In: M.S. Gazzaniga, (Ed.)
The Cognitive Neurosciences
Cambridge, MA: MIT Press (1994).

27 A Model of Visual Motion Processing in Area MT of Primates

TERRENCE J. SEJNOWSKI AND STEVEN J. NOWLAN

ABSTRACT Motion perception requires the visual system to satisfy two conflicting demands: first, spatial integration of signals from neighboring regions of the visual field to overcome noisy signals, and second, sensitivity to small velocity differences to segment regions corresponding to different objects (Braddick, 1993). We have developed a computational model for the visual processing of motion in area MT that accounts for these conflicting demands. The model has two types of units, similar to those found in area MT. One type of unit in the model integrates information about the direction motion to estimate the local velocity; these local velocity units compete among themselves to determine the most likely local velocity. A second type of unit selects regions of the visual field where the velocity estimates are most reliable; these selection units have nonclassical receptive field surrounds by virtue of competition with pools of similar units across the visual field. The output of the model is a distributed segmentation of the image into patches that support distinct objects moving with a common velocity.

The processing of motion in the primate's visual cortex begins in area V1, where cells with reliable selectivity for direction of motion are found (Maunsell and Newsome, 1987); however, these cells do not detect true velocity but instead are tuned to a limited range of spatiotemporal frequencies and exhibit spatially restricted receptive fields so that they can report only the perpendicular component of the velocity for straight edges. This so-called aperture problem is illustrated in figure 27.1. To overcome these limitations and compute true local velocity measurements, it is necessary to integrate motion responses from cells with a variety of directions and spatiotemporal frequency tunings over a wider area of the visual field (Heeger, 1987; Grzywacz and Yuille, 1990).

Neurons that respond selectively to velocity over a wide range of spatial frequencies are found in visual area MT, which receives a direct projection from area V1 (Albright, 1992; Maunsell and Newsome, 1987; Rodman and Albright, 1989). A class of cells in MT, the "pattern cells" of Movshon and colleagues (1985), respond to the direction of overall motion of plaid patterns composed of two differently oriented gratings rather than to the direction of the individual components. Psychophysical studies suggest that the perceived velocity of such patterns generally is close to the velocity that is uniquely consistent with the constraints imposed by the individual component's motions (Adelson and Movshon, 1982), although other possibilities have been suggested (Wilson et al., 1992; Rubin and Hochstein, 1993). In this chapter, we are concerned with the properties of such pattern motion cells and their responses to motion stimuli that have previously been used in psychophysical and physiological experiments.

We have proposed a computational model for the formation of local velocity responses in MT and for the combination of these local velocity estimates into a representation of the velocity of objects in the visual scene (Nowlan and Sejnowski, 1993, 1994a,b). The model departs from previous suggestions of how local velocity is estimated from visual scenes (Horn and Schunk 1981; Heeger 1987, 1992; Nagel, 1987; Grzywacz and Yuille 1990). Our model does not compute an accurate estimate of local velocity at all image locations, called the *optical flow field*, but instead evaluates the reliability of local velocity estimates and forms a coarse representation of the motion of objects in the

TERRENCE J. SEJNOWSKI and STEVEN J. NOWLAN Howard Hughes Medical Institute, The Salk Institute, and University of California, San Diego, Calif.

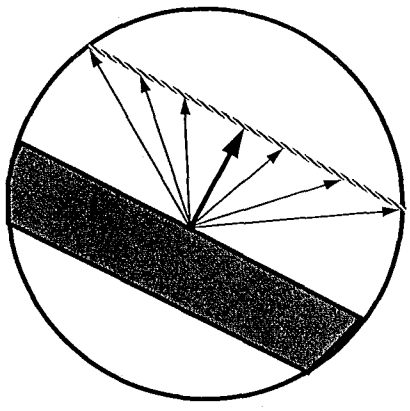

FIGURE 27.1 The aperture problem. The receptive field of a single cortical neuron is restricted to a fraction of the visual field. When a bar is introduced into the receptive field of the cell, only the velocity component orthogonal to the edge of the bar, indicated by the heavy arrow, can be measured by any local velocity mechanism. 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. The aperture problem cannot be solved without appeal to information outside the classical receptive field.

visual scene by combining only the most valid subsets of local velocity measurements. The same general strategy may be used for computing other properties of objects in other areas of cortex. Selection can be used whenever several objects must be represented at the same time within the same population of neurons.

The model

DESIGN PRINCIPLES The model provides a framework for studying how the signals carried by cortical neurons could be used to estimate the velocity of moving objects in the presence of occlusion. Our goal is threefold: First, the model should provide a computationally robust algorithm for computing the velocities of moving objects in visual scenes. Second, it should be consistent with the known physiology and anatomy of visual cortex. Third, its performance should agree with psychophysical studies from primates. The processing units in the network model are meant to capture the responses that are observed at the level of the average firing rates of neurons. We have not attempted to account for how these responses are actually synthesized by cortical neurons, nor have we replicated in detail the interactions that occur between neurons. However, the computational operations in the model are relatively simple ones, such as summation and normalization, which can be implemented by real neurons in a variety of ways (see the discussion later in this chapter).

Soft-maximization is an example of a simple computational operation that occurs at several stages of processing in our model. The purpose of soft-maximization is to enhance the firing rate of the unit that has the highest firing rate in a population and to normalize all other responses so that the total activity is a constant, regardless of the initial firing rates. This can be accomplished mathematically by the following function:

$$R_k = e^{\alpha R_k'} / \sum_i e^{\alpha R_i'} \qquad (1)$$

where R_k' are the initial firing rates and R_k are the normalized firing rates in a population of units indexed by k, and α is a constant that determines the degree of separation of the highest firing rate from all the others. The summation occurs here over all units in the population, but the operation can be performed within any subpopulation. In our model, it was applied in three separate subpopulations of units. As the amplification constant α is increased, the differences are further enhanced until, in the limit as α becomes very large, one unit fires at its maximum rate and all the rest are reduced to zero, a limit called *winner-take-all*. The value of α can be different for each use in the model, but once it has been chosen it is fixed for all stimuli. In a more sophisticated model, α could adapt dynamically. (See the discussion later for ways that this soft-maximization operation could be implemented with neural mechanisms found in the cerebral cortex.)

