# Model for the extraction of image flow

**David J. Heeger**

*General Robotics and Active Sensory Processing Laboratory, Department of Computer and Information Science, The University of Pennsylvania, Philadelphia, Pennsylvania 19104*

Received April 28, 1986; accepted April 1, 1987

A model is presented, consonant with current views regarding the neurophysiology and psychophysics of motion perception, that combines the outputs of a set of spatiotemporal motion-energy filters to extract optical flow. The output velocity is encoded as the peak in a distribution of velocity-tuned units that behave much like cells of the middle temporal area of the primate brain. The model appears to deal with the aperture problem as well as the human visual system since it extracts the correct velocity for patterns that have large differences in contrast at different spatial orientations, and it simulates psychophysical data on the coherence of sine-grating plaid patterns.

## 1. INTRODUCTION

The world that we live in is constantly in motion: An observer (either a biological organism or a computer being) who depends on visual perception to gain an understanding of his or her environment must be able to interpret visual motion. Some of the important functions of motion perception are (1) to act as an early-warning system; (2) to allow an observer to track the location of moving objects and recover their three-dimensional structure; (3) to help an observer to determine his or her own movement (egomotion) through the environment; (4) to help an observer to divide the visual field into meaningful segments (e.g., moving versus stationary and rigid versus nonrigid).

The perception of visual motion does not depend on prior interpretation or recognition of shape and form. However, it does depend on there being motion information, i.e., changes in intensity over time throughout the visual field. Without texture, a perfectly smooth moving surface yields an image sequence in which most local regions do not change over time. But in a highly textured world (e.g., natural outdoor scenes with trees and grass), there is motion information throughout the visual field.

It is generally believed that the analysis of visual motion proceeds in two stages. The first stage is the extraction of two-dimensional motion information (direction of motion, speed, displacement) from image sequences. The second stage is the interpretation of image motion. Optical flow, a two-dimensional velocity vector for each small region of the visual field, is one representation of image motion. In this paper I address the issue of extracting a velocity vector for each region of the visual field by taking advantage of the abundance of motion information in a highly textured image sequence.

Most machine-vision efforts that try to extract image flow employ just two frames from an image sequence—either matching features from one frame to the next[1] or computing the change in intensity between successive frames along the image gradient direction.[2,3] In a highly textured world neither of these approaches seems appropriate, since there may be too many features for matching to be successful and the image gradient direction may vary randomly from point to point.[4]

There have recently been several approaches to motion measurement based on spatiotemporal filtering[5–9] that utilize a large number of frames sampled closely together in time. These papers describe families of motion-sensitive mechanisms, each of which is selective for motion in different directions. To be able to use such mechanisms in computing optical flow, one must overcome two obstacles: (1) the aperture problem and (2) the fact that the filter outputs depend not solely on the velocity of a stimulus but rather on the spatial and temporal frequencies of the stimulus.

In Section 2 I review the mathematics of motion in the spatiotemporal-frequency domain. A family of motion-sensitive Gabor filters is described in Section 3, and in Section 4 a model for extracting image velocity from the outputs of these filters is derived. Section 5 reformulates the model in terms of parallel, physiologically plausible mechanisms. In Section 6 I discuss how the model deals with the aperture problem and compare its performance with that of the human visual system. Finally, in Sections 7 and 8 the model is used to simulate psychophysical and physiological data.

## 2. MOTION IN THE FREQUENCY DOMAIN

Watson and Ahumada[5,6] and Fleet and Jepson[10] have pointed out that some properties of image motion are most evident in the Fourier domain. In this section first one-dimensional motion is described in terms of spatial and temporal frequencies; then the observation is made that the power spectrum of a moving one-dimensional signal occupies a line in the spatiotemporal-frequency domain. Analogously, the power spectrum of a translating two-dimensional texture occupies a tilted plane in the frequency domain.

### A. One-Dimensional Motion

The spatial frequency of a moving sine wave is expressed in cycles per unit of distance (e.g., cycles per pixel), and its temporal frequency is expressed in cycles per unit of time (e.g., cycles per frame). Velocity, which is distance over time or pixels per frame, equals the temporal frequency

divided by the spatial frequency:

$$\mathbf{v} = \omega_t/\omega_x. \tag{1}$$

When a signal is sampled evenly in time, frequency components greater than the Nyquist frequency (1/2 cycle per frame) become undersampled, or aliased. As a consequence, if a sine-wave pattern is shifted more than half of its period from frame to frame, it will appear to move in the opposite direction. For example, a sine wave with a spatial frequency of 1/2 cycle per pixel can have a maximum velocity of one pixel per frame, and a sine wave with a spatial frequency of 1/4 cycle per pixel can have a maximum velocity of two pixels per frame; in other words, the range of possible velocities of a moving sine wave is limited by its spatial frequency.

Now consider a one-dimensional signal, moving with a given velocity $\mathbf{v}$, that has many spatial-frequency components. Each such component $\omega_x$ has a temporal frequency of $\omega_{t_1} = \omega_x\mathbf{v}$, while each spatial-frequency component $2\omega_x$ has twice the temporal frequency, $\omega_{t_2} = 2\omega_x\mathbf{v}$. In fact, the temporal frequency of this moving signal as a function of its spatial frequency is a straight line passing through the origin, where the slope of the line is $\mathbf{v}$.

### B. Two-Dimensional Motion

Analogously, two-dimensional patterns (textures) translating in the image plane occupy a plane in the spatiotemporal-frequency domain:

$$\omega_t = u\omega_x + v\omega_y, \tag{2}$$

where $\mathbf{v} = (u, v)$ is the velocity of the pattern.[6] For example, the expected value of the power spectrum of a translating random-dot field is a constant within this plane and zero outside it.

If the motion of a small region of an image may be approximated by translation in the image plane, the velocity of the region may be computed in the Fourier domain by finding the plane in which all the power resides. To extract optical flow we could take small spatiotemporal windows out of the image sequence and fit a plane to each of their power spectra. Below I present a technique for estimating velocity by using motion-sensitive spatiotemporal Gabor-energy filters to sample these power spectra efficiently (as depicted in Fig. 3 below).

### C. The Aperture Problem in the Frequency Domain

An oriented pattern, such as a two-dimensional sine grating or an extended step edge, suffers from what has been called the aperture problem (for example, see Ref. 11). For such a pattern there is not enough information in the image sequence to disambiguate the true direction of motion. At best, we may extract only one of the two velocity components, as there is one extra degree of freedom. In the spatiotemporal-frequency domain the power spectrum of such an image sequence is restricted to a line, and the many planes that contain the line correspond to the possible velocities. Normal flow, defined as the component of motion in the direction of the image gradient, is the slope of that line.

## 3. MOTION-SENSITIVE FILTERS

Fahle and Poggio[12] and Adelson and Bergen[7] have pointed out that image motion is characterized by orientation in space–time. For example, Fig. 1(a) depicts a vertical bar moving to the right over time. Imagine that we film a movie of this stimulus and stack the consecutive frames one after the next; we end up with a three-dimensional volume (space–time cube) of luminance data like that shown in Fig. 1(b). Figure 1(c) shows an $x$–$t$ slice through this space–time cube; the slope of the edges in the $x$–$t$ slice equals the horizontal component of the bar's velocity (change in position over time). The figure also depicts a linear filter that is tuned for the motion of this moving bar. Thus motion is like orientation in space–time, and spatiotemporally oriented

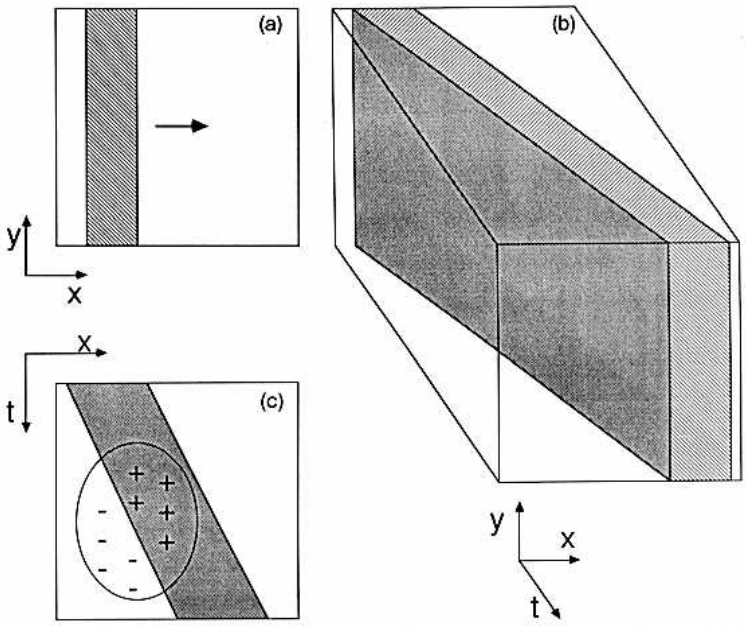



Fig. 1. Spatiotemporal orientation (redrawn from Ref. 7). (a) A vertical bar translating to the right. (b) The space–time cube for a vertical bar moving to the right. (c) An $x$–$t$ slice through the space–time cube. The orientation of the edges in the $x$–$t$ slice is the horizontal component of the velocity. Motion is like orientation in space–time, and spatiotemporally oriented filters can be used to detect it.

filters can be used to detect it. Three-dimensional (3-D) Gabor-energy filters, presented below, are such oriented spatiotemporal filters.[13]

### A. Gabor-Energy Filters

A one-dimensional sine- (or odd-) phase Gabor filter is simply a sine wave multiplied by a Gaussian window:

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-t^2}{2\sigma^2}\right) \sin(2\pi\omega t). \tag{3}$$

These filters were originally introduced by Gabor.[14] The power spectrum of a sine wave is a pair of impulses located at $\omega$ and $-\omega$ in the frequency domain. The power spectrum of a Gaussian is itself a Gaussian (i.e., it is a low-pass filter). Since multiplication in the space (or time) domain is equivalent to convolution in the frequency domain, the power spectrum of a Gabor filter is the sum of a pair of Gaussians centered at $\omega$ and $-\omega$ in the frequency domain, i.e., it is a bandpass filter. Thus a Gabor function is localized in a Gaussian window in the space (or time) domain, and it is localized in a pair of Gaussian windows in the frequency domain.

Daugman[15,16] has extended Gabor filters to a family of two-dimensional functions, an example of which is shown along with its power spectrum in Fig. 2.

An example of a 3-D (space–time) Gabor filter is

$$g(x, y, t) = \frac{1}{\sqrt{2}\pi^{3/2}\sigma_x\sigma_y\sigma_t} \exp\left[-\left(\frac{x^2}{2\sigma_x{}^2} + \frac{y^2}{2\sigma_y{}^2} + \frac{t^2}{2\sigma_t{}^2}\right)\right] \times \sin(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y + 2\pi\omega_{t_0}t), \tag{4}$$

where $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ is the center frequency (the spatial and temporal frequency for which this filter gives its greatest output) and $(\sigma_x, \sigma_y, \sigma_t)$ is the spread of the spatiotemporal Gaussian window. Three-dimensional Gabor functions look something like a stack of plates with small plates at the top and the bottom of the stack and the largest plates in the middle of the stack. The stack can be tilted in any orientation in space–time.

It is a simple matter to tune the filter to different frequencies and orientations while trading bandwidth for localization. To change the frequency tuning we independently vary $\omega_{x_0}$, $\omega_{y_0}$, and $\omega_{t_0}$. Narrowing the Gaussian window in the space–time domain broadens the bandpass window in the spatiotemporal-frequency domain and vice versa.

Gabor filters have the additional property that they can be built from separable components, thereby greatly increasing the efficiency of the computations. A new technique for computing Gabor-filter outputs from separable convolutions is presented in Appendix A. Let $k$ be the size of the convolution kernel, let $m$ be the number of images in a sequence, and let each image be $n$ pixels in size. By simplifying the complexity[17] of 3-D convolution from $O(k^3n^2m)$ to $O(kn^2m)$, separability speeds it up by 2 orders of magnitude, given a kernel size of 10 pixels.

The model presented in the following sections employs quadrature pairs of filters, odd-phase and even-phase filters of identical orientation and bandwidth. The sum of the squared output of a sine-phase filter, Eq. (4), plus the squared output of a cosine-phase filter gives a measure of Gabor energy that is invariant to the phase of the signal.

