Filter selection model for motion segmentation and velocity integration

Steven J. Nowlan* and Terrence J. Sejnowski

Howard Hughes Medical Institute, The Salk Institute, P.O. Box 85800, San Diego, California 92186-5800, and Department of Biology, University of California, San Diego, La Jolla, California 92093

Received November 4, 1993; revised manuscript received July 25, 1994; accepted July 27, 1994

We present a new approach to computing from image sequences the two-dimensional velocities of moving objects that are occluded and transparent. The new motion model does not attempt to provide an accurate representation of the velocity flow field at fine resolutions but coarsely segments an image into regions of coherent motion, provides an estimate of velocity in each region, and actively selects the most reliable estimates. The model uses motion-energy filters in the first stage of processing and computes, in parallel, two different sets of retinotopically organized spatial arrays of unit responses: one set of units estimates the local velocity, and the second set selects from these local estimates those that support global velocities. Only the subset of local-velocity measurements that are the most reliable is included in estimation of the velocity of objects. The model is in agreement with many of the constraints imposed by the physiological response properties of cells in primate visual cortex, and its performance is similar to that of primates on motion transparency.

1. INTRODUCTION

In most computational models of motion processing, analysis of motion proceeds in two stages: in the first stage the local velocity at every point in the image is computed from a sequence of two-dimensional image frames, and this high-resolution representation of local velocity is referred to as the optical flow field.[1-7] At a later stage this information is analyzed and combined into a three-dimensional interpretation of motion in the visual scene. There are two conflicting demands that must be met in the computation of the optical flow field: first, signals from neighboring regions of the visual field need to be spatially integrated for noise and the aperture problem to be overcome; and second, sensitivity to small velocity differences across space should be preserved for segmentation of regions corresponding to different objects.[8] The goal of this paper is to provide a computational solution to the problem posed by these conflicting demands for occluding and transparent moving stimuli.

The earliest cells in the primate visual system that have reliable direction and motion sensitivity are found in primary visual cortex, area V1.[9,10] However, these cells do not detect true velocity but instead are directionally selective and are tuned to a limited range of spatiotemporal frequencies.[10,11] In addition, these cells exhibit marked orientation selectivity and spatially restricted receptive fields. As a result, these early local motion responses are sensitive not just to the velocity of a moving object but also to many other features of the object, such as its spatial-frequency profile and local edge orientation (Fig. 1). The sensitivity to local edge orientation is usually referred to as the aperture problem[4,5,7,9]: within a small region of space it is possible to measure locally only the velocity component that is parallel to the local intensity gradient (i.e., perpendicular to local edge orientation).

In order to overcome these limitations and compute true local-velocity measurements, one must integrate motion responses from cells with a variety of direction and spatiotemporal frequency tunings.[2,3] There is now considerable psychophysical evidence suggesting that motion integration can be affected by a variety of figural segmentation cues such as contrast, spatial frequency, binocular disparity, color, transparency, and occlusion.[12,13] This suggests that biological visual systems integrate local motion measurements in a highly stimulus-dependent manner. The first area of the visual cortex in which motion integration appears to occur is area MT.[14,15]

Previous computational approaches to motion have attempted to combine local motion responses to produce precise local-velocity estimates at all points in an image: the optical flow field.[2,3,16-19] This is, in general, an ill-posed problem. The local-velocity estimates are often noisy and in some regions are systematically incorrect, particularly in regions of constant intensity or near motion boundaries. Therefore most of these models have relied on smoothness assumptions combined with techniques such as relaxation labeling[20-23] and finite-element regularization for estimation of optical flow.[17,18] The goal of these models is to get an estimate that is as close as possible to the true optical flow, given the assumptions under which the model operates. Despite their sophistication, these approaches have difficulty with scenes involving partial occlusion and transparency.

Rather than attempting to get a good estimate of the true optical flow at all image locations, the visual system may use a relatively simple and imperfect strategy for computing velocity that depends on estimating the validity of local velocity estimates. There is psychophysical evidence for an initial coarse velocity estimate that may be used for image segregation.[24] One can obtain a coarse representation of object motion by relying on only the most valid subsets of local-velocity measurements. In this paper we present a model that uses this approach and apply it to images of moving objects that include occlusion and transparency. We optimized our model to es-

0740-3232/94/123177-24$06.00 © 1994 Optical Society of America

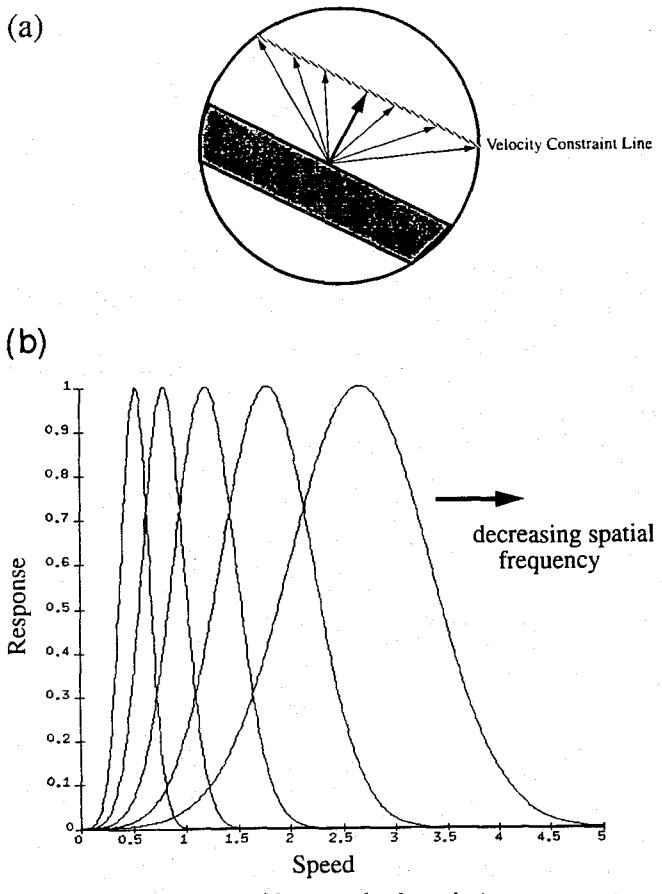

Fig. 1. (a) Aperture problem: only the velocity component orthogonal to the edge of the bar, indicated by the heavy arrow, can be measured by any local-velocity mechanism restricted to this region of the image. However, any of the velocities indicated by the other arrows, all of which terminate on the velocity constraint line, would produce the same motion within this aperture. (b) Limited spatiotemporal tuning: velocity tuning of motion-energy filters defined in Eqs. (3)–(5) as a function of speed for a moving sine-wave grating. The tuning curves are normalized by the maximum responses and are plotted for five different spatial frequencies (7.6, 5.1, 3.4, 2.3, and 1.5 cycles/deg, with the highest spatial-frequency curve at the left and the lowest spatial-frequency curve at the right). If the speed tuning were independent of speed, the curves would have peaks at the same speed. The tuning curves become broader and shift to the right as the spatial frequency of the grating decreases, indicating that these filters do not respond to pure velocity. Cells in V1 exhibit similar limited spatiotemporal-tuning bandwidth.

timate the velocity of visual targets by solving in parallel the problems of *how* to compute local-velocity estimates and *which* local velocity estimates to use. The optimized model yielded accurate velocity estimates from synthetic images containing multiple moving targets of varying size, luminance, and spatial-frequency profile. The properties of the units in the model were consistent with the responses of neurons in area MT.[25]

In Section 2 we present an overview of the model, describing the stages of processing in the model in some detail. In Section 3 we introduce the statistical assumptions underlying the model and derive the procedure for modifying the parameters in the model during training. Section 4 presents simulation results for the model on synthetic images containing a variety of segmentation and transparency phenomena. The discussion in Section 5 places the model in the context of previous models of motion processing. A preliminary version of this model was described previously.[26]

2. MODEL

A. Overview

The model is a feed-forward cascade of locally connected networks of processing units organized into two parallel processing pathways. The stages of the model form layers of units with a roughly retinotopic organization. Figure 2 schematically represents the activity in the model at one instant in time.

Processing in the model is divided into three main stages, as described in more detail in the following subsections. In the first stage local motion energy is extracted from all locations in the input image. There are 36 motion-energy measurements for each image location that represent filters tuned to four different directions and nine different combinations of spatial and temporal frequency (only a few are shown in Fig. 2). In the second stage the local velocities and the validity of each velocity estimate are computed in parallel. The local-velocity pathway combines information from motion-energy filters tuned to different directions and spatial and temporal frequencies to find the plane or planes of maximal motion energy in spatial- and temporal-frequency space. This plane of maximal motion energy provides an estimate of local velocity.[2,3] The selection pathway estimates the local validity of each velocity estimate from the local motion-energy distribution on which that estimate was based and compares it with the motion-energy distributions supporting this velocity in other regions of the image. In the final stage of the model, global estimates of the velocities of objects within the visual scene are formed by integration across subsets of the local-velocity estimates according to the relative confidence values assigned by the selection pathway.

The model computes evidence for particular velocities in an image rather than computing velocity directly. In the local-velocity pathway, velocity is represented as a distribution of activity over a set of units representing different directions and speeds of motion, and these activities represent the local evidence in favor of a particular velocity. This local evidence is required to sum to 1, so that

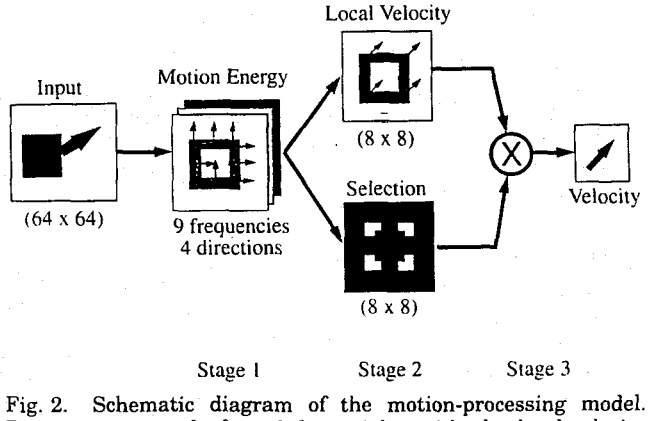

Fig. 2. Schematic diagram of the motion-processing model. Processing proceeds from left to right, with the local-velocity and selection stages operating in parallel. Shading within the boxes indicates different levels of activity at each stage. Arrows within motion-energy and local-velocity stages indicate local directions of motion-responding units. The responses shown are illustrative (see Figs. 10–12 below for more-detailed examples).

each location in the image can provide strong evidence for only one velocity or weaker evidence for multiple velocities. In a similar manner the selection pathway weights each region of an image according to the amount of support the region provides for a particular velocity relative to other regions of the image. The total amount of support for each velocity across the entire image cannot exceed 1, so the only way to produce strong global evidence for a particular velocity is for the selection pathway to focus all support for that velocity on regions of the image that provide strong evidence for that velocity. In this way the selection pathway coarsely segments an image into regions that support different velocities in a manner similar to that of the support maps proposed in Ref. 27. Note that this segmentation is not explicitly represented in the output of the model but is easily derived from signals present in the selection pathway.

In the final stage of the model the global evidence for a visual target moving at a particular velocity $V_k(t)$ is computed as a sum over the product of the outputs of the local velocity and selection pathways:

$$V_k(t) = \sum_{x,y} I_k(x, y, t) S_k(x, y, t), \tag{1}$$

where $I_k(x, y, t)$ is the local evidence for velocity k computed by the velocity pathway from region (x, y) at time t and $S_k(x, y, t)$ is the weight assigned by the selection pathway to that region. The motion-energy stage of the model was fixed; however, all later stages of the model were adaptive, and the weights were adjusted to optimize a measure of the overall performance of the model (see Section 3 below). The optimization procedure was an extension of the mixture-of-experts model.[28,29]

B. Local Motion Energy

The first stage of the model requires some localized mechanism for measuring motion parameters that is based on changes in intensity in an image over time. Two different classes of motion detectors have been proposed: spatiotemporal filters and feature matchers. The latter class of detector identifies abstract features in an image (such as corners or line conjunctions) and matches these features across successive frames of an image sequence. One can determine the velocity of the feature by dividing the (directed) distance that the feature traveled by the time between frames. Although a feature-matching scheme may be involved in some aspects of the long-range motion process,[30] the responses of direction-selective cells in early visual areas seem much better characterized by spatiotemporal filters.[31–33]

Motion detectors based on spatiotemporal filters can themselves be divided into three broad categories. Correlation-based methods[34,35] correlate a spatial response with a delayed and displaced version of the same spatial response. Motion-energy methods combine the (squared) responses of linear filters oriented in space–time.[36,37] Gradient methods[5,18] assume that intensity (Q) is conserved across a sequence of frames (i.e., $dQ/dt = 0$). This assumption allows one to derive an expression that relates intensity changes over time (across frames) to intensity changes over space and to the velocity components of motion:

$$\frac{dQ}{dt} = \frac{\partial Q}{\partial x}\frac{dx}{dt} + \frac{\partial Q}{\partial y}\frac{dy}{dt} + \frac{\partial Q}{\partial t} = 0. \tag{2}$$

Physiological implementations of all three types of spatiotemporal method have been proposed, but physiological evidence[38–41] suggests that responses in primary visual cortex best resemble the responses expected from various stages of a motion-energy model. As with neurons in primary visual cortex that are directionally selective, the limited spatiotemporal tuning of motion-energy units means that their optimal velocity (the velocity for maximal response) changes as the spatial-frequency content of the visual scene changes (see Fig. 1).

The motion-energy model is based on the observation that image motion is characterized by orientation in space–time. An intensity edge moving at a constant velocity produces a line with a particular slope in space–time. Thus an oriented space–time filter responds most strongly to intensity edges moving at a particular velocity. A single oriented space–time filter is, however, a less-than-ideal motion detector, because its output is phase sensitive. This means that the response of such a filter to a moving pattern depends on how that pattern lines up with the receptive field at each moment in time. Simply switching the polarity of the pattern (making dark regions bright and vice versa) will change the sign of the output of the filter. If a pattern such as a moving sine-wave grating passes through the receptive field, the output of the filter will tend to oscillate sinusoidally over time. Watson and Ahumada[37] showed that these oscillations can be used in some cases to compute velocity. However, a phase-invariant response is physiologically more plausible.[42]

Adelson and Bergen[36] demonstrated a particularly simple way of constructing a phase-invariant motion detector by combining the squared outputs of a pair of spatiotemporal filters with the same orientation but whose receptive fields were 90° out of phase (a quadrature pair). Adelson and Bergen called the nonlinear filter constructed in this manner a motion-energy filter, and this is the type of filter used to provide the initial motion responses in our model. Several authors have suggested that the responses of the oriented linear filters are similar to the responses of simple cells in visual cortex, whereas the outputs of the energy units are more similar to responses of complex cells.[36,40,41,43,44]

A convenient mathematical description of a quadrature pair of oriented space–time filters is sine-phase and cosine-phase Gabor filters.[45–47] One can calculate motion energy by combining the squared outputs of sine and cosine Gabor filters with the same center frequency and spread function. Filters of this form were used in previous models of velocity extraction[2,3] and have the important property that they have limited extent in both space–time and spatiotemporal frequency domains. Although a single motion-energy unit built with filters of this form is not velocity selective (see Fig. 1), a set of such units that tile the spatiotemporal frequency space can be used to extract the velocity of a pattern independent of its spatial-frequency content, as shown in Subsection 2.C below.

The spatial receptive field of a Gabor filter is similar to the spatial receptive fields of simple cells in striate cortex[2,36]; however, simple cells tend to have space–time responses that are unlike those of spatiotemporal Gabor filters. In addition, the temporal response of a Gabor signal is acausal and therefore difficult to realize physi-

cally. However, Adelson and Bergen[36] demonstrated that space–time-oriented Gabor filters can be approximated by summation of the responses of four space–time-separable filters with Gabor spatial responses and low-pass-filter temporal responses. Simple cell responses in visual cortex can also be approximated as sums of similar space–time-separable filters.[48] The use of sums of separable filters also has computational advantages, because it reduces the number of operations required for performing the filter convolutions.