The general design of the model is a cascade of stages, each consisting of an array of processing units that are locally connected and arranged in a roughly retinotopically organized map (figure 27.2). Within each stage, there are a number of layers or channels, each covering the visual field. For example, for each of several directions of movement there will be an array of neurons that together provide an overlapping map of the visual field, in the same way that orientation is represented in area V1. Processing is divided into four main stages, which are summarized here and described in greater detail in Nowlan and Sejnowski (1994a,b). The retinal and motion-energy stages were accomplished by fixed filters, each of which could be computed by a simple feedforward network of converging and diverging connections. The subsequent stages of processing were adaptive filters in the sense that the connection strengths in the model were not predeter-

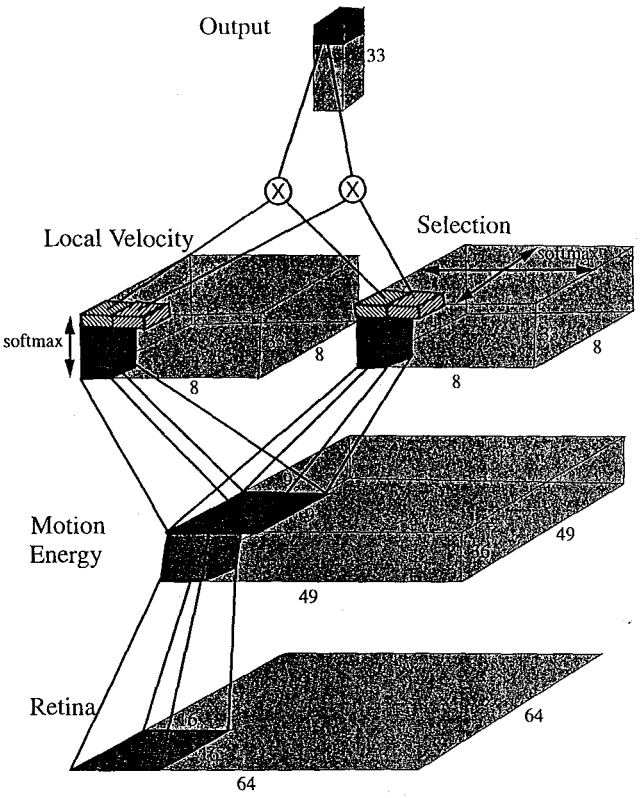

FIGURE 27.2 Architecture of a network model for a small region of area MT. The inputs to the model are "movies" containing 64 frames of 64 × 64–pixel arrays, with each pixel having a range of 256 gray levels. The movies are presented to the retina and processed sequentially by each layer. The first layer of processing, which corresponds to the primary visual cortex area V1, is composed of motion-energy units. Each motion-energy column contains 36 units that receive inputs from a 16 × 16 region of the retina for 16 successive frames of the movie. Each motion-energy unit is tuned to a different combination from four directions of motion, three spatial frequencies, and three temporal frequencies. The next layer of processing contains two independent networks of units, each receiving inputs from a 9 × 9 spatial array of motion-energy columns. Each local velocity and selection column contains 33 units that represent eight different directions and four different speeds as well as one unit for null motion. For each local velocity unit, there is a corresponding selection unit that receives inputs from the same 9 × 9 array of motion-energy columns. Because adjacent motion-energy columns have only partially overlapping receptive fields, the effective receptive field areas of the local velocity and selection columns are 81 times larger than the receptive field area of the motion-energy units. Competition occurs within each local velocity column, as indicated by the double-pointed arrow: Units that are most strongly activation are enhanced, and the weakest are suppressed by soft-maximization. For the selection units, there is a competition across space. Soft-maximization is applied between selection units having the same preferred velocity (double-pointed arrows). The information within the local velocity and selection networks is then combined to form the final estimate of velocities on the output layer, which is a single column of 33 units. The value of each output unit is computed by multiplying the outputs of corresponding pairs of local velocity and selection units and summing across a slab of these units. This output is the final estimate of the velocities of all the objects moving with the 64 × 64–pixel array. The velocity of more than one object may be represented in the array at the same time, as along as the velocities are not too similar. There were a total of 8.8 million weights in the network but, because of translational symmetries, only 138,600 of them were independent.

mined but were instead determined by an optimization procedure. Once the adaptive weights were found, they were fixed, and all the results presented here were obtained from one network with fixed connection strengths.

RETINAL PROCESSING The 64 × 64–pixel input array was roughly equivalent to an array of photoreceptors that represents the intensity at each pixel location by one of 256 gray levels. A motion stimulus consisted of a sequence of images that were processed first through the retinal stage and subsequently through spatiotemporal filters, as described in the next section. The retinal stage of processing contrast-enhanced each image with a difference-of-gaussian filter at each location of the array, which removed the constant component of intensity across the image, smoothed the noise, and enhanced the edges.

MOTION-ENERGY FILTERS In the first stage of motion processing, a distributed representation of motion was extracted that served as a model for the inputs to MT. We used the motion-energy model of Adelson and Bergen (1985), which consisted of arrays of spatiotemporal filters, each tuned to a particular direction of motion and sensitive to a particular combination of spatial and temporal frequency. Altogether, there were 36 channels of motion-energy filters tuned to four directions and nine combinations of spatial and temporal frequencies. Each filter received input from a 16 × 16 patch of the retinal output. These filters were broad and overlapping in their selectivities, and their velocity tuning depended on the spatial characteristics of the moving pattern. Soft-maximization was applied to the pool of 36 motion-energy channels to normalize

their outputs and make them report relative rather than absolute values for contrast to the next stage of processing (Albrecht and Geisler, 1991; Heeger, 1992). The properties of motion-energy filters resembled those of directionally tuned complex cells in the visual cortex of cats and monkeys (Nakayama, 1985; Maunsell and Newsome, 1987; Emerson, Bergen, and Adelson, 1992).

LOCAL VELOCITY NETWORK For a single, rigidly moving object in the visual field with no occluding or transparent objects, relatively simple averaging schemes can be used to estimate the local velocity (Heeger 1987, in press; Grzywacz and Yuille 1990). A linear weighted summation can be performed for each direction of motion and speed and the maximum taken across these channels. In our model, there were eight different best directions corresponding to equally spaced compass points and four different best velocities, for a subtotal of 32 different combinations, plus one more unit that represented zero velocity, giving a total of 33 velocity-tuned units. There was an array of 8 × 8 locations, each containing 33 velocity units, and each of these velocity units received inputs from 9 × 9 motion-energy units (see figure 27.2).

These velocity units were broadly tuned around their best direction and velocity so that a given motion stimulus produced a pattern of activity in these 33 units. The unit with the largest response, representing the most likely local velocity in that patch of the visual field, was enhanced by soft-maximization, which also reduced the weaker responses (equation 1). The output of each unit can be considered the evidence for a particular velocity in a particular region of the image. The constraint that the sum of activity in the pool of 33 neurons must equal one can be interpreted as the constraint that the total evidence across all velocities must sum to one. Using soft-maximization rather than a winner-take-all limit means that the population can represent more than one velocity in each pool.

SELECTION NETWORK For multiple moving objects and visual scenes that include occlusion and transparency, the local velocity estimates may not be accurate, and it is not at all obvious which features of the image are relevant for determining whether a local region contains reliable information. If information from several objects is within the receptive field of a pool of local velocity units, the output will be ambiguous. These locations may provide little or no unambiguous information and should be ignored as long as other parts of the object contain reliable velocity estimates. The purpose of the selection network was to identify the regions of the image that contain the most reliable estimates. The inputs to the selection network were the same motion-energy array used for the local velocity network. There was a selection unit corresponding to each local velocity unit.