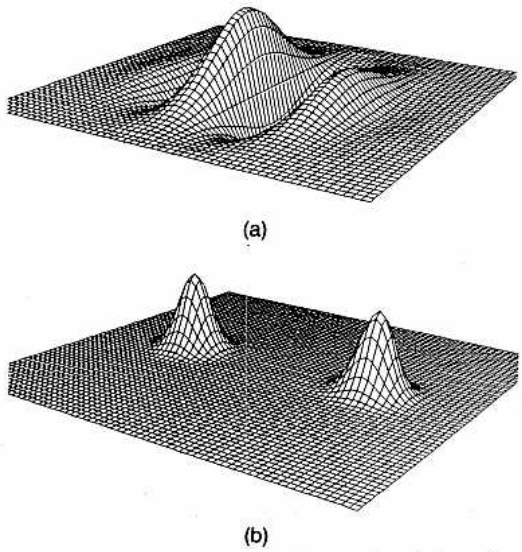



Fig. 2. Perspective views of a two-dimensional sine-phase Gabor function and its power spectrum.

The frequency response of such a Gabor-energy filter is the sum of a pair of 3-D Gaussians:

$$\begin{aligned} G(\omega_x, \omega_y, \omega_t) = {} & (1/4)\exp\{-4\pi^2[\sigma_x{}^2(\omega_x - \omega_{x_0})^2 \\ & + \sigma_y{}^2(\omega_y - \omega_{y_0})^2 + \sigma_t{}^2(\omega_t - \omega_{t_0})^2]\} \\ & + (1/4)\exp\{-4\pi^2[\sigma_x{}^2(\omega_x + \omega_{x_0})^2 \\ & + \sigma_y{}^2(\omega_y + \omega_{y_0})^2 + \sigma_t{}^2(\omega_t + \omega_{t_0})^2]\}. \end{aligned} \tag{5}$$

Equation (5) means that a motion-energy filter with center frequency $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ will give an output of $G(\omega_x, \omega_y, \omega_t)$ for a moving sine grating with spatial and temporal frequencies $(\omega_x, \omega_y, \omega_t)$. The filter will give a large output for a stimulus that has a great deal of power near the filter's center frequency and will give a smaller output for a stimulus that has little power near the filter's center frequency.

### B. A Family of Motion-Energy Filters

The model uses a family of Gabor-energy filters, all of which are tuned to the same spatial-frequency band but to different spatial orientations and temporal frequencies, i.e., $(\omega_{x_0}{}^2 + \omega_{y_0}{}^2)^{1/2}$ is constant for all the filters in one such family.

Eight of the twelve energy filters used in the present implementation have their peak response for patterns moving in a given direction—for example, one of them is most sensitive to rightward motion of vertically oriented patterns, while another is most sensitive to leftward motion. The other four filters have their peak response for stationary patterns, each with a different spatial orientation. The power spectra of the twelve filters are pairs of 3-D Gaussians (each pair of Gaussians corresponds to one filter) that are positioned on the surface of a cylinder in the spatiotemporal-frequency domain (Fig. 3): eight of them around the top of the cylinder, eight of them around the middle, and eight around the bottom.

We can build several such families of filters tuned to different spatiotemporal-frequency bands. For the current implementation I have opted to compute a Gaussian pyramid (described by Burt[18]) for each image in the sequence, and I convolve with a single family of filters at each level of the pyramid. This is essentially the same as using families

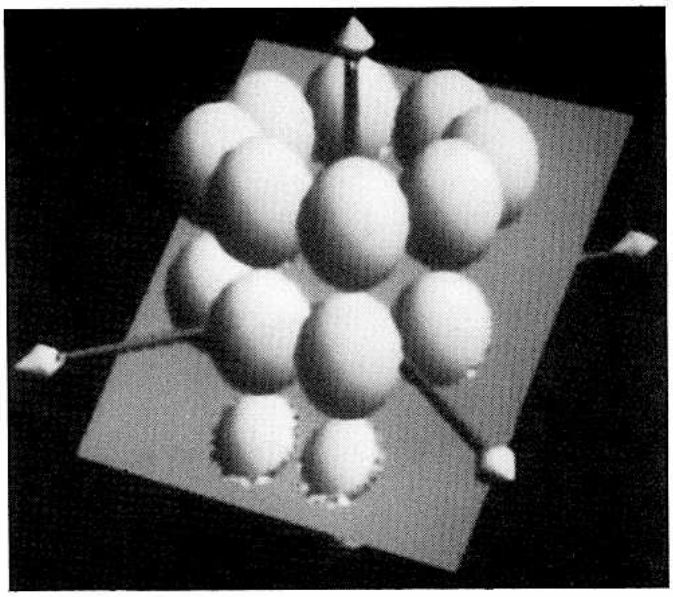


Fig. 3. The power spectra of the 12 motion-sensitive Gabor-energy filters are positioned in pairs on a cylinder in the spatiotemporal-frequency domain (temporal-frequency axis pointing up). Each symmetrically positioned pair of ellipsoids represents the power spectrum of one filter. The plane represents the power spectrum of a translating texture. A filter will give a large output only for a stimulus that has much power near the centers of its corresponding ellipsoids, and it will give a relatively small output only for a stimulus that has no power near the centers of its ellipsoids. Each velocity corresponds to a different tilt of the plane and thus to a different distribution of outputs for the collection of motion-energy mechanisms.

of filters with equal bandwidths that are spaced 1 octave apart in spatial frequency but are tuned to the same temporal frequencies.[19]

## 4. MOTION ENERGY TO EXTRACT IMAGE FLOW

Spatiotemporal bandpass filters such as Gabor-energy filters and those filters discussed in previous papers[5–10] are *not* velocity-selective mechanisms but rather are tuned to particular spatiotemporal frequencies. A single such mechanism cannot distinguish among variations in the spatial-frequency content of the stimulus, variations in its temporal-frequency content, and variations in its contrast. But an unambiguous velocity estimate may be computed from the outputs of a collection of such mechanisms.

In what follows I describe a new way of combining the outputs of a collection of motion-energy mechanisms in order to extract velocity. The role of the filters is to sample the power spectrum of the moving texture. The problem is to estimate the slope of the plane in the frequency domain that corresponds to the actual velocity. First, I derive equations for Gabor energy resulting from motion of random textures or random-dot fields. Based on these equations I formulate a least-squares estimate of velocity.

Consider an analogous two-dimensional problem: estimating the slope of a line that passes through the origin by viewing it with a finite number of circular windows. Figure 4 shows a dotted line and two circular windows. We are given a family of such windows, a finite number of them centered at known positions. The only information that we have is the number of points from the dotted line that lie within each window (in particular, we do not know the spacing between the dots). The upper window in the figure has many points within it, while the lower one has few; in other words, the line must pass close to the center of the upper window while staying far from the center of the lower one. Therefore the slope of the dotted line is nearly the same as that of the line passing directly through the center of the upper window, and it is quite different from the slope of the line passing through the center of the lower window. Notice that it is impossible to estimate the slope given only one circular window since the number of dots within a particular window depends both on the slope of the line and on the dot density.

### A. Extracting Pattern Flow

In order to extract image velocity from the outputs of motion-energy filters we replace the dotted line in Fig. 4 with a plane, and we replace the circular windows by 3-D Gaussian windows. Circular windows simply count the number of points within them. Gaussian windows count the points and weight each according to its distance from the center of the window. This is formalized by Parseval's theorem, which states that the integral of the squared values over the space–time domain is proportional to the integral of the squared Fourier components over the frequency domain:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x, y, t)|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}t$$

$$= \frac{1}{8\pi^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |F(\omega_x, \omega_y, \omega_t)|^2 \mathrm{d}\omega_x\mathrm{d}\omega_y\mathrm{d}\omega_t$$

$$= \frac{1}{8\pi^3}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(\omega_x, \omega_y, \omega_t)\mathrm{d}\omega_x\mathrm{d}\omega_y\mathrm{d}\omega_t, \quad (6)$$

where $F(\omega_x, \omega_y, \omega_t)$ is the Fourier transform of $f(x, y, t)$ and $P(\omega_x, \omega_y, \omega_t)$ is the power spectrum. Convolution with a bandpass filter results in a signal that is restricted to a limited range of frequencies. Therefore the integral of the

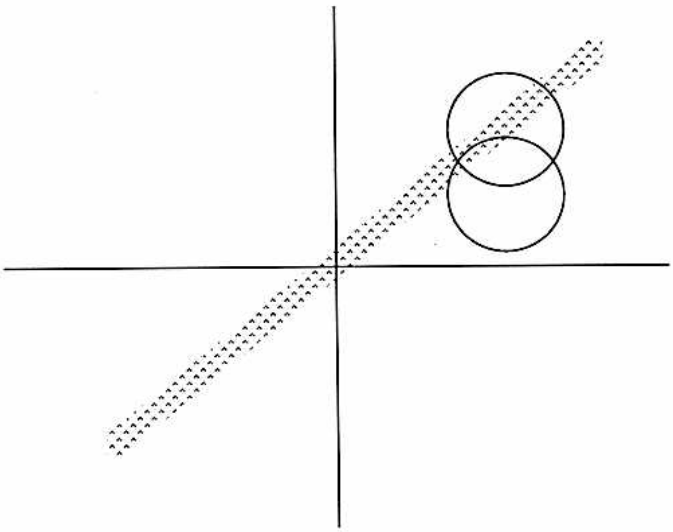


Fig. 4. A problem analogous to that of extracting velocity: estimating the slope of a line that passes through the origin by viewing it with a finite number of circular windows. The upper window has many points within it, while the lower one has very few; in other words, the line must pass close to the center of the upper window while staying far from the center of the lower one.

square of the convolved signal is proportional to the integral of the power of the original signal over this range of frequencies.

Parseval's theorem may be used to derive an equation that predicts the output of a Gabor-energy filter in response to a moving random texture. The expected value of the power spectrum of a translating random-dot field is zero, except within a plane [Eq. (2)], where it is a constant $k$. The frequency response of a Gabor-energy filter is the sum of a pair of 3-D Gaussians [Eq. (5)]. By Parseval's theorem, Gabor energy in response to a moving-random texture is twice the integral of the product of a 3-D Gaussian and a plane; by substituting Eq. (2) for $\omega_t$ in Eq. (5), multiplying by two, and integrating over the frequency domain we get

$$\mathcal{R}(u, v, k) = (k^2/2) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-4\pi^2[\sigma_x^2(\omega_x - \omega_{x_0})^2 + \sigma_y^2(\omega_y - \omega_{y_0})^2 + \sigma_t^2(u\omega_x + v\omega_y - \omega_{t_0})^2\}\mathrm{d}\omega_x\mathrm{d}\omega_y. \tag{7}$$

This integral evaluates to

$$\mathcal{R}(u, v, k) = H_4(u, v, k)\exp[-4\pi^2\sigma_x^2\sigma_y^2\sigma_t^2 H_1(u, v)],$$

$$H_1(u, v) = \frac{H_2(u, v)}{H_3(u, v)},$$

$$H_2(u, v) = (u\omega_{x_0} + v\omega_{y_0} + \omega_{t_0})^2,$$

$$H_3(u, v) = (u\sigma_x\sigma_t)^2 + (v\sigma_y\sigma_t)^2 + (\sigma_x\sigma_y)^2,$$

$$H_4(u, v, k) = \frac{k^2}{8\pi[H_3(u, v)]^{1/2}}. \tag{8}$$

For a family of Gabor-energy filters, we get a system of equations (one for each filter) in the three unknowns ($u$, $v$, $k$). The factor $H_4(u, v, k)$ that appears in each of these equations can be eliminated by dividing each equation by the sum or the average of all of them.

This results in a system of equations depending only on $u$ and $v$ that predict the output of Gabor-energy filters due to local translation. These predicted energies are exact for a pattern with a flat power spectrum. But what if the power spectrum of the pattern is not flat? In particular, what if the image contrast is different for different spatial orientations? Rather than dividing each filter output by the sum of *all* the filter outputs, we can group the filters according to their spatial orientation and normalize each spatial orientation separately.

A least-squares estimate for $u$ and $v$ minimizes the difference between the predicted and measured motion energies. Let $\mathcal{R}_i$ be the predicted motion energies given by Eqs. (8) for a family of filters: each $i$ corresponds to a filter with a different center frequency. Let $m_i$ be the observed motion energies—the outputs of that family of filters. Let $\overline{m}_i$ be the sum of the outputs of those filters that have the same preferred spatial orientation as the $i$th filter, and let $\overline{\mathcal{R}}_i$ be the corresponding sum of the predicted motion energies. A least-squares estimate of $\mathbf{v} = (u, v)$ minimizes

$$f(u, v) = \sum_{i=1}^{12}\left[\overline{m}_i\frac{\mathcal{R}_i(u, v)}{\overline{\mathcal{R}}_i(u, v)} - m_i\right]^2. \tag{9}$$

There are standard numerical methods for estimating $\mathbf{v} = (u, v)$ to minimize Eq. (9), e.g., the Gauss–Newton gradient-descent method.[20]

Alternatively, the least-squares estimate of $\mathbf{v} = (u, v)$ maximizes

$$F(u, v) = \sum_{i=1}^{12}(m_i)^2 - f(u, v) = \sum_{i=1}^{12}(m_i)^2 - \left[\overline{m}_i\frac{\mathcal{R}_i(u, v)}{\overline{\mathcal{R}}_i(u, v)} - m_i\right]^2, \tag{10}$$

where $f(u, v)$ is given by Eq. (9).[21] In Section 5 I describe a parallel technique for locating this maximum.