The filters used in the model were implemented with pairs of spatial filters that have a Gabor spatial response:

$$g_{\sin}(x, y) = \frac{1}{(2\pi)\sigma_x\sigma_y} \exp\left[-\left(\frac{x^2}{2{\sigma_x}^2} + \frac{y^2}{2{\sigma_y}^2}\right)\right] \times \sin(2\pi\omega_x x + 2\pi\omega_y y), \quad (3)$$

$$g_{\cos}(x, y) = \frac{1}{(2\pi)\sigma_x\sigma_y} \exp\left[-\left(\frac{x^2}{2{\sigma_x}^2} + \frac{y^2}{2{\sigma_y}^2}\right)\right] \times \cos(2\pi\omega_x x + 2\pi\omega_y y), \quad (4)$$

where (ω_x, ω_y) is the filter center frequency and (σ_x, σ_y) is the spread of the spatial Gaussian window. These filters were combined with bandpass temporal filters of the form

$$f_k(t) = (\omega_t t)^k \exp(-\omega_t t)\left[\frac{1}{k!} - \frac{(\omega_t t)^2}{(k+2)!}\right], \quad (5)$$

where ω_t is the filter center frequency and k determines the tuning width, with higher values of k producing more narrowly tuned filters. By appropriate choices of a pair of Gabor filters and bandpass filters, quadrature pairs of spatiotemporally oriented filters with a bandpass spatiotemporal frequency response similar to that of a Gabor filter can be constructed. The construction of four such filters, forming a leftward tuned and a rightward tuned quadrature pair, is shown in Fig. 3 (see Adelson and Bergen[36] for further details).

In the implementation of the model, a sequence of 64×64 pixel input frames with 256 gray levels per pixel location was first convolved with a difference of Gaussians filter ($\sigma_1 = 2$ pixels, $\sigma_2 = 7$ pixels). This filter was balanced to remove the dc component of the image and also provided smoothing and edge enhancement. This filter is a simplified model of retinal processing and provides a sequence of contrast images to be processed by the motion-energy filters. This sequence of contrast images was then convolved with 36 different motion-energy filters.

These filters were tuned to four different directions of motion (up, down, left, and right). For each direction of motion there were nine different filters representing all possible combinations of three temporal and three spatial frequencies. These filters were implemented as $16 \times 16 \times 16$ kernels. The filter center frequency spacings were 1 octave spatially (0.25, 0.125, and 0.0625 cycles/pixel) and 1.5 octaves temporally (0.5, 0.177, and 0.0625 cycles/frame). The filters were designed so that there was an inverse relationship between the center frequency of the filter and its receptive-field size (filters tuned to lower spatial and temporal frequencies had receptive fields with larger spatial and temporal extent). This relationship is similar to that found between the size and spatial-frequency tuning in the retina and visual cortex.[49–51] This type of relationship also permits the response of the family of filters to be approximated with use of a Gaussian pyramid.[52] A single family of filters representing the three different temporal frequencies at the highest spatial frequency were applied to three levels of a Gaussian pyramid for computing the filter responses for each direction of motion. Since the images at the highest level of the pyramid are only 1/16 the size of the original images, this procedure represents a considerable computational savings.

The outputs of the motion-energy stage of the model were organized into a grid of 49×49 receptive-field locations. (This grid is smaller than the 64×64 input image because of edge effects: filters centered near the edge of the image have part of their receptive field lying outside the image, and these motion-energy measurements are ignored.) For each of these receptive-field locations (which we will index with x and y) there were 36 raw motion-energy measurements. The strength of these motion-energy measurements was a quadratic function of the local contrast in the image. Cortical cell responses have a limited dynamic range and appear to respond to relative rather than absolute contrast.[53,54] Heeger[41] suggested that primary visual cortical responses can be better modeled as the outputs of a normalized energy mechanism. In our model we adopt this suggestion and compute the output of the motion-energy stage at each receptive-field location by a soft-maximum normalization[55]:

$$\hat{E}_i(x, y) = \frac{\exp[E_i(x, y)]}{\sum_j \exp[E_j(x, y)]}, \quad (6)$$

where $E_i(x, y)$ is one of the 36 raw motion-energy measurements at location (x, y) and $\hat{E}_i(x, y)$ is the corresponding normalized response. These normalized responses lie between 0 and 1 and are modulated by relative rather than absolute contrast. For strictly positive inputs the response of the soft maximum is quasi-linear and has saturation characteristics similar to those of the half-squared operators proposed by Heeger,[41] although the offset and gain of the soft-maximum operation are different.

C. Local Velocity

The properties of units in the selection and local-velocity pathways were determined by the optimization procedure, which is described formally in Section 3 below. The properties of these units were also determined by the constraints imposed by the architecture of the local-velocity and selection networks in the model. The organization of the local-velocity pathway was based on principles established in previous models of velocity estimation. The spatiotemporal power spectrum of a rigidly translating scene lies on a plane in the frequency domain.[2,3,37] This is illustrated in Fig. 4(a), where for illustrative purposes we show only one spatial-frequency dimension. This plane is defined by

$$\omega_t + v_x\omega_x + v_y\omega_y = 0, \quad (7)$$

where $(\omega_x, \omega_y, \omega_t)$ are the spatial- and temporal-frequency components of the pattern and (v_x, v_y) its velocity components. This equation is a restatement of intensity conservation [Eq. (2)] and can be violated in scenes containing occlusion and transparency. Never-

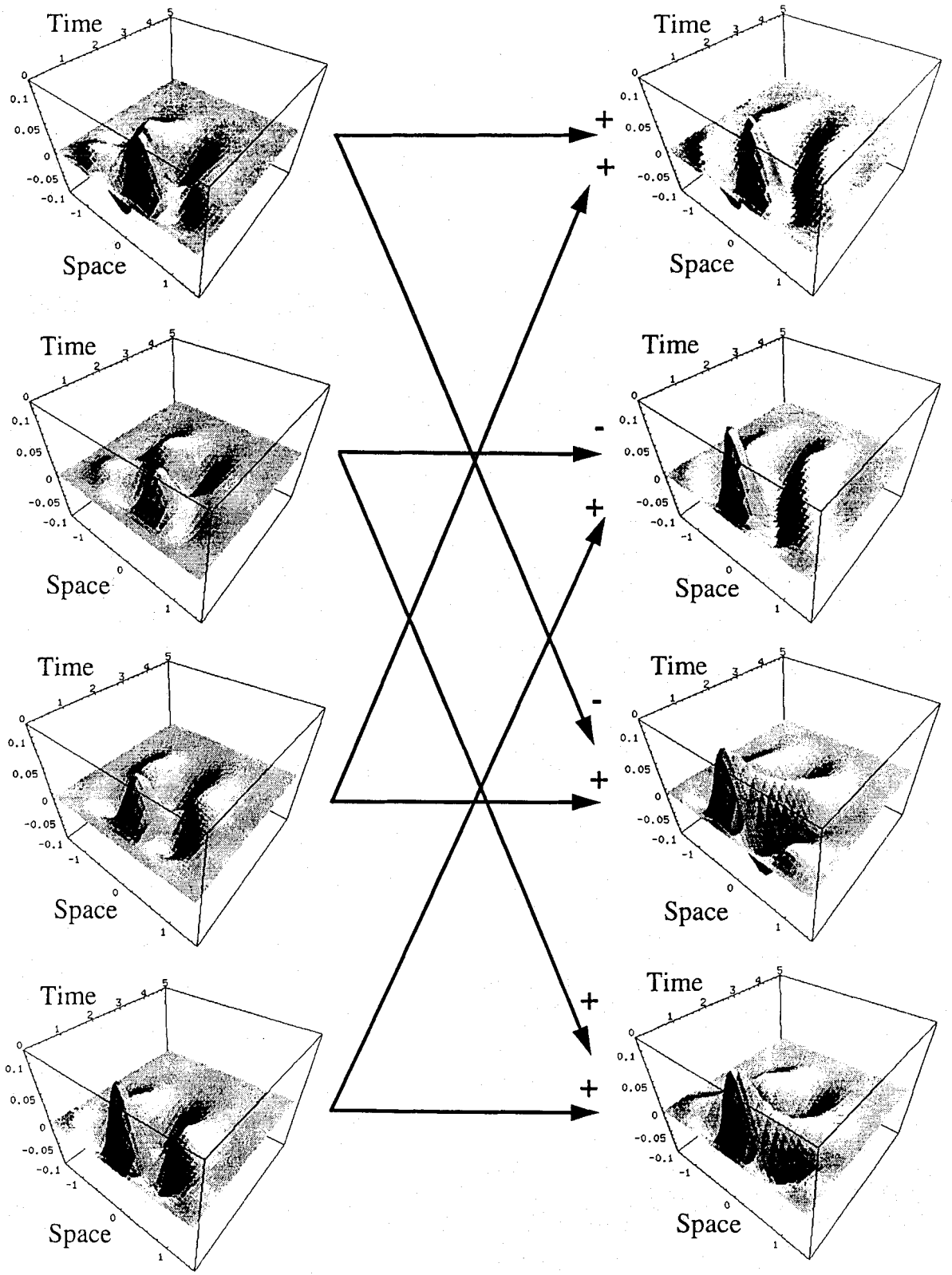

Fig. 3. Space–time-oriented filters used in the motion-energy stage of the model. The column at the left shows four filters constructed by multiplication of Eq. (3) or Eq. (4) by Eq. (5), with $k = 3$ and $k = 5$. The resulting filters are tuned in space and time but are oriented parallel to the space–time axes and are not directionally tuned. The four filters in the right-hand column are formed by addition of pairs of filters from the left-hand column, with the signs shown at the arrowheads. The resulting four filters are oriented in space–time. The top pair of filters at the right forms a nearly quadrature pair tuned to leftward motion, and the bottom pair at the right are the equivalent filters tuned to rightward motion. These are two of the quadrature pairs of filters used in the implementation of the model.

theless, it provides an important starting point for local-velocity estimation. All planes that obey Eq. (7) must pass through the origin, and different velocities correspond to planes with different angles. To measure the velocity of an object it is sufficient that one determine the angles of this plane.

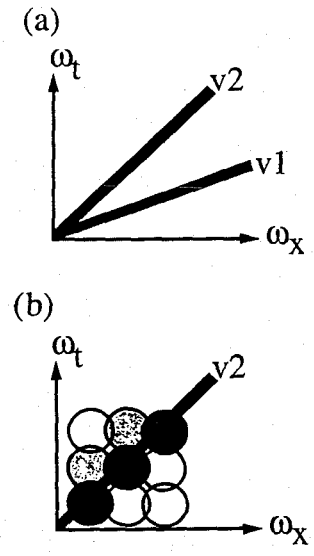

Fig. 4. Distributed representation of a moving bar by a population of spatiotemporal Gabor filters. The axes are spatial frequency (ω_x) and temporal frequency (ω_y). (a) The spatiotemporal power spectrum of a rigidly translating scene lies on a plane in frequency space. Only a single spatial dimension is shown in this figure, so the planes representing scenes translating at velocities $v1$ and $v2$ appear as the shaded bars. (b) A tiling of spatiotemporal frequency space with Gabor filters. Each circle represents a different Gabor filter, with the level of shading corresponding to the level of excitation of the filter. Filters whose centers are closest to the plane defined by $v2$ are the most strongly excited. The relative activations within this set of filters can be used to estimate the angle of the plane defined by $v2$ and hence the velocity of the bar.

The power spectrum of a single space–time-oriented Gabor filter is a pair of elliptical Gaussian windows in the spatial-frequency domain, where the width of the window is inversely related to the width of the Gabor filter in the space–time domain. Several motion-energy filters with different center frequencies can be used to tile the spatiotemporal frequency domain. Figure 4(b) shows one such tiling, where only a single spatial dimension and only one quadrant of the spatiotemporal frequency space are shown (so only one half of the power spectrum of each filter is shown). Each motion-energy filter samples the power spectrum of the scene in a small spatiotemporal band; to identify the slope of the power spectrum, one must compare the relative responses of the set of motion-energy filters (see Fig. 4).

Heeger[2] and Grzywacz and Yuille[3] proposed methods for extracting velocity measurements from a set of motion-energy filters by use of the filter responses to estimate the slope of the plane of maximal power in the frequency domain. The approach adopted in our model is closest to the "ridge" strategy proposed by Grzywacz and Yuille. In this scheme a set of velocity-tuned units is used for each image location. Each of these units receives a weighted input from a number of motion-energy units with different center frequencies:

$$I_k'(x, y) = \sum_{\omega_x,\omega_y,\omega_t} w_{k,\omega_x,\omega_y,\omega_t} \hat{E}_{\omega_x,\omega_y,\omega_t}(x, y), \qquad (8)$$

where I_k' is the total input to a unit tuned to velocity k, $\hat{E}(x, y)$ was defined in Eq. (6), and the weights are inversely proportional to the distance between the plane defined by velocity v_k and the center frequency of each motion-energy unit. I_k' will be largest for the velocity unit whose tuning most closely matches the plane of maximal motion energy. Grzywacz and Yuille[3] proposed a simple winner-take-all strategy to identify this unit.

Our model adapts the general structure for velocity estimation used by Grzywacz and Yuille but without defining in advance the weights from the motion-energy measurements to each velocity-tuned unit. These weights are instead determined by the optimization procedure. However, because we are using a structure similar to previous structures that were used to estimate velocity from motion energy, we have an *a priori* reason to believe that a set of weights can be found that will allow velocity to be estimated, at least locally, for rigid translational motions.

In our model the local-velocity pathway was organized into a grid of 8 × 8 receptive-field locations, with separate velocity estimates calculated for each grid location. At each location there was a pool of 33 velocity-tuned units (Fig. 5). These units represented motion in eight different directions, with four different speeds for each direction plus a central unit indicating no motion. The units were organized to form a log-polar representation of the local velocity at a given receptive-field location and were tuned to speeds of 0.25, 0.5, 1.0, and 2.0 pixels per frame in each direction.

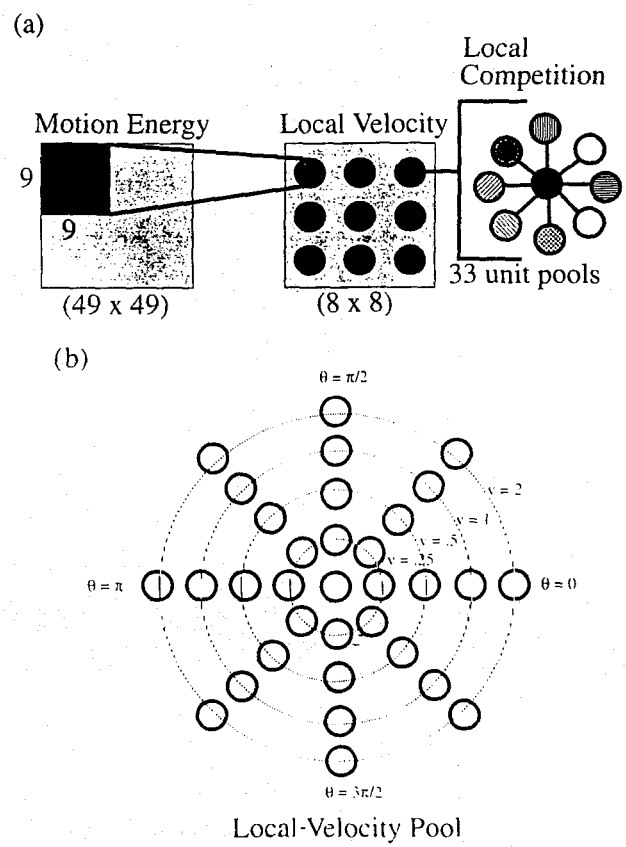

Fig. 5. Diagram showing the local-velocity pathway. (a) Local-velocity units are organized into a retinotopic grid of 8 × 8 receptive-field locations. Each local-velocity unit receives inputs from a 9 × 9 block of motion-energy measurements from overlapping regions of the image. At each location there is a pool of 33 velocity-tuned units, only 9 of which are shown. These units compete with one another locally, as described in the text. (b) A pool of 33 velocity units organized to form a log-polar representation of velocity. All eight units lying on a circle are tuned to the same speed but to different directions of motion.