Before the selection units were combined pairwise with the local velocity units, the outputs of the directionally selective units in the selection network competed spatially across the visual field. The soft-maximization operation (equation 1) was applied separately to each of the 33 selection channels. The purpose of this comparison was to identify the spatial locations containing the most reliable information and to suppress those locations containing the least reliable information. Finally, the outputs of each channel in the 8 × 8 local velocity network and the 8 × 8 selection network were multiplied, point by point, then summed to produce a final estimate of the velocity:

$$v_k(t) = \sum_{x,y} I_k(x,y,t) S_k(x,y,t) \qquad (2)$$

where $v_k(t)$ is global evidence for a visual target at time t moving at a particular velocity k, the $I_k(x,y,t)$ is the local evidence for velocity k computed by the local 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.

CREATING THE NETWORK The properties of the local velocity and selection units were not set a priori but were determined by an optimization procedure. We specified the input representation of motion to the model, the problem that we believe the system was designed to solve, and a network architecture that reflected some of the important constraints imposed by cortical physiology and anatomy (Churchland and Sejnowski, 1992). We then used an optimization procedure called *mixtures of experts* to adjust the model parameters to solve a velocity estimation problem for prototypical input patterns, as described elsewhere (Nowlan, 1990; Nowlan and Sejnowski, 1994b). The input patterns used for optimizing the performance of the network were 500 "movies" of moving objects, such as the example shown in figure 27.3. The known veloc-

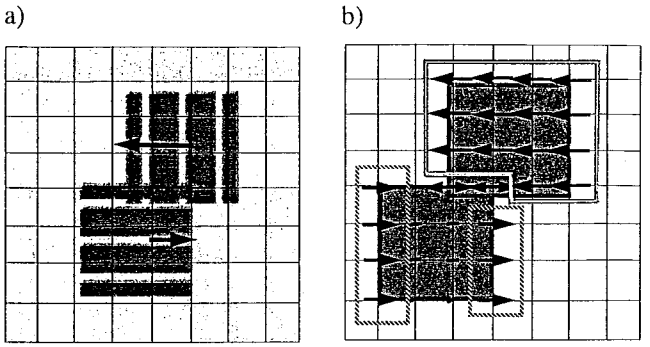

FIGURE 27.3 Responses of local velocity and selection units to partially occluding objects. (a) Input to the model consists of two blobs moving in opposite directions (one to the right at 0.25 pixels per frame and the second to the left at 0.5 pixels per frame). The arrows indicate the direction and speed of motion. (b) Hashed lines indicate the region of motion selected for the rightward-moving blob, whereas the solid lines indicate a separate region of support for the motion of the leftward blob. The local velocity estimates in the region of overlap are ambiguous, showing two directions of motion simultaneously.

ity of each object was used to modify the weights in the appropriate local velocity network, and the weights in the selection networks were changed to identify the most reliable regions.

Although the velocity and selection networks appear to be symmetrical, the optimization procedure treated the selection network differently because its role was to gate the velocity network rather than compute the velocity itself. The optimization procedure modified the input weights to the local velocity and selection units to minimize the error of the output units, but only the local velocity units were given feedback about the correct output velocity; the selection units were instead given information about which local velocity units carried the most reliable estimates. There was also a difference in the way these two networks were normalized: soft-maximization was applied to the local velocity units only within a column but, for the selection network, the soft-maximization was spatially applied separately for each velocity channel. The optimization procedure converged to sets of weights that gave robust estimates of object velocities despite occlusion; more importantly, the model performed well on psychophysical stimuli, such as those presented in the next section, which were not used during optimization (Nowlan and Sejnowski, 1993, 1994a,b).

To illustrate how the selection network handles the problem of segmentation, consider two objects moving in opposite directions (see figure 27.3a). One object is striped horizontally and moves to the right with a speed that is half that of the second object, which is striped vertically and moves to the left. The local velocity estimates and selected regions are shown in figure 27.3b. Selected regions are denoted with a dashed line if they correspond to the rightward motion of the first object or with a stippled line if they correspond to the leftward motion of the second object. In regions where the two objects overlap, the second object occludes the first object totally, and intensity in these regions is the intensity of the second object.

In figure 27.3, the local velocity estimates in the region of overlap of the two objects are marked with double arrowheads. In these regions, the activity in the local velocity pool is not concentrated in a single unit but is distributed bimodally with activity peaks corresponding to two opposing directions of motion. This ambiguous region is not included in the region of support for the motion of either object. Note that for the rightward-moving object, the selected regions correspond to only the leading and trailing edges. For the leftward-moving object, the contrast stripes are perpendicular to the direction of motion, providing strong motion signals over most of the region covered by the object. As a result, the region of support for this object is correspondingly much larger.

The properties of the selection units were determined indirectly by optimizing the network to produce the correct velocity in the final output layer of units. To a first approximation, the selection units in our network detected discontinuities in the distribution of motion-energy inputs but, because they were optimized to respond primarily to the patterns of discontinuities that characterize regions of reliable support for object motion, the algorithm that they implement is more restrictive. Thus, not all discontinuous patterns of velocity activate the selection units equally well (figure 27.4). For example, some units preferred motion end-stopping within their receptive fields. For a given input, the spatial regions selected tended to form disjointed subsets over which to integrate the local velocity field. This allowed the model to account for interpenetrating motion fields for which the assumption of spatial continuity of the velocity field is invalid. We have empirically determined the properties of the local velocity and se-

