ST filters have been discussed elsewhere (Watson and Ahumada,

CH2322-6/86/O000/0151$01.00 © 1986 IEEE

1985; Adelson and Bergen, 1985; Fleet and Jepson, 1984)

These various filters divide the incoming spatial-temporal signal into a set of energy bands. If the filters are narrowly tuned, then it will be possible to analyze the ST frequency content quite precisely; however these filters will have poor localization in space and time. In human vision it appears that the temporal frequency tuning is substantially broader than is the spatial frequency tuning. Psychophysical evidence suggests that the TF axis is broken up into only two or three bands, while the SF axis is divided into seven or more bands (Watson and Robson, 1981; Thompson, 1984; Bergen and Wilson, 1985).

Removing phase and extracting energy.

A linear filter gives an oscillating output that incorporates information about phase, amplitude and frequency content. In order to remove the phase information, and retrieve a simple measure of the local energy within the frequency band, one can use a pair of filters whose responses are 90° out of phase, i.e. a quadrature pair (Adelson and Bergen, 1985).

This procedure is illustrated in figure 3. The input stimulus is a light bar moving the right (fig. 3(a)). The luminance profile at one instant is shown below. This pattern is convolved with two rightward selective spatio-temporal filters, one even and one odd, to produce the even and odd responses $R_0(x,t)$ and $R_1(x,t)$ shown in fig. 3(b) and 3(c). Each of the responses contains phase dependent oscillations, but they may be combined in a quadrature sum to produce a rightward energy measure,

$$\mathbf{R} = (R_0^2 + R_1^2)^{1/2}$$

which is shown in fig. 3(d). Underneath each image is a slice at a single time (shown by the dashed line), indicating the amplitude of the response at each position. The quadrature sum makes use of the fact that $\sin^2+\cos^2=1$, and produces a phase-independent measure of local energy.

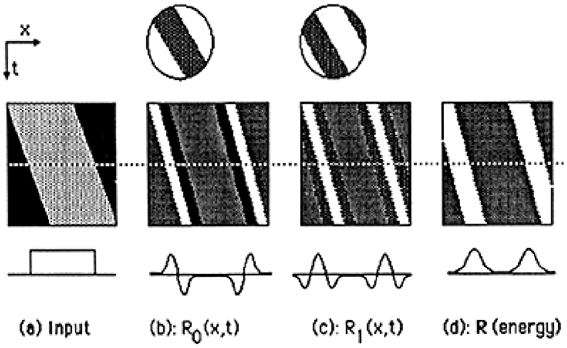

Figure 3

A visual illusion: the fluted square wave.

The energy-based analysis of motion leads to some interesting predictions about visual illusions. The "fluted square wave illusion, " illustrated in figure 4, is a recent illusion that was devised specifically as a test of such models (Adelson, 1982).

Fig.4(a) shows the luminance profile of a square wave grating that is moved to the right in discrete jumps that are 90° ($\pi/2$) of the grating period, i.e. half of a bar width. The stimulus, then, is a stroboscopic motion stimulus: the pattern appears at one position at time t_1, then jumps to the next position at t_2, and continues to jump rightward at successive frames, and there is nothing unexpected in the appearance of this stimulus: it simply looks like a square wave grating moving to the right in 90° jumps.

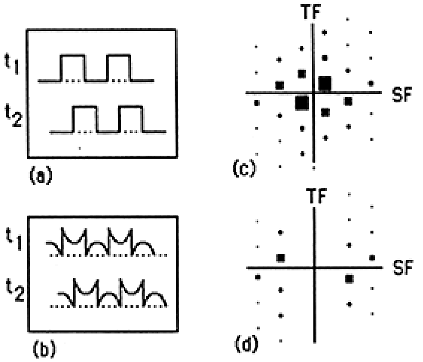

Figure 4

The spectrum of the stimulus is shown in fig. 4(c); energy is indicated by the size of the squares at each point. By far the strongest motion energy in contained in the rightward motion of the fundamental component, which shows up as the large squares near the origin. The perceived motion is that with the same speed and direction of the fundamental.