Each unit received as input a weighted combination of the motion-energy responses from a local region of the input image [Eq. (8)]. The velocity units at each receptive-field location received inputs from a 9×9 region of the motion-energy layer, so the velocity grid formed a much coarser representation of the visual scene than did the motion-energy stage of the model. Velocity units received inputs from all motion-energy units, with directional preferences within $\pm 90°$ from the preferred direction of the velocity unit. On average, each velocity unit received ~ 2100 motion-energy measurements as input. The weights between the motion-energy units and the velocity-tuned units were trained to optimize a global measure of performance of the model, as is discussed in detail in Section 4 below. All the pools of velocity-tuned units shared a common set of weights, so in the final optimized model the velocity computations at all locations in the image were identical.

The velocity at a receptive-field location was represented by the relative strengths of the inputs to each velocity-tuned unit. In the model, this relative strength was encoded directly into the activity of each velocity-tuned unit by use of a soft-maximum nonlinearity[55]:

$$I_k(x, y) = \exp[I_k'(x, y)] \Big/ \sum_j \exp[I_j'(x, y)], \qquad (9)$$

where $I_k(x, y)$ is the final state of the unit representing velocity k and $I_k'(x, y)$ is the initial state of the unit. Note that the summation is performed over all the units in a velocity pool that share the same receptive-field location. The soft-maximum nonlinearity forces the total activity across a pool of velocity units to sum to 1. If one of the velocity units receives much stronger input than any of the others, its state will be very close to 1, whereas the states of all of the other units in the pool will be close to 0. In this case, the soft-maximum nonlinearity acts very much like winner-take-all competition. If, however, the motion-energy distribution is more ambiguous, activity may be distributed across several of the velocity units. This might occur, for example, in regions of transparent motion at which more than a single motion is present locally. The ability to represent ambiguous velocities is a major advantage for this type of distributed representation.[56]

The velocity computation in our model differed from the ridge strategy proposed by Grzywacz and Yuille[3] in its use of a soft-maximum nonlinearity rather than a winner-take-all strategy and in its use of motion-energy information from several adjacent image locations. The weights in the trained model were approximately inversely proportional to the distance between the plane defined by a velocity v_k and the center frequencies of the motion-energy units, as in the model of Grzywacz and Yuille, producing fairly tightly tuned velocity units. In addition, the use of a coarser representation of visual space at the velocity level is consistent with the smaller number of cells and larger receptive-field sizes found in MT compared with primary visual cortex. The soft-maximum nonlinearity has a statistical interpretation that we exploit in Section 3 below when deriving rules for adapting the parameters of the model. For now, the activity of each unit represents *evidence* for a particular velocity in an image region, and the soft-maximum nonlinearity enforces the constraint that the total evidence across all velocities in each image region sum to 1. This ensures that any local image region can provide strong support for only a single velocity.

D. Selection and Integration

The final output of our model consisted of a pool of 33 units, which represented velocity in the same manner as each pool of local-velocity units. If $I_k(x, y)$ is viewed as the local evidence for the kth velocity hypothesis v_k, and $S_k(x, y)$ is viewed as the amount of support for this hypothesis assigned to region (x, y), then the global evidence for this hypothesis is given by Eq. (1). This global evidence was encoded by the state of output unit k. The total amount of support for each output hypothesis was constrained to total 1 (as explained below). This constraint prevented the model from producing strong global evidence for a hypothesis without strong local evidence anywhere in an image.

The support assigned to each location for the different velocity hypotheses was computed by the units in the selection layer. The input to these units was computed as a weighted sum of the activity of a subset of the motion-energy units (similarly to the computation for the velocity units):

$$S_k'(x, y) = \sum_{\omega_x, \omega_y, \omega_t} w_{k, \omega_x, \omega_y, \omega_t} \hat{E}_{\omega_x, \omega_y, \omega_t}(x, y), \qquad (10)$$

where S_j' is the total input to a unit that computes the support for velocity hypothesis k and $\hat{E}(x, y)$ was defined in Eq. (6). The weights to these units were initialized randomly, and their final values are determined by the optimization procedure. Unlike in the case of the local-velocity units, for which we had reason to believe from other models that a solution should be found by this method, we had no *a priori* expectation of an appropriate weight structure for these units.

The selection units were organized into a grid of 8×8 receptive-field locations in one-to-one correspondence with the local-velocity pools (Fig. 6). There were 33 selection units at each receptive-field location, representing the local support for each of the different global-velocity hypotheses. The selection units were spatially organized into 33 layers of units, with each layer representing the region of support for a different velocity hypothesis. Two of these selection layers are shown in Fig. 6. The top layer represents the support for the output unit for upward motion (top lollipop), and states of units in this selection layer weight the states of upward-motion units in each local-velocity pool (also on top).

The constraint on the total amount of support for each hypothesis was enforced by use of global competition among all the units in each selection layer, which was implemented with a soft-maximum nonlinearity:

$$S_k(x, y) = \exp[S_k'(x, y)] \Big/ \sum_{x', y'} \exp[S_k'(x', y')], \qquad (11)$$

where $S_k'(x, y)$ is the net input to a selection unit in layer k [Eq. (10)] and $S_k(x, y)$ is the output state of that unit. Note that, unlike in the case of local-velocity competition, which occurred between units with different velocity tuning at the same location, the summation in this case was

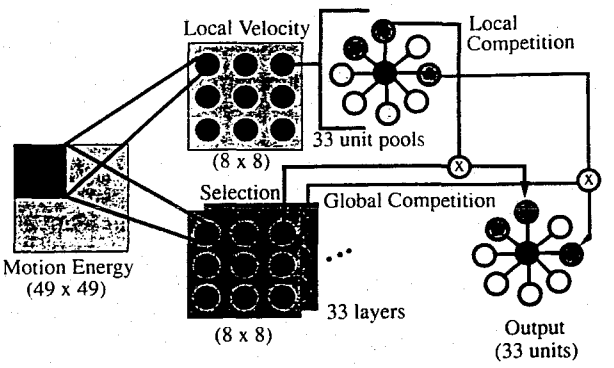

Fig. 6. Diagram showing the integration of the local velocity and selection-processing stages. The 49 × 49 pools of motion-energy filters feed the local-velocity and selection stages, which are integrated by the output units. An example of 9 of 33 local-velocity units in one local-velocity pool is shown. Different shadings for units in the integration and output pools correspond to different directions of motion. Only 2 of the 33 selection layers are shown, and the backgrounds of these layers are shaded to match their corresponding local-velocity and output units. Units in the output pool receive inputs that are the product of activity in the local-velocity and selection stages. The receptive field of an output unit is the entire 64 × 64 pixel input array.

performed for the same velocity tuning over all receptive-field locations. The constraint on the support field for each output unit, combined with the fact that the local-velocity unit states could never exceed 1, ensured that the states of the output units were constrained to be between 0 and 1 and could be interpreted as the global evidence within the image for each velocity, as stated above. The combination of global competition in the selection layers and local competition within the velocity pools means that the only way to produce strong evidence for a particular global velocity is for the corresponding selection network to focus all its support on regions that contain strong evidence for that velocity. The intersection of two different types of constraint is required for eligibility of a local-velocity estimate to contribute to the output layer.

The activity of units in the local-velocity stage of our model was compared with the known global velocities within the scene. Local-velocity estimates that were close (compared with the velocity estimates from other regions) to one of the global velocities were assigned responsibility for that velocity. Regions that were assigned a lot of responsibility for a global velocity received an error signal that caused the velocity units in that region to adjust their weights to predict the global velocity better from their current input. At the same time the selection units were optimized to predict the responsibility that should be assigned to each local region. This procedure forced the selection units to learn to predict what pattern of local motion-energy features permitted good local-velocity predictions..

During optimization the selection units became sensitive to regions that contained sufficient information to disambiguate the direction and speed of true motion. In earlier experiments with a one-dimensional version of the current model the weight patterns of selection units corresponded to edge detectors in a one-dimensional motion-energy distribution. Selection units in the trained two-dimensional model were also primarily detectors of motion discontinuities. They responded maximally in regions in which the distribution of motion-energy measurements corresponded to a border between different velocities; however, this tuning was restricted to a certain range of orientations and speeds. As a result, selection units responded strongly only to motion discontinuities that were consistent with motion at a particular speed and direction (Fig. 7). For example, selection units were usually sensitive to regions containing motion energy at several directions, since velocity predictions from these regions were less likely to suffer from the aperture problem [Fig. 7(b), example 2]. However, some patterns containing several directions of motion, such as those

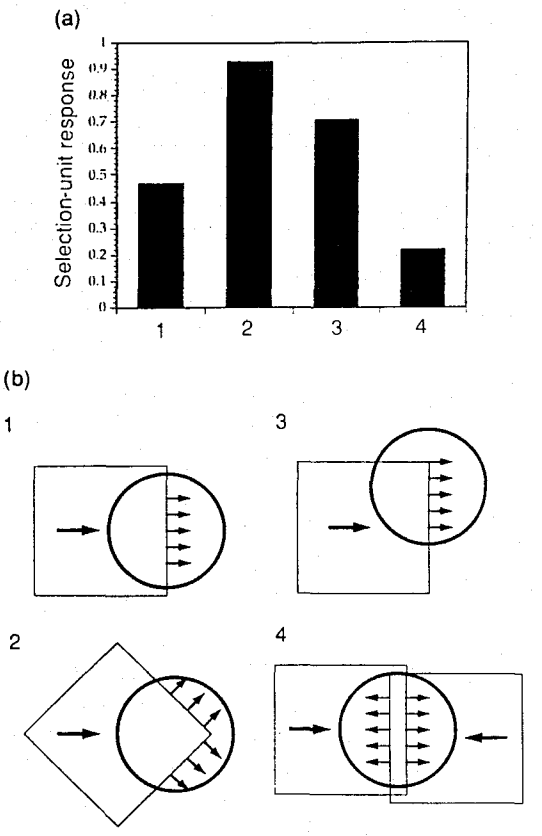

Fig. 7. Responses of a selection unit to motion discontinuities. (a) The responses of a selection unit tuned to rightward motion to four different distributions of motion-energy inputs. (b) Motion-energy distributions for the four responses are shown schematically. In examples 1–3 a square is moving in the direction indicated by the large arrow against a stationary background. In the last example, two squares, one semitransparent, move against a stationary background. In each example the circle indicates the receptive field of the selection unit, and the small arrows indicate the direction of local motion reported by the motion-energy units within this receptive field. In example 1 the receptive field is centered over an edge moving to the right and sees a uniform distribution of rightward motion-energy responses producing a moderate response (0.48) from the unit. In contrast, the response to a similar motion in example 3 is much stronger (0.74), because in this case the receptive field covers a corner region that contains a discontinuity between a region of rightward motion-energy response and a region containing no motion-energy response. The strongest response from this selection unit (0.93) is from example 2, in which the receptive field encloses a discontinuity between two orthogonal sets of local measurements. Finally, in example 4, local responses corresponding to two opposed motions (generated in this case by one transparent object moving in front of a second moving object) suppress the response of the selection unit. In this diagram the local response of a single selection unit was presented in isolation. The overall response of a selection unit is also determined by the responses of similarly tuned selection units in other image regions through competitive normalization.

produced by two objects moving in opposing directions, will suppress the selection units [Fig. 7(b), example 4]. Visual stimuli constructed so that there is a local balance between motion in opposing directions tend to cancel and are not seen as transparent.[57]

3. OBJECTIVE FUNCTION AND TRAINING

A. Mathematical Model

There are several assumptions that the statistical model embodies. First, motion in the visual scene is due to some finite number of motion processes, each of which may or may not be present in any particular scene. We are using the term process here in its statistical sense: a process is a means of generating a set of observations (in this case velocity measurements). A single process may correspond to a single moving object or to several objects all moving at the same velocity. If one of these processes is present in the scene, it may occupy not all regions of a scene but only some subset. *A priori* we do not know which subset of a scene any process will occupy, but we make the assumption that every region of the scene contains at least one motion process. One of the motion processes can be the null, or no-motion, process. We consider first a simple model that captures these constraints and gradually refine it.

Let V_k be a binary random variable that takes on the value 1 if the kth-velocity process is present in an image and has the value 0 otherwise. We assume that the presence of a particular velocity process is independent of the presence of any other process. This is a reasonable assumption if velocity processes correspond to different objects moving independently in the scene. We divide the visual scene into a finite number of regions indexed by (x, y) and define the set of regions in which process k is present as Support (V_k). Support (V_k) may be an empty set. We now define the probability that velocity process V_k is present in a visual scene as

$$P(V_k = 1) = \sum_{(x,y)} P[(x, y) \in \text{Support}(V_k)] \times P[V_k = 1 \mid v(x, y)], \quad (12)$$

where $P[V_k = 1 \mid v(x, y)]$ is the probability that process k is present, given the local velocity measured in region (x, y). Equation (12) simply states that we compute the probability that process k is present by summing the evidence for that process from all regions of the image, where the evidence is weighted by the probability that process k is present in the region. A major difficulty with the simple model just described is that each V_k can take on only the value 0 or 1, so the model can represent only a discrete set of velocities. It would be preferable to represent a continuous range of velocities. One way to accomplish this is to assume that each output unit actually represents a Gaussian bump centered at a velocity $v_0(k)$. The bump can easily be normalized so that a velocity exactly equal to $v_0(k)$ produces a maximal activity of 1, with activity smoothly decreasing as the Euclidean distance between the velocity represented and $v_0(k)$ increases. The state of a single unit of this sort cannot uniquely represent velocity [any state less than 1 could correspond to any velocity lying on a circle centered at $v_0(k)$]; however, the states of a set of such units form a distributed representation of a range of velocities (Fig. 8). The same representation is used for each pool of velocity units and for the final output of the model.

For this continuous representation it makes more sense to think of V as a continuously valued variable, and we are interested in $P[V = v_0(k)]$ or, more precisely, $P(v_0 - \epsilon \le V \le v_0 + \epsilon)$. For notational convenience we use the form $P[V = v_0(k)]$. Equation (12) then becomes

$$P[V = v_0(k)] = \sum_{(x,y)} P\{(x, y) \in \text{Support}[v_0(k)]\} \times P[V = v_0(k) \mid v(x, y)]. \quad (13)$$

In order for Eq. (13) to be a proper probability statement, some additional conditions must be met. First,

$$\int_V P[V = v_0(k)]\mathrm{d}V = 1,$$

and $P[V = v_0(k)] \ge 0$ everywhere. In addition, we would like to enforce the condition that at least one velocity process occur in each region of the image, by letting

$$\sum_k P[V = v_0(k) \mid v(x, y)] = 1, \quad (14)$$

where we are using the shorthand notation for $P[V = v_0(k)]$ defined above.