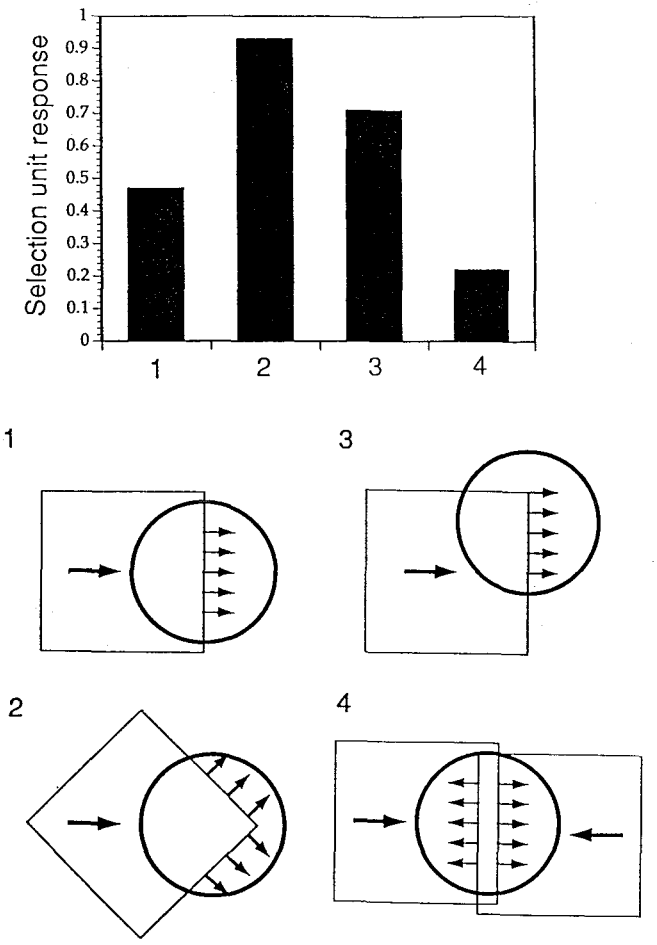

FIGURE 27.4 Responses of a selection unit to motion discontinuities. (a) The strength of the response of a selection unit tuned to rightward motion to four different distributions of motion-energy responses within the selection unit's receptive field. The four motion-energy distributions are represented schematically in (b). In the first three examples, the 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 seen in example 2, where the receptive field encloses a discontinuity between two orthogonal sets of local motion-energy measurements. Finally, example 4 shows that the selection unit is directionally tuned: Local motion-energy responses corresponding to two opposed motions (generated, in this case, by one transparent object moving in front of a second moving object) will suppress the response of a selection unit. In this figure, the responses of a single isolated selection unit are shown. The overall response of a selection unit in the network is determined by its own local response and the responses of similarly tuned selection units in other regions of the image with which a selection unit competes.

lection units by presenting the model with a variety of standard motion stimuli, such as oriented gratings.

Unit response properties

The first stage of our model is intended to correspond approximately to primary visual areas V1 and V2. The choice of normalized motion-energy responses as the representation of image motion in this first stage of the model reflects the measured response properties of simple and complex cells in mammalian primary visual cortex (Tolhurst and Movshon, 1975; Holub and Morton-Gibson, 1981; Emerson et al., 1987; McLean and Palmer, 1989). The spatial and temporal response characteristics of these filters and the broader tuning in temporal frequency compared to spatial frequency were chosen based on measured responses in primary visual cortex (Adelson and Bergen, 1985; Heeger, in press). The inverse relationship between motion-energy filter center frequencies and receptive field sizes also matches the relationship found in visual cortex (Hochstein and Shapley, 1976; Maffei and Fiorentini, 1977; Andrews and Pollen, 1979). The normalization of the motion-energy responses is suggested by the saturating contrast response curves of cells in primary visual areas (Ohzawa, Sclar, and Freeman, 1985; Heeger, 1992).

The local velocity and selection stages of the model we associate primarily with the middle temporal cortical area (MT) of visual contex, although some of the functions in the model may be occurring in the medial superior temporal area (MST) and possibly parietal cortex. Although the response properties of units in these stages were not specified in advance, a number of architectural decisions constrained the model. The localized receptive fields and roughly retinotopic organization of the units in both the first and second stages of the model are matched by the organization of visual cortical areas in the early stages of visual processing

(Maunsell and Newsome, 1987). In particular, the area of the receptive fields for units in the selection and local velocity layers of the model were 81 times larger than receptive fields in the motion-energy layer. The receptive fields in MT are, on average, 100 times larger than those found in V1 (Gattass and Gross, 1981), and the spatial scale of directional responses is proportionally larger (Mikami, Newsome, and Wartz, 1986). We have also constrained the feedforward "weights" of units in both the local velocity and selection pathways to be purely excitatory. Thus, the properties of the units in these networks are due primarily to excitatory inputs, with inhibitory effects being expressed only in the competitive renormalization used in all parts of the model.

The renormalization used in our model has been suggested previously to account for some aspects of neural responses in both V1 (Heeger, 1992) and MT (Snowden et al., 1991). The architecture proposed in our model makes two strong assumptions about the nature of this normalization in areas such as MT. The local velocity pathway requires competitive interactions among units with similar receptive field locations but different tuning properties (this is similar to the interactions proposed by Heeger, 1992, and Snowden et al., 1991). The selection pathway requires interactions among cells covering most of the visual field, but these long-range interactions occur only among cells with similar tuning properties. Neuronal mechanisms that might be used to implement the competitive soft-maximization operation are examined later in this chapter.

We explored the response characteristics of selection and local velocity units in the optimized model using drifting sinusoidal gratings. The local velocity units all exhibited very similar tuning curves, which tended to be symmetrical about the optimal direction of response and narrowly tuned. The average tuning bandwidth for these units was 53°. The tuning curves for selection units showed considerably more variation. Selection units tend, on average, to have much broader tuning curves (average tuning bandwidth, 84°). A small number of selection units showed a bimodal directional tuning curve. In addition, selection units tended to have maximal orientation responses for bars oriented close to their preferred direction of motion, whereas local velocity units showed maximal responses for bars nearly orthogonal to the preferred direction of motion. The broader directional tuning and similarity of orientation and direction tuning suggests that the selection units resemble the pattern (Movshon et al., 1985) or type II cells (Albright, 1984) found in monkeys, whereas the local velocity units are more similar to the component or type I cells. The responses of the local velocity and selection units to plaid patterns also supports this identification (Nowlan and Sejnowski, 1994a,b). The directional tunings of both local velocity and selection units are sharper than the tunings found by Albright (1984) but similar to those reported by Maunsell and Van Essen (1983).

We tested the units in the model for velocity tuning with gratings oriented optimally for each unit and spanning a range of temporal and spatial frequencies. The velocity units were tuned to a fairly narrow range of velocities over a broad range of spatial and temporal frequencies, and some cells with this type of spatiotemporal frequency response have been found in MT (Newsome, Gizzi, and Movshon, 1983). In general, the units in our model were tuned to velocity over a broader range of spatial and temporal frequencies than typically found in MT, perhaps because there were many fewer units than MT cells to represent the same range. Thus, a single local velocity unit may represent a population of cells in MT, each of which is sensitive to velocity over a narrower range of spatial and temporal frequencies.

The spatial frequency and temporal frequency sensitivities were more separable for selection units than for the local velocity units because they could be better approximated by products of purely spatial and purely temporal functions. In addition, the response of the selection unit showed a slight dip in the midrange frequencies that is not seen in velocity unit responses. Nearly all the selection units showed some degree of sensitivity of their velocity tuning to variation in spatial frequency. Many MT neurons also show some sensitivity of their velocity tuning to spatial frequency (Maunsell and Newsome, 1987), but this variation has not been systematically correlated with other types of variation (such as tuning width).

