Perception & Psychophysics
2002, 64 (8), 1191-1217

An oscillatory correlation model of visual motion analysis

ERDOGAN CESMELI
Ohio State University, Columbus, Ohio

DELWIN T. LINDSEY
Ohio State University, Mansfield, Ohio

and

DELIANG WANG
Ohio State University, Columbus, Ohio

We describe and evaluate a model of motion perception based on the integration of information from two parallel pathways: a *motion pathway* and a *luminance pathway*. The motion pathway has two stages. The first stage measures and pools local motion across the input animation sequence and assigns reliability indices to these pooled measurements. The second stage groups locations on the basis of these measurements. In the luminance pathway, the input scene is segmented into regions on the basis of similarities in luminance. In a subsequent integration stage, motion and luminance segments are combined to obtain the final estimates of object motion. The neural network architecture we employ is based on LEGION (locally excitatory globally inhibitory oscillator networks), a scheme for feature binding and region labeling based on oscillatory correlation. Many aspects of the model are implemented at the neural network level, whereas others are implemented at a more abstract level. We apply this model to the computation of moving, uniformly illuminated, two-dimensional surfaces that are either opaque or transparent. Model performance replicates a number of distinctive features of human motion perception.

A central problem in computational vision, as well as in understanding human motion perception, is the estimation of object velocity by selective integration of local motion signals. Although there have been a number of approaches suggested for solving it, several aspects of the problem have not yet been fully worked out. The present paper seeks to address some of these problems. One problem concerns the fact that moving borders and textures do not stimulate single populations of motion sensors, tuned to detect a particular object velocity. Instead, the contours and surface textures of moving objects stimulate many separate populations of local motion sensors tuned to different velocities. Furthermore, it is well known that human observers can perceive motion transparency; that is, they can perceive motion simultaneously in two or more directions when a translucent object moves in front of and across the surface of another moving object. Thus, motion integration processes must be able to appropriately partition those sensor responses that are associated with each moving surface, even though they occupy the same spatial position. Otherwise, only a single motion estimate would be associated with the regions of transparent overlap. How does one reliably compute global object velocity from these estimates? A second problem in motion integration concerns the fact that many objects consist of untextured, homogeneous surfaces. Uniform surfaces convey no information about motion. It is not clear, therefore, how these regions are labeled for motion. This problem, known as the *blank-wall problem* (Simoncelli, 1993), requires a motion integration process that will somehow "fill in" these silent regions with motion estimates that are based on sensor responses obtained in the border or textured regions of the object.

In this paper, we will describe a model of motion perception that attempts to deal with the velocity estimation and blank-wall problems described above. The model consists of two initially parallel segmentation processes or pathways: a *motion pathway* and a *luminance pathway*. The two pathways converge at an integration stage, where static luminance information is used to iteratively guide and constrain the final segmentation and velocity-labeling processes. The impetus for this design comes from a number of studies that suggest that motion integration in humans depends not only on the outputs of local motion sensors, but also on a variety of nonmotion cues that label overlapping surfaces for relative depth (Shimojo, Silverman, & Nakayama, 1989; Trueswell & Hayhoe, 1993).

This work was supported by NSF Grant SRB-9514522 to D.T.L. and by NSF Grant IIS-0081058 and AFOSR Grant F49620-01-1-0027 to D.L.W. Correspondence concerning this article should be addressed to D. T. Lindsey, Department of Psychology, Ohio State University, 203 Lazenby Hall, Columbus, OH 43210 (e-mail: lindsey.43@osu.edu).

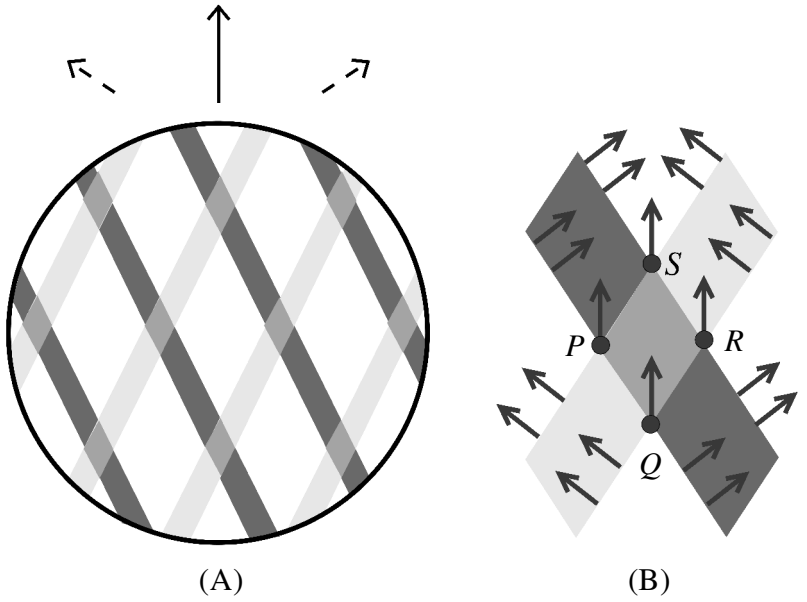

Figure 1. Perceived motion in plaids depends on the luminance of the intersection regions of gratings. (A) A sample plaid stimulus, consisting of two rectangular wave gratings moving on a white background. In the original study by Stoner, Albright, and Ramachandran (1990), the plaids were symmetrical—that is, the bars in the two gratings were assigned identical gray values. Intersection luminances usually differed from bar luminances. When intersection luminances were inconsistent with phenomenal transparency, grating components cohered, and a single, unified moving plaid was seen (solid arrow). However, when intersection luminances were consistent with transparency, plaid dissociated into two moving gratings (dotted arrows). Lindsey and Todd (1996) later employed asymmetric plaids, where the gratings have different luminances, as is shown in panel A. (B) A fragment of the plaid moving upward in panel A. If motion analysis is confined to local regions of the plaid, different regions give different estimates of velocity. Along grating boundaries, local motions are perpendicular to either of the two local boundary orientations. Intersection points, locations *P*, *Q*, *R*, and *S*, have local motions parallel to the plaid motion. These different estimates generally do not depend critically on the intersection luminances in the plaid. However, perceptual integration of estimates does depend on intersection luminances.

Among these are cues based on an analysis of the relative luminances of juxtaposed regions in a scene. Stoner, Albright, and Ramachandran (1990), for example, created moving plaid stimuli by superposition of two rectangular wave gratings moving at different orientations (see Figure 1). The gratings consisted of gray bars of identical luminance moving on a white background. Stoner et al. manipulated the luminance of the regions in the plaid where the gray bars intersect. If the luminance of the intersection regions was consistent with phenomenal transparency, the moving plaid dissociated into two gratings sliding transparently past one another. Otherwise, the grating components unified perceptually, and a single moving plaid was seen. Although models of motion integration and segmentation based solely on motion energy can account for some of the phenomena observed in transparent plaids (Wilson, Ferrera, & Yo, 1992), local motion alone cannot account for all of the phenomena (Lindsey & Todd, 1996; Stoner & Albright, 1996). Stoner and Albright (e.g., 1993) have suggested that transparent plaid stimuli reveal the action of motion integration processes that are partitioned, or "gated," across the two grating components if the luminance analysis is consistent with global phenomenal transparency. Otherwise, all local motion estimates are unified into a single estimate of plaid velocity. Subsequent studies (Lindsey & Todd, 1996; Stoner & Albright, 1996) indicate that this gating process is not all or none and that the interactions between motion and luminance in motion integration may be more subtle than originally thought. Our model attempts to capture these aspects of the motion integration process.

A block diagram of the model, illustrating the different components and the overall architecture, is shown in Figure 2 (see Cesmeli & Wang, 2000). The model consists of two initially parallel segmentation processes or pathways: one in which segmentation is based on similarities in estimated local motion velocity (indicated by the components enclosed within the dotted boundary) and another based on local static luminance information. The two pathways converge on an integration stage that iteratively in-

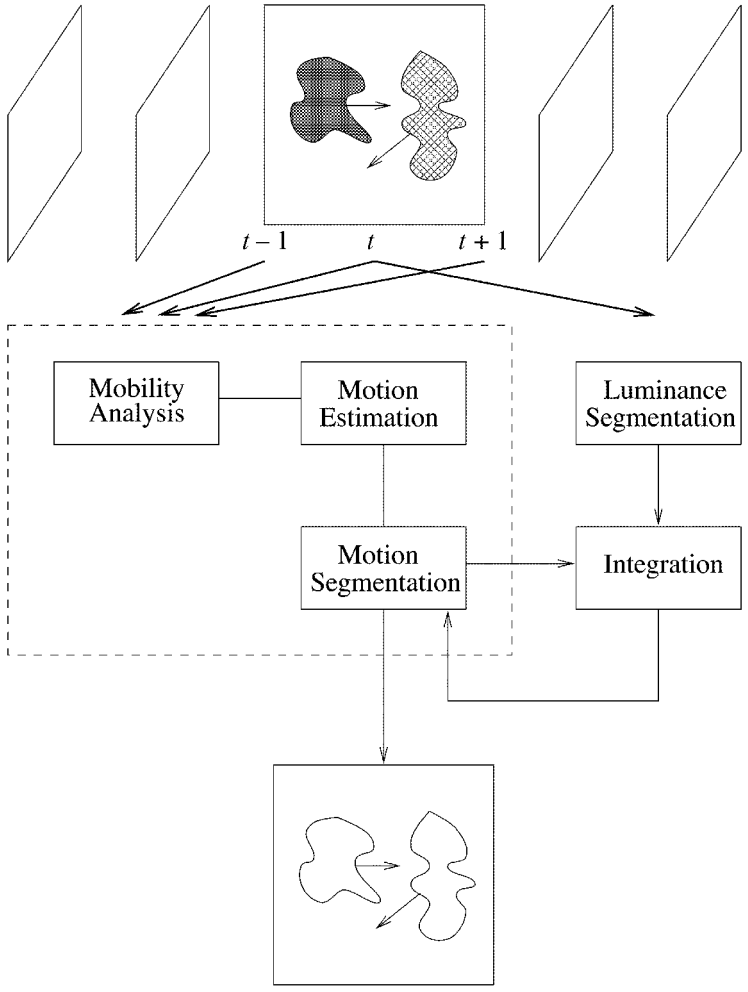

Figure 2. The flow diagram of the motion model. Processing starts from the top and proceeds downward. Following the analyses in the motion and the luminance pathways, results are combined at the integration stage to refine motion estimates. Note that final segmentation is performed in the motion network on the basis of the refined motion estimates.

teracts with the motion segmentation process. The final image segmentation and velocity labeling are done within the motion segmentation pathway.

The motion segmentation pathway of our model consists of motion estimation, motion segmentation, and mobility analysis components. Motion estimation proceeds initially by spatiotemporal block matching, a method for obtaining local motion estimates that is similar computationally to delay-and-compare methods for motion detection, described previously (Reichardt, 1961; van Santen & Sperling, 1985). In addition, the motion estimation component in our model incorporates three other functions that have recently received considerable attention in the computational vision literature and in studies of human motion perception. The first of these is the *slowness* assumption proposed by Weiss and Adelson (1997; Weiss, Simoncelli, & Adelson, 2002), in which local motion analysis seeks the slowest speed that is consistent with the motion sensor data. Another feature of the motion estimation component of the model is that it weights local estimates of image motion by the local statistics of samples obtained from the front-end sensors within some neighborhood. Estimates of local motion that exhibit high variance are weighted less heavily in the final segmentation stages of visual processing than are local estimates with low variance. Estimates with associated variances that exceed a criterion are eliminated from consideration. Qian and his colleagues (e.g., Qian, Andersen, & Adelson, 1994) have presented evidence for the existence of these kinds of processes in human motion perception. Finally, the motion estimation process employs pooling regions that adjust dynamically in size and shape to improve the reliability of local motion estimates. Dynamic tuning is accomplished by the component labeled Mobility Analysis in Figure 2.

This aspect of our model has two noteworthy features. First, these dynamic processes lead to front-end motion processes that adjust to the local orientations of contours, even though our model does not employ a front-end that is explicitly tuned to local orientation. A second feature is that these dynamic processes promote analysis at different spatial scales. The different scales are not explicitly specified but are determined by the scales of moving contours and surfaces in the visual scene being analyzed.

The luminance pathway segments the scene on the basis of local measurements of luminance. In the simulations that we will describe below, the luminance pathway assists motion segmentation in three ways. First, luminance segmentation greatly facilitates object segmentation, on the basis of differences across objects in their luminances. These results can then be used to guide motion segmentation. Black and Jepson (1996) have previously made this point. Second, luminance segmentation is important in solving the blank-wall problem mentioned earlier: Velocity estimates propagate from boundary contours, where the motion estimates are robust, to the interiors of homogeneous surfaces, on the basis of the results of the luminance segmentation analysis. Finally, luminance segmentation processes are important in determining local surface transparency and occlusion relationships at the boundaries separating potentially overlapping surfaces. This analysis is important not only in assessment of the reliability of local motion estimates, but also in that it allows the model to detect and deal with motion segmentation in phenomenally transparent moving patterns.

In the final stage of our model, motion and luminance information are integrated, the local motion estimates are refined, and final image segmentation occurs. The integration stage consists of two components. There is a process that performs occlusion analysis, where T- and X-junctions are used to determine occlusion and transparency relationships. There is also a process for iteratively refining the motion estimates and filling in all regions of the scene with estimates of object or background velocity. These results are then fed back to the motion segmentation stage for final segmentation and velocity labeling.

The two stages of segmentation described above—motion and luminance—are modeled using a multilayer LEGION network. LEGION is an acronym for locally excitatory, globally inhibitory oscillator networks (D. L. Wang & Terman, 1995). LEGION is an instantiation of oscillatory correlation as a method of solving the feature-binding problem. It has been used successfully in modeling a number of aspects of visual processing. Here, we use it to group local regions on the basis of similarities in motion and in luminance, as well as in the final motion labeling of segmented image elements of our model.

The present paper builds on previously published work in which we have demonstrated a model of image segmentation based on the integration of luminance and motion information (Cesmeli, Lindsey, & Wang, 1999; Cesmeli & Wang, 2000). The original model was intended to be used in engineering applications where the goal is to segment video-based image sequences of natural scenes. As such, the original model was designed to be relatively immune to various kinds of image noise. Although the present model retains all of the basic features of the previous version, much is new. For example, the computational strategies for the integration stage are largely new: The occlusion and transparency analyses are new, as are the slowness constraint and the certainty analysis embodied in Equations 4 and 5, below. These additional features of the present model permit an analysis of a broad range of psychophysical results in humans not previously modeled by us.

The remainder of the paper is organized as follows. First, we will provide an overview of our motion integration model. This section is divided into separate discussions of the motion segmentation pathway, the luminance segmentation pathway, and the integration stage. Then we will describe the results of simulations involving several kinds of stimuli that illustrate how the model deals with some of the issues in motion integration described above.

FUNCTIONAL DESCRIPTION OF THE MODEL

Motion Segmentation Pathway

Motion estimation. In our model, the motion estimation stage consists of sensors that provide local measurements of motion at different velocities at each point in the image array and of processes that make point-wise estimates of velocities, on the basis of the sensor measurements and the robustness of these measurements. Local measurements of motion are obtained using temporal block matching (Anandan, 1987), as is illustrated in Figure 3. Let N_B represent a two-dimensional (2-D) block of light sensors centered at location (i, j). For a temporal block matcher employing two consecutive image frames, the correlation corresponding to displacement $\mathbf{r} = (r_i, r_j)$, at location (i, j) and time t can be expressed as

$$\tilde{v}_{\mathbf{r}}(i,j,t) = \sum_{(k,l)\in N_B(i,j)} \frac{I(k,l,t)}{\left|I\left(k,l,t\right) - I\left(k-r_i, l-r_j, t-1\right) + \delta\right|}, \tag{1}$$

where $I(i, j, t)$ is the luminance at location (i, j) in the image frame t. The denominator of the above expression, as in the rest of the equations, includes a small quantity, δ, to avoid division by zero. A large correlation implies a high probability of the displacement, $\mathbf{r} = (r_i, r_j)$, at location (i, j) in time interval time t. One can view the magnitude of a correlation for a particular displacement, $\mathbf{r}$, as the response of a detector tuned to a velocity of $\mathbf{r}/t$.