## B. The Algorithm

The main steps in the computations performed by the model are (1) to convolve the image sequence with 3-D Gabor filters, (2) to compute motion energy as the squared sum of the sine- and cosine-phase Gabor filter outputs, and (3) to estimate velocity by either minimizing Eq. (9) or maximizing Eq. (10). In this section I explain the additional steps that need to be computed, and I summarize the entire algorithm.

First, Parseval's theorem, Eq. (6), relates an integral over the space–time domain to an integral over the frequency domain; since the filters are localized in both domains, convolving with a 3-D Gaussian is one way to approximate this integral. We can think of the model as computing the average image velocity within this Gaussian window.

Of course, Gaussian convolution will tend to smooth over motion boundaries and other regions where the velocity changes rapidly from point to point. Some possible solutions to this problem are (1) to use images of higher resolution and (2) to use a different method for combining information other than Gaussian convolution, e.g., relaxation labeling methods (for references, see Hummel and Zucker[22]) and finite-element regularization methods (see Ref. 23).

There are two situations for which this smoothing problem is particularly bad. First, in regions moving with high speed, we must use filters that are higher in the pyramid, i.e., of lower spatial resolution. Second, where there is a region of low image contrast adjacent to one of high contrast, the filter outputs for the high-contrast region (since they are greater on average) will bias the velocity estimates for the low-contrast region. The former situation may be controlled by incorporating eye/camera movements: an initial low-resolution estimate may be used to drive tracking eye movements, thereby decreasing the image velocity and allowing for estimates of higher spatial resolution.

Finally, a problem with Gabor filters themselves is that all but the sine-phase filters have some dc response. If an image is very bright (large mean luminance) and of low contrast, the output of the filter may be dominated by response to the dc rather than to the image-contrast signal. Clearly, this is undesirable. This difficulty can be alleviated by first subtracting the local mean luminance, e.g., by convolving with a center-surround filter that has a sharp positive center and a broad negative surround.[24]

In summary, an algorithm for extracting image flow proceeds as follows:

1. Compute a Gaussian pyramid for each image in the image sequence.
2. Convolve each of the resulting image sequences with a 3-D center-surround filter to remove the dc and lowest spatial frequencies.
3. Convolve each sequence with the separable filters described in Appendix A and compute the sine- and cosine-phase Gabor-filter outputs as linear combinations of these separable convolutions.
4. Compute motion energy as the squared sum of the sine- and cosine-phase Gabor filter outputs.
5. Convolve the resulting motion energies with a Gaussian to approximate the integral in Parseval's theorem.
6. Find the best choice of $u$ and $v$ given by Eq. (9) or (10), e.g., by employing the Gauss–Newton gradient-descent method or the parallel technique presented in Section 5.

### C. Some Results

All the results presented in this paper were produced with a single choice for each of the parameters of the model: The spatial-frequency tuning of each Gabor filter is $(\omega_{x_0}^2 + \omega_{y_0}^2)^{1/2} = 1/4$ cycle per pixel; the temporal-frequency tunings are either $\omega_{t_0} = 0$ cycle per frame (stationary filters) or $\omega_{t_0} = \pm 1/4$ cycle per frame (right–left, up–down, etc.); the standard deviation of all the spatial Gaussians is $\sigma_x = \sigma_y = 4$ (the spatial kernel size of the filters is 23 pixels), and that of the temporal Gaussians is $\sigma_t = 1$ (the temporal kernel size is 7 frames). Except for the Yosemite fly-through sequence discussed below, all the results are computed using only the lowest level of the pyramid.

Each vector in the flow fields depicted below represents a motion in a direction given by the vector's angle at a speed given by the vector's length. Errors in the velocity estimates are expressed in terms of the percentage error in each component of the actual velocity vectors.

*Translating Image Sequences*

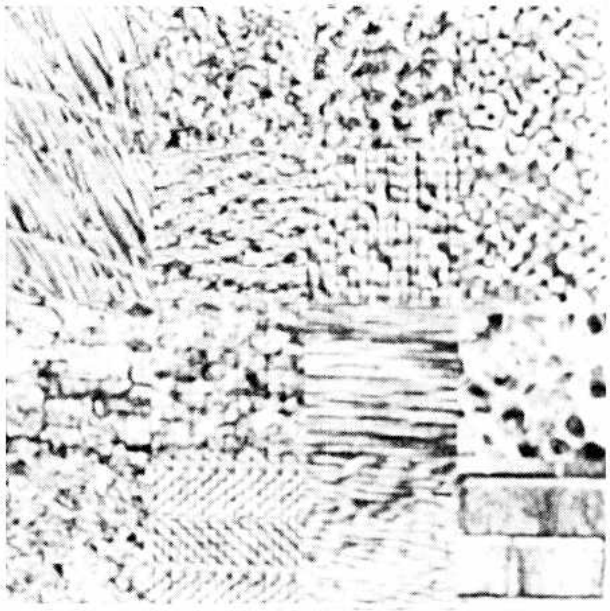
Translating image sequences were generated from a textured image by (1) enlarging the image to four times its original size, (2) shifting the image in each frame by an integral number $i$ of pixels horizontally and an integral number $j$ of pixels vertically, and (3) reducing each image in the resulting sequence to the original resolution. The final result is an image sequence with velocity $(i/4, j/4)$ pixels per frame.

The model gives accurate velocity estimates (within 10% of the actual velocities) for translating image sequences of a wide variety of textured patterns, including random-dot patterns (with dot densities ranging from 5 to 50%), images of fractal textures,[25] some sine-grating plaid patterns (discussed in Section 6), and natural textures (discussed below).

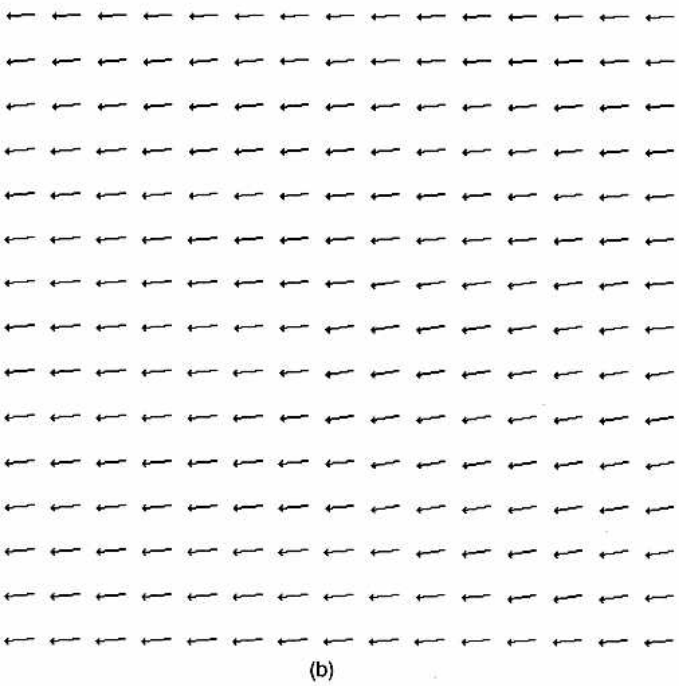
*Noise Sensitivity*

Translating random-dot image sequences were used to study the error in the velocity estimates. For image sequences with speeds ranging from 0.25 to 1.75 pixels per frame, percentage error is roughly normally distributed, with a mean of −2.9% and a standard deviation of 3.6.

Noise sensitivity was studied by adding spatiotemporal white (Gaussian) noise to translating random-dot sequences. Define the signal-to-noise ratio (S/N) to be the brightness of the image dots divided by the standard deviation of the noise. If S/N = 10, then the mean percentage error in the estimates is −4.3% and the standard deviation is 4.1. This means that when the standard deviation of the sensor noise is as much as 10% of the sensor's dynamic range, most velocity estimates are still within 10% of the actual values.



(a)



(b)

Fig. 5. (a) Fourteen natural textures (the two texture squares at the upper left are the same, and so are the two at the upper right). Each texture square was used to generate motion sequences translating 1/2 pixel per frame in each of eight directions. The velocities extracted by the model are accurate to within 10%. (b) Example flow field extracted from a motion sequence generated from the straw texture in the upper-left-hand corner of (a). The actual motion was (−0.5, 0.0). The mean of the extracted velocities is (−0.473, −0.04), and the standard deviation for both the horizontal and vertical components is 0.01.

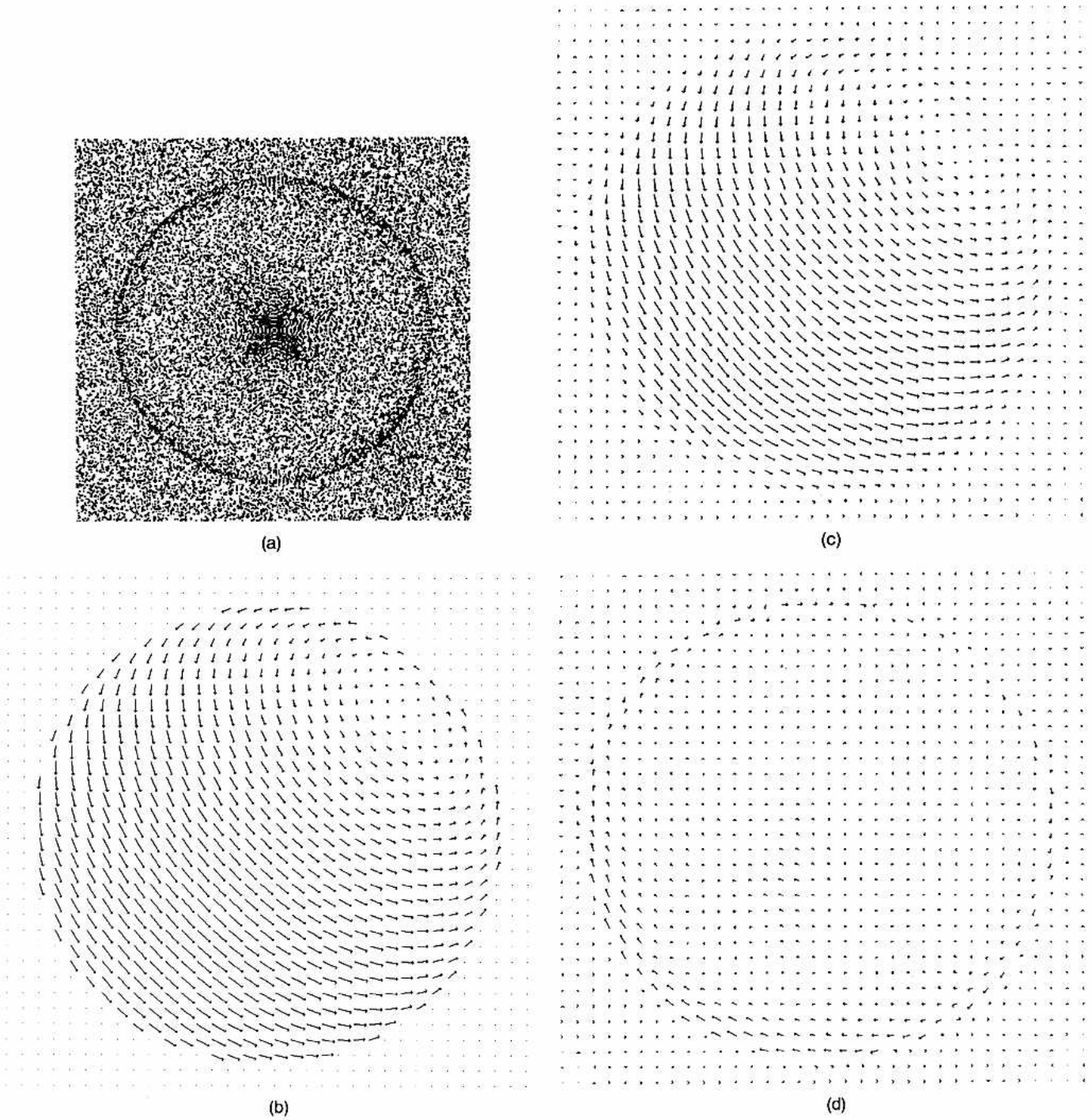



Fig. 6. A rotating random-dot sphere. (a) A frame from the motion sequence. (b) The actual flow field. (c) Flow field extracted by the model. (d) Difference between (b) and (c).