In order to enforce these constraints we define the distributed representation for a velocity $\hat{v}$ as

$$a_k = \frac{\exp[\gamma \|\hat{v} - v_0(k)\|^2]}{\sum_j \exp[\gamma \|\hat{v} - v_0(j)\|^2]}, \quad (15)$$

where a_k is the activity of unit k (in either the set of global-output units or one of the local-velocity pools). We accommodated the presence of multiple velocity processes globally by additively combining the distributed representation of each process, clipping output unit values at 1. This additivity means that there is a limit on how similar

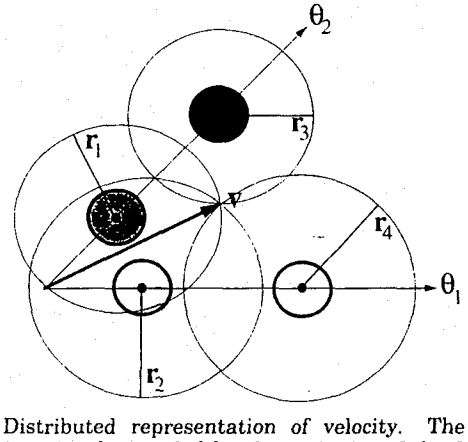

Fig. 8. Distributed representation of velocity. The velocity vector v is uniquely encoded by the activities of the four units represented by the small shaded circles. The center of each circle is located at the direction and speed of motion that produces maximum response from the corresponding unit. The activity of each unit decreases with distance from the center of the unit as $\exp(-r_i^2)$, and this activity is shown by the density of shading in each unit. The darker the shading, the more active a unit, which in this case implies that $r_3 < r_1 < r_2 < r_4$. The larger circles around these four units intersect at only one location, the tip of the velocity vector v.

the velocities of two processes can be with the two processes still being distinguishable from a single process at the average velocity of the two.

Once these local probability constraints have been enforced, Eq. (13) is valid as long as

$$\sum_{(x,y)} P\{(x, y) \in \text{Support}[v_0(k)]\} = 1. \quad (16)$$

This statement requires that the support for each global-velocity process be itself a probability distribution.

Let $I_k(x, y)$ denote the state of local-velocity unit k at location (x, y) and v_k denote the continuous-valued state of output unit k. The presence of a particular global velocity $\hat{v}$ corresponds to a particular set of values for each of the output units, and, similarly, the presence of the same velocity in a particular region of the image corresponds to a set of values for all the units in a particular local-velocity pool. We are interested in $P[V = \hat{v} \mid v(x, y)]$, but we would like to express it in terms of $I_k(x, y)$ and v_k. We accomplish this by adopting a Gaussian error model and comparing the local and the global values of corresponding units:

$$p(v_k \mid (x, y) \in \text{Support}(v_k), ME(x, y), w_I) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{-[v_k - I_k(x, y)]^2/2\sigma^2\}, \quad (17)$$

where v_k is assumed to be the value that a global unit takes on under the distributed representation of $\hat{v}$. This scheme is easily extended to account for the presence of several different global motions, because it merely corresponds to a different set of values in the distributed output representation.

If we define

$$S_k(x, y) \doteq P[(x, y) \in \text{Support}(v_k)], \quad (18)$$

then each layer of selection units defines a conditional probability distribution. We can combine this distribution with the distributions already defined for $I_k(x, y)$. These probability distributions depend on both the weights in the selection and velocity pathways and the motion-energy measurements in each region of the visual scene. If we let $ME(x, y)$ denote the set of motion-energy measurements from region (x, y) of the image and ME denote the set of motion-energy measurements from all regions of the image, we can restate Eq. (13) more formally as

$$P(V = \hat{v} \mid w_I, w_S, \text{ME}) = \prod_k \sum_{(x,y)} P[v_k \mid w_I, ME(x, y), (x, y) \in \text{Support}(v_k)] \times P[(x, y) \in \text{Support}(v_k) \mid w_S, ME], \quad (19)$$

where w_I denotes the set of adjustable weights in the local-velocity pathway, w_S is the set of selection weights, and v_k are the output unit states used to represent $\hat{v}$. Since we have a fixed number of output units, Eq. (19) is valid whether we have just one motion process or a set of motion processes present, except that the values v_k change.

The model was trained on sequences of frames from visual scenes for which the global motions of processes present in the scene were known. Let $ME_{c,t}$ denote the set of motion-energy measurements associated with frame t of sequence c, and let $\mathbf{v}_{c,t}$ denote the set of motion processes present in this frame. We seek a set of parameters so that the model performs well on the training data and generalizes to other scenes. A common method is to maximize the conditional log likelihood of our training data. If we assume that scenes in the training set are independent of one another and that the frames within each sequence are also independent, then the conditional likelihood of the training data is

$$L = \prod_c \prod_t P(\mathbf{v}_{c,t} \mid ME_{c,t}, w_I, w_S), \quad (20)$$

where $P(\mathbf{v}_{c,t} \mid ME_{c,t}, w_I, w_S)$ is defined as in Eq. (19). We search for a set of parameters w_I, w_S that maximize this expression (or, equivalently, the log of this expression). The second assumption is clearly not valid, since sequential frames are, in fact, highly correlated; however, the main concern is with the conditional independence of $P(\mathbf{v}_{c,t} \mid ME_{c,t})$ and $P(\mathbf{v}_{c,t+1} \mid ME_{c,t+1})$. Since $ME_{c,t}$ is a function of several preceding frames as well as of the current frame, this conditional-independence assumption is more reasonable. Technically, we are making a Markov-order assumption and assuming that $P(\mathbf{v}_{c,t} \mid ME_{c,t}, \mathbf{v}_{c,t-1}) = P(\mathbf{v}_{c,t} \mid ME_{c,t})$, in other words, that knowing $\mathbf{v}_{c,t-1}$ gives us no more information than knowing $ME_{c,t}$.

Combining results from Eqs. (13) and (18)–(20), we need to maximize the expression

$$\log L = \sum_c \sum_t \sum_k \log\Big(\sum_{x,y} S_k(x, y) K_\sigma \times \exp\{-[v_k - I_k(x, y)]^2/2\sigma^2\}\Big), \quad (21)$$

where K_σ is the Gaussian normalization constant. We adjusted the weights of the local-velocity and selection pathways by using a gradient-ascent procedure to find a local maximum of this likelihood function. The constant σ was set to the value 0.25 for all simulations. This value was a compromise based on the spacings between the velocity tuning of the units in the log-polar velocity representation.

Some insight into the nature of this objective function can be gained by consideration of its gradient with respect to the net input to units in the local-velocity [Eq. (8)] and selection [Eq. (10)] pathways. Define the weighted evidence for output v_k from region (x, y) as

$$\mathcal{E}[v_k \mid (x, y)] = p[v_k \mid (x, y) \in \text{Support}(v_k), ME(x, y), w_I] \times P[(x, y) \in \text{Support}(v_k) \mid ME, w_S]. \quad (22)$$

The total evidence for v_k from all regions of the image is then

$$\mathcal{E}[v_k] = \sum_{(x,y)} \mathcal{E}[v_k \mid (x, y)], \quad (23)$$

and the relative evidence for v_k from region (x, y) is

$$\mathcal{R}[(x, y) \mid v_k] = \frac{\mathcal{E}[v_k \mid (x, y)]}{\mathcal{E}[v_k]}. \quad (24)$$

The weighted evidence depends on both how well the local-velocity prediction from region (x, y) matches v_k

and how much support for v_k is assigned to region (x, y). If we assume that initially support is assigned uniformly over the entire image, this weighted evidence is dominated by the match between v_k and the local-velocity prediction. The total evidence is simply the sum of the weighted evidence from all regions and is equivalent to the expression in Eq. (19) under our continuous-probability model. The relative evidence is simply the proportion of the total evidence coming from region (x, y), and this relative evidence plays a prominent role in the gradient of our objective function. If support is initially assigned uniformly over the image, this relative evidence is a measure of how good the prediction for v_k from region (x, y) is in comparison with the predictions from all other regions of the image.

The derivative of the log likelihood with respect to the input to a unit in the selection pathway is

$$\frac{\partial \log L}{\partial S_k'(x, y)} = \sum_c \sum_t \{\mathcal{R}[(x, y) \mid v_k] - S_k(x, y)\}. \quad (25)$$

This derivative becomes zero only when $S_k(x, y)$ exactly matches the relative evidence for v_k from region (x, y):

$$\mathcal{R}[(x, y) \mid v_k] = S_k(x, y).$$

This means that during training the selection-pathway weights are adjusted so that support is assigned to the regions that relative to other regions provide the best evidence for each candidate velocity.

The derivative of the log likelihood with respect to the input to the local-velocity units has a slightly more complex form:

$$\frac{\partial \log L}{\partial I_k'(x, y)} = \sum_c \sum_t \frac{1}{\sigma} I_k(x, y) \Big\{ \mathcal{R}[(x, y) \mid v_k][v_k - I_k(x, y)] - \sum_j \mathcal{R}[(x, y) \mid v_j][v_j - I_j(x, y)] I_j(x, y) \Big\}. \quad (26)$$

The term inside the braces on the first line of Eq. (26) is zero when $I_k(x, y)$ matches v_k. This is expected, since $I_k(x, y)$ was intended to be a local prediction for v_k. Note, however, that this error term is weighted by the relative evidence, so $I_k(x, y)$ is forced to predict v_k only if it already is a better predictor for v_k than predictions from other regions of the image. The term on the second line of Eq. (26) appears because of the constraint in Eq. (14). This term is essentially the same error term that appears inside the brace on the first line of Eq. (26), but summed over all the other candidate velocities in the local-velocity pool, and it serves to balance the error from all the local candidate velocities in proportion to the relative evidence for each candidate in this local region. The first term on the right-hand side, $I_k(x, y)$, scales the rest of the expression down when $I_k(x, y)$ is close to zero (meaning that there is little evidence from the motion-energy measurements for the velocity that this unit represents).

B. Architecture and Training

The model described above corresponds to regions of visual cortex responsible for processing moving stimuli in a small patch of the visual field. There were 86,436 units in the motion-energy stage of the model and a further 4,224 units in the local-velocity and selection stages. Only the weights in the local-velocity and selection pathways of the model, which included approximately 8.8×10^6 adjustable parameters, were adaptive. However, as discussed in Section 2, units at each receptive-field location in both the local-velocity and the selection pathways share common sets of weights, so the total number of free parameters in the model was reduced to 138,600.

The system was trained by use of 500 image sequences containing 64 frames each (giving a total of 32,000 training cases). We used a conjugate gradient optimization procedure to adjust the weights in both the selection and the local-velocity pathways to find a (local) minimum of Eq. (21). We generated the training image sequences by randomly selecting one or two visual targets for each sequence and moving these targets through randomly selected trajectories (see Section 4 for some examples). The targets were rectangular patches that varied in size, texture, and intensity. Several examples of training inputs are shown in Fig. 9. The training set contained examples of both transparent [Figs. 9(b) and 9(e)] and occluding [Figs. 9(d), 9(f), and 9(g)] interactions between pairs of objects.

Target widths and heights ranged from 4 to 20 pixels, and targets contained randomly oriented textures with spatial-frequency content from 0.05 to 0.3 cycle/pixel and 128 different gray levels. We formed textures by combining square-wave gratings of varying phase and frequency, and targets typically contained multiple spatial frequencies. All the motion trajectories began with the objects stationary, and then one or both objects were rapidly accelerated to constant velocities that were maintained for the remainder of the trajectory. Targets moved in one of eight possible directions, at speeds ranging from 0 to 1.75 pixels/frame. In training sequences containing multiple targets, the targets were permitted to overlap (targets were assigned to different depth planes at random), and the upper target was treated as opaque in some cases and as partially transparent in other cases. We selected target sizes, textures, and velocities at random, using a uniform distribution for all parameters and selection without replacement for multiple targets. The initial positions of the targets were also selected randomly but were constrained to lie within the central two thirds of the input window. The system was trained until the response of the system on the training sequences deviated by less than 1% on average from the desired response.

During both training and testing the evaluation of the model was divided into two phases. In the first phase the motion-energy filters were convolved with a sequence of images, and in the second phase the outputs of the selection and local-velocity units and the final output of the model were computed. Because the motion-energy stage of the model was not adaptive, only the second phase of evaluation had to be performed repeatedly during optimization of the model parameters. All simulations were performed with use of 128 nodes of an Intel ipsc hypercube operating in parallel. Computation of the Gaussian pyramid and of all 36 motion-energy convolutions required $\sim$ 19.2 s for a sequence of 64 images (0.3 s/image). A set of 500 sequences of 64 frames would require 160 min to process. Evaluation of the local-

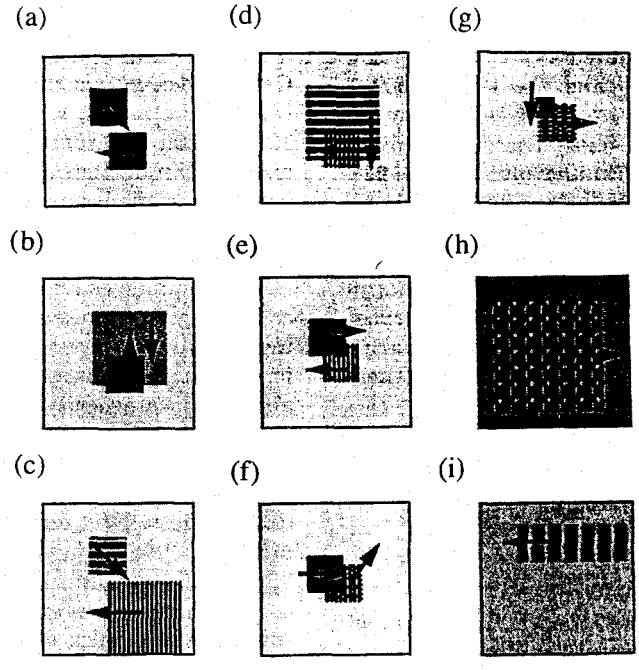

Fig. 9. Nine examples of images from the set of image sequences used to train the model. In each example the arrows indicate the direction of movement of the object in the scene, and the length of the arrows represents the object's speed. (b), (e), (g), Examples of transparent interactions between two objects, in which the lower object is partially visible through the upper object. These transparent interactions are computed on the basis of the transmittance and reflectivity of the objects in the scene. (d), (f), Examples of nontransparent interactions, in which the nearer object completely occludes the object behind it. These examples span the range of object sizes and relative contrasts used in the training set.

velocity and selection pathways on the same 64-image sequence required ~ 32 s, and the combined forward evaluation and computation of the gradient for all parameters during the learning phase required 80 s per 64-frame sequence. One pass through a set of 500 image sequences during training required slightly over 11 h, and convergence of the optimization procedure required 20 passes through the training set. The training of the entire model took approximately 10 days.

The performance of the trained system was tested with a separate set of 50 test-image sequences. These sequences contained 10 novel visual targets with different random combinations of width, height, texture, intensity, and trajectories generated in the same manner as for the training sequences. The responses on this test set remained within 2.5% of the desired response, with the largest errors occurring at the highest velocities. Several of these test sequences were designed so that targets contained edges oriented obliquely to the direction of motion, demonstrating the ability of the model to deal with aspects of the aperture problem. In addition, only small, transient increases in error were observed when two moving objects intersected, whether these objects were opaque or partially transparent.

The log-polar representation of velocity at both the output and the local-velocity stages of the model may account for the fact that larger errors were observed at higher velocities. Because the nominal velocities represented were spaced farther apart for units representing higher velocities, the same absolute difference in activity level translated into a larger difference in represented velocity. The error metric used in training did not take into account the effect of the log-polar velocity representation.

4. RESULTS

We show first, by considering several examples of image sequences, how the local-velocity estimation and selection pathways work in parallel. The first three examples were part of the test sequence described in Section 3 and are qualitatively similar to the sequences used for training the model. In the figures for these examples, (Figs. 10–12) we show one frame of an input to the model and a representation of the local-velocity estimates and how these estimates are grouped by the selection pathway. Local-velocity estimates are shown in these figures by arrows, with the direction of motion indicated by the direction of the arrow and the speed by the length of the arrow. Velocity estimates that fall below the detection threshold (see below) are indicated by short horizontal lines with no arrowheads. Selected velocity estimates are surrounded by dashed or solid boxes. The activity of the later stages would be based not on just this one input frame but also on several preceding frames. For all examples, velocities of objects will be given in polar coordinates, with 0° defined as rightward motion and angles measured in a counterclockwise direction.