We apply an adaptive temporal block matcher to three consecutive frames, instead of two, to improve accuracy (Singh, 1991). The correlation corresponding to displacement $\mathbf{r} = (r_i, r_j)$ is

$$v_{\mathbf{r}}(i,j,t) = \tilde{v}_{\mathbf{r}}(i,j,t) + \tilde{v}_{\mathbf{r}}(i,j,t+1), \tag{2}$$

where $\tilde{v}_{\mathbf{r}}(i, j, t)$ and $\tilde{v}_{\mathbf{r}}(i, j, t + 1)$ are the correlations at (i, j) between the image frames at t and $t - 1$ and those at

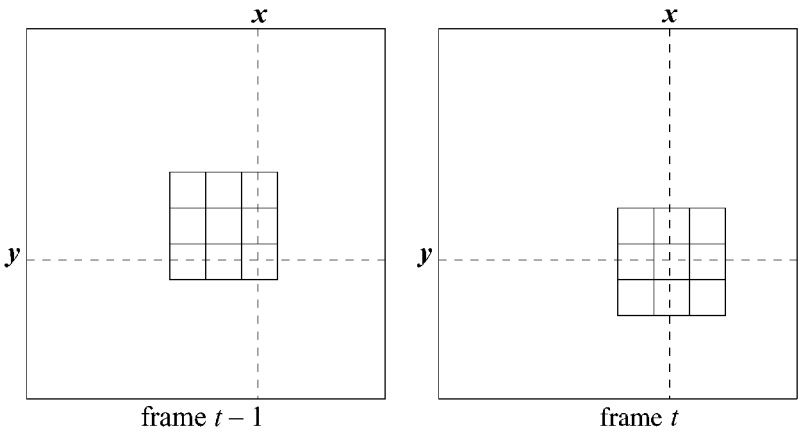

Figure 3. Luminance within the regions indicated by two 3 × 3 neighborhoods, N_B, in two consecutive frames are cross-correlated by temporal block matching to detect a diagonally rightward and downward motion of 1 pixel per frame. We use the term *block* to refer to the small regions of visual space being analyzed across successive time intervals. Mechanisms that perform the block-matching operation might be said to have *receptive fields* that indicate the blocks over which a given mechanism operates.

$t + 1$ and t, as given in Equation 1. We have a total number of $L = (2R + 1)^2$ different displacements, hence velocities, corresponding to a set of displacements varying from $-R$ to R in the x- and y-directions. In order to make motion estimation resistant to the effects of dynamic luminance noise, we apply block matching not only to neighborhoods of luminance, but also to local spatial correlation surfaces. A spatial correlation surface (SCS) is obtained by cross-correlating luminance within a block, N_B, at a location with that of a neighboring location in the same image frame. Repeating cross-correlation for all neighbors in N_S around the same location, its SCS is constructed, whereby a relative luminance distribution is assessed. SCS for all locations including the neighboring frames are obtained similarly. Subsequently, both luminance within N_B and SCS of size N_S are matched across time for different displacements, to obtain the temporal correlation surface at a location. Two SCSs are matched using Equation 2, and substituting SCSs for a block luminance. This results in $c_{\mathbf{r}}(i, j, t)$ for displacement $\mathbf{r} = (r_i, r_j)$ at location (i, j, t). Consequently, the temporal correlation at (i, j, t) for displacement $\mathbf{r} = (r_i, r_j)$ is

$$\hat{V}_{\mathbf{r}}(i, j, t) = \upsilon_{\mathbf{r}}(i, j, t) + c_{\mathbf{r}}(i, j, t), \tag{3}$$

Furthermore, we include a bias favoring "slow" motions (Weiss & Adelson, 1997; Weiss et al., 2002). As is shown in Figure 4A, for a location, P, along a straight border, there are multiple new locations, P', along the new position of the border. Correspondingly, the cross-correlation surface has multiple peaks, as is illustrated in Figure 4B. In contrast, perceived motion is unique and appears to be in the direction perpendicular to the orientation of the border. In other words, among multiple possible estimates, the smallest velocity is perceived. In order to capture this preference, we multiply the cross-correlation surface, $\hat{V}_{\mathbf{r}}(i, j, t)$, obtained using Equation 3 with a 2-D Gaussian surface centered at zero velocity or origin (Weiss, 1998):

$$V_{\mathbf{r}}(i, j, t) = \frac{1}{2\pi} \exp\left(\frac{-\|\mathbf{r}^2\|}{2\sigma^2} \right) \hat{V}_{\mathbf{r}}(i, j, t). \tag{4}$$

Owing to the shape of the Gaussian surface, the magnitudes of the peaks decrease as they get farther away from the origin. Consequently, the resulting correlation surface has its maximum at a single peak that corresponds to a perceptually preferred estimate, as is shown in Figure 4C. Note, however, that the magnitude of the single peak is not drastically different from those of the others; the peak magnitudes vary smoothly.

The displacement resulting in the maximum correlation at a location yields the local motion estimate at that location. An estimate is obtained at each location that satisfies a reliability criterion, which will be discussed next.

The shape of a cross-correlation surface depends on the underlying luminance structure at a location. A correlation surface might have a less pronounced maximum, as in the example of a straight border in Figure 4, implying a large uncertainty in the estimate. However, when a location is near a curved border or in a textured region, its correlation surface has a sharp peak, indicating a small uncertainty in the estimate. In order to capture these variations, we determine a certainty measure for each estimate on the basis of the shape of its correlation surface. We define a 2-D coordinate system at the maximum of a correlation surface, $V_{\mathbf{e}}$, corresponding to the estimate $\mathbf{e} = (e_i, e_j)$, as is shown in Figure 5. The *speed axis* passes through the origin and $\mathbf{e}$. The direction axis is perpendicular to the speed axis at $\mathbf{e}$, as is shown in Figures 5A and 5B. Note that a one-dimensional sharp maximum along an axis represents a better selectivity along the axis. Certainty in the estimation of local

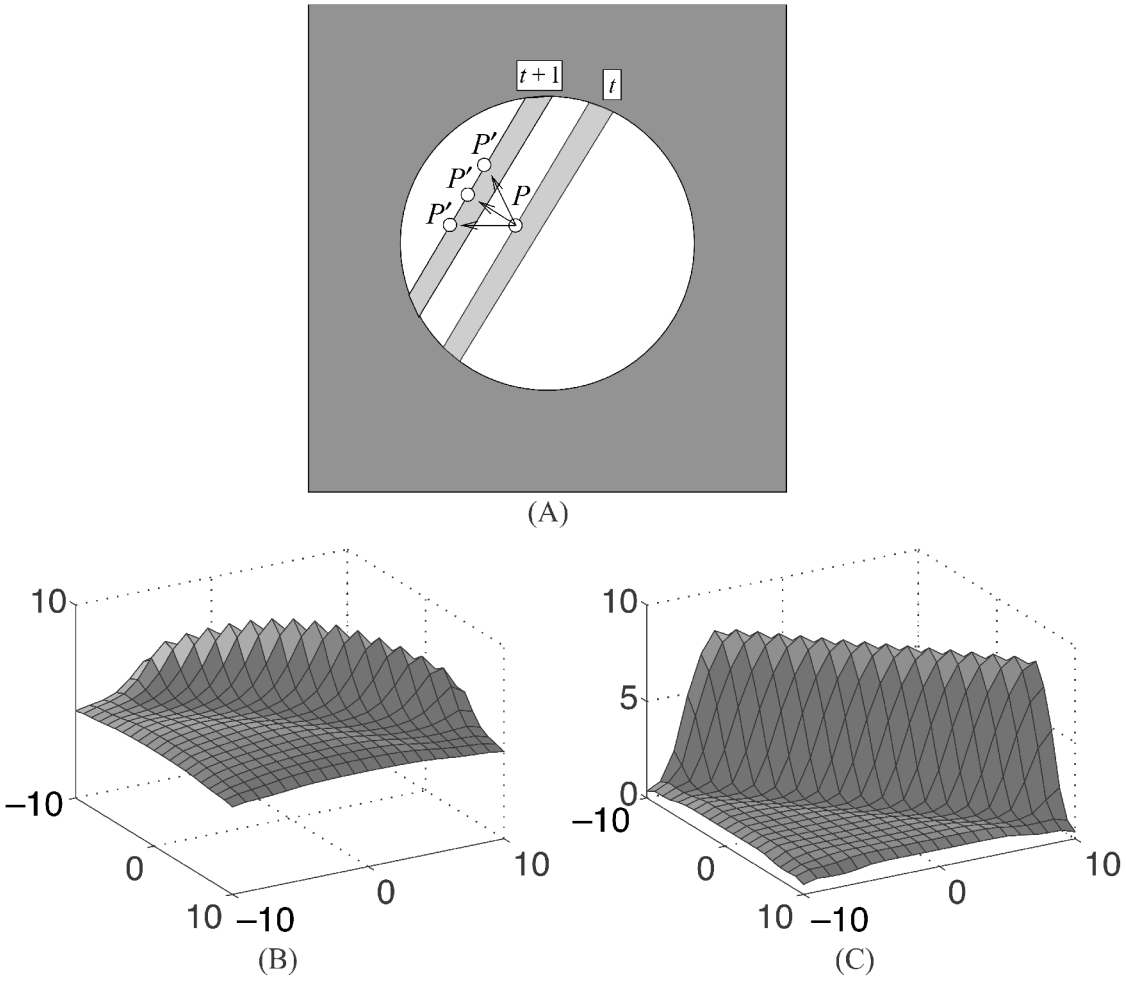

Figure 4. (A) The aperture problem. Two positions of a gray region are shown at times t and $t + 1$. Location P along the leading edge of the gray region at time t might appear to move any location along the region's leading edge at time $t + 1$. In the figure, only a subset of possible new locations, P's, is shown. The location closest to P is generally perceived to be the new location. Next shown is the cross-correlation surface for the location P, (B) without or (C) with the *slowness* assumption. Horizontal axes are velocity components along x- and y-directions.

motion is based on the sharpness of the maximum along the direction axis. The sharper the maximum is, the better the direction selectivity is and, thus, the more certain the estimate is assumed to be. Dropping (i, j, t) from the expressions for convenience, the certainty, ω, of the estimate $\mathbf{e}$ is given by

$$\omega_{\mathbf{e}} = \frac{\left(V_{\mathbf{e}} - V_{\mathbf{e}-\mathbf{k}}\right)\left(V_{\mathbf{e}} - V_{\mathbf{e}+\mathbf{k}}\right)}{2V_{\mathbf{e}} - V_{\mathbf{e}-\mathbf{k}} - V_{\mathbf{e}+\mathbf{k}}}. \tag{5}$$

Here, $\mathbf{e} - \mathbf{k}$ and $\mathbf{e} + \mathbf{k}$ are the nearest neighboring displacements to $\mathbf{e}$ along the direction axis, corresponding to correlations $V_{\mathbf{e}-\mathbf{k}}$ and $V_{\mathbf{e}+\mathbf{k}}$, respectively, as is shown in Figure 5B. According to Equation 5, the larger $\omega_{\mathbf{e}}$ is, the more certain estimate $\mathbf{e}$ is assumed to be. Figure 6A shows gratings moving upward behind a rectangular aperture. The cross-correlation surfaces at locations P and Q have quite different shapes around their peaks, corresponding to their local motion estimates, as is depicted in Figures 6B and 6C. According to Equation 5, $\omega_P < \omega_Q$, and hence, the estimate at Q is assigned a larger certainty than that of P.

Mobility analysis and dynamic spatial pooling. Mobility analysis is used for two purposes. The first is to dynamically adjust the size and shape of the correlation blocks, N_B, employed by the motion sensors. Basically, mobility analysis seeks to adjust the correlation blocks so that they will be optimized for the analysis of regions of the scene that are undergoing changes in luminance over time. Regions of uniform luminance or regions that are static have low mobility values, whereas those that contain motion information give high mobility values. Noisy regions will give intermediate mobility values. The second reason for mobility analysis is to restrict analysis of motion to those areas that contain motion information. The

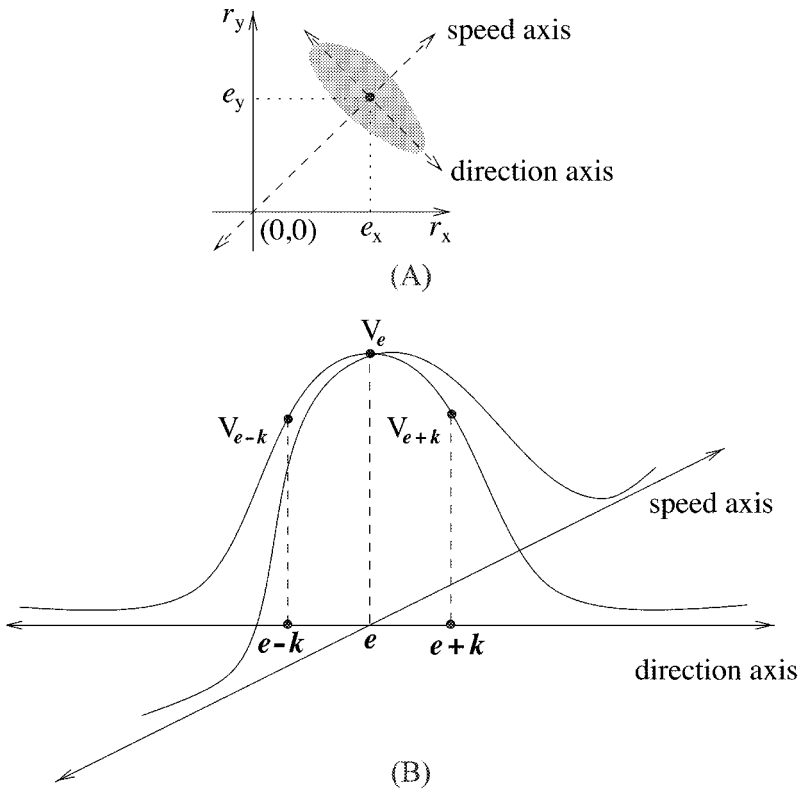

Figure 5. The definition of our estimate certainty. (A) Shaded ellipsoid is the top view of a crosscorrelation surface. The speed axis passes through the origin and the estimate e = (e_i, e_j) that corresponds to the maximum cross-correlation, V_e. The direction axis is perpendicular to the speed axis and intersects the latter at the estimate, e. (B) A side view of the cross-correlation surface. V_{e-k} and V_{e+k} correspond to the nearest neighboring displacements of the estimate, e, along the direction axis.

correlation blocks do not grow in size without bound. They are allowed to grow up to a certain size, in order to increase mobility. If a threshold mobility has not been reached by a predetermined size, those regions are excluded from the motion estimation process. For example, in scenes containing homogeneous surfaces bounded by luminance borders, mobility analysis confines analysis to the borders that are undergoing changes in position over time and adjusts the shape of the correlation blocks so that they are aligned with those borders.

We use a mobility measure similar to that described by Irani, Rousso, and Peleg (1994). The mobility value at each location and time (i, j, t) is given by

$$M(i,j,t) = \frac{\sum_{(k,l)\in N_M(i,j)} |I(k,l,t) - I(k,l,t-1)|}{\sum_{(k,l)\in N_M(i,j)} I(k,l,t)}, \quad (6)$$

where N_M is a mobility neighborhood. Note that changes are normalized by the summated local luminance. A large value at a location indicates a high probability of motion at that location and in its vicinity. Otherwise, the presence of motion is uncertain, since noise could also cause a temporal change in luminance. Thus, we do not obtain estimates at locations with small mobility values. Note that our reliability criterion serves two purposes. First, locations with sufficient motion evidence are identified. Second, pooling neighborhoods, N_B, at these locations are modified so that their estimates are more accurate.