*Images of Natural Textures*

Image sequences were generated from each of the 14 natural textures shown in Fig. 5(a). A sample flow field, shown in Fig. 5(b), was extracted from an image sequence of the straw texture in the upper-left-hand corner of Fig. 5(a). The model correctly estimates the velocity (to within 10%) for every one of these textures. This is particularly impressive for the straw texture in the upper-left-hand corner, the brick texture in the lower-right-hand corner, and the texture second from the lower-right-hand corner of 5(a) because they have such strong spatial orientations. The model is capable of recovering accurate velocity estimates for these textures since it normalizes each spatial orientation separately. Conversely, if we normalize the filter outputs isotropically, i.e., by dividing each motion energy by the sum of all of them, then the estimates for these three textures are erroneous.

*A Rotating Sphere*

Figure 6(a) shows one frame of a random-dot image sequence of a sphere rotating about an axis through its center in front of a stationary background. Figure 6(b) shows the actual flow field for this image sequence, 6(c) shows the flow

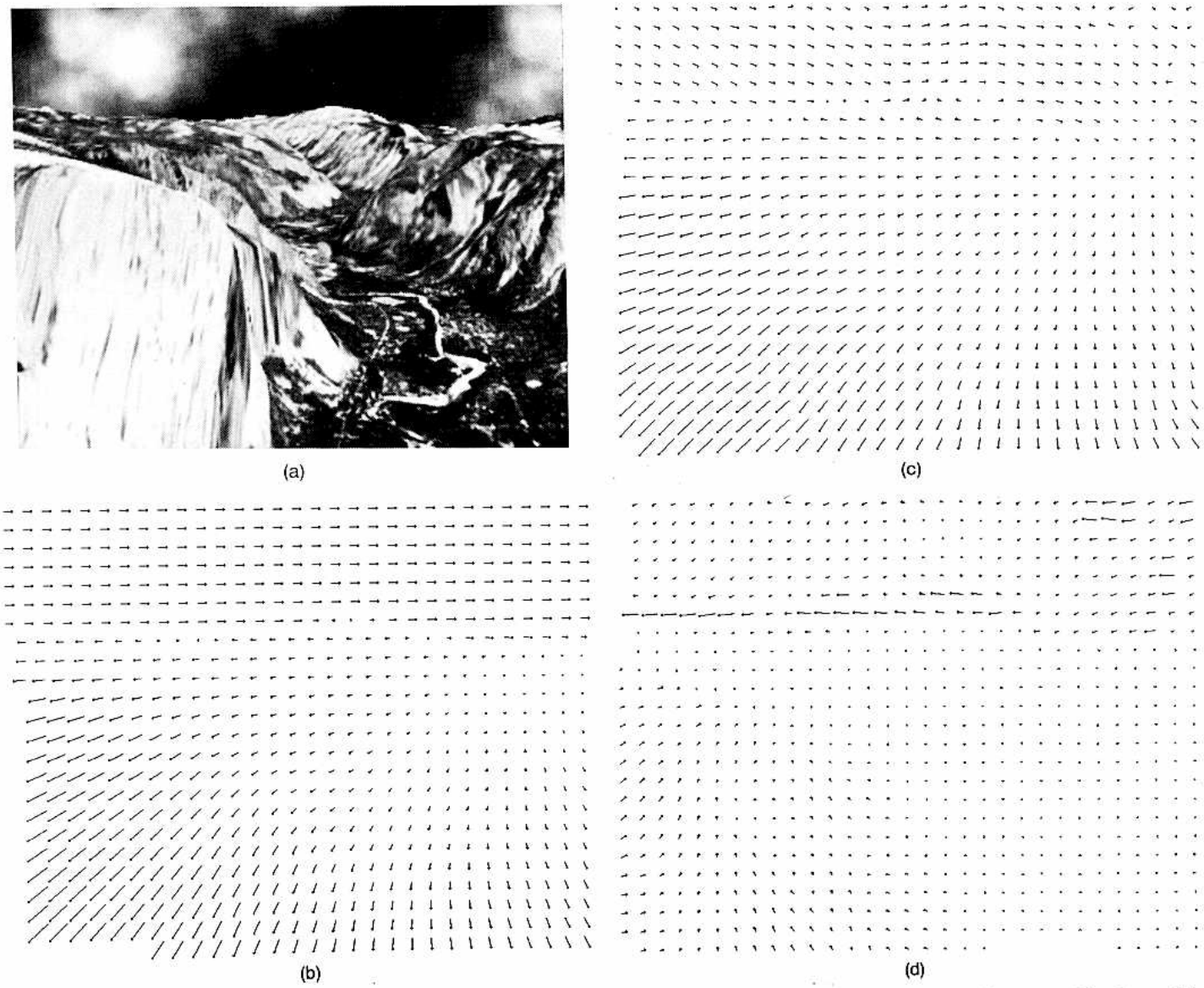



Fig. 7. (a) One frame of an image sequence flying through Yosemite valley. (b) The actual flow field. (c) Flow field extracted by the model. (d) Difference between (b) and (c).

field extracted by the model, and 6(d) shows the difference between them. The effect of the Gaussian smoothing is clearly evident, as there are errors along the motion boundary.

*A Realistic Example*

Figure 7(a) shows one frame of a computer-generated image sequence of a flight through Yosemite valley. Each frame was generated by mapping an aerial photograph onto a digital-terrain map (altitude map). The observer is moving toward the horizon. The clouds in the background were generated with fractals (see Mandlebrot[26] and recent SIGGRAPH[27] conference proceedings for definitions and references) and move to the right while changing their shape over time.

Figure 7(b) shows the actual flow field for this image sequence, Fig. 7(c) shows the flow field extracted by the model,[28] and Fig. 7(d) shows the difference between them. The effect of Gaussian smoothing is evident along the boundary at the horizon. Small errors are also evident on the face of El Capitan (in the lower left) since this image region moves with high speed (see the discussion in Subsection 4.B) and in the cloud region since the clouds change shape over time while moving rightward.

## 5. A PARALLEL IMPLEMENTATION

Electrophysiological studies of the middle temporal (MT) area[29] in macaque and owl monkeys reveal cells that are velocity tuned. Here we reformulate the last step of the model in terms of parallel, physiologically plausible, velocity-tuned mechanisms, and in Section 8 the model is compared with physiology. First, I explain how to build velocity-tuned units (analogous to the velocity-tuned cells of area MT) by combining the outputs of motion-energy filters (recall that the motion-energy filters are not themselves velocity tuned since they confound spatial-frequency, temporal-frequency, and image contrast).

Step 6 in the algorithm in Subsection 4.B is to find the maximum of a two-parameter function, Eq. (10). One way to locate this maximum is to evaluate the function in parallel at a number of points (say, on a fixed square grid[30]) and to

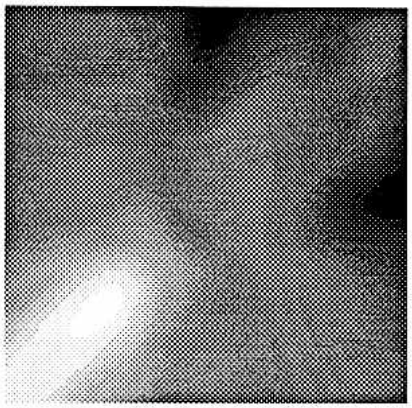


Fig. 8. Distribution of outputs of velocity-tuned units for a moving random-dot field moving leftward and downward 1 pixel per frame. Each point in the image corresponds to a different velocity; for example, $\mathbf{v} = (0, 0)$ is at the center of the image, $\mathbf{v} = (2, 2)$ at the top right-hand corner. The maximum in the distribution of outputs corresponds to the velocity extracted by the model. Units in the brighter regions have positive outputs, and units in the darker regions have negative (inhibited) outputs.

pick the largest result. In the context of the model each point on the grid corresponds to a velocity. Thus, evaluating the function for a particular point on the grid gives an output that is velocity tuned.

For each velocity $\mathbf{v} = (u, v)$, $F(u, v)$ in Eq. (10) is a measure of how closely $\mathbf{v}$ approximates the true velocity—in other words, for a fixed $u$ and $v$, $F(u, v)$ is tuned to a particular velocity. Local image velocity may be encoded as the maximum in the distribution of the outputs of a number of such velocity-tuned units, each tuned to a different $\mathbf{v}$. The units tuned to velocities close to the true velocity will have relatively large outputs (small difference between the predicted and measured motion energies), while those tuned to velocities that deviate substantially from the true velocity will have small outputs.

For a fixed velocity, the predicted motion energies $\mathcal{R}_i(u, v)$ defined by Eqs. (8) are fixed constants; denote them by $w_{ij}$, where each $i$ corresponds to a different motion-energy filter and each $j$ corresponds to a different velocity. We may rewrite Eq. (10) for a fixed $\mathbf{v}$ as

$$F_j = \sum_{i=1}^{12} (m_i)^2 - \left( \overline{m}_i \frac{w_{ij}}{\overline{w}_{ij}} - m_i \right)^2, \tag{11}$$

where $F_j$ is the response of a single velocity-tuned unit and $w_{ij}$ and $\overline{w}_{ij}$ are constant weights corresponding to the $i$th motion energy for the $j$th velocity. A mechanism that computes a velocity-tuned output from the motion-energy measurements performs the following simple operations:

1. A linear stage, a weighted summation given by $[\overline{m}_i(w_{ij}/\overline{w}_{ij}) - m_i]$.
2. A nonlinear stage, squaring.[31]
3. A second linear stage, the summation over $i$.

An example of the outputs of a set of velocity-tuned units is shown in Fig. 8, which displays a map of velocity space, with each point corresponding to a different velocity. The brightness at each point is the output of a unit tuned to that velocity; therefore the maximum in the distribution of outputs corresponds to the velocity extracted by the model.

## 6. DEALING WITH THE APERTURE PROBLEM

In this section I use a class of moving stimuli known as sine-grating plaids in order to test the model's capability of solving the aperture problem, and I compare the model's performance with that of the human visual system. I also develop a curvature measure that enables the model to recognize when there is an ambiguous velocity estimate resulting from the motion of a strongly oriented pattern (such as a single grating); in such cases, the model may choose the normal-flow velocity.

### A. Sine-Grating Plaids

A sine-grating plaid is the sum of two moving gratings and may be seen as a single coherent plaid motion. The gratings are combined not as the vector sum or the vector average of the two component normal-flow velocities but rather as the

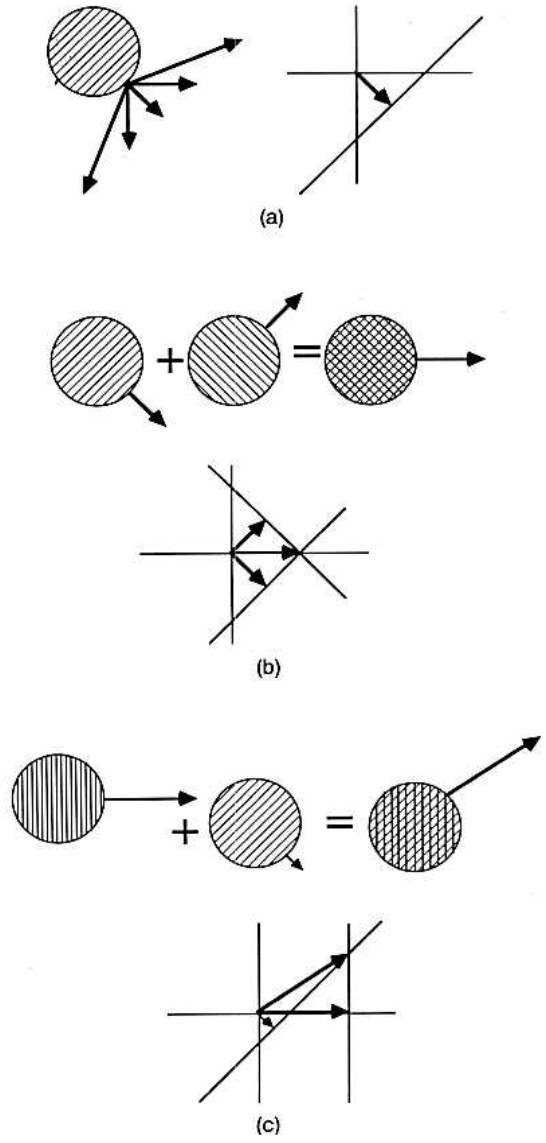



Fig. 9. The perceived motion of two moving gratings is the intersection of the perpendiculars to the two velocity vectors. (a) A single moving grating: the diagonal line indicates the locus of velocities compatible with the motion of the grating. (b), (c) Plaids composed of two moving gratings. The lines give the possible motions of each grating alone. Their intersection is the only shared motion and corresponds to what is seen. (Redrawn from Ref. 32.)