But suppose we remove the fundamental from the grating pattern, leaving the fluted square wave shown in fig. 4(b). Again let the pattern jump to the right in steps that are 90° of the grating period. The spectrum of this stimulus is shown in fig. 4(d). Now the strongest component is that of the third harmonic. And its motion energy is mainly to the left (this is because it moves in jumps that are 270° of its period, which is the same as -90°). And indeed when we look at the jumping pattern, it seems to move to the left, not to the right.

Note that a simple matching model would not predict this illusion. If one matches a feature, such as an edge, to its nearest neighbor from one frame to the next, then one will extract the "true," rightward motion, rather than the illusory leftward motion.

From energy to velocity.

At a given SF, the value of an energy measure is a function of both the velocity and the contrast of the stimulus pattern. Figure 5 shows one way of deriving a velocity estimate that is invariant with stimulus contrast. Fig. 5(a) shows that the spectrum of a moving pattern occupies a diagonal line in the spatio-temporal frequency domain. The slope of this line is inversely related to velocity. The position of the line within a given SP band may be determined by comparing the outputs of a set of ST energy detectors with different TF tunings. As shown in the figure, we can have three detectors, here labeled R, L, and S (sensitive to rightward, leftward, and static energy in the same SF band); their relative outputs can be used to extract a velocity estimate.

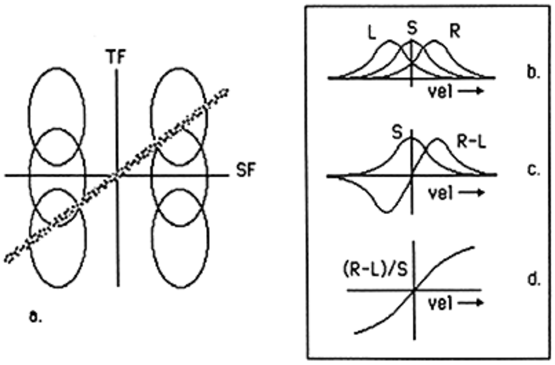

Figure 5

One such estimate is shown in figures 5(b) to 5(d). The sensitivity of each detector varies with the velocity of the input, as shown in 5(b). Let the motion selective detectors be used for an opponent energy measure, **R-L**, as in figure 5(c). Opponent energy measures are particularly interesting because they can be directly extracted by Reichardt detectors (Reichardt, 1961; Poggio and Reichardt, 1980; van Santen and Sperling, l984, 1985; Adelson and Bergen, 1985). Then a monotonic estimate of velocity is the ratio $v = (R\text{-}L)/S$, as shown in fig. 5(d). This estimate is invariant with contrast.

Relationships between ST energy and gradient methods.

Gradient methods, like energy methods, deal with low-level intensity information rather than with feature matching. But gradient methods are quite differently motivated and give direct velocity estimates without the intermediate stage of ST energy analysis. Nonetheless certain versions of gradient estimates turn out to be equivalent to energy based estimates.

As described by Pennema and Thompson (1979), velocity may be estimated as a ratio of temporal and spatial derivatives:

$$v = I_t/I_x$$

Pointwise velocity estimates are noisy and unreliable in regions of low spatial gradient. One solution is to reject velocity estimates when the gradient drops below some threshold. A more elegant approach is to take a weighted sum over a small patch of image, where the weighting (confidence) is proportional to I_x^2:

$$v = \Sigma_x(v_e w)/\Sigma_x(w) = \Sigma_x(I_t/I_w)I_x^2/\Sigma_x(I_x^2)$$

where $V_e = I_x/I_t$ is a velocity estimate at a point, and $w = I_x^2$ is a weighting that measures confidence.