Initially, a small correlation block is assumed to be sufficient for spatial pooling. When the sum of the mobility values, M_s, and the average mobility, M_a, within the block are large enough, matching is performed using Equation 3. When M_s is small, the block is expanded in the direction that maximally increases M_s. The expansion continues until either both M_s and M_a are sufficiently large or the block size reaches an upper limit. If the upper size limit is reached, estimation is not performed. Otherwise, the smallest block satisfying the former condition is used for estimation. We use the selected block, N_B, for both luminance matching and SCS construction. Figure 7 illustrates the adaptive changes in N_B for an image composed of a horizontally moving rectangular region and a background, where both regions are homogeneous (Figure 7A). Among the three consecutive input frames, only the middle one is depicted. Figure 7B shows the mobility

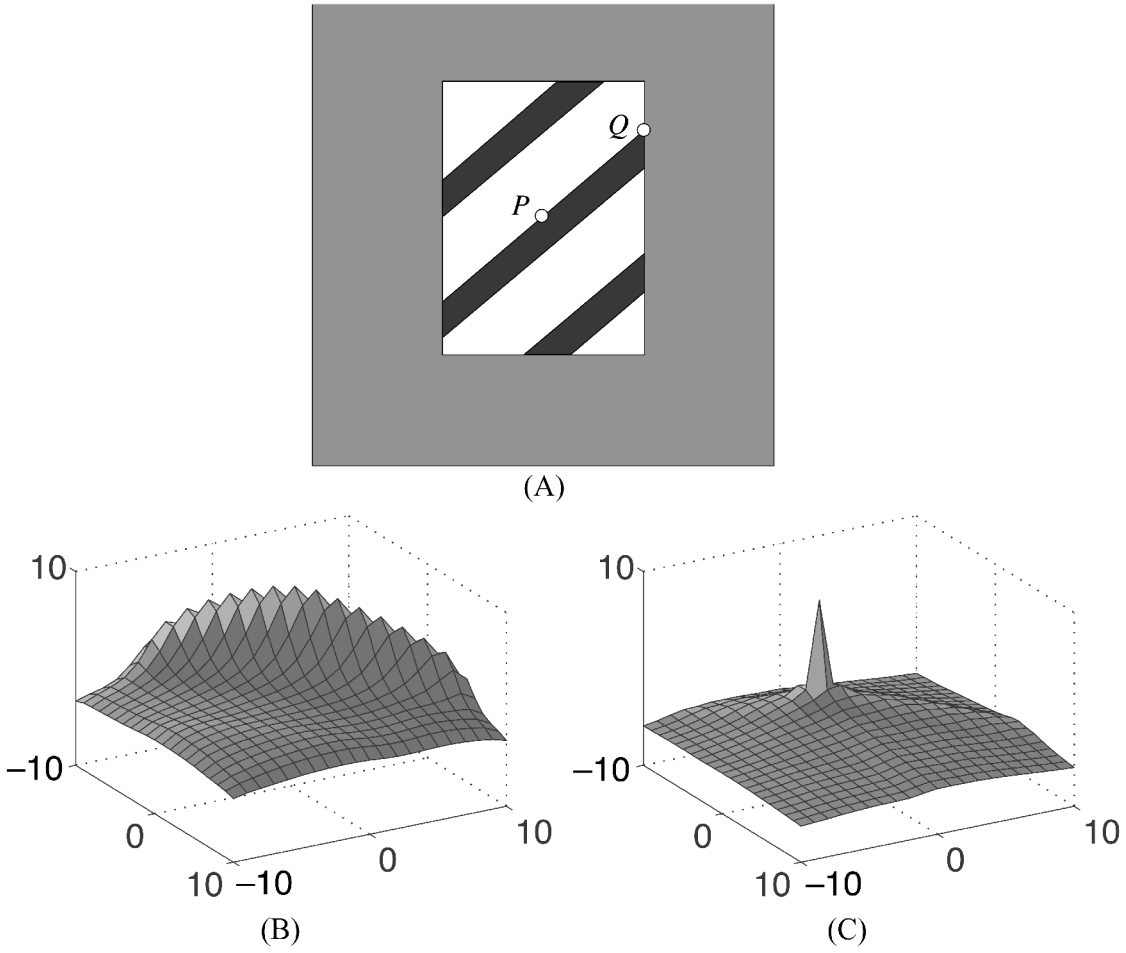

Figure 6. The shape of a cross-correlation surface at a location depends on its underlying luminance structure. (A) Gratings are moving vertically upward behind a rectangular aperture. Note that location *P* is less spatially constrained than location *Q*. Thus, the cross-correlation surface at *P* shown in panel B has a less pronounced maximum than that of *Q* depicted in panel C.

image, where the intensity at a location is its mobility value. In this image, the evolution of the blocks is depicted at a location near a boundary. Initially, $N_B = 5 \times 5$, and the stopping condition is not met. Since boundaries and surface markings of moving regions have large mobility values, N_B usually extends to them. For the location considered in the image, N_B expands to cover a considerable portion of the boundary. Initially, expansion is isotropic, and mobility is very low, since the starting point is away from the moving edge. However, when the edge is encountered, expansion proceeds in a direction parallel with the moving contour. When a location does not produce large mobility values in its N_M (e.g., a location away from boundaries in Figure 7B), M_s and/or M_a become small, because of the normalization in the denominator of Equation 3. Eventually, N_B reaches the upper limit, mobility fails to reach threshold, and motion is not estimated at that location.

Motion segmentation. Motion segmentation is performed by a multilayer network, schematized in Figure 8A. Each layer represents the spatial distribution of motion at one particular velocity. Figure 8A depicts four layers/velocities. Each of the layers corresponds to a separate LEGION network. The multilayer architecture allows multiple velocities to be active at any given time, and our model therefore can handle motion transparency, either in the case in which objects partially overlay or in the case in which one object completely overlaps another.

LEGION employs the idea of oscillatory correlation, where oscillator phases encode the bindings of local features with global objects. In this network, the nodes consist of relaxation oscillators. When stimulated, these oscillators alternate periodically between a brief active phase and a somewhat longer silent phase. The short duty cycle of the active phase means that multiple objects can be represented by multiplexing the active phases of multiple oscillator populations within the overall relaxation cycle of the network. The oscillators are locally coupled—for example, four nearest neighbors—with excitatory

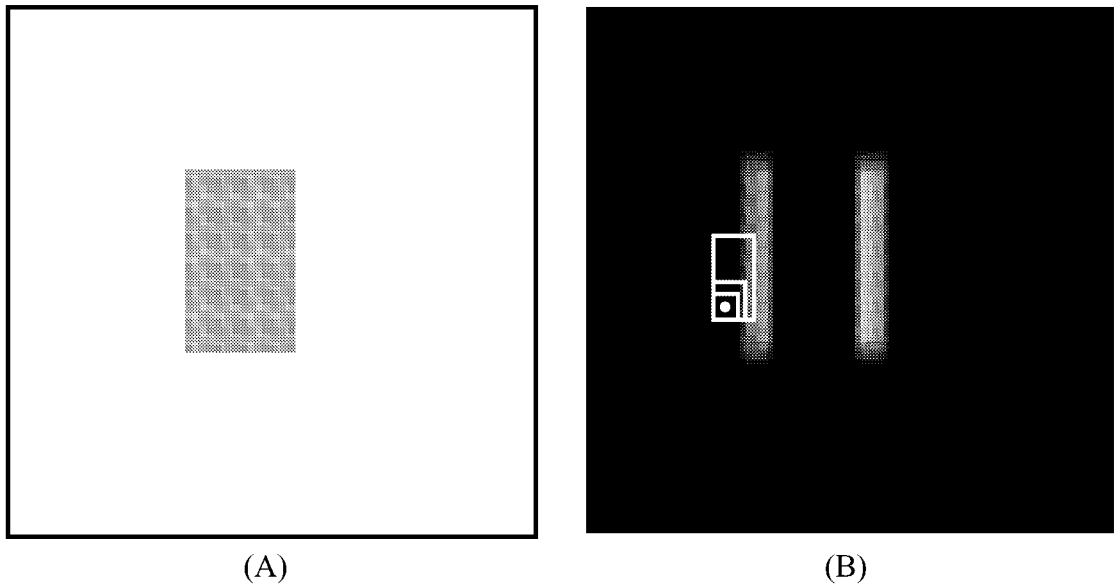

Figure 7. An example for adaptive spatial pooling. (A) The middle frame of an input sequence in which the gray rectangular region is moving to the right. (B) The mobility image for the frame in panel A, where intensities are the corresponding mobility values. The evolution of the pooling blocks is depicted as white rectangles for a location near the boundary (white dot). Note that the neighborhood extends along the boundary to encompass locations with large mobility values.

weights, whereas there is a single global inhibitory input to all oscillators within, as well as across, layers. Terman and Wang (1995; D. L. Wang & Terman, 1995) have shown that LEGION networks have the property that groups of oscillators stimulated by the same temporal input tend to rapidly synchronize their oscillations to one another, owing to local excitation, whereas oscillator groups stimulated by different inputs tend to rapidly desynchronize, owing to global inhibition. Thus, in this network, different moving objects give rise to different (although possibly overlapping) populations of network nodes oscillating at the same frequency, but in different phases relative to one another.

A single oscillator, (i, j), of a LEGION network is defined as a feedback loop between an excitatory unit x_{ij} and an inhibitory unit y_{ij}:

$$\frac{dx_{ij}}{dt} = 3x_{ij} - x_{ij}^3 + 2 - y_{ij} + S_{ij} + \rho \qquad (7a)$$

$$\frac{dy_{ij}}{dt} = \epsilon\left\{\alpha\left[1 + tanh\left(x_{ij}/\beta\right)\right] - y_{ij}\right\}. \qquad (7b)$$

Here, S_{ij} denotes coupling, α and β are system parameters, and ρ is the variance of a Gaussian noise term. The parameter ϵ is chosen to be a small positive number so that Equation 7 defines a relaxation oscillator. ϵ induces two time scales for the relaxation oscillators: a slow scale corresponding to the period between active and silent phases and a fast time scale when the oscillator is active and is rapidly oscillating between two fixed states.

The coupling term, S_{ij}, includes a variable called *lateral potential* and coupling from neighboring oscillators and a global inhibitor:

$$S_{ij} = W\,H\left[\sum_{kl \in N(i,j)} W_{ij,kl} H(x_{kl}) - \theta\right] + W_p H(p_{ij} - 0.5) - W_z H(z - 0.5), \qquad (8)$$

where $W_{ij,kl}$ is the connection weight from oscillator (k, l) to oscillator (i, j), H is the Heaviside step function, W_p and W_z are the weights for the lateral potential and the global inhibition, respectively, and N represents a local coupling neighborhood—for example, four nearest neighbors. W and θ are local coupling contribution and local similarity threshold, respectively. The lateral potential is introduced for each oscillator, to distinguish a homogeneous region from a noisy one:

$$\frac{dp_{ij}}{dt} = (1 - p_{ij})H\left[\sum_{kl \in N_p(ij)} H(x_{kl}) - \theta_p\right] - \epsilon p_i, \qquad (9)$$

where N_p is the potential neighborhood, which is larger than N. The potential of an oscillator, which is initially set to 1, continuously decays. Only when an oscillator is active and has a number of active neighbors greater than θ_p in its N_p does its potential rise to 1. Oscillators become either *leaders* or *followers*. Leaders are those that maintain high potential, and the rest constitute followers. Groups of strongly coupled oscillators are candidates for forming segments. However, owing to the potential term, only groups that have a

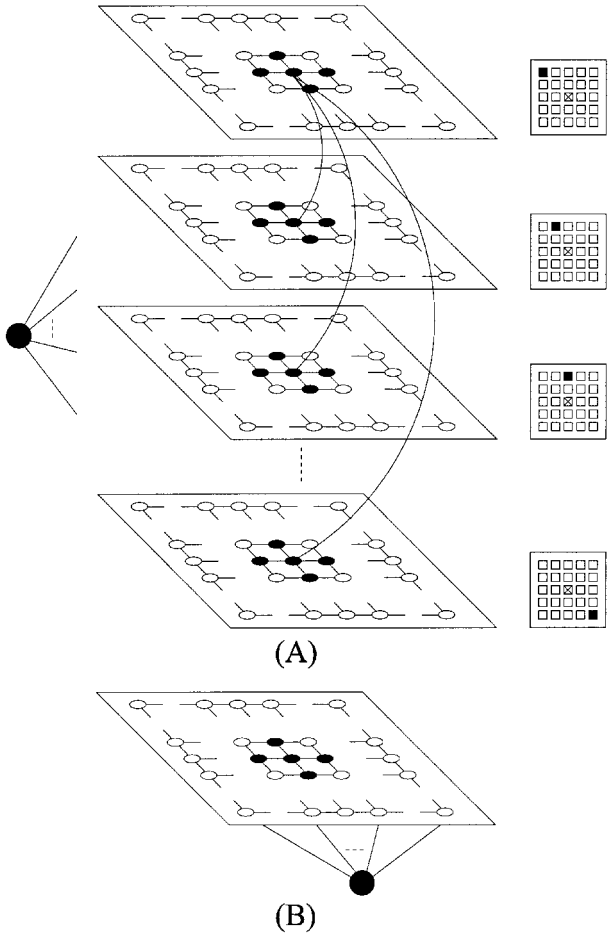

Figure 8. Neural network architectures in our motion model. Small ellipses represent oscillators, and the big black circle is the global inhibitor. The global inhibitor is connected to all oscillators in each pathway. (A) Multilayer motion network in which corresponding to each velocity layer is a LEGION network. Each layer is tuned to a particular velocity indicated by a filled square in a small rectangular grid representing the set of velocities varying from −2 to 2 in the *x*- and *y*-directions. Units interact vertically within velocity columns and horizontally within velocity layers. Vertical interactions are shown only for one oscillator. (B) Luminance network. Each oscillator is locally coupled with its four nearest neighbors (reproduced from Cesmeli & Wang, 2000).

leader can form segments. The potential neighborhood is chosen so that only relatively large homogeneous regions can produce leaders. Since noisy fragments tend to be small, they are unable to produce leaders; moreover, they tend to be isolated and, thus, cannot be recruited by other active oscillators. As a result, oscillators corresponding to a noisy fragment will stop oscillation after a brief initial period. These oscillators together form the *background.*

Whether an oscillator becomes active at any moment also depends on the activity of the global inhibitor, z, which is defined by

$$\frac{dz}{dt} = H\left[\sum_{ij} H\left(x_{ij}\right) - 1\right] - z. \tag{10}$$

When no oscillator is active, z decays to 0; otherwise, it rises to 1. A leader can jump only when $z < .5$. Otherwise, an oscillator can become active only if it has strong couplings with currently active oscillators.

The coupling weight between two oscillators, (i, j) and (k, l), on a velocity layer $\mathbf{r}$ is determined by their temporal correlations for the corresponding displacement per time interval Δt. Dropping time t from the expressions for convenience, the coupling weight is given by

$$W_{\mathbf{r},ij,kl} = \frac{V_{\mathbf{r}}(i,j) + V_{\mathbf{r}}(k,l)}{|\, V_{\mathbf{r}}(i,j) - V_{\mathbf{r}}(k,l)\,|}. \tag{11}$$

When oscillators have similar correlations for a particular displacement, they are strongly coupled in the corresponding velocity layer. At locations without estimates, couplings are set to zero. We replace S_{ij} in Equation 8 by

$$S_{\mathbf{r},ij} = H(\tilde{S}_{\mathbf{r},ij} - \theta_M),$$

where θ_M is a threshold and

$$\begin{aligned}\tilde{S}_{\mathbf{r},ij} = W\, H\left[\sum_{kl \in N(i,j)} W_{\mathbf{r},ij,kl} H\left(x_{kl}\right) - \theta\right] - W_z H(z - 0.5) \\ + W_p H\left(p_{\mathbf{r},ij} - 0.5\right) H\left(M_{ij} - 0.25\right) \\ H\left[\sum_{q=1}^{L} H\left(V_{\mathbf{r},ij} - V_{\mathbf{q},ij}\right) - L\right].\end{aligned} \tag{12}$$

Note that the major change is the two terms multiplying the potential to extend the definition of leadership. The first term ensures that only locations with large mobility values can become leaders. The second term allows for only a single leader within each velocity column. Owing to the inhibitory interaction within each column, the oscillator with the largest correlation becomes a winner. Provided that a winner has sufficiently high mobility and potential, it becomes a leader and starts forming its segment. A leader recruits oscillators on its layer through local couplings when they have similar correlations. As in the single-layer LEGION network, recruited oscillators become active simultaneously (synchronization). Because of the global inhibition across all layers, leaders form their segments at different times (desynchronization). Periodic activity of segments continues as long as the input stays the same.