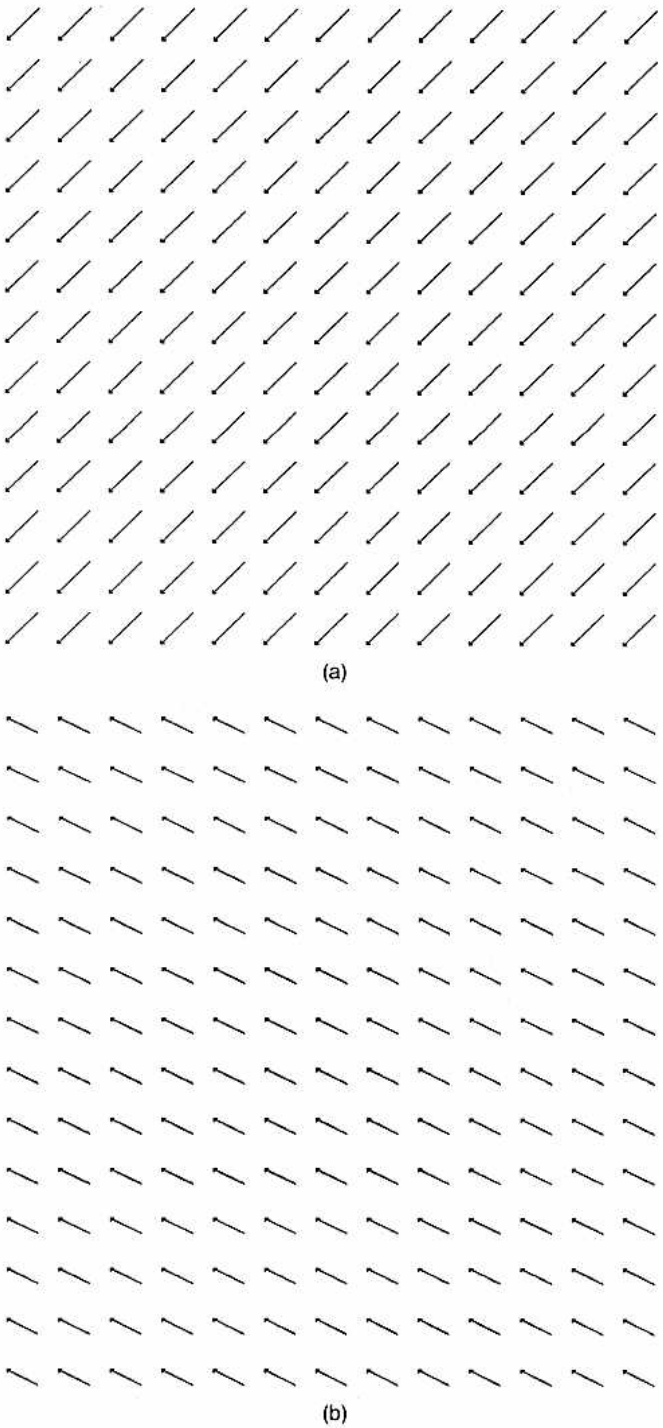



Fig. 10. (a) Flow field extracted by the model for a plaid pattern made up of a sine grating moving leftward 1 pixel per frame plus a sine grating moving downward 1 pixel per frame. The combined motion extracted by the model is 1 pixel leftward and 1 pixel downward in each frame. (b) Flow field for a plaid pattern made up of a sine grating moving leftward 1 pixel per frame plus a sine grating moving downward and leftward 1/4 pixel each frame. The counterintuitive combined motion is leftward 1 pixel per frame and *upward* 1/2 pixel per frame as shown in the flow field extracted by the model. The spatial frequency of the gratings for both (a) and (b) was 0.25 cycle pixel$^{-1}$.

intersection of the perpendiculars to the two velocity vectors. Figure 9(a) depicts a single grating moving behind an aperture; the arrows represent flow vectors, and the diagonal line represents the locus of velocities compatible with the grating's motion. There is an infinite number of such compatible motions any of which will result in exactly the same stimulus. Figure 9(b) shows a plaid composed of two orthogonal gratings moving at the same speed; the intersection of the perpendiculars to the two normal-flow velocities (the intersection of the two constraint lines) is the only shared motion and corresponds to what is seen. Figure 9(c) shows a plaid composed of two oblique gratings, one moving slowly and the other more rapidly; one grating moves rightward and the other moves downward and rightward, but the pattern moves *upward* and rightward.

The model recovers the correct pattern-flow velocity for a number of such plaids.[32] Examples of flow fields extracted by the model for plaids made up of gratings with equal contrasts and spatial frequencies are shown in Fig. 10. The combined motion extracted by the model in both Figs. 10(a) and 10(b) is accurate to within 5%.

Adelson and Movshon[33] studied the phenomenon of coherence by varying the angle between the two gratings, their relative contrasts, and their relative spatial frequencies. They found that for a range of relative angles, contrasts, and spatial frequencies the two gratings are seen as a single coherent plaid motion and that beyond this range the two gratings look like separate motions, one moving past the other. The phenomenon of coherence tests the human visual system's ability to solve the aperture problem: Given the ambiguous motion of a single moving grating, how much additional information is needed from the second grating to give an unambiguous coherent percept?

The model is capable of extracting the correct pattern-flow velocity for plaids that have large differences in contrast; e.g., for plaids made up of orthogonal gratings, velocity estimates are accurate to within 10% for contrast ratios of greater than 32:1. This is comparable with human performance.[34] As the contrast difference between the two component gratings gets larger than this, the model begins to tilt the extracted velocity vector toward the higher-contrast grating. Although the perceived velocity of plaids has not yet been measured precisely,[35] Adelson[34] notes that observers also see the direction of motion tilt toward the higher-contrast grating when the relative contrast difference is large.

If the model is to withstand large contrast ratios, it is crucial that the spatial bandwidths of its filters be less than their temporal bandwidths; in the frequency domain, this means that the filters are oblong hotdog-shaped (longer in $t$ than in $x$ and $y$) instead of spherical in shape. As an illustrative example, consider a plaid made up of rightward- and upward-moving gratings. The idea of normalizing the filter outputs separately for each spatial orientation is that the up–down filters should give the same responses relative to one another regardless of the contrast of the rightward grating. If the filters were spherical in shape, then the response of the downward filter would be dominated by the rightward grating (the impulse from a rightward grating is closer than that from an upward grating to the center frequency of the downward filter). This would be bad because we want the relative responses of the up and down filters to be unaffected if the contrast of the rightward grating is varied. But, since the filters are oblong in shape, the response of the downward filter is dominated by the grating moving upward for a wide range of relative contrasts.

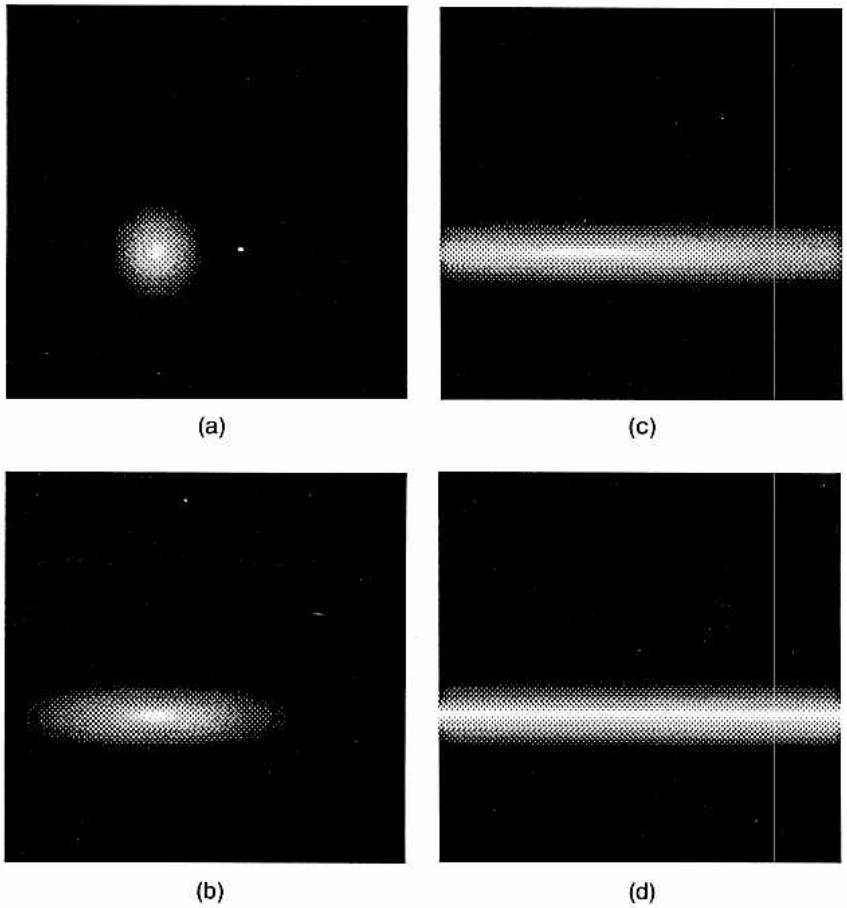



Fig. 11. Distribution of outputs of velocity-tuned units for sine-grating plaids made up of orthogonal gratings. The gratings moved 1 pixel frame$^{-1}$ leftward and downward, and their spatial frequency was 0.25 cycle pixel$^{-1}$. (a) The two component gratings had the same contrast. The maximum in the distribution of outputs corresponds to the velocity extracted by the model. (b) One grating had twice the contrast of the other grating. (c) One grating had four times the contrast of the other grating. (d) One grating had zero contrast; the aperture problem is evident, as there is a ridge of maxima. Each velocity-tuned unit along this ridge has the same output (to within 1 part in 100,000).

### B. Recognizing Ambiguity

We can think of the outputs of the velocity-tuned units as forming a surface in velocity space: the height of the surface at each velocity is given by the output of a unit tuned to that velocity. As the contrast of one of the gratings is decreased relative to the contrast of the other, the peak in this surface gets broader in one direction. This becomes evident if one compares Figs. 11(a)–11(d). In Fig. 11(a), the two component gratings are of equal contrast, so the peak is symmetrical. When the contrast ratio is increased, as in Figs. 11(b) and 11(c), the peak elongates in one direction. Eventually, as shown in Fig. 11(d), the peak turns into a ridge.

When there is an unambiguous peak we can extract the correct pattern-flow velocity, but how do we know if there is a ridge or a peak? Intuitively, it is a peak if it falls off sharply in all directions and it is a ridge if it stays constant in one direction. We know from differential geometry (for example, see Ref. 36) that a surface can be characterized locally by its maximum and minimum curvatures. If the minimum curvature of a surface is small or zero at a point while the maximum curvature is large, then the surface looks like a ridge. If both curvatures are large, then it looks like a peak.

The minimum curvature of the surface at the peak divided by the height of the peak is a measure of whether a moving pattern gives an unambiguous velocity estimate. The minimum curvature can be computed at any point on the surface of velocity-tuned outputs from the first and second derivatives of Eq. (10). Figure 12(b) shows a plot of the curvature measure as the relative contrast of a plaid's component gratings is varied; the curvature measure decreases monotonically with contrast for a wide range of test contrasts. We may pick a value to act as a threshold: if the curvature measure is above this value we pick the pattern flow given by the location of the peak, and if it falls below this value we may pick the normal-flow vector[37] or we may choose any other velocity along the ridge (a familiar example of when people see motion other than in the normal-flow direction is the barberpole illusion).

## 7. SIMULATING PSYCHOPHYSICS

In this and the next section, I use the model to simulate psychophysical and physiological data. For the most part, this simulation merely demonstrates that the model is consistent with some of the experimental results on biological motion perception. The emphasis in future research will be to compare the predictions made by this model with those made by other image-flow models and to test those predictions with further experiments.