There is another important qualitative difference in the responses of local velocity and selection units that was apparent when gratings were either restricted to the receptive field region or presented across the entire visual field. The responses of selection units were strongly suppressed when the grating was presented to the entire visual field rather than just to the receptive field of a unit (figure 27.5a). The responses of local

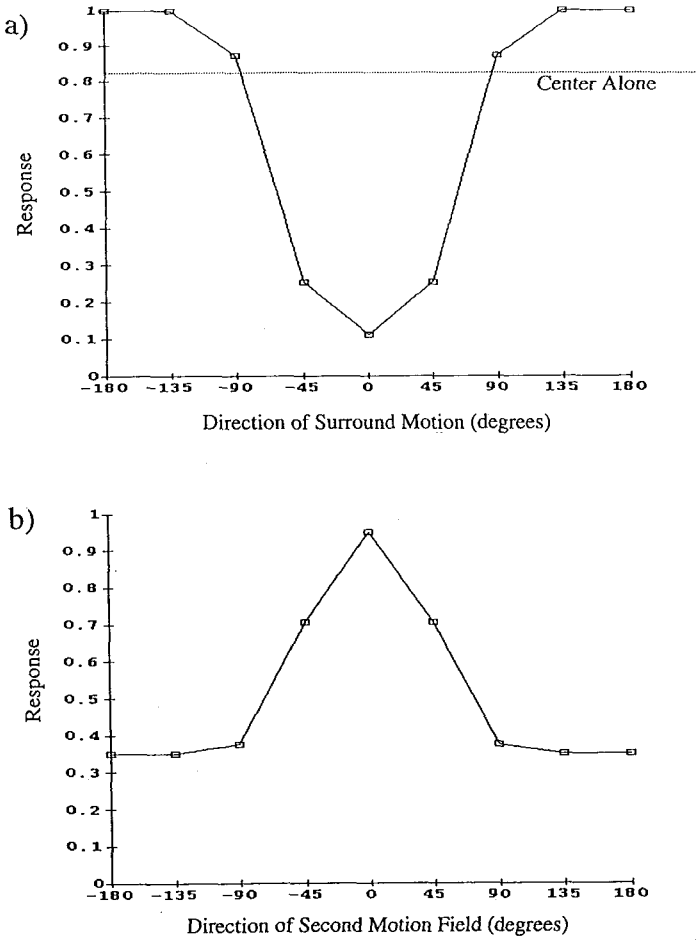

FIGURE 27.5 Interactions between units in the selection and velocity networks. (a) Response of selection unit to an optimal sinusoidal grating when a second grating is presented outside the receptive field of the selection unit. The dashed line shows the response to the grating within the receptive field alone. The response is displayed as a function of the orientation of the surround grating, measured relative to the orientation of the grating within the receptive field. When the surround motion matches the receptive field motion, there is strong inhibition of responses. This inhibition decreases as the surround and receptive field motions differ more in direction, becoming weak facilitation when the difference in orientation is greater than 90°. (b) Response of local velocity unit to a dynamic transparent random-dot stimulus. The stimulus consisted of two fields of random dots moving within the receptive field. The first field was chosen so its motion produced a maximal response from the unit. The direction of the second field relative to the first was adjusted, and the response of the unit was plotted as a function of the direction of this second field of dots. The presence of a second motion always suppressed the response of the unit, with this suppression saturating for motions orthogonal to the preferred direction of motion of the unit (at which point the local velocity pool has a bimodal activity distribution).

velocity units, in comparison, were weakly facilitated by whole-field presentation. The suppression of response due to motion in regions surrounding the selection unit's receptive field was strongest when the surrounding motion matched the optimal motion for eliciting a response in the selection unit's receptive field. The degree of suppression decreased as the surround motion differed more from the optimal motion for the unit (see figure 27.5a). In fact, motion in a direction more than 90° away from optimal for the selection unit tended to facilitate the response of the selection unit. This is a consequence of the shift in balance that occurs in the soft-maximization operation at different spatial locations. This type of modulation of the receptive field response by the nonclassical surround in MT was first reported by Allman, Miezin, and McGuinnes, 1985 (see the discussion later).

The substructure of the receptive fields in MT were studied by Snowden and colleagues (1991), whose stimuli consisted of random dots moving within the classical receptive field. One field of dots always moved in the preferred direction of the cell, whereas a second field, if present, moved in a different direction, producing two distinct motions simultaneously within the cell's receptive field. We presented similar stimuli to the local velocity and selection units in our model. The local velocity units exhibited responses that are qualitatively very similar to the responses observed by Snowden's group (figure 27.5b). The presence of a second motion in the receptive field always suppressed the response of a local velocity unit, with the degree of suppression increasing until the second motion was orthogonal to the preferred direction of the unit. These response characteristics are explained by the local competition within each pool of velocity units.

Transparent plaids

Moving plaid patterns have been used to study how the visual system integrates multiple motion signals into a coherent motion percept (Adelson and Movshon, 1982; Ramachandran and Cavanaugh, 1987; Welch, 1989; Stoner, Albright, and Ramachandran, 1990; Wilson, Ferrera, and Yo, 1992). These stimuli consist of two independently moving gratings that are superimposed (figure 27.6a). When human observers are presented with either grating alone, they always reliably report the motion of the grating. However, if the two gratings have similar properties and are superim-

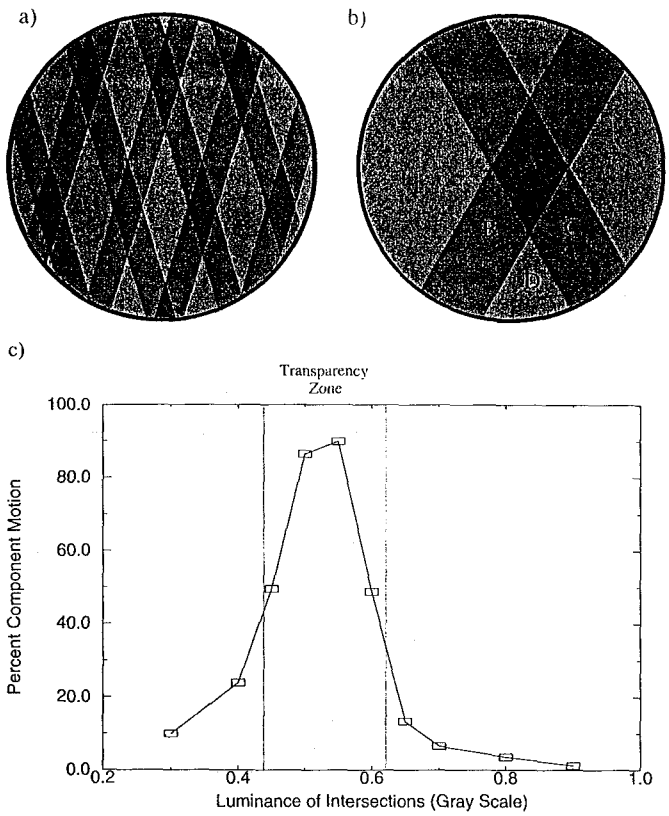

FIGURE 27.6 Plaid pattern stimuli used to test the model on transparency. (a) The plaid patterns were composed of two square-wave gratings of dark bars on a lighter background (Stoner, Albright, and Ramachandran, 1990). (b) Detail of pattern in (a). In the experiments, the intensity of region A was varied, whereas the intensity of regions B, C, and D were held constant. Depending on the luminance of A, the observer would see either two drifting gratings or a single coherent pattern motion. (c) Responses of model to similar plaid patterns. There was a transparency zone, in which the model responses were consistent with two separate component motions. Outside this transparency zone, the model reported only a single pattern motion, consistent with human perception.

posed, the two independent motions of the gratings seem do disappear; most observers see the two gratings cohere and form a single pattern moving in a direction different from either of the gratings alone. Cells that respond in the direction of pattern rather than component motion are found in MT, which suggests the importance of MT in producing the percept of coherent motion (Movshon et al., 1985; Rodman and Albright, 1989).