Lucas and Kanade (1979), in their analysis of stereo matching, point out that this is equivalent to the least squares estimate of velocity,

$$v = \Sigma_x I_x/I_t / \Sigma_x(I_x^2)$$

This estimator, when fully developed, turns out to be equivalent to an opponent energy estimator of velocity, as we will now show.

Figure 6 shows the steps of the gradient computation we will consider:

1) Start with an input image $I_0(x,t)$ that is continuous in both space and time.

2) Convolve it with a spatio-temporal Gaussian, $G(x,t)$, to remove the high spatial and temporal frequencies:

$$I_G(x,t) = I_0(x,t)^* G(x,t)$$

3) Compute the spatial and temporal derivatives of that image,

$$I_x(x,t) = d/dx(I_G(x,t))$$
$$I_t(x,t) = d/dt(I_G(x,t))$$

4) Create a confidence image, by squaring the local spatial derivative:

$$I_C(x,t) = (I_x(x,t))^2$$

5) Create a product image, by multiplying the local spatial and temporal derivatives:

$$I_P(x,t) = I_x(x,t) \bullet I_t(x,t)$$

6) Create spatially weighted sums on the above two images, i.e. convolve them with a spatial Gaussian weighting function, $G_s(x)$, so that

$$I'_C(x,t) = I_C(x,t)^*G_S(x)$$
$$I'_P(x,t) = I_P(x,t)^*G_S(x)$$

7) Finally, compute the velocity nela as a the ratio of these weighted sums:

$$V_{est}(x,t) = I'_P(x,t)/I'_C(x,t)$$

The last panel shows the estimated velocities as arrows where the length of the arrow is indicates velocity, and the brightness indicates confidence.

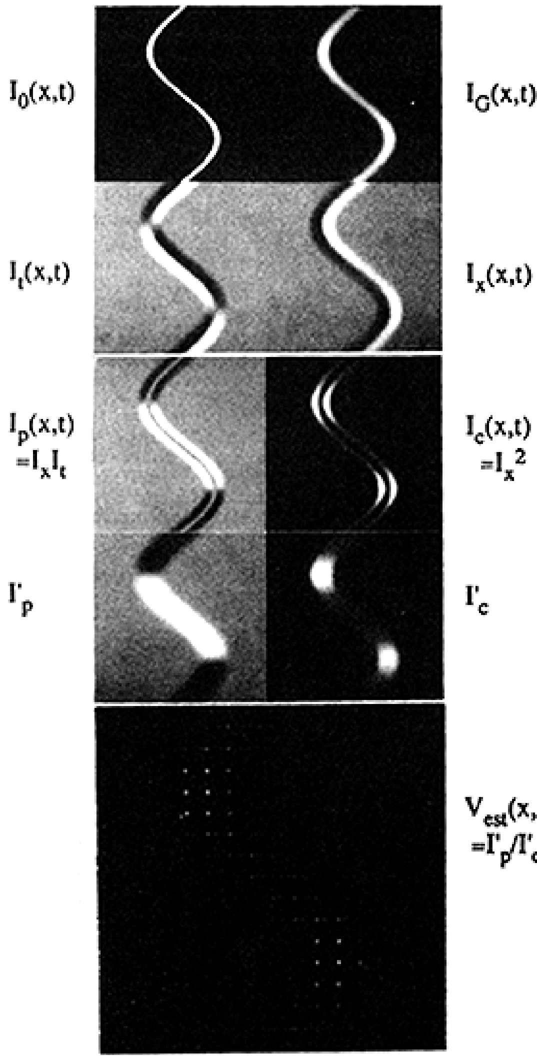

Figure 6

To discover the connection to energy models, consider the following. The images I_x and I_t were derived by differentiating a Gaussian blurred version of thc original image sequence. But the combined operations of blurring and differentiating can be captured in a single linear filters, viz:

$$K_x(x,t) = d/dx(G(x,t))$$

$$K_t(x,t) = d/dt(G(x,t))$$

Now the numerator of the gradient estimator is the local sum over the product image I_tI_x. But observe that