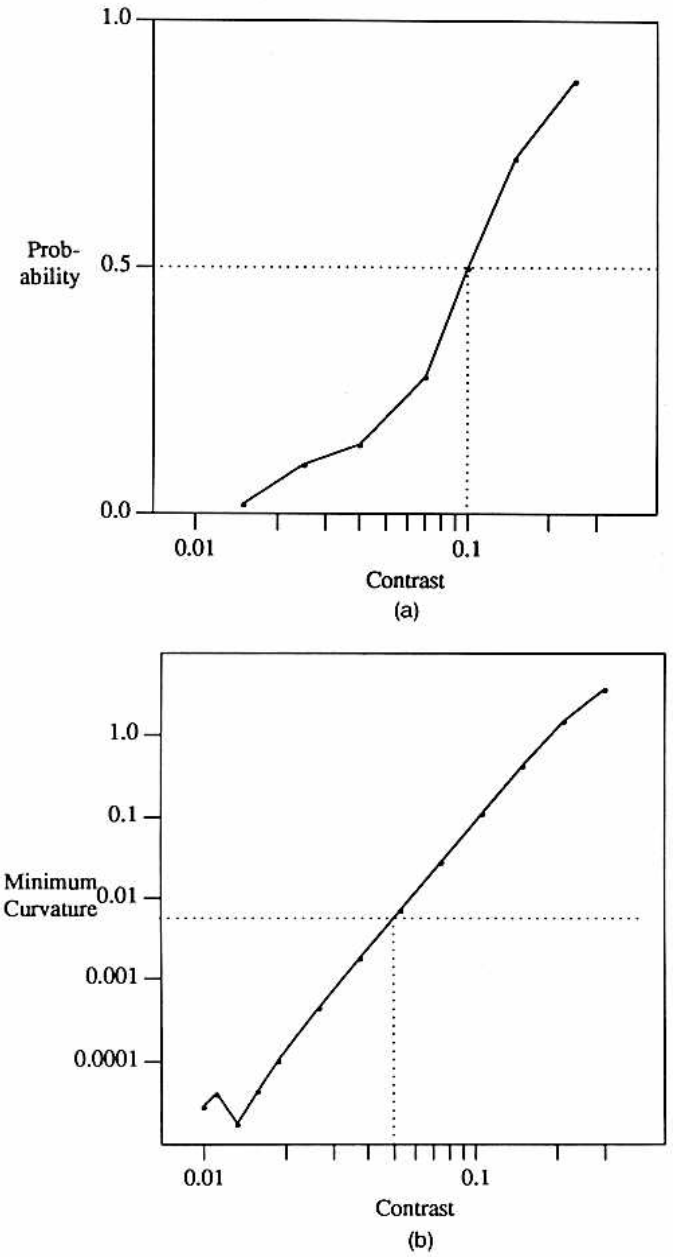



Fig. 12. The influence of contrast on the coherence of sine-grating plaids. (a) One grating had a fixed contrast of 0.3, while the other was of variable contrast. The two gratings moved at an angle of 135°, both had a spatial frequency of 1.6 cycles deg$^{-1}$, and both moved at 3 deg sec$^{-1}$. The plot shows the probability that the observer judged the two gratings to be coherent. The dotted lines indicate the test-grating contrast needed to attain threshold (50% probability) coherence. Subject, EHA. (Replotted from Ref. 32.) (b) One grating had a fixed contrast of 0.3, while the other was of variable contrast. The two gratings moved at an angle of 120°, both had a spatial frequency of 0.25 cycle pixel$^{-1}$, and their speeds were chosen so that the coherent plaid moved at a speed of 2/3 pixel frame$^{-1}$. The plot shows the curvature measure as the contrast of the test grating was varied. The dotted lines indicate the test-grating contrast needed to attain threshold (0.006 curvature) coherence.

In this section the curvature measure presented above is used to simulate the psychophysical data on the coherence of sine-grating plaids. Figure 12(a) plots the psychometric function for coherence (probability of coherence) as the contrast of one of the component gratings is reduced. Figure 12(b) shows a plot of the curvature measure as the relative contrast of the two component gratings is varied. In each case we may pick a threshold value (e.g., 50% probability, 0.006 curvature). Then we may vary the angle between the two component gratings or we may vary their relative spatial

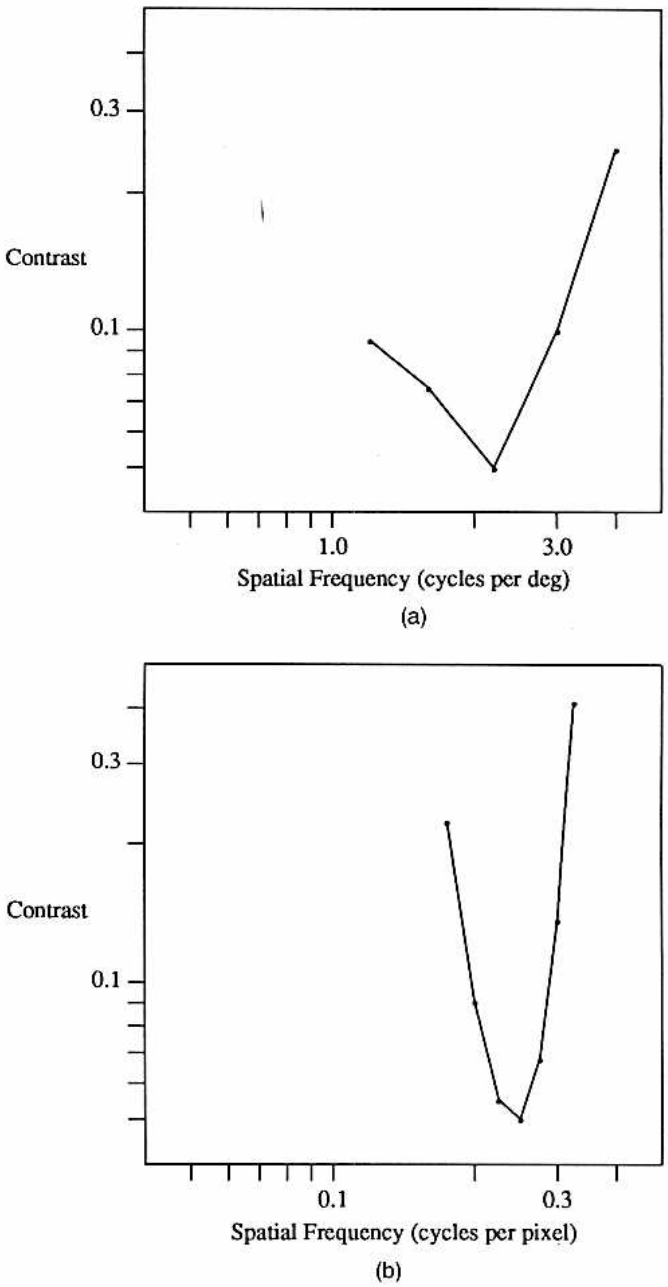



Fig. 13. The influence of spatial frequency on the coherence of sine-grating plaids. (a) One grating had a fixed contrast of 0.3, while the other was of variable contrast. The two gratings moved at an angle of 135 deg, and both moved at 3 deg sec$^{-1}$. The test grating was of variable contrast and variable spatial frequency. The plot shows the threshold contrast for coherence for a range of test spatial frequencies when the first grating was fixed at 2.2 cycles deg$^{-1}$. Subject, PA. (Replotted from Ref. 32.) (b) One grating had a fixed contrast of 0.3 and a fixed spatial frequency of 0.25 cycle pixel$^{-1}$, while the other was of variable contrast and spatial frequency. The two gratings moved at an angle of 120 deg, and their speeds were chosen so that the coherent plaid moved at a speed of 2/3 pixel per frame. A fixed value was chosen as the threshold value for the curvature measure. This value was chosen in order to match the psychophysical data in (a) for the case when the fixed grating and the test grating were of equal spatial frequency. For each test grating, the plot shows the contrast needed at that spatial frequency for the curvature measure to attain that value.

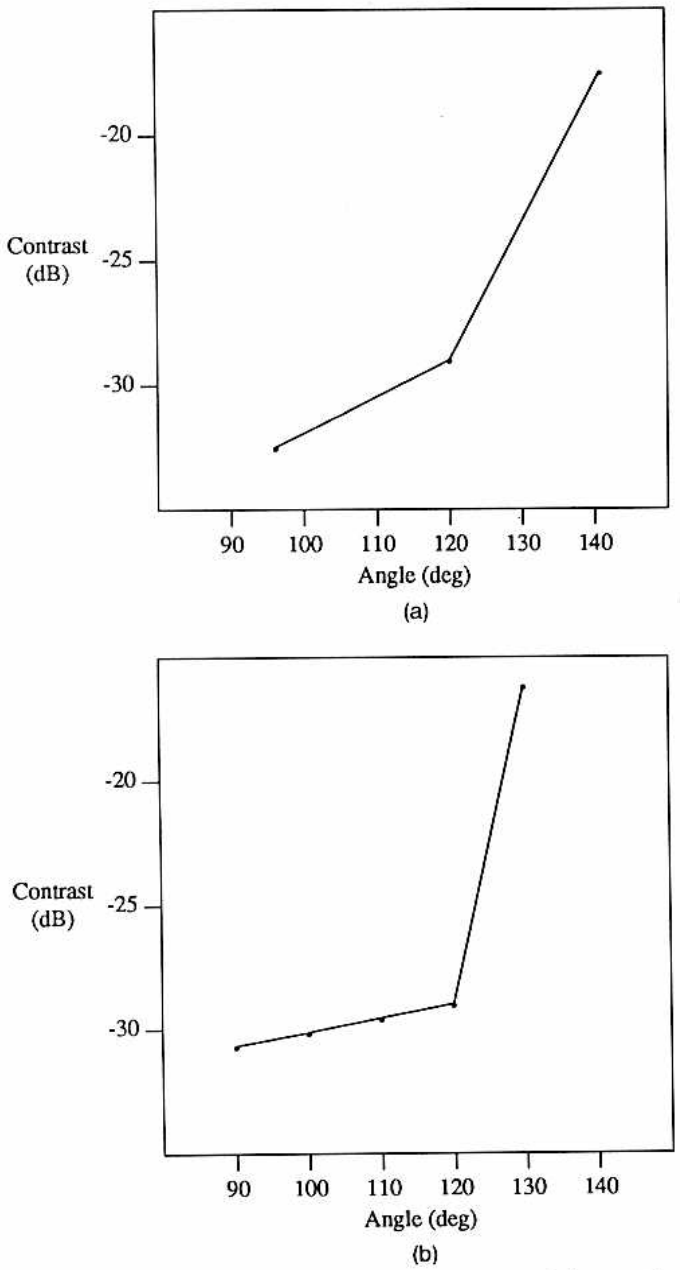



Fig. 14. The influence of angle on the coherence of sine-grating plaids. (a) One grating had a fixed contrast of 0.3, while the other was of variable contrast. The spatial frequency of one grating was fixed at 2.4 cycles deg$^{-1}$, and that of the second grating was fixed at 1.2 cycles deg$^{-1}$. As the angle between the two gratings varied, their speeds were chosen so that the coherent plaid moved at a fixed speed of 7.5 deg sec$^{-1}$. The plot shows the threshold contrast for coherence for a range of angles. Subject, EHA. (b) One grating had a fixed contrast of 0.3, and both had a fixed spatial frequency of 0.25 cycle pixel$^{-1}$. The speed of the gratings was chosen so that the coherent plaid moved at a fixed speed of 2/3 pixel frame. A fixed value was chosen as the threshold value for the curvature measure. [This value was chosen in order to match the psychophysical data in (a) for an angle of 120 deg.] For each angle, the plot shows the test-grating contrast needed for the curvature measure to attain that value.

frequencies, and for each test case we measure the contrast that is needed to attain those threshold values.

In this way Adelson and Movshon[33] measured the threshold elevation of coherence for plaids made up of gratings with different spatial frequencies, plotted in Fig. 13(a). As the frequencies of the two gratings were made different, the tendency to cohere was reduced and the contrast needed for coherence was increased. Figure 13(b) was generated by choosing a threshold value for the curvature measure; the plot shows the contrast elevation needed at each relative spatial frequency for the curvature measure to attain that value. Comparison of Figs. 13(a) and 13(b) indicates that the model's mechanisms are tuned to a somewhat narrower band of spatial frequencies than are the mechanisms of the human visual system.

Figure 14(a) shows the effect on coherence of varying the angular separation between the two gratings. As the angle was increased from 90° the tendency to cohere was reduced and the contrast needed for coherence was increased. The simulated data, plotted in Fig. 14(b), are similar to those plotted in Fig. 14(a), although the rate of increase is somewhat different.

The plots in Figs. 13 and 14 are promising. There are several parameters of the model that may be adjusted with the hope of matching the psychophysical data exactly[38]: (1) the spatial bandwidths of the motion-energy filters—broader spatial bandwidth should make the plot in Fig. 13(b) broader; (2) the ratio of the temporal bandwidths to the spatial bandwidths—decreasing this ratio should make the plot in Fig. 14(b) steeper; and (3) the nature of the nonlinearity—for example, squaring accentuates the contrast difference more than absolute value and should tend to make the plot in Fig. 13(b) narrower and the plot in Fig. 14(b) steeper.