Stoner, Albright, and Ramachandran (1990) have found that the percept of coherent pattern motion can be affected by altering the luminance of the region of intersection of the two gratings. Some cells in MT also tend to respond to either the direction of pattern or component motion, depending on the luminance of this intersection region (Albright, 1992). The stimuli used by Stoner, Albright, and Ramachandran consisted of two square-wave gratings with thin bars on a lighter background (figure 27.6a). The luminance of the intersection regions was varied, whereas the luminance of the bars and the background was left the same. Stoner and colleagues (1990) manipulated the luminance of the intersection regions and showed that there is a range of luminances for which humans see one transparent grating lying on top of the other and reliably report the presence of the two grating motions rather than the coherent pattern motion.

We presented the model with a series of plaid patterns consisting of two square-wave gratings in which the luminance of the intersection region of the plaid pattern was varied systematically through a series of values that spanned the transparency zone. The response of the output units in the model varied nonlinearly and nonmonotonically with the luminance of the intersection region. As shown in figure 27.6c, the performance of the model closely matched the psychophysical results reported by Stoner, Albright, and Ramachandran (1990). We analyzed the network to determine how it made its decision and found that when it reported component motion, the total support for motion was concentrated on the portions of the gratings that were outside the intersection regions but, during pattern motion, support for motion was concentrated on the regions surrounding the intersections of the two gratings (Nowlan and Sejnowski, 1993). This is another example of how the selection network responds to patterns of motion-energy discontinuities (see figure 27.4).

Discussion

In our model of motion processing in MT, the conflicting demands for spatial averaging and spatial segmentation were satisfied by two separate networks, one that computes the local velocity estimate and a second that selects regions where reliable velocity estimates are possible. The outputs of these two networks were combined multiplicatively to produce reliable estimates of the velocities of objects without assuming spatial continuity. In our model, there was only a single population

of output units representing the velocities in a small patch of the visual field. In MT, there is an array of such units that can, in turn, serve as the input to other areas that represent nonuniform flow fields, such as expansion and rotation, which are preferred by neurons in MST (Saito et al., 1986; Duffy and Wurtz, 1991).

ROBUST ESTIMATION Many models of motion processing perform spatial averaging by assuming the spatial continuity of the velocity field (Marr and Ullman, 1981; Adelson and Movshon, 1982; Hildreth, 1984; Heeger, 1987; Grzywacz and Yuille, 1990). The problems for this approach posed by occlusion and transparency in motion stimuli were overcome in our model by the selection network, which estimated not the local velocity but rather the confidence with which the local velocity could be measured. Because each spatial region is considered to be independent, parts of different objects that interpenetrate are not averaged. The local regions of support for each velocity are then combined by a global competitive mechanism. Our selection pathway can be regarded as a feedforward mechanism for computing regions of support for robust velocity estimation (Li, 1985).

The selection units respond primarily to motion-energy gradients. Velocity gradients have been used by Koch, Wang, and Mathur (1989) to determine the boundaries between regions with different uniform velocities and by Smith and Grzywacz (1993) to determine where to apply a winner-take-all operation. In our model, the selection units form a pattern recognition network that weights motion-energy patterns in a graded fashion according to their degree of robustness; the spatial competition imposed by soft maximization ensures that the most reliable regions gain the strongest support (see figure 27.4).

SELECTIVE ATTENTION Selection may represent a fundamental aspect of cortical processing that occurs with many preattentive phenomena (Bergen and Julesz, 1983; Treisman, 1988). The same mechanisms that are used to implement the covert, preattentive form of selection in our model could also be used for overt attentional processing. Top-down influences could enhance the probability that a selection is made to a particular property of the input. Attentional modulation of single-unit responses has not been reported in MT, but other motion areas such as MST may be better candidates. Moran and Desimone (1985) have observed in area V4 that the response of a neuron to its preferred stimulus is reduced if the monkey attends to a nonpreferred stimulus but only if the nonpreferred stimulus is presented within the receptive field for the neuron. This is evidence for the type of local competition that is required within our local velocity network.

The inhibitory effects of conflicting motion signals within the receptive field of local velocity units in our model are similar to effects found by Snowden and colleagues (1991) for some MT cells. The suppression reached a maximum when the second field of dots was roughly orthogonal to the preferred direction of the cell and was relatively constant after that. The soft-maximization operation used in the model produced precisely this type of suppression; increasing the number of dots in the nonpreferred direction causes the slope of the response as a function of dot density in the preferred direction to decrease, but the response always saturates at the same level. The competition among the local velocity units in the model makes the prediction that a second random-dot pattern moving in a cell's preferred direction, but at a speed significantly different from the optimal speed for the cell, will also have a suppressive effect on the cell's response to an optimal stimuli.

NONCLASSICAL SURROUNDS The inhibitory effects of surround motion on selection units in our model are very similar to inhibitory surround effects that have been observed in many cells in MT (Allman, Miezin, and McGuinnes, 1985; Tanaka et al., 1986). The effect of whole-field motion on the responses of selection units in our model derives from the competition among the selection units representing the same candidate velocity across all regions of the visual field. When evidence for a particular velocity is present in all regions of the image equally, very little support needs to be assigned to any one region. Thus, the presence of similar motion in surrounding regions tended to suppress the selection units. On the other hand, if support for the velocity were concentrated in a small region of the image, the selection units in these regions had much stronger responses.

Recently, Born and Tootell (1992) have used the 2-deoxyglucose technique to identify the spatial organization of two broad classes of cell in MT based on how these cells responded to whole-field motion. *Band*

cells responded to some directions of whole-field motion, whereas *interband* cells showed no strong response to any direction of whole-field motion. If we identify the interband cells of Born and Tootell with our selection units, the model makes the strong prediction that long-range intrinsic connections will occur primarily between groups of interband cells. The long-range horizontal connections between pools of selection units may modulate local inhibitory circuits.

SOFT-MAXIMIZATION A soft-maximization operation can be implemented by local mutual inhibition between the units in the population (Feldman and Ballard, 1982; Heeger, 1992), but other circuits that are faster and more reliable also can perform the task (Grzywacz and Yuille, 1990). An effective way to control the gain of the soft-maximization operation, the value of α required in equation 1, is to control local amplification. Recurrent excitation within networks of cortical pyramidal neurons appears to amplify inputs and the degree of amplification is controlled by inhibitory circuits (Douglas, Martin, and Whitteridge, 1989). We have assumed that the average firing rate of a neuron carries the output signal and that the neurons fire asynchronously. Synchronization of the spike firing among a pool of cells would also enhance their impact on mutual postsynaptic targets and, in principle, could be used to implement a fast soft-maximization operation (Steriade, McCormick, and Sejnowski, 1993).