The use of the distributed representation of velocity in both the output and the local-velocity stages of the model means that one must be careful in interpreting the output of the model. The ideal response to a motion of 0.5 pixel/frame to the right is an activity of 1.0 in the single unit representing this velocity, and the ideal response to a motion of 0.75 pixel/frame to the right corresponds to an activity of 0.37 in two units (representing velocities of 0.5 and 1.0 pixel/frame to the right). The output representation is designed so that a single veloc-

ity corresponds to activity in at most four units simultaneously, and these units would correspond to nearest neighbors according to Fig. 5(b). When there is more than a single motion in a scene, the ideal response is computed by linear superposition. Thus a scene containing motions of 0.5 pixel/frame upward and 0.5 pixel/frame to the right has an ideal response with an activity of 1.0 in two different units.

In order to determine whether a particular motion was seen by the model, we compared the output of the model with the ideal response for some particular velocity. To do this efficiently we used a simple thresholding procedure, in which units had to be within 75% of their ideal values to be considered active. We could quickly determine the minimum threshold that is consistent with any ideal responses in which a unit could participate, by examining the activations of a unit and all its nearest neighbors. Units that fell below this threshold were immediately discarded from further consideration, and units that were above threshold were then interpreted in terms of superpositions of ideal responses.

A. Single Objects

Consider first a square object of uniform intensity moving at 0.5 pixel/frame with a heading of 45° [denoted by (0.5, 45°)]. Figure 10(a) shows the input stimulus, and the grid superimposed upon this stimulus indicates the coarseness of the representation of this space at the local-velocity and selection stages of the model. Figure 10(b) shows the local-velocity estimates that were above threshold for this input. Each velocity estimate is represented by an arrow indicating the magnitude and direction of the estimate. In addition, dashed boxes enclose the estimates that make the main contribution to the calculation of overall velocity of the object. The contribution of each velocity estimate to the overall velocity is the product of the activity of that velocity unit and the activity of the corresponding selection unit, and both of these must be large for an estimate to contribute to the overall output of the model. Only a single global motion is present in this scene, and only the support for this motion is illustrated.

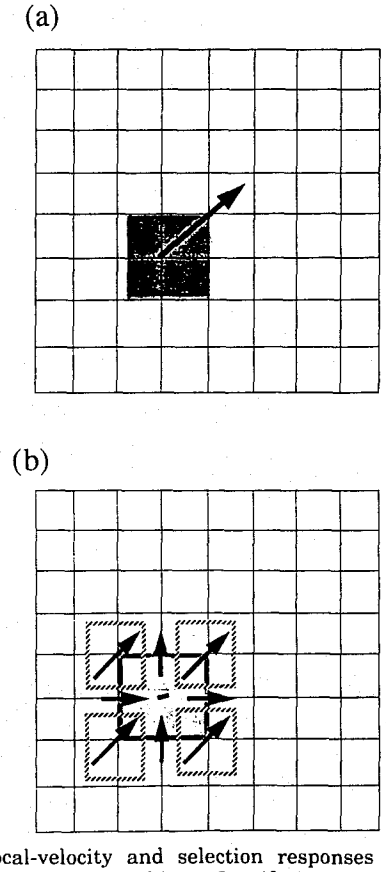

Fig. 10. Local-velocity and selection responses to a moving square. (a) The square object of uniform intensity moves at 0.5 pixel/frame at a heading of 45°, as indicated by the arrow. Intersections of grid lines correspond to locations of centers of local-velocity and selection-receptive fields. (b) Representation of the outputs of the local-velocity and selection stages of the model for the input shown in (a). The arrows at grid points represent the outputs of the local-velocity pool at the respective receptive-field locations. Regions enclosed by dashed lines correspond to local-velocity measurements selected for integration at the output stage.

This simple example reveals some important aspects of processing in the model. The central region of the object contains no contrast variations, so, even though this region is moving along with the rest of the object, velocity estimates centered in this region show no velocity. Similarly, the velocity estimates near the center of the object's edges see only contrast edges of a single orientation. As a result, these estimates can represent only the component of motion that is orthogonal to these edges. The only local-velocity estimates that reflect the true motion of the object are those estimates from regions near the four corners of the object. In these regions the presence of edges at two orientations is sufficient to disambiguate the true velocity. By examining the patterns of activity in the local-velocity and selection networks, we can see how these regions were selected.

Consider the calculation of global evidence for three idealized hypotheses—(0.35, 0°), (0.5, 45°), and (0.35, 90°)—corresponding to the true motion in the scene and the rightward and upward components of that motion. For convenience, consider the contributions of the velocity and selection units from only the nine receptive-field locations labeled with arrows in Fig. 10(b), and assume that units explicitly representing these particular velocity hypotheses exist in the model (the activities that we report for these idealized units are interpolated from real unit activities in the model). For the receptive-field locations corresponding to the four corners of the object, the velocity unit representing (0.5, 45°) had activity of 0.9, and the units representing (0.35, 0°) and (0.35, 90°) each had activity 0.05. (Values have been rounded to two significant digits for convenience.) For the receptive-field locations in the centers of the two vertical edges, the velocity unit representing (0.35, 0°) had activity 0.95, and units for (0.35, 90°) and (0.5, 45°) were both essentially at zero activity. Similarly, at the centers of the two horizontal edges, the velocity unit representing (0.35, 90°) had activity 0.95, and the velocity units for the other hypotheses had practically no activity. The velocity units of all three idealized hypotheses in the central receptive-field location had zero activity; this region will not be considered further, as it can make no contribution to the global activity.

Now consider the selection-unit activities for the hypothesis (0.5, 45°). The four corner receptive-field locations had selection activity of 0.25, and all other locations had selection activity of zero. The global evidence for

(0.5, 45°) is simply the sum of the velocity-unit responses from the four corners, each weighted equally:

$$v_{(0.5,45°)} = 4 \times 0.25 \times 0.9 = 0.9.$$

The ideal response for (0.5, 0°) corresponds to an activation of 1.0 in the appropriate unit, so the response for this hypothesis is well above its threshold value of 0.75. The selection activity for the hypothesis (0.35, 0°) was distributed across six receptive-field locations: the four corner regions and the middles of the two vertical edges. The selection activity in each of the corner locations was 0.2, and in the middle of the edges it was 0.1. This bias in favor of the corner regions was due to the sensitivity of selection units to velocity discontinuities that was discussed above. This is a sensible bias, and, if the motion in the scene really were to the right, the corner regions would provide unambiguous evidence of this motion. As a result of this distribution of selection activities, quite a bit of the support for the hypothesis (0.35, 0°) was concentrated in regions in which there was little local evidence for this velocity, and as a result the global evidence became quite weak:

$$v_{(0.35,0°)} = 2 \times 0.1 \times 0.95 + 4 \times 0.2 \times 0.05 = 0.23.$$

This global evidence actually maps into activity levels of 0.15 for the unit representing (0.25, 0°) and 0.13 for the unit representing (0.25, 0°). The corresponding thresholds for these two units are 0.5 and 0.4, respectively, so both of these units had activity well below threshold for this hypothesis. The calculation of support for the hypothesis (0.35, 90°) was nearly identical, with the centers of the horizontal edges replacing the centers of the vertical edges in contributing to the global evidence of 0.23.

It is apparent in this first example that the computation of global velocity even in the simplest case, which is dominated by velocity estimates from the corners of the object, is the result of fairly complex interaction between velocity and selection-unit activities. The weighted combination of velocity estimates provides a good overall estimate of the true velocity of the object. The activity of the selection units indirectly segments the scene into a region of coherent motion; however, it is important to note that the region is disjoint in this case and does not correspond to the entire object. The segmentation would be the same if the original scene contained only the corners of the object rather than the entire object. For the object to be perceived as a single object, some other process would be required to fill in the region corresponding to the object (much the way a Kanisza triangle can be filled in on the basis only of the presence of its corners). This other process could use texture, color, or intensity cues to fill in this surface, and we make no attempt to model such a process.[58]

The second example was also a single object moving at (0.5, 45°) (Fig. 11). In this example, however, the object did not have uniform intensity but was instead filled with a pseudorandom texture. As a result, all regions of the object contained both local contrast variations and edges at several different orientations. The local-velocity estimates from all regions of the object were largely in agreement in this case: the velocity unit with maximal response in all 16 receptive-field locations enclosed by the dashed box in Fig. 11(b) corresponded to (0.5, 45°), with the activity in this unit ranging between 0.68 and 0.92. The distribution of selection-unit activities for the velocity hypothesis (0.5, 45°) was also much more uniform over these 16 receptive-field locations, since most receptive-field locations contained local regions of motion discontinuity that were due to the textured pattern within the receptive field. There was still a slight bias in favor of the true corners of the object, which received selection-unit activities near 0.08, with interior regions of the object having selection activities ranging between 0.05 and 0.06. The global evidence for the velocity (0.5, 45°) in this case was 0.93, and the evidence for the two component velocities (0.35, 0°) and (0.35, 90°) was 0.25 and 0.2, respectively. Thus, although the pattern of selection for this example was quite different from that of the first example, the final global evidences were quite similar. In the second example the segmentation performed by the model corresponded well with what would be identified as a single moving object.

The velocity estimates for the object were similar in the first two examples, yet the output of the selection stage of the model was quite different. In the first example the region of support for the object velocity was concentrated over a quite-small portion of the object, and a large weight was assigned to each selected velocity measurement. In the second example the region of support encompassed the entire object, with much less weight assigned to any one velocity measurement. Although it is intuitively more

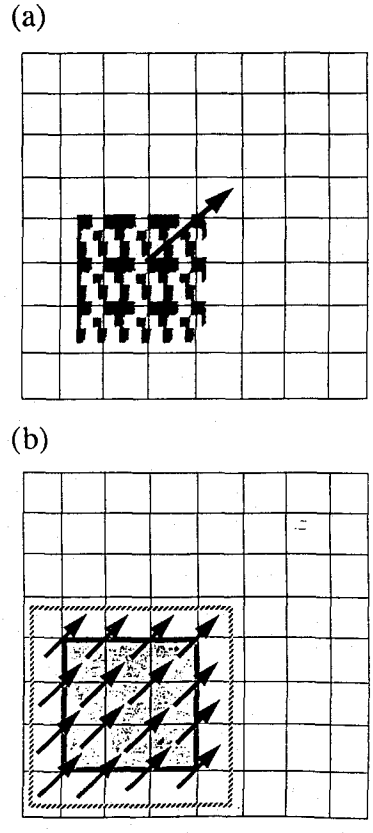

Fig. 11. Local-velocity and selection responses to a moving textured object. (a) The input to the model is a square object with a pseudorandom intensity moving at 0.5 pixel/frame at a heading of 45°, as indicated by the arrow. (b) The outputs of all local-velocity pools are in close agreement, and the entire object region is selected (indicated by the dashed box).

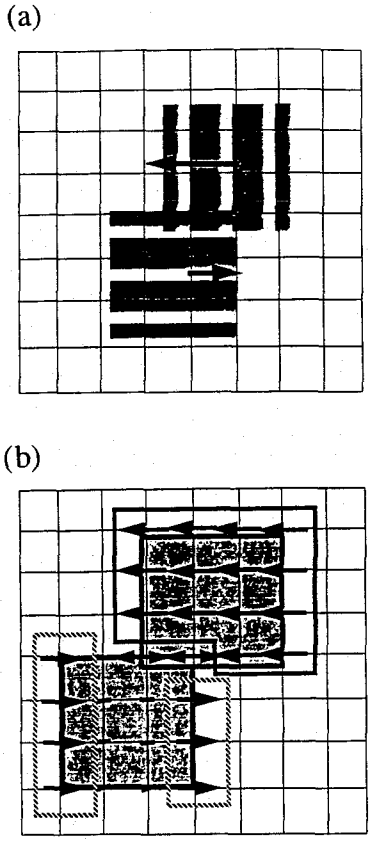

Fig. 12. Local-velocity and selection responses to partially occluding objects. (a) The input to the model consists of two objects moving in opposite directions (one to the right at 0.25 pixel/frame and the second to the left at 0.5 pixel/frame). Each arrow indicates the direction and speed of motion of the object. (b) The dashed outlines indicate the region of motion selected for the rightward-moving object, and the solid outline indicates a separate region of support for the motion of the leftward-moving object. The local-velocity estimates in the region of overlap are ambiguous, showing two directions of motion simultaneously.

appealing, the segmentation in the second example was no better than that in the first for estimating the velocity of the object. It could be argued that, because the region of support for the second object was spread over more measurements, it was more tolerant to noise in individual measurements. However, noise is a problem only if it is injected at the velocity-estimation stage. Noise injected into the image at earlier stages produces local contrast variations in the image. These contrast variations serve much the same role as the texture in the second example, causing the region of support to spread over more of the portion of the image covered by the object. As local-velocity estimates become noisier, there is generally a corresponding spread in the region of support, which reduces sensitivity to the noise.

B. Occluding Objects

The third example included two objects moving in opposite directions (Fig. 12). One object was striped horizontally and moved at a velocity of (0.25, 0°). The second was striped vertically and moved at (0.5, 180°). The leftward-moving object was on top of the rightward-moving object (i.e., closer to the observer) and was composed of a material that reflects but does not transmit light. As a result, in regions where the two objects overlapped, the second object occluded the first object totally, and the intensity in these regions was the intensity of the second object. The local-velocity estimates and the regions contributing strongly to the global velocity are shown in Fig. 12(b). Regions contributing to rightward motion are denoted with dashed outlines, and those contributing to leftward motion are denoted with solid outlines. (Recall that separate regions of support are computed for different candidate global velocities.)

In this example the region of overlap of the two objects is marked with double arrow heads to indicate that the activity in the local-velocity pool was not concentrated in a single unit but was distributed bimodally, with activity peaks corresponding to two opposing directions of motion. For example, the velocity pool of the receptive field at the top-right corner of the lower object has a net activity of 0.3 for velocity (0.25, 0°) and of 0.4 for velocity (0.5, 180°). In addition, the activities of the selection units for both (0.25, 0°) and (0.5, 180°) were relatively weak (0.02 and 0.013, respectively). As a result, this ambiguous motion region made little contribution to the overall evidence for either motion in the scene. The global evidence for (0.25, 0°) in this scene was 0.86, and the evidence for (0.5, 180°) was 0.91.

Note also that for the leftward-moving object there were contributions to the global-velocity estimate from all regions of the object, whereas these contributions were concentrated at the leading and trailing edges of the rightward-moving object. This difference was due to the effects of local edge orientation in the two objects. For the leftward-moving object the contrast stripes were oriented perpendicular to the direction of motion, providing strong local motion signals over most of the region covered by the object. The selection-unit activity still tended to be concentrated on the top and bottom edges of the object, because these were the regions with motion discontinuities (between regions of rightward motion and regions of no motion). However, because these edge regions were only ~1.5 times as strong as the center regions, the center regions still made a significant contribution to the overall global motion estimate. For the rightward-moving object the contrast stripes were oriented parallel to the direction of motion, and motion-energy detectors near the center of these stripes experienced no contrast variation and hence failed to respond to the motion of the object. As a result, strong local motion signals for rightward motion were found only along the leading and trailing edges of this object. The selection activity for rightward motion was concentrated near the three corners of the rightward-moving object (if we ignore the one ambiguous corner), because of the presence of motion discontinuities in these regions. However, since the discontinuities here were again only between regions of rightward and zero motion, these corners were weighted only ~1.5 times as strongly as the middles of the leading and trailing edges.