$$I_tI_x = (I^*K_t)(I^*K_x),$$

which may be rewritten as

$$= [(I*K_t + I*K_x)^2 - (I*K_t - I*K_x)^2]/4$$

and thus we have,

$$4I_xI_t = (I*K_t + I*K_x)^2 - (I*K_t - I*K_x)^2$$

We now have two new filters, the sum and difference filters, which we may label,

$$K_R = K_t - K_x$$
$$K_L = K_t + K_x$$

so that the equation may be written,

$$4I_xI_t = (I*K_L)^2 - (I*K_R)^2$$

Thus, the product image looks suspiciously like an opponent energy image. The suspicion is confirmed when we observe that the kernels,

$$K_R = d/dtG(x,t) - d/dxG(x,t)$$
$$K_L = d/dtG(x,t) - d/dxG(x,t)$$

are Gaussian derivatives, tilted diagonally in the (x,t) domain. The kernels, along with their spectra, are illustrated in figure 7.

Figure 7(a) shows the kernel K_t; fig. 7(b) shows K_x. These are the kernels that are used in the gradient computation. But their product can be expressed as the difference of the squares of the two ST oriented kernels, K_R and K_L, shown in figures 11(c) and 11(d). The corresponding spectra are shown in fig. 7(e-h).

Thus the numerator in the gradient procedure is a local sum over an opponent motion energy. The denominator, $\Sigma(I*K_x)^2$, is a local sum over static energy. Thus the gradient procedure, as formulated here, is equivalent to an opponent energy estimate of velocity. To be specific, we have a correspondence of the form

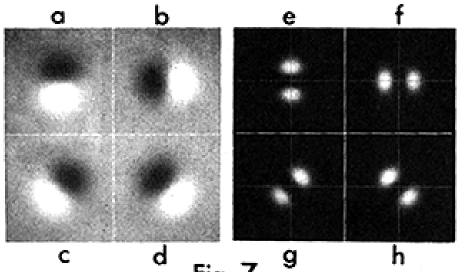

Fig. 7

$$\frac{\sum_x (I*K_t)(I*K_x)}{\sum_x (I*K_x)^2} = \frac{(R-L)}{S}$$

where **R**, **L**, and **S** are energy measures.

Furthermore, observe that the pre-filter function G(x,t) used in the above derivation, could be replaced by any ST kernel. For example the prefilter could be bandlimited in space and/or time, thus achieving a greater specificity for ST energy.

Figure 8 shows the kernels that are involved if one starts with a prefilter of form

$$G_{xt}(x,t) = d/dx\ (d/dt\ G(x,t))$$

where G(x,t) is a spatio-temporal Gaussian. The spatial and temporal derivatives then become,

$$G_{xxt} = d^2/dx^2(d/dt\ G(x,t))$$
$$G_{xtt} = d^2/dt^2(d/dt\ G(x,t))$$

These kernels are shown in figure 8(a) and (b). The sum and difference kernels are shown in 8(c) and (d). Observe that these new kernels are beautifully oriented, and closely resemble Gabor functions. The spectra of these kernels are shown in figs. 8(c-h).

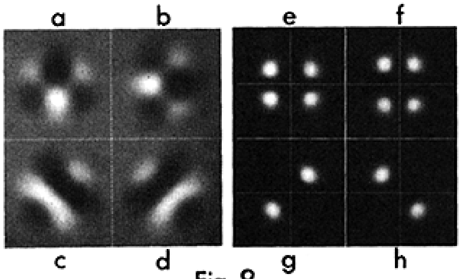

Fig. 8

Seeing two motions in one place:

It is possible to see two different motions in the same place at the same time. This is especially true if the one motion involves high spatial frequencies and the other involves low spatial frequencies (Adelson and Movshon, 1982). To a lesser extent it is true when the two motions involve different temporal frequencies.