RANDOM DOTS We have also applied our model of motion processing to dynamic random-dot displays (Nowlan and Sejnowski, 1994a,b), where there are no regions of consistent motion and the velocity of neighboring dots can be very different (Braddick, 1974; Morgan and Ward, 1980; Nakayama and Tyler, 1981; Williams and Sekuler, 1984). Newsome, Britten, and Movshon, (1989) computed psychometric functions for the ability of a monkey to identify correctly the direction of motion as a function of the percentage of dots moving coherently and showed that these psychometric functions were closely matched by "neurometric" functions computed from the responses of single MT neurons. Some neurons carried signals that were as reliable as the behavior of the monkey. When we used similar dynamic random-dot stimuli in our model, the percent of correct responses plotted as a function of coherence level for the model was qualitatively similar to the psychometric response curve from neurons in MT. The threshold for the model depended on the size of the coherent region. Nonetheless, the model was able to process this type of motion display properly even though it was not used to optimize the original model.

Predictions

In this chapter, we have introduced a novel strategy for computing the reliability of local velocity estimates. The purpose of the selection units, although they are velocity-tuned, is not to represent the local velocity. Likewise, the fact that a neuron is velocity-tuned is not sufficient evidence to conclude that the neuron's function is to represent local velocity, which suggests that some neurons in the visual cortex may be more concerned with grouping information than with representing the information itself. A similar algorithm may be used in other regions of sensory cortex to assess the importance of information processing within local neighborhoods. For example, selection networks could be used in binocular vision to assess the reliability of stereoscopic correspondences and also to group nearby regions of the visual field that contain parts of the same object based on similar binocular disparities, even when there are transparencies.

Selection in our model occurred in two stages. First, the selection network at each spatial location rapidly computed a selection value for each broadly tuned velocity unit. This feedforward operation was followed by soft-maximization normalization across spatial locations for each velocity. The intrinsic horizontal axonal system within neocortex is the most likely substrate for this process. We predict that one function of these intrinsic collaterals is to compare the relative activity levels within different columns. The physiological effects of these collaterals should, in some circumstances, be the suppression (rather than enhancement) of activity in neighboring columns.

The selection network provides a partial solution to the problem of image segmentation. Previous attempts to segregate figure from ground have implicitly assumed that objects were spatially continuous and that the first step was to find a bounding contour. Our approach to segmentation does not make this assumption; the selection network may group information that is spatially separated by intervening ground or by other objects. This leaves open the problem of how motion is integrated with other properties of the object. Integration could be achieved by referencing each selected esti-

mate back to a high-resolution spatial map, such as area V1, which predicts that the cortical feedback projections carry information about segmentation and that neural correlates of segmentation should be observed in single neurons in area V1. Visual stimuli to test this hypothesis would allow comparison of the same image over the receptive field of a neuron while it is part of either a figure or the background.

In contrast to previous approaches that have attempted to construct a motion flow field throughout space or for all parts of an object (Marr, 1982), we attempt only to represent explicitly those selected parts that are particularly salient and unambiguous. Separate regions that share the same local velocity will be grouped and assigned to one object. Such reduced representations may also have advantages for indexing object representations in the ventral processing stream. With fewer salient features to match, the combinatorial problem of finding the correct match is greatly simplified. Reduced representations may also be helpful in learning the causal relationships between representations as salience is already a part of the representation (Ballard and Whitehead, 1990; Churchland, Ramachandran, and Sejnowski, 1994).

REFERENCES

ADELSON, E. H., and J. R. BERGEN, 1985. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. [A]* 2:284–299.

ADELSON, M., and J. A. MOVSHON, 1982. Phenomenal coherence of moving visual patterns. *Nature* 300:523–525.

ALBRECHT, D. G., and W. S. GEISLER, 1991. Motion sensitivity and the contrast-response function of simple cells in the visual cortex. *Visual Neurosci.* 7:531–546.

ALBRIGHT, T. D., 1984. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* 52:1106–1130.

ALBRIGHT, T. D., 1992. Form-cue invariant motion processing in primate visual cortex. *Science.* 255:1141–1143.

ALLMAN, J., F. MIEZIN, and E. MCGUINNES, 1985. Stimulus-specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu. Rev. of Neurosci.* 8:407–430.

ANDREWS, B. W., and T. A. POLLEN, 1979. Relationship between spatial frequency selectivity and receptive field profile of simple cells. *J. Neurophysiol. (Lond.)* 287:163–176.

BALLARD, D. H., and S. D. WHITEHEAD, 1990. Active perception and reinforcement learning. *Neural Computation* 2: 409–419.

BERGEN, J. R., and B. JULESZ, 1983. Rapid discrimination of visual patterns. *IEEE Trans. Systems Man Cybern.* 13:857.

BORN, R. T., and R. B. H. TOOTELL, 1992. Segregation of global and local motion processing in primate middle temporal visual area. *Nature* 357:497–500.

BRADDICK, O. J., 1974. A short-range process in apparent motion. *Vision Res.* 14:519–527.

BRADDICK, O. J., 1993. Segmentation versus integration in visual motion processing. *Trends Neurosci.* 16:263–268.

CHURCHLAND, P. S., V. S. RAMACHANDRAN, and T. J. SEJNOWSKI, 1994. A critique of pure vision. In *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. Davis, eds. Cambridge, Mass.: MIT Press.

CHURCHLAND, P. S., and T. J. SEJNOWSKI, 1992. *The Computational Brain.* Cambridge, Mass.: MIT Press.

DOUGLAS, R. J., K. A. C. MARTIN, and D. WHITTERIDGE, 1989. A canonical microcircuit for neocortex. *Neural Computation* 1:480–488.

DUFFY, C. J., and R. H. WURTZ, 1991. Sensitivity of MST neurons to optic flow stimuli: II. Mechanisms of response selectivity revealed by small-field stimuli, *J. Neurophysiol.* 65:1346–1359.

EMERSON, R. C., J. R. BERGEN, and E. H. ADELSON, 1992. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Res.* 32:203–218.

EMERSON, R. C., M. C. CITRON, W. J. VAUGHN, and S. A. KLEIN, 1987. Nonlinear directionally selective subunits in complex cells of cat striate cortex. *J. Neurophysiol.* 58:33–65.

FELDMAN, J., and D. BALLARD, 1982. Connectionist models and their properties. *Cogn. Sci.* 6:205–254.

GATTASS, R., and C. G. GROSS, 1981. Visual topography of striate projection zone (MT) in the posterior superior temporal sulcus of the macaque. *J. Neurophysiol.* 46:621–638.

GRZYWACZ, N. M., and A. L. YUILLE, 1990. A model for the estimation of local image velocity by cells in the visual cortex. *Proc. R. Soc. Lond. [Biol.]* 239:129–161.

HEEGER, D. J., 1987. Model for the extraction of image flow. *J. Opt. Soc. Am. [A]* 4:1455–1471.

HEEGER, D. J., 1992. Normalization of cell responses in cat striate cortex. *Visual Neurosci.* 9:181–198.

HILDRETH, E. C., 1984. *The Measurement of Visual Motion.* Cambridge, Mass.: MIT Press.