Comparison of the first and third examples illustrates some of the complexity that may arise from the interactions between the local-velocity and the selection networks in the model. In the first example, velocity measurements near the centers of object edges did not make a significant contribution to the evidence for the overall motion of the object. In contrast, for both moving objects in the third example, regions near the centers of the leading and trailing edges of the objects did make a significant contribution to the evidence for overall motion

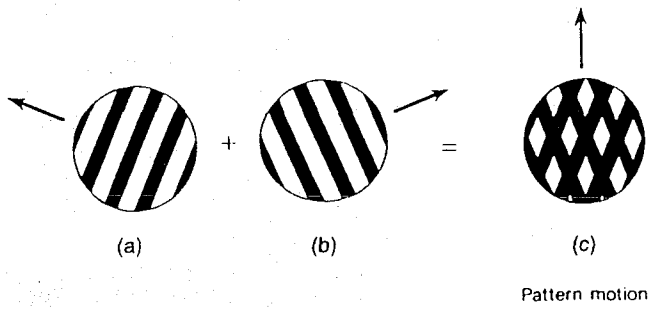

Fig. 13. Construction of plaid patterns. (a), (b) Two superimposed drifting gratings. When either grating is presented alone, its direction of motion is reliably reported (arrows). (c) The two moving gratings in (a) and (b) are superimposed. They cohere, and the pattern appears to move in a direction different from the direction of motion of either component grating (upward arrow).

of the object. This difference was due to the multiplicative interaction between selection and velocity units and the different way in which competition was organized in the two pathways. In both examples, corners were selected most strongly, although the presence of two orthogonal motion directions caused the corner regions to be selected somewhat more strongly in the first example. However, in the third example there were far more regions that supported the global motion of each object, and as a result the competition across locations in the selection pathway produced a relatively diffuse distribution of selection activity. In the first example there were only four locations supporting the global motion of the object, so each of these received considerably more weight. This difference in selection activity can still account only for part of the difference in the two examples. The local-velocity distributions in the corner regions were also very different between the two examples. In the first example the local-velocity distributions in corner regions were very different from those near the center of object edges, whereas in the third example the velocity distributions in the centers and at the corners of leading and trailing edges were nearly identical. It is the multiplicative combination of these two effects that accounts for the dramatic difference in the overall result obtained in these two examples.

C. Transparent Plaids

In addition to testing visual stimuli similar to those in the training set, we tested the model with a variety of stimuli that have been used in psychophysical studies of human motion perception. These simulations, and the corresponding psychophysical and physiological results, are discussed in detail in a companion paper.[25] Here we only briefly summarize some of the results, focusing primarily on the ability of the model to deal with more complex cases of transparency and occlusion. Many computational motion models have difficulty with transparency, which is common in natural scenes. Transparency occurs whenever two different motion vectors appear simultaneously at some location in the visual scene, and it can occur whenever transparent or partially transparent objects pass in front of each other.

A well-studied class of psychophysical stimuli for transparency is the plaid pattern.[7,12,59] These stimuli consist of two independently moving gratings that are superimposed (Fig. 13). When human observers are presented with only one of the gratings [Figs. 13(a) and 13(b)] they always reliably report the motion of the grating in the direction perpendicular to their orientation. When two gratings are superimposed [Fig. 13(c)], rather than seeing the two independent motions of the gratings, most observers see the two gratings cohere and form a single pattern moving in a direction different from that of either of the component gratings. However, under certain conditions, rather than fusing the two gratings into a single coherent motion, a human observer reliably reports seeing both motions simultaneously, as if one grating were sliding above or below the second. This perceptual state is called motion transparency, since one grating appears to be partially transparent, allowing the second grating to be seen through it.

Any computational model that integrates or averages over large homogeneous regions tends to combine the motion of the two component gratings into a coherent percept, especially if the components have similar spatial frequencies. However, Stoner *et al.*[12] showed that human observers can see transparent motion even when the components of a plaid have identical spatial frequencies. This tendency to see transparent motion can be affected by simply an alteration in the luminance of the region of intersection of the two gratings. Our model is capable of qualitatively duplicating these results; when the luminance of the intersection region is altered, the model either will report a single velocity corresponding to the pattern motion or will report the velocity of both components of the pattern independently.[25] In addition, noncoherent motion is seen by the model over a broader range of luminance when the angle between the gratings is 135° rather than only 90°. This result agrees with psychophysical results obtained with human subjects.[12]

The coherence of plaid patterns can also be affected by the spatial frequencies of the two components.[59,60] If the component frequencies are different by more than an octave, human observers see two separate motions rather than a single coherent motion. We explored this effect in the model by using a plaid pattern in which the luminance of both gratings was held constant (Fig. 14) and in which the frequency of one component was held fixed while the frequency of the second component was varied systematically. To measure the coherence of the response of the model, we defined a measure of the percent of the response corresponding to component motion:

$$PC = 50\left(1 - \frac{a_P}{I_P} + \frac{a_C}{I_C}\right), \tag{27}$$

where PC stands for percent component motion, I_P indicates the ideal response to pattern motion, a_P indicates the actual amount of pattern response, I_C indicates the ideal response to component motion, and a_C indicates the actual response to component motion.

Since a_P, I_P, a_C, and I_C are strictly positive and less than 1 and $a_C \leq I_C$ and $a_P \leq I_P$, PC lies between 0 and 100. We computed the ideal component response by presenting the gratings to the model independently and adding together all the responses of the output units above a threshold of 10/33 to get a single scalar value. This threshold was approximately 10 times the resting state of the output units and was calculated by the technique described above for estimating thresholds for

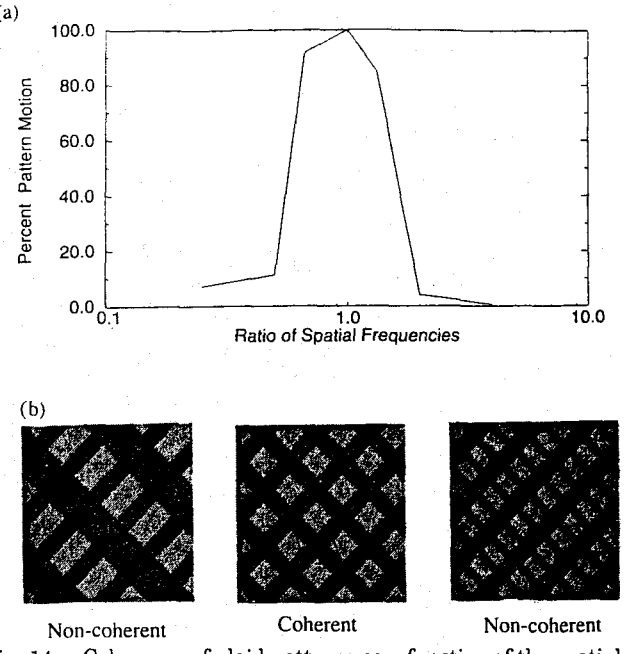

Fig. 14. Coherence of plaid patterns as a function of the spatial frequency of component gratings. The input to the model consisted of two component gratings, and the spatial frequency of one of the gratings was varied. (a) Percent of pattern motion in the output of the model as a function of the ratio of the spatial frequencies of the two component gratings. When the component spatial frequencies differed by an octave or more, the model always reported the two component motions rather than the pattern motion. (b) Examples of two noncoherent plaid patterns and one coherent plaid pattern. In the leftmost pattern the variable grating has one half the spatial frequency of the fixed grating. In the rightmost pattern the variable grating has twice the spatial frequency of the fixed grating. These patterns mark the boundaries at which the model fails to detect coherent pattern motion.

output responses. We computed the actual component response by adding together the activity in all the output units that were above threshold in the ideal case (typically near four units). The ideal pattern was a pattern of intersecting bars (with intersection regions having the same luminance as the bars) that looked like a plaid pattern when it was viewed statically. The entire pattern was moved in the intersection-of-constraints direction of the original plaid stimuli. We computed the ideal pattern response by adding together the activity of all the output units above threshold, and we computed the actual pattern response by adding together the activity in the output units that had been above threshold in the ideal pattern response (typically near two units).

PC is not the most direct measure of the response of the model, but it captures two important effects that are present in the psychophysical experiments with which we wish to compare the model. First, the *PC* measure takes into account the effects of random variation; units in the model are noise free, so the response to any particular input pattern is deterministic. A simpler winner-take-all strategy based on the strength of the output of units tuned to component motion versus the output of units tuned to pattern motion would always pick the larger response even when the two responses were nearly identical in value. In the presence of random variation, however, when the two responses are nearly equal in value we would expect each response to win roughly half the time. Similarly, when the pattern and component responses are nearly equal, *PC* is near 50%. The psychophysical experiments are also choice experiments: the subject makes a perceptual judgment between two categories on the basis of their similarity to the ideal categories. Similarly, the *PC* measure compares the actual responses with ideal responses in order to decide which category is most similar. This is especially important for these plaid stimuli, because we did not train the model with similar stimuli and therefore had no *a priori* expectation of what the model response corresponding to these stimuli should be. For example, the activity of units in the ideal pattern response is generally weaker than that of the component response, so simply using unit response values would lead to misleading results.

When the two component gratings had the same spatial frequency, the output of the model was coherent pattern motion. When the difference in the ratio of the component spatial frequencies was 1 octave or greater, the output of the model was component motion rather than coherent motion, as shown in Fig. 14. This is similar to human performance on the same stimuli. The failure of human observers to see coherence in plaids with significantly different spatial-frequency components has been used as an argument in favor of the independent processing of spatial-frequency channels in early stages of the human visual system. The failure of the model to report coherent motion in plaids can also be traced to a failure to integrate across very different spatial frequencies in the local-velocity stage of the model. The range of spatial frequencies for which plaid patterns would cohere corresponds very well to the range of frequencies over which an individual local-velocity unit responds.

Plaid-pattern coherence may also be affected by the relative contrast of the two component gratings.[7] As the contrast difference between the two component gratings increases, human subjects have a stronger tendency to see transparent rather than coherent motion. We have also replicated these experiments on the model. The initial plaid configuration was reliably seen as coherent by the model when the two component gratings had equal contrast; the contrast of one of the two gratings was then varied, and the model's response was assayed with the *PC* measure defined above. As the ratio of the contrast of the two components was increased, the *PC* measure increased, until for a sufficiently large contrast ratio the model always reliably reported only component motion (Fig. 15). The model was somewhat sensitive to relative (but not absolute) contrast, so the lower-contrast component tended to produce a weaker output than did the higher-contrast component, but both responses were well above threshold for a broad range of contrasts. Because the model was exposed during training to scenes with objects of different contrast, the selection units learned to compensate at least partially for relative contrast differences in a scene: lower-contrast objects have more concentrated support, which provides a larger multiplier for the outputs of the local-velocity pathway.

D. Dynamic Random-Dot Stimuli

The primate visual system is sensitive to coherent motion in the presence of a background of dynamic noise. This sensitivity has been studied with use of dynamic random-dot displays. Newsome and Pare[61] have shown that rhesus monkeys are able to identify reliably the direc-

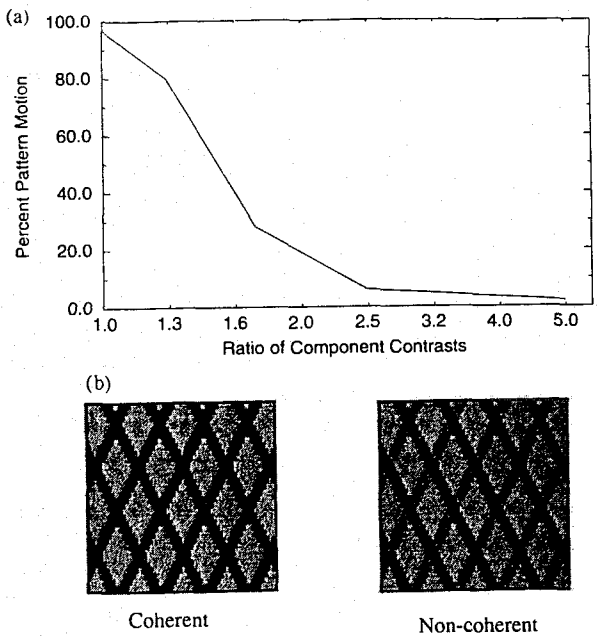

Fig. 15. Coherence of plaid patterns as a function of the contrast ratio of component gratings. The input to the model consisted of two component gratings, and the contrast of one of the gratings was varied. (a) Percent of pattern motion in the output of the model as a function of the ratio of the contrasts of the two component gratings. When the component contrasts differed by more than 1 octave, the model always reported the two component motions rather than the pattern motion. (b) Examples of the boundary between coherent and noncoherent plaid patterns. In the pattern at the left, the ratio of component contrasts is 1.3, and this pattern was coherent 83% of the time. In the pattern at the right, the ratio of component contrasts is 2.5, and this pattern resulted in noncoherent output 94% of the time.

tion of motion in a field of randomly moving dots when fewer than 5% of the dots in the field move in a coherent direction.

We used identical dynamic random-dot displays to examine the sensitivity of our model to coherent motion against a background of dynamic noise.[25] With the selection layer functioning, the model was able to identify the direction of coherent motion reliably with 4% of the dots in the field moving coherently. However, with the selection pathway disabled, the same level of performance needed 18% of the dots in the field in order to move coherently. The selection layer greatly improved the sensitivity of the model to coherent motion in the presence of dynamic noise.

The improvement in sensitivity to coherent motion was due to the role of the selection units in finding regions of the image for which motion estimates are most reliable (Fig. 16). The selection layer tended strongly to select only regions containing several pairs of dots that supported the same motion. These regions tended to contain a higher proportion of coherently moving dots than occur on average within the entire image, so averaging only over these regions greatly improved the overall signal-to-noise ratio. In the example shown, the signal in the coherent direction was 1.8 times as strong as the motion signal in any other direction when it was averaged over the entire image. However, when averaged over only the selected regions, the signal in the coherent direction was 7.5 times as strong as the signal in any other direction.

E. Barber Pole

The ability of the model to integrate over only selected regions of an image can improve the model's response in noisy environments, but it can also lead to biased results when the selected regions contain misleading information. An example of such a failure of the model appears in the barber-pole illusion (Fig. 17). The black bands move to the right and are viewed though a vertical rectangular aperture. As each band reaches the edge of the aperture, it disappears and is replaced by a new band on the left-hand side. This is equivalent to having the black stripe continuously wrapped around a cylinder spinning to the right. However, the same display could also be produced by vertical movement of the bands or by movement at any angle except parallel to the bands (because of the aperture problem).

As shown in Fig. 17(b), diagonal motion upward and to the right was observed in the middles of the bands, as a result of the aperture effect. However, at the ends of the bands the local velocity was strongly upward, as a result

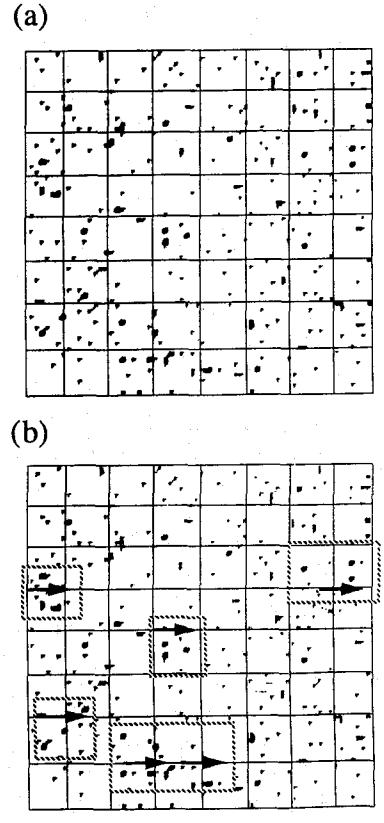

Fig. 16. Local-velocity and selection responses to a dynamic random-dot display. (a) One frame of the input to the model consisting of approximately 250 randomly placed dots. Most of the dots were replaced in random locations in the next frame; however, a subset of the dots (darkened) was moved a fixed displacement to the right in the next frame. (b) The dashed squares surround regions that were selected for integration in this frame. Arrows within the regions indicate the direction of the local-velocity estimate for the region. The selected regions had denser-than-average numbers of coherently moving dots, and averaging over only these selected regions greatly improved the signal-to-noise ratio compared with averaging over the entire field.