Note that the selection of a single leader within each column does not deprive the network of representing multiple motions at a location. An oscillator can become active if it is a leader or a follower recruited by a leader on its layer. Since there could be several followers in the velocity column of a leader, more than one oscillator at a single location can become active, representing different motions. This ability has a key role in the representation of motion transparency.

The complete activity of all layers is captured in an output network in which each unit has the summated activity of the corresponding velocity column. Since, at most, one oscillator is active in each column at any time, the output network displays segments in the order they become active.

Luminance Segmentation Pathway

In the parallel luminance pathway, the middle frame of the sequence analyzed in the motion pathway is processed in terms of its luminance distribution. In this pathway, where also a LEGION network is employed, as is depicted in Figure 8B, we assume that each region is approximately homogeneous. The coupling weight between two oscillators, (i, j) and (k, l), is defined as

$$W_{ij,kl} = \frac{I(i,j) + I(k,l)}{|\, I(i,j) - I(k,l)\,|}. \qquad (13)$$

When locations have relatively similar luminances, a strong coupling weight results. We employ the network model given in Equation 8 where S_{ij} is replaced by $H(S_{ij} - \theta_B)$ and θ_B is a threshold. Strongly coupled oscillators with at least one leader become synchronized. Oscillators corresponding to regions with different luminances become active at different times (desynchronization). When oscillators correspond to textured regions, they tend not to have strong couplings and, thus, leaders. Since they do not form segments, textured regions are distinguished from homogeneous ones, a property that has a significant functional role in the subsequent integration stage.

Integration Stage

So far, the motion and the luminance pathways perform segmentation independently. Lacking spatial luminance variations, homogeneous regions have low mobility values and, thus, estimates only along their moving boundaries. Conversely, locations in textured regions are not grouped by the luminance network, owing to their luminance variations. Thus, one purpose of the integration stage is to facilitate segmentation of these regions of the visual scene. In addition, the integration stage performs general occlusion analysis, eliminates unreliable motion estimates, and fills in the unlabeled regions with more reliable motion estimates. The results are then fed forward to the motion segmentation stage, described above, for final segmentation and velocity labeling.

Occlusion analysis. Our occlusion analysis, first, classifies a given motion segment into one of the three categories. In the first category, all locations in the segment belong to textured regions. In the second category, they belong to untextured (homogeneous) regions. In the final category, locations are from both textured and homogeneous regions. Note that, since the luminance pathway segregates textured regions from homogeneous ones, the classification step utilizes the segmentation results in both pathways.

When all the locations in a motion segment belong to textured regions, estimates in this segment are not changed. When the majority of the locations in a motion segment belong to multiple homogeneous regions, an occlusion relationship among these regions is obtained, using T- and X-junction analyses. We adopt a simple template-matching technique to detect junctions, since our goal is to illustrate the usefulness of occlusion analysis in the integration of local motion estimates. More sophisticated techniques have been described previously (see Liu & Wang, 2000; Parida, Geiger, & Hummel, 1998).

Our model detects T-junctions by applying a set of 24 templates to the luminance segmentation result. Templates are composed of three regions whose borders form a T-junction with various orientations and look like the examples shown in Figure 9A. A T-junction detected among three homogeneous regions uniquely determines the occluding opaque region and the two occluded regions (see, e.g., Cavanagh, 1987)). As is illustrated in Figure 9B, T-junction analysis provides important information during motion integration. In that figure, regions B_1, B_2, and B_3 form a T-junction at two different points along the border between B_2 and B_3. The relative positions of the regions at the T-junctions indicate that B_3 occludes B_2 and B_1. This implies that local motion estimates in B_1 and B_2 along the border of B_3 are likely to be unreliable, owing to the motion of B_3.

X-junction analysis indicates occlusion relationships among regions where the one lying on the top is transparent (Beck, Pradzny, & Ivry, 1984; Metelli, 1974). Our model detects X-junctions, using luminance segments and their luminances. Unlike T-junctions, analysis of X-junctions involves four neighboring regions and also the relationship among their luminances (see Figure 10). Our model first applies a set of templates to the luminance segmentation result, to locate X-junctions. When an X-junction is detected, the luminance relationships among the four regions defining the junction are analyzed, subject to two constraints that determine whether the junction exists because of a transparent occlusion or because of four juxtaposed opaque regions. Let us consider the example in Figure 10B, where regions B_1, B_2, B_3, and B_4 form two X-junctions. When B_1 and B_3 are opaque and B_2 and B_4 are the regions forming a transparent surface, the luminances of these regions satisfy the following two constraints (Beck et al., 1984; Metelli, 1974):

(1) the *order constraint*, in which the sign of the luminance change across the border of B_1 and B_3, $I_{B1} - I_{B3}$, is the same as that of B_2 and B_4, $I_{B2} - I_{B4}$—in other words, $(I_{B1} - I_{B3})(I_{B2} - I_{B4}) > 0$;

(2) the *magnitude constraint*, in which regions forming a transparent surface, B_2 and B_4, have a luminance change across their common border smaller than that of opaque regions, B_1 and B_3—namely, $(I_{B1} - I_{B3}) > (I_{B2} - I_{B4})$.

Having detected junctions satisfying these constraints, region B_2 is inferred to transparently overlap with regions B_3 and B_1. Thus, estimates in B_1, B_2, and B_3 along the border of B_4 are assumed to be unreliable. As a result, B_4 is considered to be part of both B_2 and B_3, a condition that has a role in the motion filling-in step. Note that, when no T- or X-junction is detected among homogeneous regions, estimates in these regions stay intact.

When locations in a motion segment are from both homogeneous and textured regions, another technique is uti-

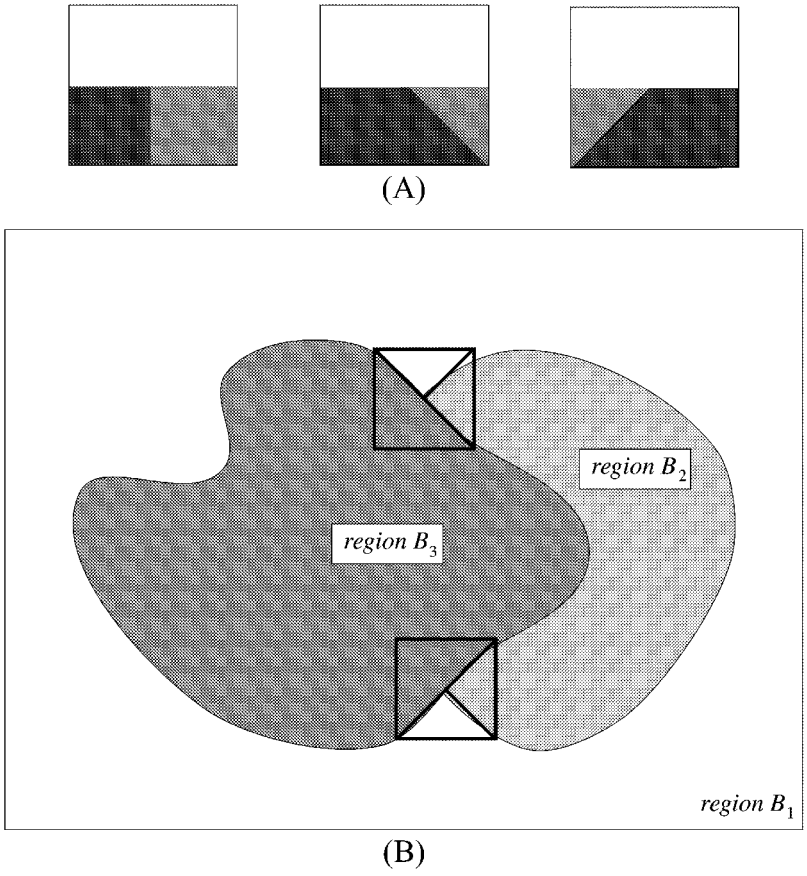

Figure 9. T-junction detection. (A) A subset of T-junction templates used in our model. (B) A schematic illustration of T-junction detection. On the basis of the detected T-junctions, region B_3 is inferred to occlude regions B_1 and B_2.

lized to resolve the occlusion relationship. Our model obtains the occlusion relationship between textured and homogeneous regions by determining motion distributions (histograms) in the homogeneous regions. For each neighboring homogeneous region of a textured region, two motion distributions are obtained within the luminance segment of the homogeneous region. The first one includes estimates at all locations in the luminance segment. The second one considers only locations along the border of the textured region. When the peak location and its magnitude in the first distribution correspond to a similar peak in the second, the textured region is assumed to occlude the homogeneous region. Accordingly, estimates in the homogeneous region that are along the border of the textured region are assumed to be unreliable. When the location of the peaks is the same but their magnitudes are different, the two regions are assumed to be moving together, and thus, estimates in the homogeneous region stay intact. This process is repeated for the other neighboring homogeneous regions of the textured region.

In the first step of the integration stage, the estimates that are classified to be unreliable on the basis of the occlusion relationship among surfaces are eliminated.

Motion filling-in. Because our reliability criterion has a local estimate, inner locations of homogeneous regions do not have estimates. Following the removal of unreliable estimates, remaining ones within each luminance segment, B, interact iteratively and result in a segment velocity, $\mathbf{r}_B$, given by

$$\mathbf{r}_B^{\tau} = \sum_{(i,j)\in B} \omega^{\tau}(i,j)\mathbf{r}(i,j)\,/\Omega_B^{\tau} \tag{14a}$$

and

$$\omega^{\tau+1}(i,j) = \frac{\omega^{\tau}(i,j)}{\Omega_B^{\tau}}\left(1+\frac{\mathbf{r}(i,j)\cdot\mathbf{r}_B^{\tau}}{\sqrt{\|\mathbf{r}(i,j)\|\,\|\mathbf{r}_B^{\tau}\|}}\right). \tag{14b}$$

Here, $\mathbf{r}_B^{\tau}$ and Ω_B^{τ} are the segment velocity and the sum of certainties, $\omega^{\tau}(i,j)$, in B at iteration step τ, respectively. $\mathbf{r}(i,j)$ is the local estimate at location (i,j). $\mathbf{a}\cdot\mathbf{b}$ is the dot product between vectors $\mathbf{a}$ and $\mathbf{b}$, and $\|\mathbf{a}\|$ is the magnitude of $\mathbf{a}$. In Equation 14a, $\mathbf{r}_B^{\tau}$ is determined by weighing the estimates in B by their certainties. In Equation 14b, $\omega^{\tau+1}(i,j)$ increases when $\mathbf{r}(i,j)$ and $\mathbf{r}_B^{\tau}$ have a similar direction. Finally, when $\omega^{\tau+1}(i,j)$'s in B do not change, $\mathbf{r}_B^{\tau}$ is assumed to have converged to its final value, $\mathbf{r}_B$. As a result, $\mathbf{r}_B$ is filled in at all locations in B.

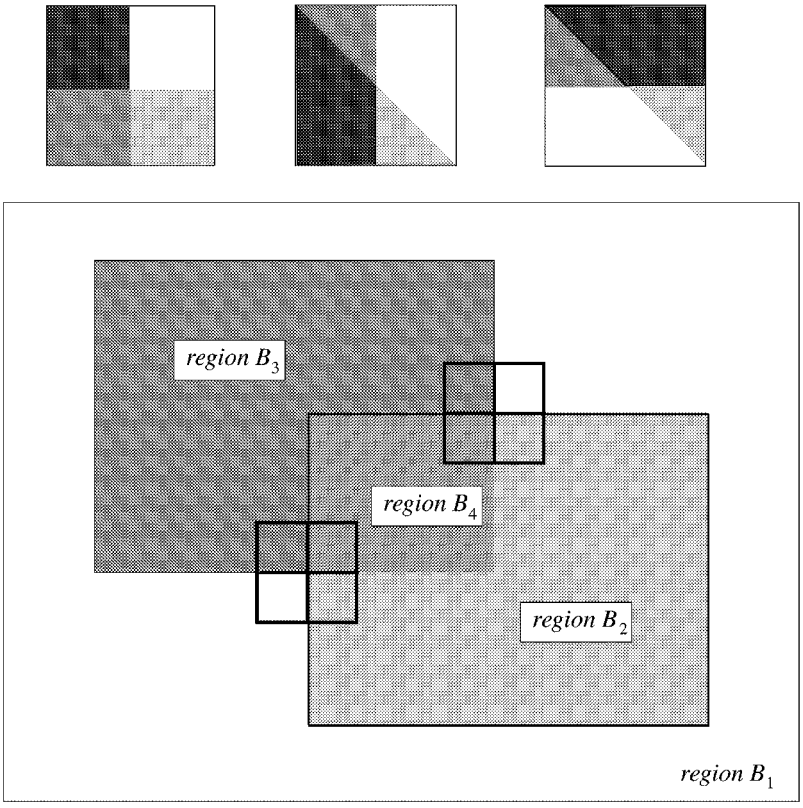

Figure 10. X-junction detection. (A) A subset of X-junction templates used in our model. (B) A schematic illustration of X-junction detection. According to the detected X-junctions, region B_2 occludes regions B_3 and B_1 and results in an overlapping region B_4.

Note that some locations in a scene may initially assume large certainties, as in the case of T-junctions. Owing to the interaction mechanism in Equation 14, different locations vote to support their motion estimates. Thus, locations that might have large certainties do not by themselves define the segment velocity of a luminance segment. It is the collective decision of locations, possibly with different certainties, that determines a segment velocity. The motion interaction in Equation 14 takes place in all luminance segments. Since textured regions, which already have reliable estimates, do not form luminance segments, the motion interaction does not take place in these regions.

Prior to Equation 14, we assign estimates of zero velocity to locations without estimates in luminance segments along the image borders. This assignment is consistent with the tendency of large homogeneous regions touching image borders to appear stationary.

Following the integration stage, couplings in the motion network are updated on the basis of the refined estimates, and the final segmentation result is obtained.

SIMULATIONS AND RESULTS

The details of our simulations, including parameter values, may be found in the Appendix.