HOCHSTEIN, S., and R. M. SHAPLEY, 1976. Quantitative analysis of retinal ganglion cell classifications. *J. Physiol. (Lond.)* 262:237–264.

HOLUB, R. A., and M. MORTON-GIBSON, 1981. Response of visual cortical neurons of the cat to moving sinusoidal gratings: Response-contrast functions and spatiotemporal integration. *J. Neurophysiol.* 46:1244–1259.

HORN, B. K. P., and B. G. SCHUNK, 1981. Determining optical flow. *Artif. Intell.* 17:185–203.

KOCH, C., H. T. WANG, and B. MATHUR, 1989. Computing motion in the primate's visual system. *J. Exp. Biol.* 146: 115–139.

LI, G., 1985. Robust regression. In *Exploring Data, Tables,*

Trends and Shapes, D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds. New York: Wiley.

MAFFEI, L., and A. FIORENTINI, 1977. Spatial frequency rows in the striate visual cortex. *Vision Res.* 17:257–264.

MARR, D., 1982. *Vision*. New York: W. H. Freeman.

MARR, D., and S. ULLMAN, 1981. Directional selectivity and its use in early visual processing. *Proc. R. Soc. Lond. [Biol.]* 211:151–180.

MAUNSELL, J. H. R., and W. T. NEWSOME, 1987. Visual processing in monkey extrastriate cortex. *Annu. Rev. Neurosci.* 10:363–401.

MAUNSELL, J. H. R., and D. C. VAN ESSEN, 1983. Functional properties of neurons in the middle temporal visual area (MT) of the macaque monkey: I. Selectivity for stimulus direction, speed and orientation. *J. Neurophysiol.* 49:1127–1147.

MCLEAN, J., and L. A. PALMER, 1989. Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat. *Vision Res.* 29:675–679.

MIKAMI, A., W. T. NEWSOME, and R. H. WURTZ, 1986. Motion selectivity in macaque visual cortex: II. Spatiotemporal range of directional interactions in MT and V1. *J. Neurophysiol.* 55:1328–1339.

MORAN, J., and R. DESIMONE, 1985. Selective attention gates visual processing in the extrastriate cortex. *Science* 229: 782–784.

MORGAN, M. J., and R. WARD, 1980. Conditions for motion flow in dynamic visual noise. *Vision Res.* 20:431–435.

MOVSHON, J. A., E. H. ADELSON, M. S. GIZZI, and W. T. NEWSOME, 1985. The analysis of moving visual patterns. In *Pattern Recognition Mechanisms*, C. Chagas, R. Gattass, and C. Gross, eds. New York: Springer-Verlag, pp. 117–151.

NAGEL, H. H., 1987. On the estimation of optical flow: Relations between different approaches and some new results. *Artif. Intell.* 33:299–324.

NAKAYAMA, K., 1985. Biological image motion processing: A review. *Vision Res.* 25:625–660.

NAKAYAMA, K., and C. W. TYLER, 1981. Psychophysical isolation of movement sensitivity by removal of familiar position cues. *Vision Res.* 21:427–433.

NEWSOME, W. T., K. H. BRITTEN, and J. A. MOVSHON, 1989. Neuronal correlates of a perceptual decision. *Nature* 341: 52–54.

NEWSOME, W. T., M. S. GIZZI, and J. A. MOVSHON, 1983. Spatial and temporal properties of neurons in macaque MT. *Invest. Ophthalmol. Vis. Sci.* 24:106.

NOWLAN, S. J., 1990. *Competing experts: An experimental investigation of associative mixture models.* (Tech. Rep. No. CRG-TR-90-5). University of Toronto, Toronto, Canada: Department of Computer Science.

NOWLAN, S. J., and T. J. SEJNOWSKI, 1993. Filter selection model for generating visual motion signals. In *Advances in Neural Information Processing Systems*, Vol. 5, S. J. Hanson, J. D. Cowan, and C. L. Giles, eds. San Mateo, Calif.: Morgan Kaufmann, pp. 369–376.

NOWLAN, S. J., and T. J. SEJNOWSKI, 1994a. Model of motion processing in area MT of primates. *J. Neurosci.* in press.

NOWLAN, S. J., and T. J. SEJNOWSKI, 1994b. Filter selection model for motion segmentation and velocity integration. *J. Opt. Soc. Am.* in press

OHZAWA, I., G. SCLAR, and R. D. FREEMAN, 1985. Contrast gain control in the cat's visual system. *J. Neurophysiol.* 54: 651–667.

RAMACHANDRAN, V. S., and P. CAVANAUGH, 1987. Motion capture anisotropy. *Vision Res.* 27:97–106.

RODMAN, H. R., and T. D. ALBRIGHT, 1989. Single-unit analysis of pattern-motion selective properties in the middle temporal visual area (MT). *Exp. Brain. Res.* 75:53–64.

RUBIN, N., and S. HOCHSTEIN, 1993. Isolating the effect of one-dimensional motion signals on the perceived direction of moving two-dimensional objects. *Vision Res.* 33:1385–1396.

SAITO, H., M. YUKIE, K. TANAKA, K. HIKOSAKA, Y. FUKADA, and E. IWAI, 1986. Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *J. Neurosci.* 6:145–157.

SMITH, J. A., and N. M. GRZYWACZ, 1993. A local model for transparent motions based on spatio-temporal filters. In *Computation and Neural Systems*, J. Bower and F. Eeckman, eds. Norwell, Mass.: Kluwer Academic.

SNOWDEN, R. J., S. TREUE, R. G. ERICKSON, and R. A. ANDERSEN, 1991. The response of area MT and V1 neurons to transparent motion. *J. Neurosci.* 11:2768–2785.

STERIADE, M., D. MCCORMICK, and T. J. SEJNOWSKI, 1993. Thalamocortical oscillations in the sleeping and aroused brain. *Science.* 262:679–685.

STONER, G. R., T. D. ALBRIGHT, and V. S. RAMACHANDRAN, 1990. Transparency and coherence in human motion perception. *Nature* 344:153–155.

TANAKA, K., H. HIKOSAKA, H. SAITO, Y. YUKIE, Y. FUKADA, and E. IWAI, 1986. Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *J. Neurosci.* 6:134–144.

TOLHURST, D. J., and J. A. MOVSHON, 1975. Spatial and temporal contrast sensitivity of striate cortical neurons. *Nature* 257:674–675.

TREISMAN, A., 1988. Features and objects: The fourteenth Bartlett memorial lecture. *Q. J. Exp. Psychol. [A]* 40:201.

WELCH, L., 1989. The perception of moving plaids reveals two motion-processing stages. *Nature* 337:734–736.

WILLIAMS, D. W., and R. SEKULER, 1984. Coherent global motion percepts from stochastic local motions. *Vision Res.* 24:55–62.

WILSON, H. R., V. P. FERRERA, and C. YO, 1992. A psychophysically motivated model for two-dimensional motion perception. *Visual Neurosci.* 9:79–97.