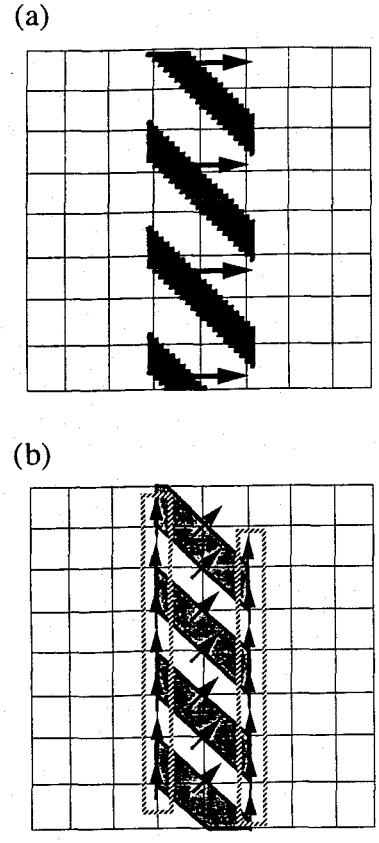

Fig. 17. Local-velocity and selection responses to the barber-pole illusion. (a) The input to the model was a set of diagonal stripes that move at 1 pixel/frame to the right (indicated by arrows). When elements of the stripes reach the right-hand edge of the pole they disappear and are replaced on the left-hand side so that the diagonal stripes are maintained at a constant size and angle. (b) Only the regions corresponding to the edges of the pole were selected for integration at the output stage (indicated by dashed lines). In these regions the changes in location of contrast edges are locally consistent with upward motion even though the true motion in the scene was rightward.

of the edge effects. The model selected the regions corresponding to the edges of the poles more strongly and as a result reported upward motion in the output units. Although the upward motion reported by the model is an incorrect bias, it nonetheless corresponds to what humans normally perceive when presented with the same stimulus.

F. Circle

The model was originally trained exclusively with rectangular objects, and an interesting question was whether the model could generalize correctly to the motion of curved surfaces. We explored this issue by showing the model examples of circles undergoing simple translational motion (Fig. 18). This circle has been antialiased in an attempt to reduce the effects of the fairly coarse input representation used by the model, but pixellation effects along the edges of the circle are apparent.

The model does indeed produce a correct response to translations of circles. In general, the proper integration of motion signals along a curved boundary is a challenging problem.[62] The coarse representation of direction and the broad directional tuning of units in the local-velocity and selection pathways actually simplifies the processing of curved contours considerably, obviating the need for fine directional integration and effectively smoothing curved contours. The support in the model is concentrated among three directional hypotheses: up-rightward, rightward, and down-rightward [Fig. 18(b)]. The strongest support is for the rightward hypothesis, which corresponds to the final output of the model. The regions near the centers of the leading and trailing edges of the circle are more strongly selected than the regions closer to the top and bottom of the circle, because these middle regions contain fairly strong motion signals for several directions of motion, whereas the edge regions contain at best signals for one direction of motion and for no motion [compare Fig. 7(b), examples 2 and 3].

5. DISCUSSION

A. Limitations

The model of motion processing presented here handles many visual stimuli, including transparency, size and contrast variation, and some special cases such as the barber-pole illusion. However, some aspects of visual perception

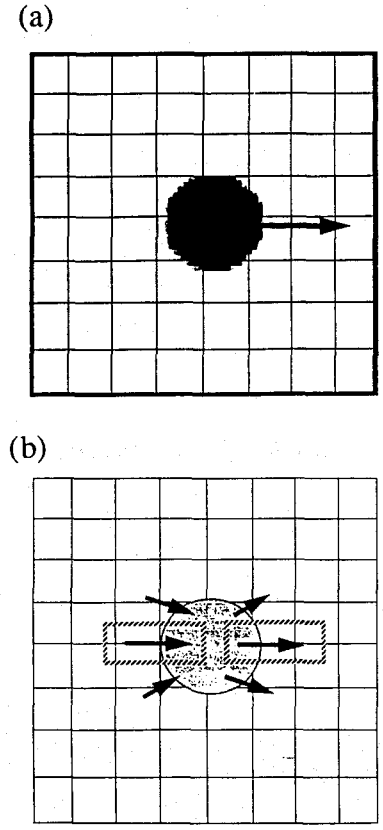

Fig. 18. Local-velocity and selection responses to a translating circle. (a) The input to the model is an antialiased circle that moves to the right at 1 pixel/frame. Although the input is antialiased, pixellation effects are clearly visible around the edges of the circle and are caused by the low resolution of the input grid. (b) The regions selected most strongly for integration correspond to the leading and trailing edges of the circle (enclosed by dashed lines). Although at the resolution of the integration and selection layers all regions around the edge of the circle contain motion signals distributed over several directions, the directions and velocities of these motions are similar enough to be averaged by the integration pool, and the resulting integration-pool activity corresponds to motion in one dominant direction, as shown by the arrows along the edges of the circle.

are not captured by the current model. Perhaps the most significant limitation of the current model is that it does not deal well with temporal integration of motion in the computation of velocity. There is clear evidence that the human visual system uses temporal integration[63] in its velocity computation. There are a number of non-Fourier motion stimuli[13,64,65] that require some form of temporal integration in order to see the motion in these stimuli. We synthesized some of these stimuli, and the model does indeed fail to match the psychophysical responses of humans to these stimuli. The model would also fail to replicate the experiments by Watamaniuk and McKee,[66] which showed that human subjects can detect the motion of a single dot in the midst of a random-dot cinematogram if the dot has a fixed, repeatable trajectory. The detection of such a weak signal in a field of dynamic noise would be nearly impossible without some form of temporal integration. It should be possible to extend the current model to perform both spatial and temporal selection and integration, and this is an important direction for future research.

The current model is also limited in its ability to interpret nonrigid motions, or flow fields associated with nontranslational three-dimensional motions. For example, we presented spinning, nontranslating stimuli to the model and found that the model has difficulty interpreting this type of motion field. Similar problems were seen with the radial motion fields created by approaching or receding objects and by nonrigid stimuli such as those used by Nakayama and Silverman.[67] Some of these difficulties can be traced to the final output stage of the model, which has only one set of units to represent the entire visual field. This lack of spatial resolution makes it inherently difficult for the model to represent some of these more-complex motion fields. One interesting direction for future development would be to extend the model so that the final output representation retained spatial information and could represent different motions in different regions of the retinal image. This extension might help the model to deal better with the motion fields produced by nonrigid and nontranslational motions.

B. Relation to Other Models

Many traditional computational vision models have been strongly influenced by the assumption that the purpose of the visual system is to create a veridical representation of the visual scene in the real world.[6,68] In the motion domain this implies first that the observer create an accurate representation of the local two-dimensional velocity at all points in the input image and then use this optical flow field to infer the three-dimensional motion in the visual scene.

The selection model of motion processing differs from most previous models in that it does not assume that the flow field is spatially continuous; rather it assumes that the early stages of motion processing coarsely segment an image into regions of coherent motion and provide an estimate of the two-dimensional velocity for each of these regions.[24] Although this representation does not contain as much information as a true flow field, it may be adequate for the majority of visual tasks, such as tracking and segmentation, in conjunction with other cues that require motion information. The coarse representation of coherent motion is easier to compute and is more robust to noise than a fine representation of flow fields.

The coarse representation of coherent motion does make some types of computation more difficult. In particular, as noted above, the complex flow fields created by certain types of nontranslational three-dimensional motions and nonrigid motions are represented more easily by a fine representation of optical flow. However, our coarse representation permits the rapid and effective computation of the velocities of multiple objects under a wide variety of conditions and is adequate for many tasks faced by a navigating creature. In this respect our model is related in spirit to active perception/animate vision, which attempts to represent only that information needed for the task at hand.[68–70] Advantages that our model enjoys include a natural way to include attention, avoidance of smoothing across boundaries, and partial segmentation by robust velocity features.

All models of motion processing make assumptions that can lead to systematic errors when these assumptions are violated. The local-motion computation used in our model is very similar to that used in previous models[2,3] and relies on assumptions of rigid translation that are often violated in real scenes. Many previous models have tried to deal with these types of systematic error by trying to fix or fill in bad estimates, using information from surrounding estimates. An important example of this type of approach is the use of smoothing or regularization techniques to integrate information from larger regions of an image and remove systematic errors caused by problems such as the aperture effect.[17,18,22] A common problem with these regularization techniques is that they assume that all data in a region of integration are homogeneous. When this assumption is violated, the techniques can oversmooth the data, producing poor overall results. Such oversmoothing is particularly severe near object boundaries (for example, the boundary between a moving object and a stationary background). Smoothing across such boundaries blends data from two nonhomogeneous regions and tends to blur object boundaries as well as producing poor velocity estimates near the boundaries.

One way around the problem of oversmoothing is to segment an image into homogeneous motion regions while the estimation of smooth image motion is being determined. The most popular paradigm for doing joint regularization and segmentation has employed line processes in conjunction with some sort of grid-based regularization method,[17,20–23] although alternative techniques have been proposed.[71] Line processes are Boolean fields that are usually set when the squared difference between two adjacent velocity estimates exceeds some threshold. Once a line process is set, smoothing or averaging across the locations joined by the line process is not permitted. Line processes can be successfully combined with regularization to produce good estimates of optical flow when boundaries between objects are quite smooth and simple. However, in natural scenes involving partial occlusion and transparency, smoothing must often be performed across noncontiguous regions with ill-defined boundaries. In these situations line-process techniques can perform very poorly.[27]

In contrast to most previous models of motion processing, rather than attempting to fix bad estimates, our model adopts the strategy of trying to ignore bad esti-

mates and concentrating only on good ones. Regularization methods begin with the assumption that nearby velocity measurements are similar unless there is explicit evidence to the contrary. Our model assumes instead that local-velocity measurements are spatially independent but that all the velocity measurements in the visual scene are generated from a small number of motion processes (which normally correspond to distinct objects in the visual scene). In particular, if a velocity estimate differs from surrounding estimates, it may be assigned to a different process than are the surrounding estimates, but no attempt is made to regress this estimate toward the surrounding estimates, as would occur with most smoothing techniques. By assigning to a particular process only a subset of local-velocity measurements and then integrating all the velocity measurements for that process, we come up with a fairly robust estimate of the velocity of that process. Within the statistical literature, assigning measurements to different processes is referred to as computing regions of support for a process and is a standard approach for robust estimation.[72] The selection pathway in our model can be regarded as a feed-forward mechanism for computing regions of support for robust velocity estimates.

Our model was especially constructed to handle motion transparency. Other recent models have also dealt explicitly with transparent objects, and certain aspects of these models have much in common with our model. Jasinchi *et al.*[62] proposed a three-stage model in which local-velocity components, normal to contours (and feature velocities), are computed, and then each of these local-velocity components contributes to a two-dimensional velocity space. Separate velocity histograms are computed for each region, on the basis of the number of votes that each bin in the velocity space receives. For two motion patterns one perceives motion coherence, transparency, or a mixture of both types of motion, depending on whether the velocity histogram is unimodal, bimodal, or trimodal, respectively. Similarly, in our model, transparency and mixed percepts are indicated by the presence of multiple activity peaks in the pool of output units. In addition, the local-velocity spaces used by Jasinchi *et al.* accumulate evidence for particular velocities in a small region, much the way local-velocity units in our model accumulate the local evidence for specific velocities from small regions. A major difference between the two models is that multimodality is actually suppressed at the local-velocity level in our model and appears primarily only at the output level of the model. Thus we represent transparent phenomena at a much coarser resolution. One advantage of the finer representation of transparency used by Jasinchi *et al.* is the ability to model the effect of contour curvature on transparency, an effect that we do not model.

Another recent motion-transparency model, by Smith and Grzywacz,[73] uses highly local computations, and, like the model of Jasinchi *et al.*,[62] is capable of representing transparency at a local level. This model was designed explicitly with plaid patterns in mind and is less general than our model or that of Jasinchi *et al.* The earlier stages of the Smith–Grzywacz model are very similar to the early stages of our model, both using normalized motion-energy responses from local regions in order to make the models less sensitive to intensity or contrast scaling. The local-velocity computation in this model is based on a weighted sum of the motion-energy responses, followed by a winner-take-all strategy, and is again very similar to the local-velocity calculation in our model. The interesting feature of the Smith–Grzywacz model is that certain regions, corresponding to regions with the motion-energy gradient above a certain threshold, are selectively excluded from the summation process. This mechanism is related to the selection mechanism that occurs in our model, although here the selection process is operating at the motion-energy level instead of at the velocity level as in our model. It is interesting to note that, in an earlier one-dimensional version of our model, the selection units learned to become detectors of large gradients in the motion-energy distribution. This phenomenon is similar to the selection criterion used by Smith and Grzywacz to prevent the accumulation of certain motion-energy measurements.

C. Segmentation

One important limitation of the current model is that the final output stage of the model has no spatial resolution. However, this final output stage can be regarded as an artifice necessary for training the model; the interesting representations exist at the level of the selection and local-velocity networks. The selection network provides a partial solution to the problem of image segmentation. Logically, units in the selection network can be divided into layers, with each layer representing essentially a different velocity. Each of these layers is organized retinotopically, and by examining which units are active in one of these layers we can determine in which regions of the image there are signals that support a particular velocity. In this sense we can segment the original input scene coarsely into regions that support different velocities.

Most previous models for segregating figure from ground have implicitly assumed that objects were spatially continuous, and the first step was to find a bounding contour. In many scenes the initial estimate of a contour is incomplete, and smoothing or regularization is used to complete the contour. Our approach to segmentation does not make this assumption: the selection network may group information that is spatially separated by intervening objects. This can be an advantage in situations involving partial occlusion and transparency, in which boundaries may be ill defined. However, the selection network provides at best only a partial solution to object segmentation. As Fig. 10 illustrates, the regions selected may correspond to only a portion of an object. In addition, the selection network will tend to group distinct objects moving at the same velocity.

Some process in addition to the selection network would be necessary for true object segmentation to be performed. We suggest that this process may operate by combining information from multiple cues and modalities, the signals from the selection network providing one source of cues. In some cases a mechanism for completing a partial contour may still be necessary, but in many cases signals from several modalities, such as color, texture, and motion, may provide sufficient information for completion of an object contour. In such cases attempts to complete a contour based on information from a single modality would be computationally wasteful.

D. Attention

The proposed model suggests a way to integrate perception and attention. The active mechanism in the model for selecting subsets of unit responses over which to integrate performs a preattentive segregation of the motion pathway. The same mechanism could be used to attend to motion actively in restricted regions of the visual field by a top-down bias of the selection network. This is less invasive than a direct bias to the local-velocity units. An additional advantage of this active mechanism is that it could be used at all levels of cortical representation, at early as well as at later stages of processing. Selection may be a fundamental aspect of cortical organization that could provide a unification of preattentive vision with attention.

The specific mechanism that we have proposed for the selection pathway is closely related to mechanisms proposed in attentional models. The type of renormalization nonlinearity appearing in Eqs. (5), (8), and (9) can be implemented by iteration of the following state equation for a finite period:

$$\frac{dy_i}{dt} = y_i\left(x_i - \sum_j x_j y_j\right), \tag{28}$$

where x_i is the net input and y_j the output of a unit.[77] This equation is easily implemented in a network with inhibitory lateral interactions and has been proposed as a mechanism for many attentional phenomena.[74-76] The time course of evolution of this equation appears to match many aspects of visual search,[77] suggesting a further link between the selection process proposed in our model and more-classical attentional models.

E. Selection

Identifying regions of support is a form of outlier rejection. Imagine that samples from some function are contaminated by noise, including some severe systematic errors. Such severely contaminated points are referred to as outliers and can be considered the result of some other process entirely. (In the motion-processing domain, severely contaminated estimates of velocity could come from regions of constant intensity or from regions containing contrast variations at just a single orientation.) The problem is to estimate the function from the samples without contamination from the outliers. One way that one can do this is first to estimate the function by using all the data points and then to see which samples are far away from their estimated functional values. Samples that are too far from their estimated values can be thrown out, and then the function can be reestimated with use of the remaining data. After a few iterations no more points exceed the threshold, and the process yields a final estimate of the function. The success of this technique depends on picking the correct threshold for deciding which data to throw out.