Overlapping Opaque Rectangles

Figure 11A shows an input scene in which two vertical rectangles are moving toward each other at the speeds of 3 and 2 pixels per frame, respectively. The horizontal one is moving downward with a speed of 2 pixels per frame. In the motion pathway, first, the mobility image is obtained using Equation 6 and is shown in Figure 11B. Next, local motions and their certainties are estimated at locations with large mobility values, as is depicted in Figures 11C and 11D. The initial motion segmentation in Figure 11E is based on these estimates. Note that missing and erroneous estimates in occluded regions along occluding boundaries cause inaccurate segmentation. However, the segmentation result in the luminance pathway, as shown in Figure 11F, matches well with the input scene. In the subsequent integration stage, the occlusion analysis is performed using the segmentation results in the two pathways. Having detected T-junctions, as depicted in Figure 11G, it is found that the larger vertical rectangle occludes the background and the horizontal rectangle, which, in turn, occludes the background and the smaller vertical rectangle. Figures 11H and 11I show the remaining estimates and their certainties after removing the ones in the occluded regions along the occluding boundaries. Finally, the motion interaction takes place in luminance

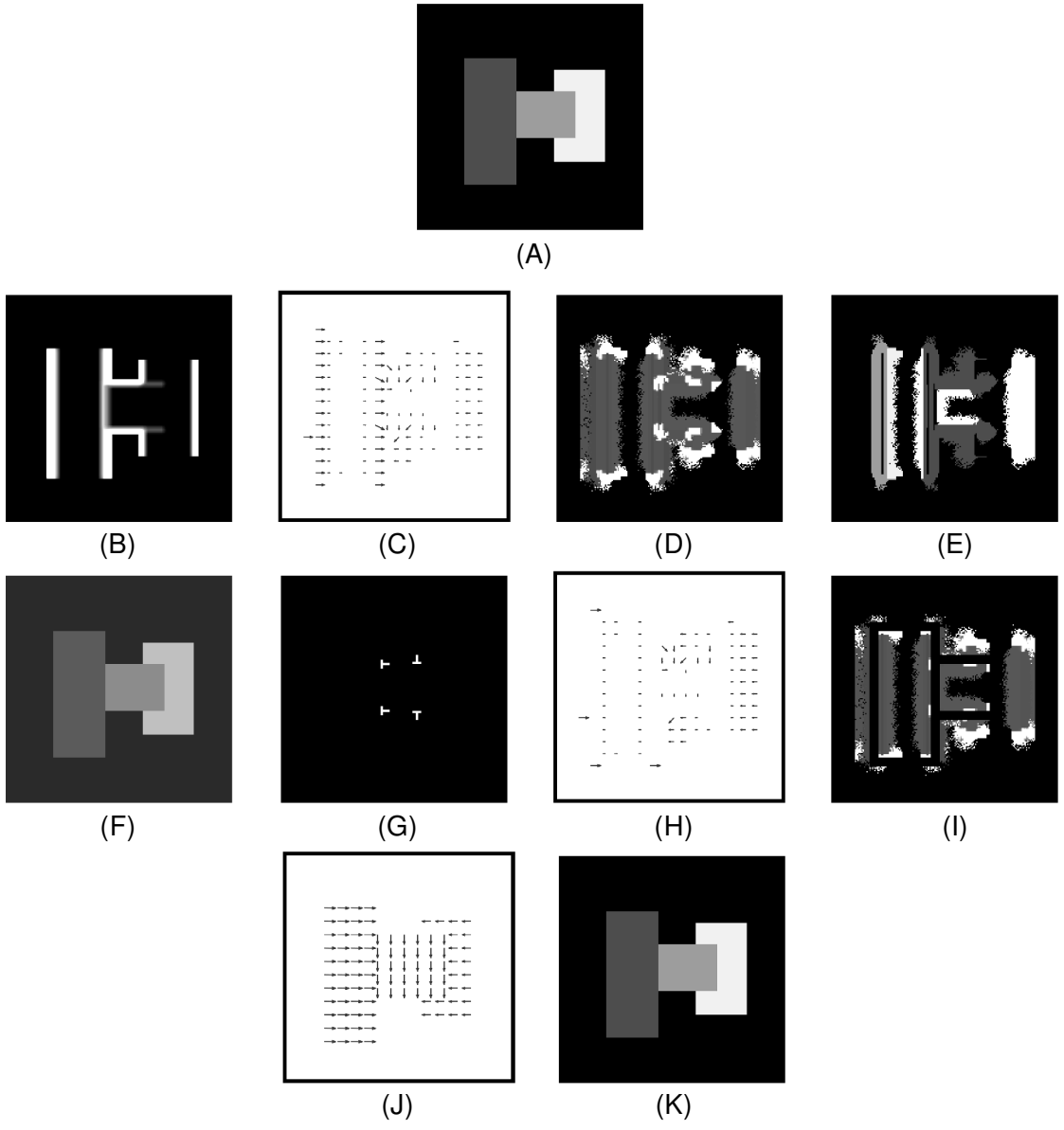

Figure 11. (A) An input scene composed of three opaque homogeneous regions moving in different directions. Gray levels indicate surface luminances. (B) Mobility image. Higher mobilities appear whiter. (C–D) Estimates and certainties. Higher certainties appear whiter. (E) Initial motion segmentation. Gray levels label recovered velocities. (F) Luminance segmentation. Gray levels label recovered segments. (G) In the occlusion analysis, only T-junctions are detected. On the basis of T-junction analysis, motion estimates in occluded regions are eliminated if they fall along borders of occluding regions. (H–I) Motion estimates and certainties, following T-junction analysis. (J) Segment velocities are determined and filled in within luminance segments. (K) Final motion segmentation based on the refined estimates. As in panel F, different segments are assigned different velocity labels. Gray levels indicate the three differently labeled regions recovered by the model.

segments, filling in their locations with their resulting segment velocities, as illustrated in Figure 11J. The final segmentation result, based on the refined estimates, is shown in Figure 11K and compares well with the regions and their motions in the input scene. Note that the homogeneous background is assigned zero velocity, owing to the introduction of estimates of zero velocity along the image borders.

The input scene in Figure 12A is composed of two square regions. The upper region moves in the rightward and downward direction, whereas the lower one has a leftward and downward motion. Both regions have the speed of 2 pixels per frame. In the center of the scene, the regions overlap transparently. Similar to the example in Figure 11, the scene is processed, and segmentation results are obtained in the two pathways. Figures 12B and 12F

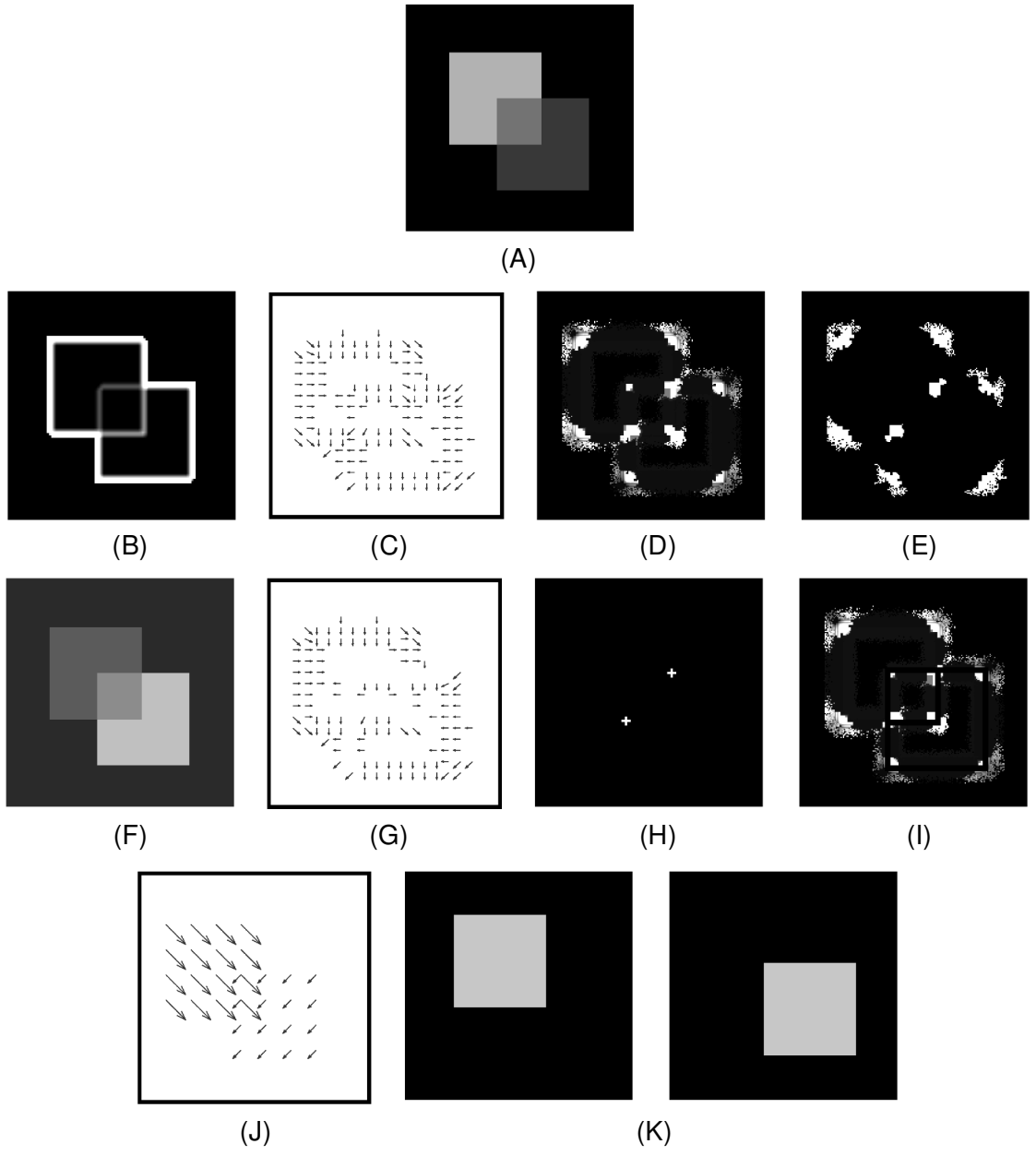

Figure 12. (A) An input scene composed of transparent homogeneous regions. (B) Mobility image. (C–D) Estimates and certainties. (E) Initial motion segmentation. (F) Luminance segmentation. (G) In the occlusion analysis, only X-junctions are detected. (H–I) Estimates and certainties in occluded regions along the borders of occluding regions are eliminated on the basis of the occlusion relationship obtained. (J) Segment velocities are filled in within luminance segments. Note that the common region is assigned two different motions. (K) Final motion segmentation based on the refined estimates. Owing to transparency, resulting segments overlap in the center of the region and, hence, are shown in two separate images.

show the mobility image, motion estimates and their certainties, initial motion segmentation, and the luminance segmentation results, respectively. In the occlusion analysis, X-junctions are detected (Figure 12G), and the lower square region is determined to overlap with the upper one transparently. As a result, the overlapping area between the two is inferred to belong to both regions. Figures 12H–12I show the remaining estimates and their certainties, following the elimination step. Subsequently, segment velocities are determined by considering the overlapping area as part of both regions. These velocities are filled in at the locations, as shown in Figure 12J. For visual clarity, estimates in the lower square region are artificially scaled down. Note that locations in the overlap-

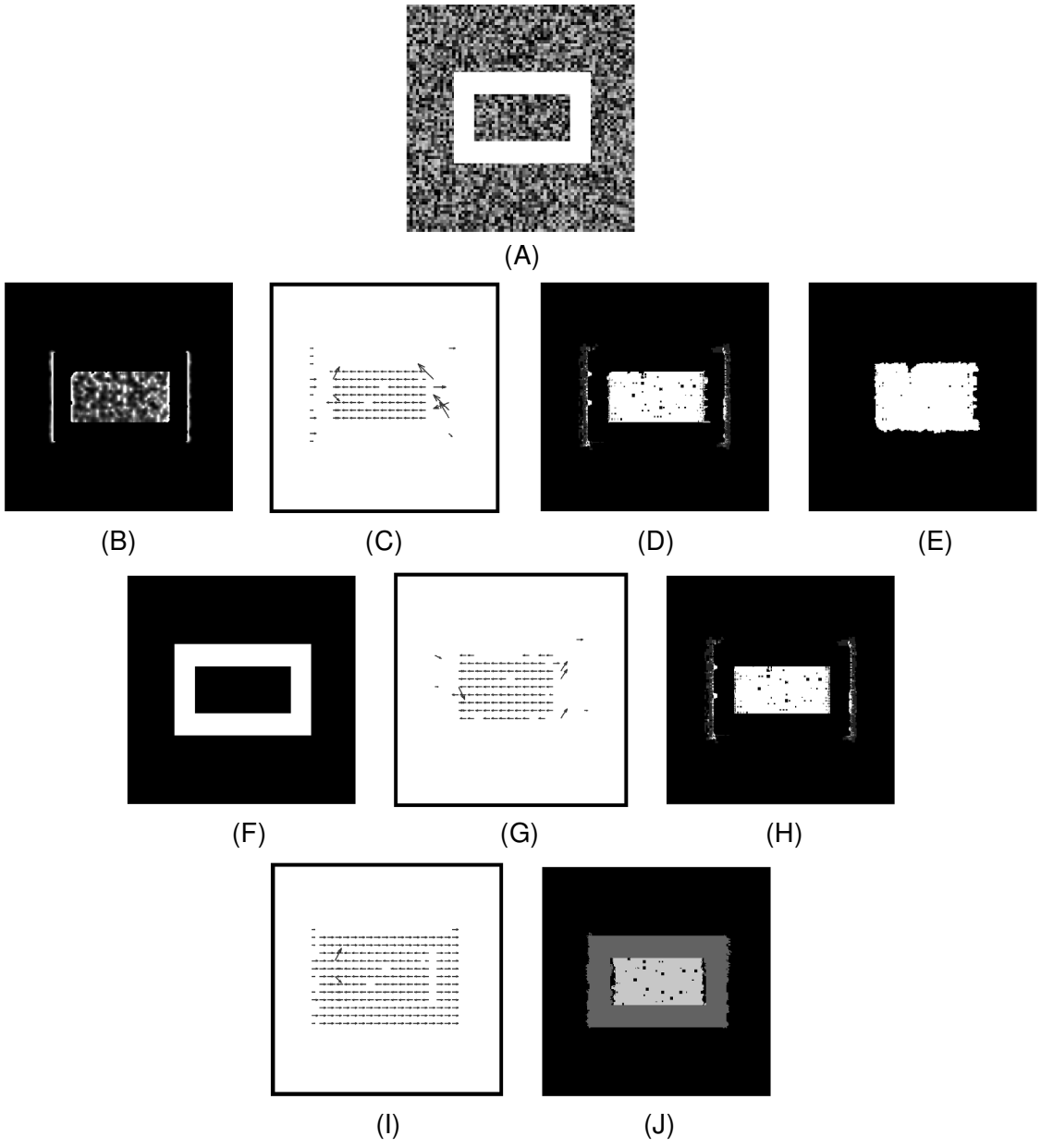

Figure 13. (A) An input scene composed of textured and homogeneous regions. (B) Mobility image. (C–D) Estimates and certainties. (E) Initial motion segmentation. (F) Luminance segmentation. (G–H) On the basis of the occlusion relationship obtained, unreliable estimates and their certainties are eliminated. (I) Common velocity of the luminance segment is filled in within its locations. (J) Final motion segmentation based on the refined estimates.

ping area are assigned both velocities. Thus, the final result includes two overlapping square segments, as depicted in Figure 12K.