Thus humans do not always extract a single overall velocity flow field, but can simultaneously extract two or more, corresponding to different ranges of spatial and temporal frequency.

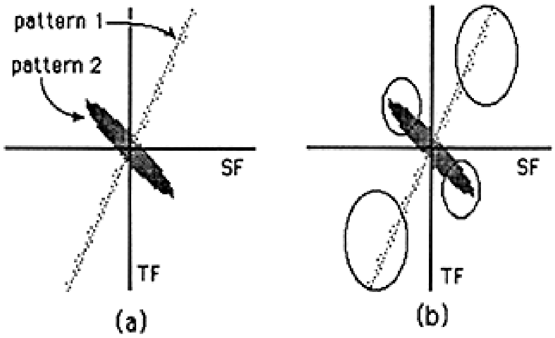

Figure 9

Machines can do the same if they extract motion in numerous finite ST energy bands (cf. Fleet and Jepson, l984). An example is shown in figure 9(a), which shows the ST spectrum of a scene that includes two different moving patterns --- for example a man walking to the left in a snowstorm that is blowing to the right. The man's ST energy is concentrated in the low frequencies, while the snowstorm spreads out into the high frequencies. Motion analyzers tuned to different domains, as in figure 9(b), can offer separate information about these different frequency bands, and thus produce two or more flow fields simultaneously.

C-flicker and opponent motion: two axes of ST energy.

To complete the picture of ST energy analysis, one must consider a second sort of ST energy: counterphase flicker (c-flicker). When leftward and rightward gratings (of the same SF and TF) are added together the leftward and rightward sensations vanish, leaving only the sensation of a flickering grating. Figure 10(a) illustrates the space of stimuli that can be generated by adding together leftward and rightward gratings with various contrasts.

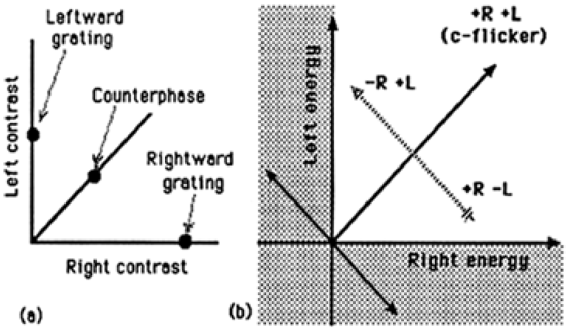

Figure 10

It appears that the visual system transforms the original axes into new perceptual axes as shown in figure 1(b).

One axis is the (R-L) opponent energy axis, the other is the (R+L) c-flicker axis. The white area shows the region of physically realizable simuli the grey area shows stimuli that would involve negative energy and are therefore unrealizable.

Although c-flicker is rarely discussed as a significant perceptual variable, humans can be quite sensitive to it. In some conditions subjects can detect c-flicker perturbations whose contrast is as small as 0.1% (Bergen and Adelson, 1985). This great sensitivity suggests that c-flicker may be perceptually useful.

C-flicker will, of course, be present in "busy" stimuli such as fluttering leaves or boiling water. But it will also be strong at dynamic occlusion boundaries, where one object is moving over another. Thus c-flicker may be a helpful cue in the segmentation of dynamic scenes.

Conclusions

In recent years, a number of models for human motion processing have been proposed in which the early stages contain filters that select out restricted bands of the spatio-temporal spectrum. In the space-time domain, the filter kernels have a spatio-temporal "orientation" that matches the ST orientation of moving stimuli. In the ST frequency domain, the power spectra of these kernels are pairs of blobs along a diagonal.

By combining the outputs of units in quadrature one can obtain a non-oscillating measure of local motion energy. A measure of static energy can be similarly obtained. The ratio of opponent and static energy, **(R-L)/S**, can be used as a measure of velocity within a given frequency band. Gradient methods for machine vision, while motivated differently, turn out to perform a similar computation.