## 8. COMPARING THE MODEL WITH PHYSIOLOGY

Figure 15 depicts the correspondence between the computations performed by the model and the stages of the visual motion pathway of the primate brain. The model's computations are simply a series of linear steps (weighted sums) alternating with point nonlinearities. In this section, I compare the model with some of the known functional properties of cells in the visual motion pathway.

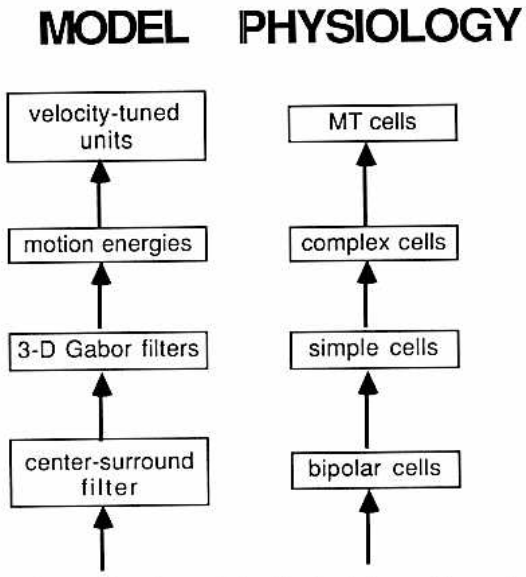



Fig. 15. Comparing the model with physiology. The model's computations are simply a series of linear steps (weighted sums) alternating with point nonlinearities.

### A. Some Differences

The receptive fields of cells in area MT are much larger than their counterparts in striate cortex (V1). Newsome *et al.*[39] report that they are 5–20 times larger in area, and Movshon[40] reports that they are as much as 50 times larger. The receptive fields of the model's velocity-tuned units are only about four times as big as the Gabor-energy filters.

The nonlinear stages in the model are different from those found in physiology. A future implementation of the model will replace squaring with a more biologically plausible S-shaped nonlinearity, such as that measured from photoreceptor cells (for example, see Ref. 41).

The model is implemented as a series of independent families of motion-energy filters arranged in an orderly manner; each filter has one of four spatial orientations and one of three temporal-frequency tunings, and the families of filters have equal bandwidths and are spaced 1 octave apart in spatial frequency. Cells in V1 are distributed more haphazardly. Some interesting experiments involve studying populations of cells in order to determine whether they are restricted to a single band of temporal-frequency tunings and whether their bandwidths and frequency tunings vary inversely with each other.

For many image sequences, the speed of image motion is faster in the periphery of the visual field then in the center (e.g., when one is walking down a hallway or through a forest). Cells in visual cortex generally have larger receptive fields and lower spatial-frequency tunings at greater eccentricities, i.e., cells with receptive fields near the fovea are well suited for estimating slow speeds, while those farther out are suited for high speeds. The model, conversely, currently has units tuned to each spatial-frequency band at every image location.

### B. Similarities with Striate Cortex

Two-dimensional Gabor filters are a physiologically plausible model for the two-dimensional receptive-field structure in striate cortex. Recent neurophysiological experiments have shown that Gabor functions may constitute a better model of cortical simple-cell structure than previously proposed receptive-field models.[42,43] Electrophysiological studies[44] also indicate that the space–time receptive-field structure of simple cells is similar to that of 3-D Gabor filters and to that of the Adelson–Bergen[7] and Watson–Ahumada[6] filters. Moreover, experiments by Emerson *et al.*[45] suggest that the space–time receptive-field structure of complex cells is well modeled by motion energy.

In principle, the model could be built by using filters with an even more biologically plausible temporal response (space–time Gabor filters are noncausal), but the straightforward analytical form of Gabor filters (Gaussians in the frequency domain) is what simplified the task of evaluating the integral in Eq. (7). This integral is used to compute the weights, $w_{ij}$ and $\overline{w}_{ij}$, in Eq. (11); different filters would yield different weights. For a particular filter it may not be possible to derive an analytical formula for computing these weights. In such cases the weights might be "learned" by implementing the model as a neural network with an iterative learning rule (for references, see Rummelhart[46]).

As was discussed in Section 6, the spatial bandwidths of the model's filters must be less than the temporal bandwidths. For example, the filter that is most sensitive to leftward motion of vertically oriented gratings gives a larger output for rightward grating motion than for upward or downward grating motions. One could measure experimentally whether this is true for direction-selective striate cells.

### C. Similarities with Middle Temporal Area

Both MT cells and the model's units increase their outputs as image contrast is increased. Note, however, that the location of the peak output in the distribution of velocity-tuned units does not change as image contrast is varied.

MT cells are inhibited (for example, see Ref. 47) by motion opposite their preferred velocity (the same speed as the latter but in the opposite direction). The model's velocity-tuned units may give positive or negative outputs; interpreting the negative outputs as inhibition, the model units are similarly inhibited by motion opposite their preferred velocity.

Most MT cells are more-or-less spatiotemporally separable,[40] meaning that they prefer the same temporal frequency for a range of spatial frequencies (or vice versa). The model units are similarly spatiotemporally separable. This is simply because the axes of the elliptical-Gaussian windows of the 3-D Gabor filters in the model are parallel to the $x$, $y$, and $t$ axes, i.e., these Gaussians are separable in space–time. Though the 3-D Gabor filters are not themselves spatiotemporally separable, the Gabor-energies and the velocity-tuned units are.

Felleman and Kaas[48] report that the majority of MT cells respond best to a particular velocity, with marked attenuation for speeds greater than or less than the preferred. Some neurons fail to show significant response attenuation even at the lowest test velocity (low-pass speed tuning), while others fail to attenuate at the highest velocities (high-pass speed tuning).

The model units have speed-tuning curves that resemble MT speed tunings. For each stimulus the model encodes velocity as the peak in the distribution of outputs of the velocity-tuned units. A speed-tuning curve plots the output of a single unit for a variety of stimuli. Consider a pattern translating with speed $s = \sqrt{u^2 + v^2}$. The output of the unit that corresponds to the location of the peak depends on two terms, $\sum_{i=1}^{12}(m_i)^2$ and $-\sum_{i=1}^{12} f(u, v)$, from Eq. (10). The second term will be maximized for speed $s$, but the first term may be maximized for some other speed. Thus, in general, the unit that corresponds to speed $s$ will be tuned to a speed other than $s$. In other words, the units generally offer the most information about speeds different from their tuning speeds. The result of this is that some of the model units have sharply peaked speed-tuning curves, while others (those that correspond to slower speeds) are low-pass speed tuned. There are no units in the current implementation of the model that are high-pass speed tuned.

Movshon *et al.*[49] have classified most of the cells that they have probed into two types by observing their responses to sine-grating plaid stimuli. The first type, called pattern-flow cells, are tuned to the velocity of the pattern as a whole, exhibiting their peak response when the combined plaid pattern moves at the preferred velocity regardless of the motion of the two component sine-gratings. The second type, called component-flow cells, yield their peak response when either of the two component sine-gratings moves at the preferred velocity.

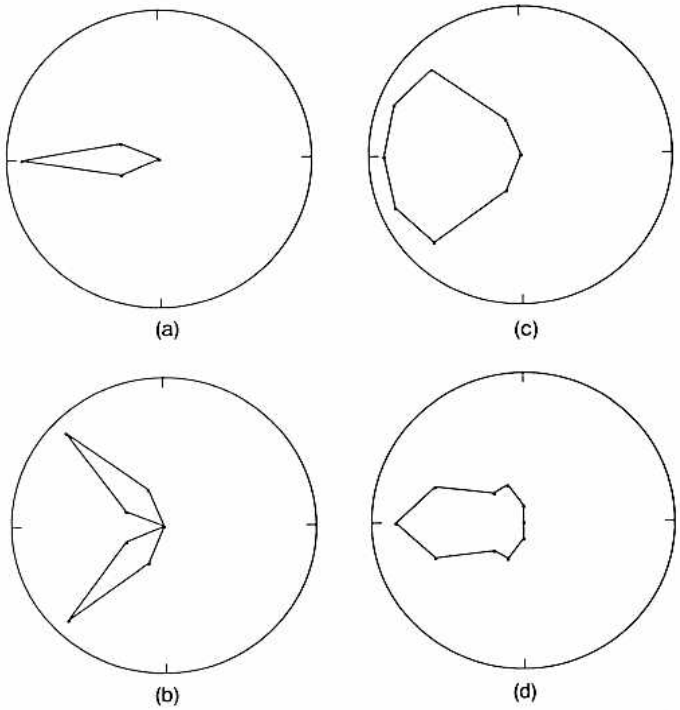



Fig. 16. Direction tuning of component- and pattern-flow model units. (a) Response of a typical component-flow unit as a function of direction of motion for moving gratings that were matched to the unit's preferred speed and spatial frequency. (b) Direction tuning of the same component-flow unit for sine-grating plaids; the tuning curve has two lobes, indicating that the unit responds when either of the two component sine-gratings move at the preferred velocity, similar to component MT cells. (c) Direction tuning of a pattern-flow unit for gratings. (d) Direction tuning of the same pattern-flow unit for plaids; the single lobe indicates that the unit responds to the combined pattern motion regardless of the motion of the component gratings, similar to pattern MT cells.

In the context of the model, the outputs after operation 2 of the algorithm in Section 5 correspond to component-flow cells and the outputs after operation 3 correspond to pattern-flow cells. Direction-tuning curves for a typical component-flow unit are shown in Figs. 16(a) and 16(b). For a single moving grating, the unit has a single preferred direction of motion corresponding to the normal-flow velocity of the grating. It has two peaks for a plaid, each of which corresponds to the normal-flow velocity of one of the component gratings. Direction-tuning curves for a typical pattern-flow unit are shown in Figs. 16(c) and 16(d). The cell has a single preferred direction of motion for single moving gratings as well as for plaids.

Finally, the model units have different direction- and speed-tuning curves for different stimuli. In particular, both the model units and the MT cells[40] have sharper direction-tuning curves for moving random-dot fields than for gratings. Allman[50] confirmed that moving random dots are the optimal stimulus for MT cells. Random dots are optimal for the model units as well, since the model was derived for such random textures.

## 9. SUMMARY

In this paper a model is presented for computing local image velocity that is consonant with current views regarding the neurophysiology and psychophysics of motion perception. The power spectrum of a moving texture occupies a tilted plane in the spatiotemporal-frequency domain. The model uses 3-D (space–time) Gabor filters to sample this power spectrum and, by combining the outputs of several such filters, the model estimates the slope of the plane (i.e., the velocity of the moving texture). The output velocity is encoded as the peak in a distribution of velocity-tuned units that behave much like cells of the MT area of the primate brain.

The model appears to solve the aperture problem as well as that of the human visual system, since it extracts the correct velocity for patterns having large differences in contrast at different spatial orientations (>32:1 contrast ratio for some patterns). It is capable of recognizing when there is an ambiguous velocity estimate resulting from the motion of a strongly oriented pattern (such as a grating), and in such cases it chooses the normal-flow velocity. In addition, the model may be used to simulate psychophysical data on the coherence of sine-grating plaid patterns.

The model gives accurate estimates of two-dimensional velocity for a wide variety of test cases, including realistic images, sequences generated from images of natural textures, and some sine-grating plaid patterns. The model may prove to be an interesting framework for future research in the psychophysics and neurophysiology of motion perception as well as in computer vision.