Homogeneous Region Surrounded by Textured Region

In Figure 13, we consider a scene composed of a homogeneous region surrounded by textured regions. In a stationary textured background, a homogeneous rectangular annulus is moving to the right while the inner textured region is moving in the opposite direction, as is shown in Figure 13A. Similar to the preceding examples, first, the mobility image is obtained in the motion pathway and is shown in Figure 13B. Next, local motions and their certainties are estimated at locations with large mobility values, as depicted in Figures 13C and 13D. As is shown in Figure 13E, the initial motion segmentation is based on these estimates and is not accurate, since the annular re-

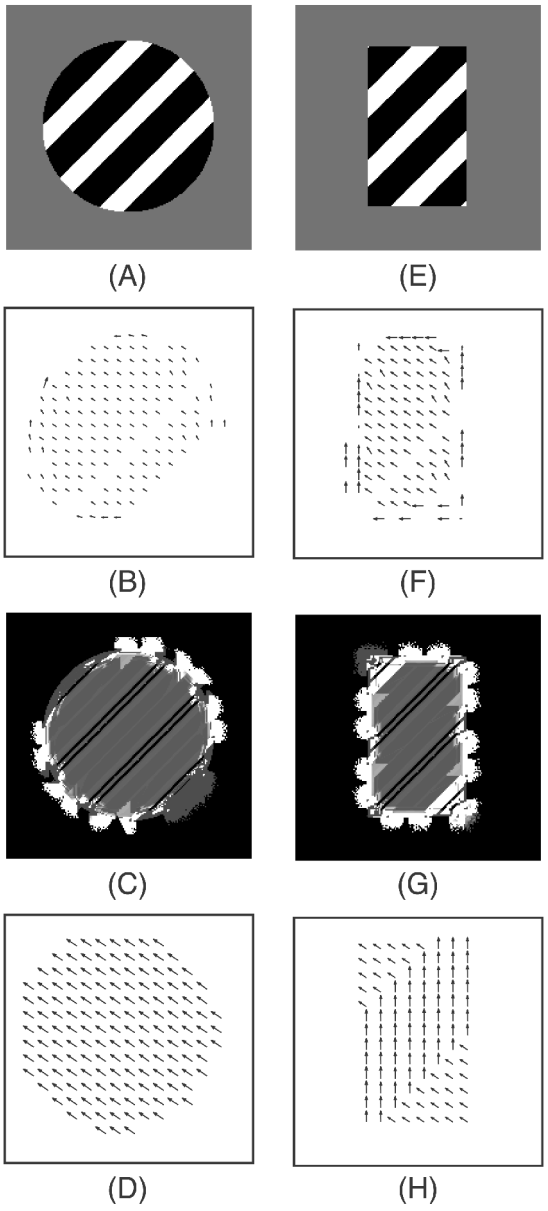

Figure 14. The barber pole illusion. Gratings are moving vertically upward, when the aperture is (A–D) circular and (E–H) rectangular. (B and F) Initial local motion estimates. (C and G) Their certainties. (D and H) Final motion distributions. Despite having the same motion, gratings appear to move differently depending on the aperture shape.

gion includes erroneous estimates along the boundary of the inner region. However, the segmentation result in the luminance pathway, as shown in Figure 13F, captures the annular region well, owing to its homogeneous luminance. In the subsequent integration stage, the occlusion analysis is performed using these segmentation results. Since the motion segment includes locations from both textured and homogeneous regions, two types of motion distributions are obtained in the homogeneous annular region. On the basis of these distributions, the inner textured region and the annular region are determined to be moving independently. As a result, estimates in the annular region along the boundary of the inner textured region are assumed to be unreliable and are eliminated. Figures 13G and 13H show the remaining estimates and their certainties. Finally, the motion interaction takes place in the luminance segment and fills in its locations with the resulting segment velocity, as is illustrated in Figure 13I. Note that estimates in the textured regions stay the same. The final segmentation result, based on the refined estimates, is shown in Figure 13J and compares well with the regions and their motions in the input scene.

The Barber Pole Illusion

An intriguing visual illusion occurs when gratings move behind an aperture (Wallach, 1935). Namely, the perceived direction of motion of a grating can be modified merely by changing the shape of the aperture. Gratings in Figures 14A and 14E move vertically upward behind a circular and a rectangular aperture, respectively, at a speed of 5 pixels per frame. When the aperture is circular, local estimates are, in general, in a direction perpendicular to the orientation of gratings, as is shown in Figure 14B. In contrast, estimates within the rectangular aperture show variations in direction, as is depicted in Figure 14F. Estimates closer to the aperture border are in a direction parallel to the border orientation, whereas the others are in a direction perpendicular to the grating orientation. Certainties in both aperture types are similar as shown in Figures 14C and 14G. Estimates closer to the aperture borders have larger certainties than those of inner locations. Following the initial segmentations, our model detects T-junctions in the occlusion analysis.

When we assume that the aperture borders are intrinsic boundaries of the gratings by ignoring T-junctions, we obtain the distributions of segment velocities in Figures 14D and 14H, corresponding to the circular and the rectangular apertures, respectively. In the case in which the aperture is circular, gratings appear to move in a direction perpendicular to their orientation. However, when the aperture is rectangular, except for the locations closer to the upper-left and the lower-right corners of the aperture, gratings appear to move parallel to the longer axis of the aperture—namely, vertically upward—mimicking the barber pole illusion (Wallach, 1935). Note that motion distributions at locations closer to the upper-left and lower-right corners are also consistent with human perception. When we assume that the aperture borders are extrinsic, the gratings appear to move in a direction perpendicular to their orientation in both cases. The model results obtained with both intrinsic and extrinsic borders are consistent with studies of human motion perception (Shimojo et al., 1989; Trueswell & Hayhoe, 1993).

Plaids

A square plaid can be created by the spatial repetition of a simple tile pattern, as is shown in Figure 15A. A characteristic parameter of gratings and, hence, plaids is their *duty cycle*, which is defined as the ratio $l_1/(l_1 + l_2)$. In gen-

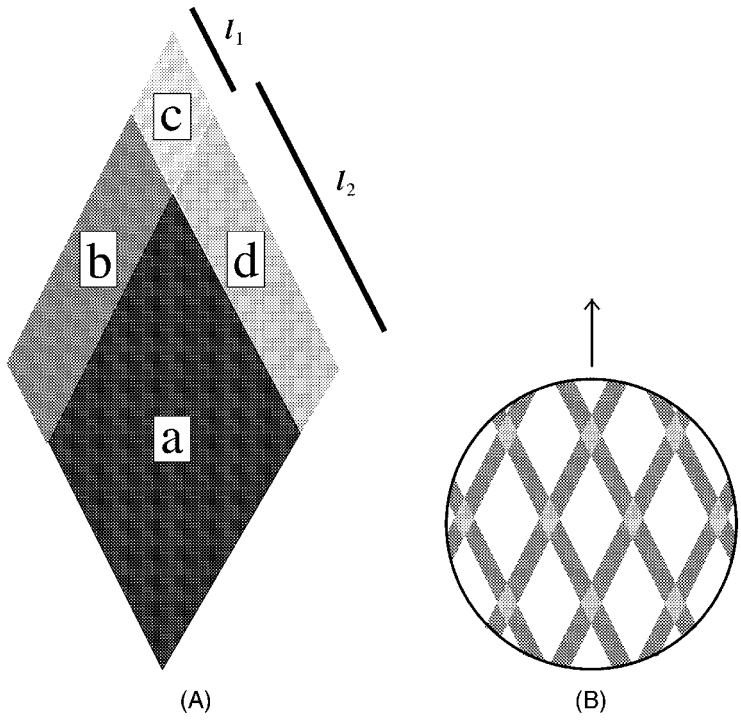

Figure 15. (A) A generic tile to form a square plaid. Each region is a parallelogram (e.g., region *b* has sides of length l_1 and l_2). (B) One frame of a symmetric plaid motion sequence. In our results, plaids move vertically upward behind a circular aperture.

eral, all four regions—namely, *a*, *b*, *c*, and *d*—can have different luminances. When plaids have the same luminance in regions *b* and *d*, we refer to them as *symmetric plaids.* Otherwise, they are called *asymmetric plaids* (see Figures 18A–20A).

Recently, an interesting phenomenon has been observed with plaids, where their perceived motion changes in a predicted way under the influence of the luminance in region *c*, L_c, (Lindsey & Todd, 1996; Stoner et al., 1990). In their studies, Lindsey and Todd (1996) and Stoner et al. reported that plaids are more likely to be perceived to have noncoherent motion when region *c* appears to be formed by the transparent overlapping of regions *b* and *d*, as is described in the introduction of this paper. In order to investigate this phenomenon, we apply our model to plaid stimuli composed of three consecutive frames. A frame of a symmetric plaid sequence, for instance, is shown in Figure 15. In our results, plaid (coherent) motion is 5 pixels per frame in the upward direction. When we apply our model, the reliability criterion allows estimates only along the grating boundaries (cf. Figure 1). Owing to the homogeneity in regions, the entire scene is segmented into a collection of parallelograms, such as regions, *a*, *b*, *c*, and *d*, in the luminance pathway. Next, the initial motion segmentation is obtained, and the occlusion analysis is performed. Depending on the presence of X-junctions, unreliable estimates are eliminated. The inclusion of detected T-junctions formed by the gratings and the aperture border does not affect our results, and so we ignore them in the subsequent analysis. Because of the spatial regularity in the segments and the proximity between regions having the same luminance, we introduce a small extension into the motion interaction mechanism. This extension does not affect earlier results in this paper and is not implemented for the sake of simplicity. According to the extension, instead of considering a single segment, local motion estimates in a set of segments having the same luminance are allowed to interact. For instance, all segments corresponding to region *a*s are treated as a single segment in Equation 14. Having obtained refined motion estimates in all luminance segments, we define a probability measure in terms of area, in order to compare the model output with the psychophysical data. For this probability, the total area of the regions having their refined motion estimates in noncoherent motion directions is calculated and normalized by the area of the circular aperture. Accordingly, when the majority of the segments have noncoherent motion, their total area becomes large, and so does the probability of noncoherent motion for the input plaid stimulus.

We first analyze a set of symmetric plaids. Three examples from this set are shown in Figure 16A. Here, $L_a = 255$, $L_b = L_d = 100$, and from left to right, L_c is 20, 60, and 120, respectively. In all plots, as in Figure 16B, the vertical axis

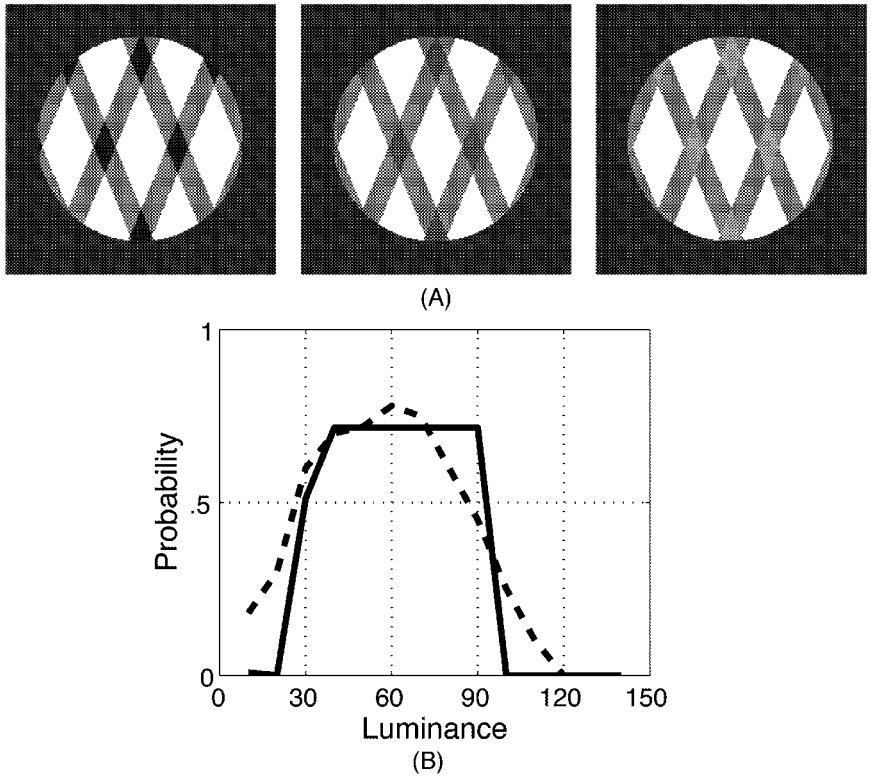

Figure 16. (A) Three different symmetric plaids where $L_a = 255$, $L_b = L_d = 100$, and, from left to right, L_c is 20, 60, and 120, respectively. (B) Resulting final motion types. The vertical axis is the probability of noncoherent motion, whereas the horizontal axis is L_c varying between 10 and 140. The solid line represents the model result, whereas the dotted line is the psychophysical data from Stoner, Albright, and Ramachandran (1990). Consistent with the psychophysical data, when L_c falls into the range of transparent interpretation ($39 \leq L_c < 100$), noncoherent motion is favored.

shows the probability of noncoherent motion perception. In Figure 16B, the horizontal axis represents the luminance in region c, L_c. For each point in the plot, we use the corresponding L_c in the input plaid and apply our model to calculate the probability. In all plots, including this one, the solid line corresponds to the model output. The dotted line in Figure 16B represents the corresponding psychophysical data from the study by Stoner et al. (1990). Examining the plot, we observe that a noncoherent motion is more frequently perceived when $L_b^2/L_a \leq L_c \leq L_a$—namely, $39 \leq L_c \leq 100$ with our values. This result is quantitatively consistent with the psychophysical data. In general, the model results in three major groups of local motion estimates in plaids. The first one is composed of estimates in the coherent motion direction, whereas the other two are those in the two noncoherent motion directions (cf. Figure 1A). Estimates in the intersection regions of gratings are dominated by the former group, whereas those in the remaining regions are the mixture of all groups. Especially, estimates at locations in the vicinity of the intersection regions are in the coherent motion direction. Thus, when an X-junction is detected, estimates at these locations are determined to belong to an occluded surface and are eliminated, reducing the total number of locations supporting the coherent motion. As a result, the probability of noncoherent motion perception in occluded regions increases. However, in the absence of X-junctions, no estimate is eliminated, and thus, estimates in the coherent motion direction are more likely to determine the overall motion perception.

In the second set, we consider the variations in the duty cycle of symmetric plaids. In Figure 17A, plaids with duty cycles of 30%, 50%, and 70% are shown from left to right, respectively. There is no psychophysical data available in the literature assessing the role of the duty cycle quantitatively. However, it has been observed that an increase in the duty cycle decreases the probability of noncoherent motion perception (Stoner & Albright, 1996). In Figure 17B, the horizontal axis is the duty cycle, whereas $L_a = 255$, $L_b = L_d = 100$, and $L_c = 60$. Our result is consistent with the qualitative observation as shown in Figure 17B. Recall that there are three groups of local motion estimates in plaids and that estimates in the intersection regions mostly support the coherent motion perception. As a result of increasing duty cycles, the intersection regions become larger, and so does the support for the coherent

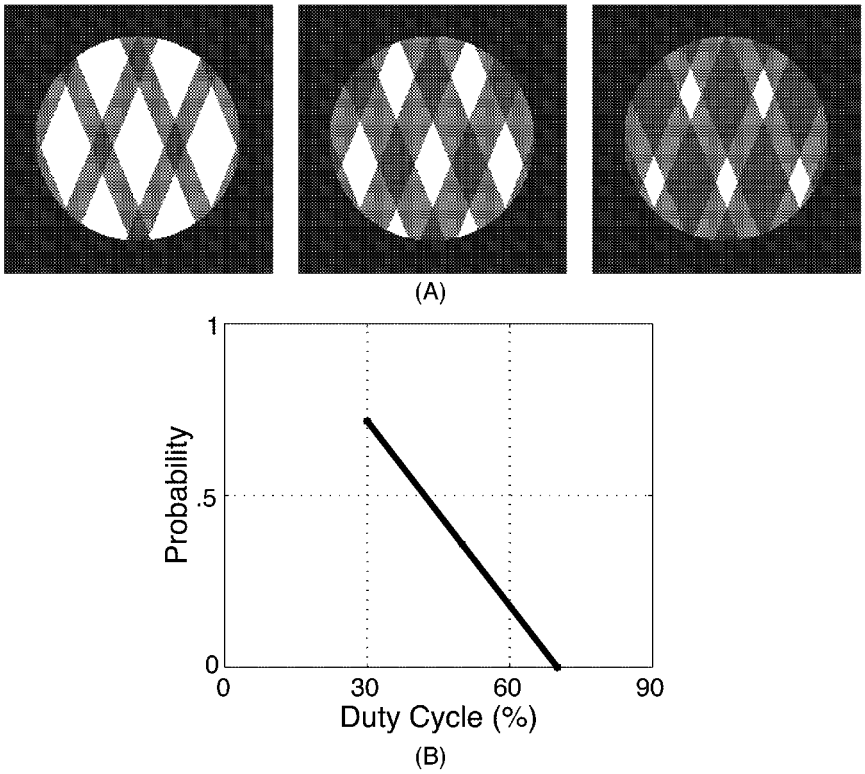

Figure 17. (A) Symmetric plaids with three different duty cycles of 30%, 50%, and 70% are shown from left to right, respectively. (B) The model results in a decreasing probability of noncoherent motion as the duty cycle increases.

motion, leading to a decrease in the probability of noncoherent motion perception.