The primary perceptual axes for dynamic scenes are opponent motion (i.e. **R-L**) and counterphase flicker (i.e. **R+L**). Humans are highly sensitive to small variations in c-flicker but the utility of this infomation is rarely discussed. Since c-flicker can be particularly strong at dynamic occlusion boundaries, we suggest that it may be useful in segmenting dynamic scenes.

References:

1) E. H. Adelson, "Some new lllusions, and some old ones, analyzed in terms of their Fourier components," *Invest. Ophth. Vis . Sci. Suppl.* **22**, 144 (1982).

2) E. H. Adelson and J. R. Bergen, "Spatio-temporal energy models for the perception of motion, " *J. Opt. Soc. Am.*, **A2**, 284-299 (1985).

3) E. H. Adelson, and J. A. Movshon, "Phenomenal coherence of moving gratings," *Nature*, **300**, 523-525 (1982).

4) S. M. Anstis, "The perception of apparent movement, " *Phil. Trans. R. Soc., Lond Ser. B*, **290**, l53-168 (1980).

5) J. R. Bergen and E. H. Adelson, "Mechanisms for the detection of motion and flicker," Annual meeting of the Optical Society of America, (1985).

6) J. R. Bergen, and H. R. Wilson, "Prediction of flicker sensitivities from temporal three pulse data" *Vis. Res*., (1985).

7) M. Fahle and T. Poggio, "Visual hyperacuity spatio-temporal interpolation in human vision," *Proc. R. Soc. Lond. Ser. B*, **213**, 451-477 (1981).

8) Fennema and W. Thompson, "Velocity determination in scenes containing several moving objects," *Comp. Graph. and Image Proc*., **9**, 301-315 (1979).

9) D. J. Fleet and A. D. Jepson, "A cascaded approach to the construetion of velocity selective meehanisms," RBCV Tech. Rept., TR-85-6, Dept. of Comp. Science, University of Toronto.

10) D. Gabor, "Theory of communication," *J. Inst. Elect. Eng*., **93**, 429-457 (1946).

11) B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proceedings of Image Understanding Workshop, 121-130 (1981).

12) J. A. Movshon, I. Thompson, and D. Tolhurst, "Receptive field organization of complex cells in the cat's striate cortex ," *J: Physiol., Lond*., **283**, 79-99 (1978).

13) T. Poggio and W. Reichardt, "On the representation of multi-input systems: Computational properties of polynomial algorithms," *Biol. Cyber*., **37**, 167-186 (1980).

14) W. Reichardt, "Autocorrelation, a principle for the evaluation of sensory information by the central nervous system," in *Sensory Communication*, W. A. Rosenblith ed. (Wiley, New York 1961)

15) J. Ross and D. Burr, "The psychophysics of motion," in *Vision, Brain, and Cooperative Communication*, M. Arbib and A. Hanson, eds. (Bradford, Amherst, Mass., 1985)

16) J. P. van Santen and G. Sperling, "Temporal covariance model of human motion -perception," *J. Opt. Soc. Am*., **A1**, 451-473 (1984).

17) J. P. van Santen and G. Sperling, "Elaborated Reichardt detectors," *J. Opt. Soc. Am*., **A2**, 300-321, (1985).

18) C. F. Stromeyer III, R. E. Kronauer, J. C. Madsen, and S. A. Klein, "Opponent mechanisms in human vision," *J. Opt. Soc. Am*., **A1**, 876-884 (1984).

19) P. Thompson, "The coding of the velocity of movement in the human visual system," *Vis. Res.*, **24**, 41-45 (1984).

20) A. B. Watson and A. Ahumada, Jr., "Model of human visual-motion sensing," *J. Opt. Soc. Am.*, **A2**, 322-342, (1985).

21) A. B. Watson and J. G. Robson, "Discrimination at threshold: labelled detectors in human vision," *Vis. Res*., **21**, 1115-1122 (1981).

22) H. Wilson, "A model for direction selectivity in threshold motion perception," *Biological Cybernetics*, (1985).