## APPENDIX A: GABOR FILTERS FROM SEPARABLE COMPONENTS

To convolve a two-dimensional image by a horizontally oriented sine-phase Gabor filter, we may convolve each image row by a one-dimensional sine-phase Gabor filter, then convolve each column of the resulting image by a one-dimensional Gaussian. This appendix outlines a new technique for building 3-D Gabor filters of any orientation and with elliptical Gaussian windows of any aspect ratio from linear combinations of separable filters by making use of the following trigonometric identities:

$$\begin{aligned}\sin(\omega_{t_0} + t\omega_{x_\sigma}x + \omega_{y_\sigma}y) = {} & \sin(\omega_{t_0}t)\cos(\omega_{x_0}x)\cos(\omega_{y_\sigma}y) \\ & - \sin(\omega_{t_0}t)\sin(\omega_{x_0}x)\sin(\omega_{y_\sigma}y) \\ & + \cos(\omega_{t_0}t)\sin(\omega_{x_0}x)\cos(\omega_{y_\sigma}y) \\ & + \cos(\omega_{t_0}t)\cos(\omega_{x_0}x)\sin(\omega_{y_\sigma}y),\end{aligned} \tag{A1}$$

$$\begin{aligned}\cos(\omega_{t_0} + \omega_{x_\sigma}x + \omega_{y_\sigma}y) = {} & \cos(\omega_{t_0}t)\cos(\omega_{x_0}x)\cos(\omega_{y_\sigma}y) \\ & - \cos(\omega_{t_0}t)\sin(\omega_{x_0}x)\sin(\omega_{y_\sigma}y) \\ & - \sin(\omega_{t_0}t)\sin(\omega_{x_0}x)\cos(\omega_{y_\sigma}y) \\ & - \sin(\omega_{t_0}t)\cos(\omega_{x_0}x)\sin(\omega_{y_\sigma}y).\end{aligned} \tag{A2}$$

Let $G_s(t, \sigma_t, \omega_{t_0})$ be a one-dimensional sine-phase Gabor function as given by Eq. (3), and let $G_c(t, \sigma_t, \omega_{t_0})$ be the corresponding cosine-phase filter. Using Eq. (A1), the output of an arbitrarily oriented 3D (space–time) sine-phase Gabor filter may be computed by doing the following separable convolutions:

1. Convolve the image sequence in time by $G_s(t, \sigma_t, \omega_{t_0})$, next each image row by $G_c(x, \sigma_x, \omega_{x_0})$, and then each column by $G_c(y, \sigma_y, \omega_{y_0})$.
2. Convolve the image sequence in time by $G_s(t, \sigma_t, \omega_{t_0})$, next each image row by $G_s(x, \sigma_x, \omega_{x_0})$, and then each column by $G_s(y, \sigma_y, \omega_{y_0})$.
3. Convolve the image sequence in time by $G_c(t, \sigma_t, \omega_{t_0})$, next each image row by $G_s(x, \sigma_x, \omega_{x_0})$, and then each column by $G_c(y, \sigma_y, \omega_{y_0})$.
4. Convolve the image sequence in time by $G_c(t, \sigma_t, \omega_{t_0})$, next each image row by $G_c(x, \sigma_x, \omega_{x_0})$, and then each column by $G_s(y, \sigma_y, \omega_{y_0})$.
5. Subtract the result of Step 2 from the sum of the results of Steps 1, 3, and 4. Note that if $\sigma_x$, $\sigma_y$, and $\sigma_t$ are not equal, the Gaussian window will be elliptical, but the axes of the ellipsoid will always be parallel to the $x$, $y$, and $t$ axes.

## ACKNOWLEDGMENTS

Special thanks are due to Ted Adelson for motivating this research and for providing the psychophysical data on coherence of sine-grating plaids, to Sandy Pentland for his invaluable help and advice, to Tony Movshon for dialogues on physiology, to David Marimont for his mathematical insight, and to Mark Turner for introducing me to Gabor filters. I especially want to thank Jack Nachmias and Grahame Smith for their detailed comments on earlier drafts of this paper and Lynn Quam for generating the Yosemite fly-through image sequence.

This research is supported at the University of Pennsylvania by contracts ARO DAA6-29-84-k-0061, AFOSR 82-NM-299, NSF MCS-8219196-CER, NSF MCS 82-07294, AVRO DAABO7-84-K-FO77 and NIH 1-RO1-HL-29985-01, at SRI International by contracts NSF DCR-83-12766, DARPA MDA 903-83-C-0027, DARPA DACA76-85-C-0004, and by the Systems Development Foundation. It was performed in part while the author was with the University of Pennsylvania and in part while he was with the Artificial Intelligence Center. SRI International, 333 Ravenswood Avenue, Menlo Park, California 94025.

## REFERENCES AND NOTES

1. S. T. Barnard and W. B. Thomson, "Disparity analysis of images," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-2,** 333–340 (1980).
2. B. K. P. Horn and B. G. Schunk, "Determining optical flow," Artif. Intell. **17,** 185–203 (1981).
3. J. K. Kearney and W. B. Thompson, "An error analysis of gradient-based methods for optical flow estimation," IEEE Trans. Pattern Analy. Mach. Intell. **19,** 229–244 (1987).
4. An error analysis of gradient-based methods[3] confirms that a major problem with the approach is that large errors are made where the image is highly textured, where there is the greatest amount of motion information!
5. A. B. Watson and A. J. Ahumada, "A look at motion in the frequency domain," Tech. Rep. 84352 (NASA-Ames Research Center, Moffett Field, Calif., 1983).
6. A. B. Watson and A. J. Ahumada, "Model of human visual-motion sensing," J. Opt. Soc. Am. A, **2,** 322–342 (1985).
7. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," J. Opt. Soc. Am. A, **2,** 284–299 (1985).
8. J. P. H. van Santen and G. Sperling, "Elaborated Reichardt detectors," J. Opt. Soc. Am. A, **2,** 300–321 (1985).
9. D. J. Fleet, "The early processing of spatio-temporal visual information," M.S. thesis (University of Toronto, Toronto, Canada, 1984; available as Tech. Rep. RBCV-TR-84-7).
10. D. J. Fleet and A. D. Jepson, "A cascaded filter approach to the construction of velocity selective mechanisms," Tech. Rep. RBCV-TR-84-6, Department of Computer Science (University of Toronto, Toronto, Canada, 1984).
11. E. C. Hildreth, "Computations underlying the measurement of visual motion," Artif. Intell. **23,** 309–355 (1984).
12. M. Fahle and T. Poggio, "Visual hyperacuity: spatio-temporal interpolation in human vision," Proc. R. Soc. London Ser. B **213,** 451–477 (1981).
13. In their earlier paper Watson and Ahumada[5] also employ Gabor filters but not Gabor energy. In their later work[6] they abandon Gabor filters. Adelson and Bergen[7] do not use Gabor filters, but they do compute motion energy.
14. D. Gabor, "Theory of communication," J. Inst. Electr. Eng. **93,** 429–457 (1946).
15. J. G. Daugman, "Two-dimensional analysis of cortical receptive field profiles," Vision Res. **20,** 846–856 (1980).
16. J. G. Daugman, "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 (1985).
17. Complexity is defined as the order of magnitude, $O(\ )$, of the number of operations required for a computation.
18. P. Burt, "Fast algorithms for estimating local image properties," Comput. Vision Graphics Image Process. **21,** 368–382 (1983).
19. Filters higher in the pyramid achieve their peak response for patterns with lower spatial frequency but with the same temporal frequency. Thus the lower-frequency filters have their greatest outputs for patterns moving at greater velocities. Psychophysical evidence (see Watson and Ahumada[6] for references) suggests that human motion channels exhibit such a relationship between spatial frequency and velocity. This makes sense from a computational viewpoint since patterns containing only high spatial frequencies may move at only low velocities, whereas patterns containing only lower spatial frequencies may move at greater velocities (see the discussion in Subsection 2.A on sampling and temporal aliasing).
20. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization* (Academic, New York, 1981).
21. Equation (10) subtracts $f(u, v)$ from a contrast-dependent value, $\sum(m_i)^2$. This was an arbitrary choice, and a number of other contrast-dependent values might be substituted. Further investigation may indicate which, if any, of these measures most closely models the physiology.
22. R. A. Hummel and S. W. Zucker, "On the foundations of relaxing labelling processes," IEEE Trans. Pattern Analy. Mach. Intell. **5,** 267–287 (1983).
23. D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," IEEE Trans. Pattern Anal. Mach. Intell. **8,** 413–424 (1986).
24. If the stimulus has uncorrelated random phase, then the dc problem may be alleviated by using only sine-phase filters: a phase-independent motion energy can be computed from sine-phase filters alone by averaging their squared outputs within appropriately sized windows.
25. Brownian fractal functions (see Mandelbrot[26] for definitions and references) are characterized by similarity across scales and have an expected power spectrum that falls off as $P(\omega) = \omega^{-\beta}$ for some constant $\beta$. Fractals may be used to generate natural-looking textures.
26. B. B. Mandlebrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).
27. Special Interest Group in Computer Graphics, Institute of Electrical and Electronics Engineers.
28. Since the image velocities in the Yosemite fly-through image sequence are as high as 5 pixels per frame, we must use three levels from the pyramid. In future research, I hope to develop a rule for automatically combining estimates from the different levels. For now, I simply pick the level that is most appropriate for a given image region: the level-zero estimate is chosen if the actual velocity is between 0 and 1.25 pixels per frame, the level-

one estimate is chosen if it is between 1.25 and 2.5 pixels per frame, and the level-two estimate is chosen if it is between 2.5 and 5.0 pixels per frame. In the Yosemite fly-through image sequence, there are regions of low contrast adjacent to high-contrast regions (e.g., the face of El Capitan and the cloud region are of low contrast). This exacerbates the smoothing problem discussed in Subsection 4.B. For this image sequence, contrast was first equalized by computing the zero crossings (see Hildreth[11] for references) of each image. The model was then applied to the resulting zero-crossing image sequence.

29. Area MT is also known as V5.
30. The maximum of Eq. (10) can be located to any precision by using a finer or coarser grid. Also, the grid need be only of limited extent, since bandpass filtering limits the range of possible velocities (as discussed in Subsection 2.A).
31. Preliminary investigation indicates that absolute value may be substituted for squaring.
32. The model does not always recover the correct pattern-flow velocity for sine-grating plaids; e.g., for plaids made up of gratings with equal contrasts, equal spatial frequencies, and equal speeds, the model's estimates are in error (correct direction of motion but wrong speed) when the spatial frequencies of the gratings are not equal to the spatial-frequency tuning of the filters or when the angle between the gratings differs from 90 deg. The model might be more robust with respect to these factors if it utilized more motion-energy filters tuned to a greater number of orientations and spatial frequencies.
33. E. H. Adelson and J. A. Movshon, "Phenomenal coherence of moving visual patterns," Nature **300,** 523–525 (1982).
34. E. H. Adelson, Media-Technology Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 (personal communication).
35. An important psychophysical experiment is to measure the perceived velocity of plaids with various relative angles, contrasts, and spatial frequencies.
36. M. P. doCarmo, *Differential Geometry of Curves and Surfaces* (Prentice-Hall Inc., Englewood Cliffs, N. J., 1976).
37. The normal-flow direction is perpendicular to the direction of minimum curvature, and the normal-flow speed is the dot product of any position along the ridge with the normal-flow direction.
38. Different subjects were used to collect the data in Figs. 13(a) and 14(a). Thus the data in these two plots are inconsistent with each other, requiring that different curvature thresholds be used to generate Figs. 13(b) and 14(b). The eventual goal is to simulate all the data for one subject with one choice of parameters.
39. W. T. Newsome, M. S. Gizzi, and J. A. Movshon, "Spatial and temporal properties of neurons in macaque mt," Invest. Ophthalmol. Vis. Sci. Suppl. **24,** 106 (1983).
40. J. A. Movshon, Department of Psychology, New York University, New York, New York 10003 (personal communication).
41. K. Naka, and W. A. H. Rushton, "S-potentials from luminosity units in the retina of fish (cyprinidiae)," J. Physiol. **188,** 587–599 (1966).
42. S. Marcelja, "Mathematical description of the response of simple cortical cells," J. Opt. Soc. Am. A, **70,** 297–1300 (1980).
43. J. McLean-Palmer, J. Jones, and L. Palmer, "New degrees of freedom in the structure of simple receptive fields," Invest. Ophthalmol. Vis. Sci. Suppl. **26,** 265 (1985).
44. J. McLean-Palmer, Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania 19104 (personal communication).
45. R. Emerson, M. Citron, W. Vaughn, and S. Klein, "Substructure in directionally selective complex receptive fields of cat," Invest. Ophthalmol. Vis. Sci. Suppl. **27,** 16 (1986).
46. D. E. Rummelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT, Cambridge, Mass., 1986).
47. A. Mikami, W. T. Newsome, and R. H. Wurtz, "Mechanisms of direction and speed selectivity in the middle temporal visual area (mt) of the macaque monkey," Invest. Ophthalmol. Vis. Sci. Suppl. **24,** 107 (1983).
48. D. J. Felleman and J. H. Kaas. Receptive-field properties of neurons in middle temporal visual area (mt) of owl monkeys. J. Neurophysiol. **52,** 488–513 (1984).
49. J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome, "The analysis of moving visual patterns," in *Experimental Brain Research Supplementum II: Pattern Recognition Mechanisms*, C. Chagas, R. Gattass, and C. Gross, eds. (Springer-Verlag, New York, 1985), pp. 117–151.
50. J. Allman (Division of Biology, California Institute of Technology, Pasadena, California 91125), in his talk at the 1986 Symposium on Computational Models in Human Vision at the Center for Visual Sciences, University of Rochester.