In order to investigate further the role of luminance in the motion perception of plaids in more detail, we apply the model to the set of asymmetric plaids used by Lindsey and Todd (1996). These plaids are formed by two narrow gratings having the duty cycle of 30%. The gratings are at the angles of $\pm 22.5°$ with the plaid motion direction. Following the procedure used by Lindsey and Todd, we use two different palettes—namely, P_1 and P_2—in the construction of these plaids. The luminances in P_1 were 0, 128, 192, and 255, and those in P_2 were 0, 64, 128, and 255. In addition, we have three different categories based on the number of possible transparent interpretations they allow. The first one (Category I) has a luminance configuration that allows two different transparent interpretations (Figure 18A). In Category I, both of the narrow gratings satisfy the transparency constraints, to appear on top of the other regions. The second category (Category II) has only one transparent interpretation, in which the grating with the darker luminance appears to be in the front (Figure 19A). In the final category (Category III), transparency is not observed (Figure 20A).

We specify configurations in each category with a four-letter string—for example, *abdc*, *dacb*. Each letter corresponds to one luminance, and luminances are in ascending order in a string. For each category, there is a prototype configuration. According to each prototype, luminances from the palettes are reordered. For example, in Category I, the prototype is *abdc*, and the reordered luminances are 0, 128, 255, and 192 for P_1 and 0, 64, 255, and 128 for P_2. Next, reordered luminances are assigned to the regions starting from region *a* and visit other regions clockwise in Figure 15A. Having defined the prototype configuration, three more configurations are obtained by rotating luminance assignments one region at a time in the clockwise direction. For example, the next configuration for Category I, using P_1, has the luminances of 192, 0, 128, and 255 for regions *a*, *b*, *c*, and *d*, respectively. Since the letters in the string should be in ascending order, this configuration uniquely corresponds to *bcad*. Similarly, the prototype of Category II is *abcd*, and that of Category III is *acbd*. Since there are three categories, where each of them has four different configurations for each palette, and there are two palettes, we have a total of 24 different plaid stimuli.

Having obtained the input sets, three consecutive frames from each plaid sequence are fed into our system. Performing the area analysis for the probability calculation of noncoherent motion perception, we obtain the plots given in Figures 18–20. In each plot, the horizontal axis is the luminance configuration. The left and the right plots in

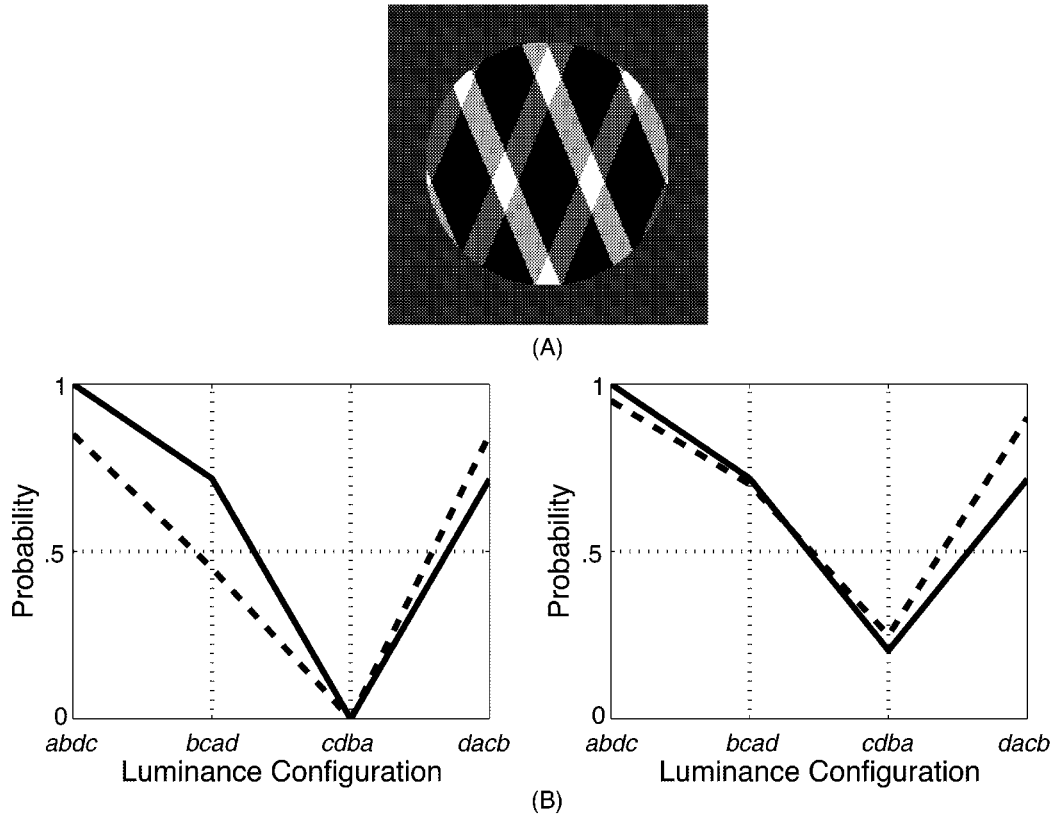

Figure 18. (A) An example plaid stimulus (*abdc*) from Category I, which is obtained using P_1. (B) The left and right plots correspond to P_1 and P_2, respectively. The vertical axis is the probability of noncoherent motion perception, and the horizontal axis is the luminance configuration. The solid line represents the model result, whereas the dotted line is the psychophysical data from Lindsey and Todd (1996). Our model captures the dependence of noncoherent motion perception on the luminance configuration in Category I.

each figure correspond to P_1 and P_2, respectively. In all plots, the dotted line is the psychophysical data from Lindsey and Todd (1996).

Note that in Category I, there are two different possible transparency interpretations. Among them, we select the interpretation that assumes the brighter grating to be in the front. This decision is consistent with human perception and with the depth-from-contrast cue presented in Egusa (1982). According to this cue, the regions that contrast more with their background tend to be seen as closer in depth than do regions with lesser contrast. By comparing the plots in Figure 18B, we observe that the model result is in good agreement with the data. Except for the configuration *cdba*, all the configurations have a relatively high probability of resulting in noncoherent perception. In fact, the model is able to capture the difference between P_1 and P_2 as reflected in the model response to the configuration *cdba*.

Similarly, in Category II—except, perhaps, for the configuration *abcd* with P_1—the model predicts the general characteristics of the data as shown in Figure 19B. Finally, in Category III, the model captures the configuration dependence in the data, as is depicted in Figure 20B, although the model overestimates the probability of noncoherent motion for configuration *acbd*. This mismatch between data and simulation for this configuration suggests that the model may underestimate the importance of static transparency cues in plaid motion analysis. As is shown in Figure 20A, the highest contrast border in this configuration corresponds to the nonoverlapping portion of one of the plaid bars. The high contrast of this region apparently provides a powerful diagonal motion component (suggesting noncoherent motion) that effectively overrides the results of X-junction analysis (suggesting coherent motion). One way in which the results could have been improved would have been to adjust the threshold θ_s for the sum of mobility values or M_s, which determines the spatial pooling neighborhood, N_B, or both. However, we chose to conduct all simulations with the same set of model parameters; there were no free parameters to manipulate across the various experiments simulated in this paper.

Despite these imperfections in the model, it can be seen that it is able to capture the general behavior in the psychophysical data for plaid motion. Instead of producing a response curve composed solely of points indicating a probability of one or zero, it also generates probabilities

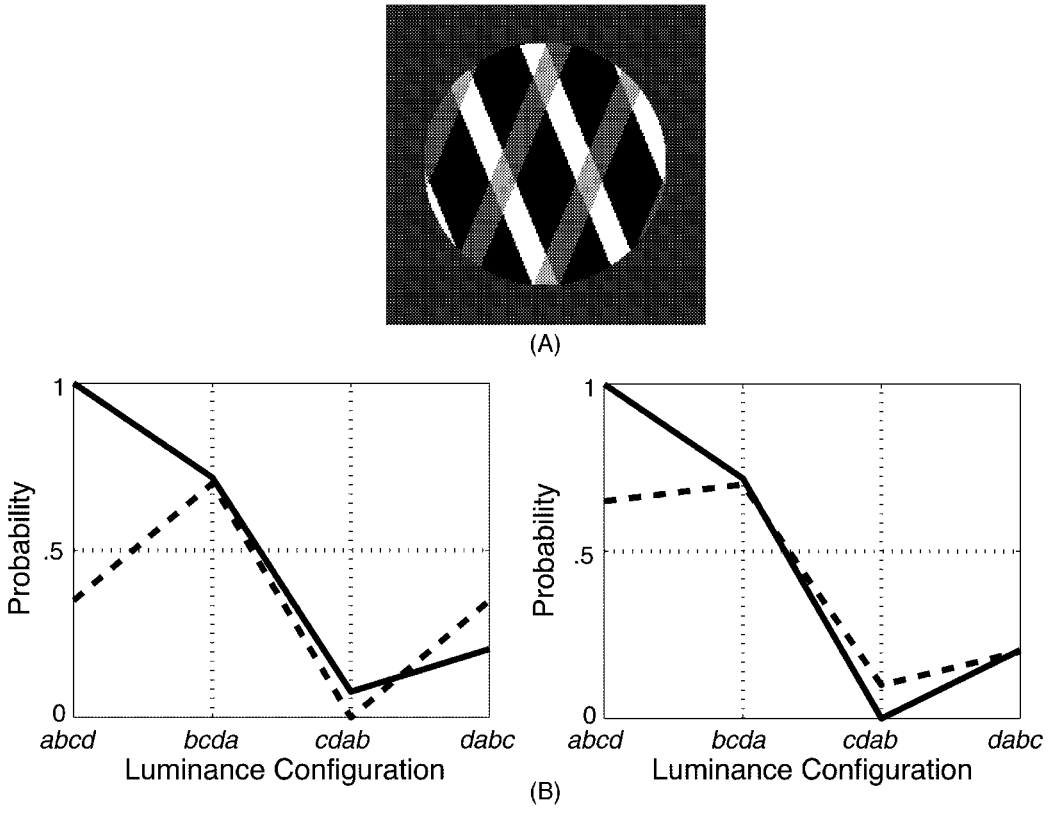

Figure 19. (A) An example plaid stimulus (*abcd*) from Category II, which is obtained using P_1. (B) The left and right plots correspond to P_1 and P_2, respectively. The vertical axis is the probability of noncoherent motion perception, and the horizontal axis is the luminance configuration. The solid line represents the model result, whereas the dotted line is the psychophysical data from Lindsey and Todd (1996). Our model, in general, captures the dependence of noncoherent motion perception on the luminance configuration. However, it overestimates the probability for the configuration *abcd*, as compared with that of the data in Category II.

of intermediate values, and the resulting response curve generally conforms to the data obtained psychophysically from human subjects.

DISCUSSION

Over the past 15 years, a number of models of motion perception have been proposed for the recovery of object motion. These models generally have focused on the use of local motion signals as the principal source of information in the solution to this problem (e.g., Giese, 1999; Hildreth, 1983; Simoncelli & Heeger, 1998; Wilson et al., 1992; Wilson & Kim, 1994). We have previously presented such a model, based on a neural network, that has proved successful in the recovery of object motion from movies of natural scenes (Cesmeli & Wang, 2000). In those scenes, surfaces are textured, and local motion estimates often provide reasonable first approximations to actual object motions. However, as we pointed out in the introduction of the present paper, there are a number of aspects of motion perception that cannot be dealt with without incorporating an additional analysis of the surface properties of objects whose patterns of motions are being determined. In the present paper, we have focused on two of these aspects: the so-called blank-wall problem and motion in overlapping transparent surfaces. Our general approach to this problem has been to segment surface information early in the visual processing chain and to integrate this information with that obtained from low-level motion sensors in a parallel channel similar to one we have described previously. Our approach has been motivated not only by its intrinsic plausibility, but also by a number of lines of evidence suggesting that the human visual system performs an analysis of object surfaces early in the visual processing chain (Nakayama, He, & Shimojo, 1995).

Our analysis of multiple overlapping rectangles moving in different directions in the Overlapping Opaque Rectangles section above (see Figure 11) illustrates the benefits of integrating luminance and motion information when object surfaces are uniform in luminance. Figure 11E shows that motion segmentation alone leads to very poor recovery of the three individual moving surfaces. However, using information from the luminance segments not only facilitates filling-in of homogeneous regions with ve-

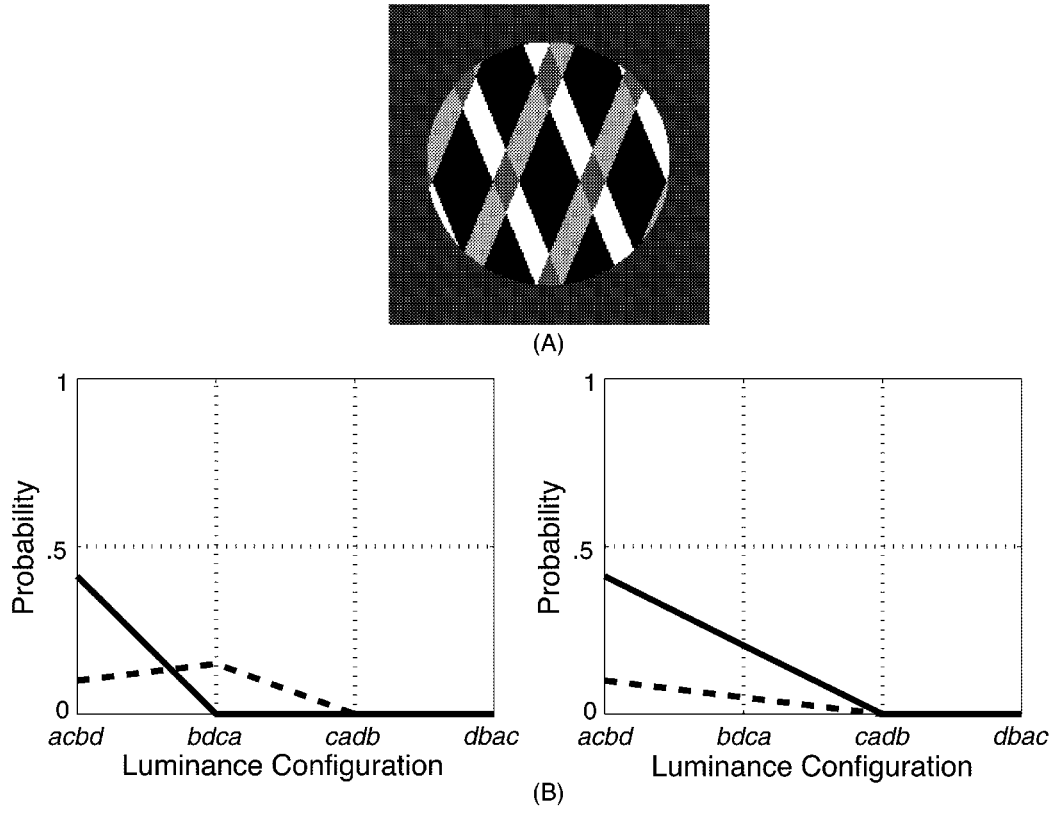

Figure 20. (A) An example plaid stimulus (*acbd*) from Category III, which is obtained using P_1. (B) The left and right two plots correspond to P_1 and P_2, respectively. The vertical axis is probability of noncoherent motion perception, and the horizontal axis is the luminance configuration. The solid line represents the model result, whereas the dotted line is the psychophysical data from Lindsey and Todd (1996). Our model, in general, captures the shape of the probability of noncoherent motion perception in Category III.

locity estimates obtained at the region boundaries; the luminance segments also support analysis of occlusion relationships among overlapping moving objects, and this analysis facilitates the assignments of velocities to both locally occluding and occluded regions of these objects. The analysis of occlusion relationships also allows the model to deal with motion transparency, as in the case of the two overlapping rectangles illustrated in Figure 12.

Another important feature of the present model is dynamic pooling and the use of certainty measures of local motion in these pooling processes. The need for some form of spatiotemporal pooling of motion responses has been appreciated for a long time. Direct psychophysical evidence for pooling has come from studies of the ability of humans to perceive the average speed and direction of dynamic random dot arrays in which dots assume a range of speeds and/or directions (e.g., Hiris & Blake, 1995; Watamaniuk & Duchon, 1992; Watamaniuk & Sekuler, 1992; Watamaniuk, Sekuler, & Williams, 1989; Williams & Sekuler, 1984).