This iterative technique is a form of M estimation. The samples that are used in the final estimate of the function are the support for that function. M estimation has been proposed for solving a variety of problems in computer vision.[78-80] The function being estimated is often referred to as a hypothesis, since it is an attempt to explain all the data within its support. Clearly, M estimators are highly sensitive to the rejection threshold that is used. In addition, if too many of the original samples come from a contaminating or outlier process, the M estimator tends to break down and converge on a contaminated estimate.[72] In a visual scene containing many moving objects, the number of velocity measurements from a particular object may be a very small proportion of the total number of measurements in the scene, so breakdown of the estimator can be a severe problem.

To overcome some of these difficulties, Darrell and Pentland[27] suggested an approach in which multiple hypotheses competed to include samples within their regions of support. In their model each hypothesis corresponded to an object in the visual scene, and, because the number of objects was not known *a priori*, a complex relaxation scheme was proposed for computing both the optimal number of hypotheses (or planes of motion) and the velocity of each plane. Local-velocity estimates were first assigned to a large number of hypotheses, and an overall velocity was estimated for each hypothesis. A point could be excluded from the region of support of a hypothesis either because it deviated too much from the predicted velocity of the hypothesis or because it could be explained better by another hypothesis. If the estimated velocities of two hypotheses began to converge, those hypotheses could be merged into a single larger hypothesis. Assignment of velocity estimates to hypotheses and merging of hypotheses were repeated until the relaxation procedure converged.

The method of assigning support in our model is similar to that proposed by Darrell and Pentland[27] in that multiple hypotheses compete to include local-velocity estimates in their regions of support. However, our approach is conceptually simpler and differs from their model in two important respects. In our model the hypotheses correspond not to distinct objects but rather to distinct *velocities*, and the number of hypotheses is always fixed. In addition, our decision to include a particular local-velocity estimate in the region of support for a hypothesis is not based directly on the difference between the local estimate and the hypothesis velocity. Instead, the selection pathway computes this assignment of support in a noniterative fashion on the basis of the same set of motion-energy measurements that are used to compute the local-velocity estimate. However, as we noted in Section 3, the assignment of support computed by the selection pathway is based indirectly on the difference between the local-velocity estimate and all the candidate global velocities.

F. Conclusion

In almost all the examples of moving objects that we have tested, the regions of support for each object were a small fraction of all the local-velocity measurements available. This sparse representation has several advantages in addition to improving robustness. The reduced representation is more compact than the local-velocity representation and can be used as the input to further levels of processing that represent nonuniform velocity fields. The selected regions tend to be located at corners and at terminators that are natural segmentation boundaries for complex objects, such as articulated limbs. Finally, sparseness could allow hyperacuity judgments to be made in distributed representations.[81]

We intend to report in future papers on extensions of the selection model to the cueing of invariant motion stimuli[13] including non-Fourier motion[64,65] and to nonuniform velocity flow fields.[67] The same approach can also be applied to other problems in vision, such as stereopsis, in which similar problems arise with occlusion and transparency.

ACKNOWLEDGMENTS

We are grateful to Gene Stoner, Thomas Albright, and Stephen G. Lisberger for discussions on motion processing and to Ning Qian, Norberto Grzywacz, Paul Viola, J. Anthony Movshon, and Thomas Albright for comments on this paper. We thank the San Diego Supercomputer Center for computer use on the Intel hypercube

*Present address, Synaptics Inc., 2698 Orchard Parkway, San Jose, California 95134.

REFERENCES

1. J. J. Gibson, *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston, Mass., 1966).
2. D. J. Heeger, "Model for the extraction of image flow," J. Opt. Soc. Am. A **4,** 1455–1471 (1987).
3. N. M. Grzywacz and A. L. Yuille, "A model for the estimation of local image velocity by cells in the visual cortex," Proc. R. Soc. London Ser. B **239,** 129–161 (1990).
4. E. C. Hildreth, *The Measurement of Visual Motion* (MIT Press, Cambridge, Mass., 1984).
5. D. Marr and S. Ullman, "Directional selectivity and its use in early visual processing," Proc. R. Soc. London Ser. B **211,** 151–180 (1981).
6. D. Marr, *Vision* (Freeman, New York, 1982).
7. M. Adelson and J. A. Movshon, "Phenomenal coherence of moving visual patterns," Nature (London) **300,** 523–525 (1982).
8. O. J. Braddick, "Segmentation versus integration in visual motion processing," Trends Neurosci. **16,** 263–268 (1993).
9. K. Nakayama, "Biological image motion processing: a review," Vision Res. **25,** 625–660 (1985).
10. J. H. R. Maunsell and W. T. Newsome, "Visual processing in monkey extrastriate cortex," Ann, Rev. Neurosci. **10,** 363–401 (1987).
11. R. A. Andersen and R. M. Siegel, "Motion processing in primate cortex," in *Signal and Sense: Local and Global Order in Perceptual Maps*, G. M. Edelman, W. E. Gall, and W. M. Cowan, eds. (Wiley, New York, 1989).
12. G. R. Stoner, T. D. Albright, and V. S. Ramachandran, "Transparency and coherence in human motion perception," Nature (London) **344,** 153–155 (1990).
13. T. D. Albright, "Form-cue invariant motion processing in primate visual cortex," Science **255,** 1141–1143 (1992).
14. T. D. Albright, "Direction and orientation selectivity of neurons in visual area MT of the macaque," J. Neurophysiol. **52,** 1106–1130 (1984).
15. J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome, "The analysis of moving visual patterns," in *Pattern Recognition Mechanisms*, C. Chagas, R. Gattass, and C. Gross, eds. (Springer-Verlag, New York, 1985), pp. 117–151.
16. D. J. Heeger and E. P. Simoncelli, "Model of visual motion sensing," in *Spatial Vision in Humans and Robots*, L. Harris and M. Jenkin, eds. (Cambridge U. Press, London, 1992).
17. C. Koch, H. T. Wang, and B. Mathur, "Computing motion in the primate's visual system," J. Exp. Biol. **146,** 115–139 (1989).
18. B. K. P. Horn and B. G. Schunk, "Determining optical flow," Artif. Intell. **17,** 185–203 (1981).
19. H. H. Nagel, "On the estimation of optical flow: relations between different approaches and some new results," Artif. Intell. **33,** 299–324 (1987).
20. S. W. Zucker, Y. G. Leclerc, and J. L. Mohammed, "Continuous relaxation and local maximum selection: conditions for equivalence," IEEE Trans. Patt. Anal. Mach. Intell. **PAMI-3,** 117–128 (1981).
21. R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labeling processes," IEEE Trans. Patt. Anal. Mach. Intell. **PAMI-3,** 267–287 (1983).
22. P. K. Kienker, T. J. Sejnowski, G. E. Hinton, and L. E. Schumacher, "Separating figure from ground with a parallel network," Perception **15,** 197–216 (1986).
23. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," IEEE Trans. Patt. Anal. Mach. Intell. **PAMI-6,** 721–741 (1984).
24. M. J. Bravo and S. N. J. Watamaniuk, "Evidence for two speed signals: a course local signal for segregation and a precise global signal for discrimination," Vision Res. (to be published).
25. S. J. Nowlan and T. J. Sejnowski, "Filter selection model of motion processing in area MT of primates," J. Neurosci. (to be published).
26. S. J. Nowlan and T. J. Sejnowski, "Filter selection model for generating visual motion signals," in *Advances in Neural Information Processing Systems 5*, S. J. Hanson, J. D. Cowan, and C. L. Giles, eds. (Morgan Kaufmann, San Mateo, Calif., 1993), pp. 369–376.
27. T. Darrell and A. Pentland, "Robust estimation of a multilayered motion representation," in *Proceedings of the IEEE Workshop on Visual Motion* (Institute of Electrical and Electronics Engineers, New York, 1991).
28. S. J. Nowlan, "Competing experts: an experimental investigation of associative mixture models," Tech. Rep. CRG-TR-90-5 (Department of Computer Science, University of Toronto, Toronto, 1990).
29. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," Neural Computation **3,** 79–87 (1991).
30. O. J. Braddick, "A short-range process in apparent motion," Vision Res. **14,** 519–527 (1974).
31. H. Ikeda and M. J. Wright, "Spatial and temporal properties of 'sustained' and 'transient' neurones in area 17 of cat's visual cortex," Exp. Brain Res. **22,** 363–383 (1975).
32. D. J. Tolhurst and J. A. Movshon, "Spatial and temporal contrast sensitivity of striate cortical neurons," Nature (London) **257,** 674–675 (1975).
33. R. A. Holub and M. Morton-Gibson, "Response of visual cortical neurons of the cat to moving sinusoidal gratings: response-contrast functions and spatiotemporal integration," J. Neurophysiol. **46,** 1244–1259 (1981).
34. 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), pp. 303–318.
35. J. P. H. van Santen and G. Sperling, "Elaborated Reichardt detectors," J. Opt. Soc. Am. A **2,** 300–321 (1985).
36. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," J. Opt. Soc. Am. A **2,** 284–299 (1985).
37. A. B. Watson and A. J. Ahumada, "Model of human visual-motion sensing," J. Opt. Soc. Am. A **2,** 322–342 (1985).
38. R. C. Emerson, M. C. Citron, W. J. Vaughn, and S. A. Klein, "Nonlinear directionally selective subunits in complex cells of cat striate cortex," J. Neurophysiol. **58,** 33–65 (1987).
39. J. McLean and L. A. Palmer, "Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat," Vision Res. **29,** 675–679 (1989).
40. R. C. Emerson, J. R. Bergen, and E. H. Adelson, "Directionally selective complex cells and the computation of motion energy in cat visual cortex," Vision Res. **32,** 203–218 (1992).
41. D. J. Heeger, "Normalization of cell responses in cat striate cortex," Vis. Neurosci. **9,** 181–197 (1992).
42. D. B. Hamilton, D. G. Albrecht, and W. S. Geisler, "Visual cortical receptive fields in monkey and cat: spatial and temporal phase transfer function," Vision Res. **29,** 1285–1308 (1989).
43. J. McLean, S. Raab, and L. A. Palmer, "Contribution of linear mechanisms to the specification of local motion by simple cells in areas 17 and 18 of cat," Vis. Neurosci. **11,** 271–294 (1994).
44. J. McLean and L. A. Palmer, "Organization of simple cell

responses in the three-dimensional (3-D) frequency domain," Vis. Neurosci. **11,** 271–294 (1994).
45. D. Gabor, "Theory of communication," J. Inst. Electr. Eng. **93,** 429–457 (1946).
46. J. G. Daugman, "Two-dimensional analysis of cortical receptive field profiles," Vision Res. **20,** 846–856 (1980).
47. 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).
48. R. C. Reid, R. E. Soodak, and R. M. Shapley, "Linear mechanisms of directional selectivity in simple cells of cat striate cortex," Proc. Natl. Acad. Sci. (USA) **84,** 8740–8744 (1987).
49. S. Hochstein and R. M. Shapley, "Quantitative analysis of retinal ganglion cell classifications," J. Physiol. (London) **262,** 237–264 (1976).
50. L. Maffei and A. Fiorentini, "Spatial frequency rows in the striate visual cortex," Vision Res. **17,** 257–264 (1977).
51. B. W. Andrews and T. A. Pollen, "Relationship between spatial frequency selectivity and receptive field profile of simple cells," J. Physiol. (London) **287,** 163–176 (1979).
52. P. Burt, "Fast algorithms for estimating local image properties," Comput. Vision Graphics Image Process. **21,** 368–382 (1983).
53. J. G. Robson, "Linear and nonlinear operations in the visual system," Invest. Ophthalmol. Vis. Sci. Suppl. **29,** 117 (1988).
54. A. B. Bonds, "Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex," Vis. Neurosci. **2,** 41–55 (1989).
55. J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neuro-computing: Algorithms, Architectures and Applications*, F. Fougelman-Soulie and J. Herrault, eds. (Springer-Verlag, London, 1989).
56. D. H. Ballard, G. E. Hinton, and T. J. Sejnowski, "Parallel visual computation," Nature (London) **306,** 21–26 (1983).
57. N. Qian, R. A. Andersen, and E. H. Adelson, "Transparent motion perception as detection of unbalanced motion signals I: Psychophysics," J. Neurosci. (to be published).
58. L. H. Finkel and P. Sajda, "Object discrimination based on depth-from-occlusion," Neural Computation **4,** 901–921 (1993).
59. V. S. Ramachandran and P. Cavanagh, "Motion capture anisotropy," Vision Res. **27,** 97–106 (1987).
60. L. Welch, "The perception of moving plaids reveals two motion-processing stages," Nature (London) **337,** 734–736 (1989).
61. W. T. Newsome and E. B. Pare, "A selective impairment of motion perception following lesions of the middle temporal visual area (MT)," J. Neurosci. **8,** 2201–2211 (1988).
62. R. Jasinchi, A. Rosenfeld, and K. Sumi, "Perceptual motion transparency: the role of geometrical information," J. Opt. Soc. Am. A **9,** 1865–1879 (1992).
63. S. P. McKee and L. Welch, "Sequential recruitment in the discrimination of velocity," J. Opt. Soc. Am. A **2,** 243–251 (1985).
64. J. D. Victor and M. M. Conte, "Coherence and transparency of moving plaids composed of Fourier and non-Fourier gratings," Percept. Psychophys. **52,** 403–141 (1992).
65. H. R. Wilson, V. P. Ferrara, and J. Kim, "A psychophysically-motivated model for two-dimensional motion perception," Vis. Neurosci. **9,** 79–97 (1992).
66. S. N. J. Watamaniuk and S. P. McKee, "Why is a trajectory more detectable in noise than correlated signal dots?" Invest. Ophthalmol. Vis. Sci. **34,** 1364 (1993).
67. K. Nakayama and G. H. Silverman, "The aperture problem—I. Perception of nonrigidity and motion direction in translating sinusoidal lines," Vis. Res. **28,** 739–746 (1988).
68. P. S. Churchland, V. S. Ramachandran, and T. J. Sejnowski, "A critique of pure vision," in *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. Davis, eds. (MIT Press, Cambridge, Mass., 1994), pp. 23–60.
69. R. Bajcsy, "Active perception," Proc. IEEE **76,** 996–1005 (1988).
70. D. H. Ballard, "Animate vision," Artif. Intell. **48,** 57–86 (1991).
71. A. L. Yuille and N. M. Grzywacz, "A mathematical analysis of the motion coherence theory," Int. J. Computer Vis. **3,** 155–175 (1989).
72. G. Li, "Robust regression," in *Exploring Data, Tables, Trends and Shapes*, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds. (Wiley, New York, 1985).
73. J. A. Smith and N. M. Grzywacz, "A local model for transparent motions based on spatio-temporal filtering," in *Computation and Neural Systems 1992*, F. H. Eeckman and J. M. Bower, eds. (Kluwer, New York, 1993).
74. K. P. Hadeler, "On the theory of lateral inhibition," Kybernetic **14,** 161–165 (1974).
75. C. Koch and S. Ullman, "Shifts in selective attention: towards the underlying neural circuitry," Human Neurobiol. **4,** 219 (1985).
76. P. A. Sandon, "Simulating visual attention," J. Cog. Neurosci. **2,** 213–231 (1990).
77. P. A. Sandon, "Logarithmic search in a winner-take-all network," in *Proceedings of the International Joint Conference on Neural Networks* (Institute of Electrical and Electronics Engineers, New York, 1991), pp. 454–459.
78. P. J. Besl, J. B. Birch, and L. T. Watson, "Robust window operators," in *Proceedings of the Second International Conference on Computer Vision* (Institute of Electrical and Electronics Engineers, New York, 1988).
79. R. M. Haralick and H. Joo, "2D-3D pose estimation," in *Proceedings of the Ninth International Conference on Pattern Recognition* (Institute of Electrical and Electronics Engineers, New York, 1988).
80. P. Meer, D. Mintz, and A. Rosenfeld, "Robust regression methods for computer vision: a review," Intl. J. Computer Vision **6,** 60–70 (1991).
81. S. R. Lehky and T. J. Sejnowski, "Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity," J. Neurosci. **10,** 2281–2299 (1990).