Evidence for pooling also has come from studies that show large improvements in motion detection thresholds when stimulus size and duration are increased (e.g., Fredericksen, Verstraten, & van de Grind, 1993, 1994a, 1994b, 1994c; Lindsey & Todd, 1998; van Doorn & Koenderink, 1982, 1984).

A number of previous models have employed fixed pooling regions (e.g., Black & Anandan, 1996; Black & Jepson, 1996; Enkelmann, 1988; Horn & Schunck, 1981). Other models have employed cooperative and competitive interactions among front-end motion sensors differing in spatiotemporal tuning and location (e.g., Chey, Grossberg, & Mingolla, 1997; Giese, 1999; Grzywacz & Yuille, 1995; Marshall, 1991; Nowlan & Sejnowski, 1995; J. Y. A. Wang & Adelson, 1994; R. Wang, 1997; Weber & Malik, 1997; Weiss & Adelson, 1996; Wilson & Kim, 1994). Few models of motion processing have used reliability measures of local motion estimation to guide pooling. One approach similar to the present one is that described by Nowlan and Seijnowski (1994, 1995). In their model, however, reliability measures are based on training of a network; we compute these measures explicitly from the responses of low-level motion sensors.

In the model we have presented, dynamic pooling occurs at two different sites: very early in motion processing, as estimates of local motion are being determined, and again at the level of motion integration, as motion and luminance information are being integrated (see Figure 2).

Pooling at the early stage of local motion estimation dynamically adjusts the sizes and shapes of the correlation blocks employed by motion sensors. In this way, motion estimation is tuned to the sizes and shapes of the image regions being analyzed. As we noted earlier, an interesting and important consequence of this pooling strategy is the appearance of classical orientation tuning of motion sensors in image regions containing objects with straight contours, even though the sensors are not themselves intrinsically orientation tuned.

In calculating flow fields for gratings moving within rectangular apertures, we have shown how sensor responses and certainty measures of these responses interact at the integration stage of the model. There are two sources of sensor responses to these stimuli: those obtained from the grating bars and those obtained from the intersections of grating bars with the boundaries of the aperture, which we referred to as *edge terminators*. The grating bars generate large responses from many motion sensors, and the resulting distributions of velocities are broad, owing to the aperture problem. Motion estimation of the edge terminators, on the other hand, does not suffer from the aperture problem, and although these features generate modest responses, the distribution of responses is considerably narrower than those obtained for the grating bars. Responses from these terminators are, therefore, weighted relatively more heavily than those from the grating bars. The computed flow field for a given grating/aperture configuration will depend on interactions among the relative responses to the grating edges and the edge terminators and the relative reliabilities of these various measurements. When the grating aperture is circularly symmetric, motion estimates from both grating contours and line terminators yield estimates of motion distributed around a direction perpendicular to the grating orientation, and the calculated flow field reflects this average direction. However, when the aperture is rectangular in shape with an aspect ratio that deviates markedly from 1, the high-reliability motion estimates obtained from the line terminators augment one another and, because of their relatively high weights, dominate the pooling process, and the computed flow field is directed parallel to the long side of the aperture, except near the corners, where motion estimates from the line terminators from adjacent sides of the aperture give conflicting velocity estimates. In these corner regions, the calculated flow field is perpendicular to the grating bars, a result that is consistent with human motion perception (Wallach, 1935). Similar results have been obtained from a model of motion reported by Chey et al. (1997). In that model, line terminators are treated as special features that guide attentive motion perception. In our model, the line terminators are accorded special status only by virtue of their providing high-reliability local estimates of object motion.

Our calculations of model responses to plaids have illustrated how all three principal features of the present model—dynamic pooling, motion reliability estimation, and motion–luminance integration—interact. X-junction analysis performed on the output of the luminance segmentation pathway is used to detect the static transparency cues and, if transparency is detected, the implied depth order of the surfaces being analyzed. Edge intersections in the regions of putative transparent occlusion generate motion responses with modest/high reliability; grating edges produce high-total-response/low-reliability motion responses. Each of these aspects of the low-level analysis—X-junction detection and edge and intersection motion—interact to predict the likelihoods of transparent motion perception as functions of the relative grating luminances and grating duty cycles. In general, the predicted relationships are remarkably close to those obtained from human observers, especially considering the simplicity of the motion model we employ.

In developing our model of motion perception, we have been motivated by the goal of providing a neural network account of motion perception. Such an approach is intended to mimic interactions occurring among neurons in humans and, therefore, to provide a biologically plausible simulation of human motion perception. Nonetheless, the principal constraint in the development of the present version of the model has been the perceptual data on human subjects and the computational adequacy of our model in accounting for these phenomena. Therefore, our approach to modeling motion perception is distinctly different from the approaches of others who have based their models solely on the properties of neurons known to be sensitive to motion (e.g., Simoncelli & Heeger, 1998). Thus, although we claim that the basic functions of our model components mimic basic processes involved in human motion perception, we do not claim a one-to-one correspondence between model structures and functions, on the one hand, and biological structures and functions, on the other. In this regard, several aspects of the present model require further comment. First, unlike most modern, biologically oriented models of human motion perception, we employ low-level motion sensors based on temporal block matching, rather than orientation-tuned motion energy detectors. Block matching was chosen because it is less intensive computationally than motion energy measurement, and we do not think that our approach to modeling human motion perception depends critically on our choice of low-level local motion sensors. Neither block matching nor local motion energy computations are sufficient for the computation of object motion, which requires additional pooling processes. Our model focuses on the processes involved in pooling low-level motion signals derived from multiple sensor populations, rather than on the specifics of how these low-level signals are obtained.

Second, our image segmentation pathway is based on local analyzers that measure static local luminance. An alternative approach has been described by Grossberg and Mingolla (1985, 1987). In their scheme, surface luminance is recovered by two interacting subprocesses: one that recovers object boundaries and another that integrates

surface features within those regions enclosed by those boundaries. This scheme has been successful in accounting for a wide variety of phenomena involving the perceptual organization of luminance- and color-defined boundaries. Although we expect that these additional features could allow our model to deal with a broader range of stimuli than we have examined in this paper, we do not expect these additions to alter the conclusions of the present study.

Third, we have not included a module for computing second-order motion in this paper. Such a module could be incorporated following standard models proposed by others. In these models, a pointwise nonlinearity is applied to the image data, and the results are low-pass filtered. First-order motion analysis is then performed on the filtered images (e.g., Lu & Sperling, 1995).

Fourth, all of our simulations involved analysis of three-frame animation sequences, which we chose for computational convenience. In general, the corresponding psychophysical data that we modeled were often obtained from longer animation sequences; for example, Lindsey and Todd's (1996) results simulated in the preceding section of our paper were obtained from 45-frame sequences. We could have extended the input integration time (as well as reduce the input gain) in our model to more closely match the time scale of the original experiments. Clearly, some aspects of plaid motion processing depend on stimulus duration, particularly when stimulus contrast is low. For example, Yo and Wilson (1992) have shown that the perceived direction of coherent motion in plaids is duration dependent, presumably owing to different latencies in the processing of first- and second-order motion components in the stimulus. The results of our simulations suggest that the model, in some cases, may slightly underestimate the contribution of static transparency cues in moving plaid perception. Perhaps, in human subjects, this contribution depends on stimulus duration in a way not captured by the present model.

Fifth, our model was designed to be relatively immune to the effects of dynamic noise of the kind encountered in high-quality video recordings of natural scenes (see Cesmeli & Wang, 2000). To this end, the motion estimation and mobility analysis components of the model attach greater weights to those regions of the scene that show low variance in motion signals across limited ranges of space and time. Because spatiotemporal integration is limited, the model's immunity to dynamic noise will be as well. In comparing this aspect of the model with human psychophysical performance, we note that while motion detection shows remarkable noise immunity, the perception of object boundaries between an object and its background does not show the same immunity to dynamic noise (see, e.g., Beverley & Regan, 1984), presumably because of tradeoffs in spatiotemporal integration that determine noise immunity and spatiotemporal resolution.

The final aspect of our model that requires some comment is our decision to base the neural network portions of the model on oscillatory correlation. The advantages of this approach are that not only does it support cooperative and competitive interactions among the various elements in the model that lead to image and motion segmentation and transparency, it also explicitly gives a solution to the binding problem—that is, the unification of responses in many different neuronal elements into groups of neurons that represent objects. In our model, groups of neurons representing an object fire in synchrony, and different objects are represented by different groups of neurons that desynchronize from each other. Although the relevance of oscillatory correlation to solving the binding problem is debatable, there is a growing body of neurobiological evidence that supports our approach (Buzsaki, Llinas, Singer, Berthoz, & Christen, 1994; Castelo-Branco, Goebel, Neuenschwander, & Singer, 2000; Eckhorn et al., 1988; Gray, Konig, Engel, & Singer, 1989; Keil, Mueller, Ray, Gruber, & Elbert, 1999; Livingstone, 1996). Furthermore, our approach is consistent with psychophysical results showing that spatiotemporal synchrony gives rise to perceptual organization (Elliot & Muller, 1988; Lee & Blake, 1999; Usher & Donnelly, 1998).

In conclusion, we have described and evaluated a model of motion perception based on the integration of motion and luminance information obtained in two parallel segmentation pathways. This approach allows the model to deal not only with situations involving textured objects and backgrounds, but also with a variety of situations involving uniformly luminant surfaces.

REFERENCES

Anandan, P. (1987). *Measuring visual motion from image sequences.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Beck, J., Pradzny, K., & Ivry, R. (1984). The perception of transparency with achromatic colors. *Perception & Psychophysics*, **35**, 407-422.

Beverley, K. I., & Regan, D. (1984). Figure–ground segregation by motion contrast and by luminance contrast. *Journal of the Optical Society of America A*, **1**, 433-442.

Black, M. J., & Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision & Image Understanding*, **63**, 75-104.

Black, M. J., & Jepson, A. D. (1996). Estimating optical flow in segmented images using variable- 45 order parametric models with local deformations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **18**, 972-986.

Buzsaki, G., Llinas, R., Singer, W., Berthoz, A., & Christen, Y. (1994). *Temporal coding in the brain.* Berlin: Springer-Verlag.

Castelo-Branco, M., Goebel, R., Neuenschwander, S., & Singer, W. (2000). Neural synchrony correlates with surface segregation rules. *Nature*, **405**, 334-337.

Cavanagh, P. (1987). Reconstructing the third dimension: Interactions between color, texture, motion, binocular disparity and shape. *Computer Vision, Graphics, & Image Processing*, **37**, 171-195.

Cesmeli, E., Lindsey, D. T., & Wang, D. L. (1999). Integration of static luminance and motion analyses: A step towards solving the blank wall problem. *Investigative Ophthalmology & Visual Science*, **40**, S423.

Cesmeli, E., & Wang, D. L. (2000). Motion segmentation based on motion/brightness integration and oscillatory correlation. *IEEE Transactions on Neural Networks*, **11**, 935-947.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction. *Journal of the Optical Society of America A*, **14**, 2570-2594.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W.,

MUNK, M., & REITBOECK, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics*, **60**, 121-130.

EGUSA, H. (1982). Effect of brightness on perceived distance as a figure–ground phenomenon. *Perception*, **11**, 671-676.

ELLIOT, M. A., & MULLER, H. J. (1988). Synchronous information presented in 40-Hz flicker enhances visual feature binding. *Psychological Science*, **9**, 277-283.

ENKELMANN, W. (1988). Investigations of multigrid algorithms for the estimation of optical flow fields in image sequences. *Computer Vision, Graphics, & Image Processing*, **43**, 150-177.

FREDERICKSEN, R. E., VERSTRATEN, F. A. J., & VAN DE GRIND, W. A. (1993). Spatio-temporal characteristics of human motion perception. *Vision Research*, **33**, 1193-1205.

FREDERICKSEN, R. E., VERSTRATEN, F. A. J., & VAN DE GRIND, W. A. (1994a). An analysis of the temporal integration mechanism in human motion perception. *Vision Research*, **34**, 3153-3170.

FREDERICKSEN, R. E., VERSTRATEN, F. A. J., & VAN DE GRIND, W. A. (1994b). Spatial summation and its interaction with the temporal integration mechanism in human motion perception. *Vision Research*, **34**, 3171-3188.

FREDERICKSEN, R. E., VERSTRATEN, F. A. J., & VAN DE GRIND, W. A. (1994c). Temporal integration of random dot apparent motion information in human central vision. *Vision Research*, **34**, 461-476.

GIESE, M. A. (1999). *Dynamic neural field theory for motion perception*. Norwell, MA: Kluwer Academic.

GRAY, C. M., KONIG, P., ENGEL, A. K., & SINGER, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, **338**, 334-337.

GROSSBERG, S., & MINGOLLA, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception & Psychophysics*, **38**, 141-171.

GROSSBERG, S., & MINGOLLA, E. (1987). Neural dynamics of surface perception: Boundary webs, illuminants, and shape from shading. *Computer Vision, Graphics, & Image Processing*, **37**, 116-165.

GRZYWACZ, N. M., & YUILLE, A. L. (1995). Theories of the visual perception of local velocity and coherent motion. In T. V. Papathomas (Ed.), *Early vision and beyond* (pp. 231-252). Cambridge, MA: MIT Press.

HILDRETH, E. C. (1983). *The measurement of visual motion*. Cambridge, MA: MIT Press.

HIRIS, E., & BLAKE, R. (1995). Discrimination of coherent motion when local motion varies in speed and direction. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 308-317.

HORN, B. K. P., & SCHUNCK, B. (1981). Determining optical flow. *Artificial Intelligence*, **17**, 185-203.

IRANI, M., ROUSSO, B., & PELEG, S. (1994). Computing occluding and transparent motions. *International Journal of Computer Vision*, **12**, 5-16.

KEIL, A., MUELLER, M. M., RAY, W. J., GRUBER, T., & ELBERT, T. (1999). Human gamma band activity and perception of a gestalt. *Journal of Neuroscience*, **19**, 7152-7161.

LEE, S. H., & BLAKE, R. (1999). Visual form created solely from temporal structure. *Science*, **284**, 1165-1168.

LINDSEY, D. T., & TODD, J. T. (1996). On the relative contributions of motion energy and transparency to the perception of moving plaids. *Vision Research*, **36**, 207-222.

LINDSEY, D. T., & TODD, J. T. (1998). Opponent motion interactions in the perception of transparent motion. *Perception & Psychophysics*, **60**, 558-574.

LIU, X., & WANG, D. L. (2000). Perceptual organization based on temporal dynamics. In S. A. Solla, T. K. Leen, & K. R. Muller (Eds.), *Advances in neural information processing systems 12* (pp. 38-44). Cambridge, MA: MIT Press.

LIVINGSTONE, M. (1996). Oscillatory firing and interneuronal correlations in squirrel monkey striate cortex. *Journal of Neurophysiology*, **75**, 2467-2485.

LU, Z. L., & SPERLING, G. (1995). The functional architecture of human visual-motion perception. *Vision Research*, **35**, 2697-2722.

MARSHALL, J. A. (1991). Self-organizing neural networks for perception of visual motion. *Neural Networks*, **3**, 45-74.

METELLI, F. (1974, Month). The perception of transparency. *Scientific American*, **230**, 90-98.

NAKAYAMA, K., HE, J. J., & SHIMOJO, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science* (pp. 1-70). Cambridge, MA: MIT Press.

NOWLAN, S. J., & SEJNOWSKI, T. J. (1994). Filter selection for motion segmentation and velocity integration. *Journal of the Optical Society of America A*, **11**, 3177-3200.

NOWLAN, S. J., & SEJNOWSKI, T. J. (1995). A selection model for motion processing in area mt of primates. *Journal of Neuroscience*, **15**, 1195-1214.

PARIDA, L., GEIGER, D., & HUMMEL, R. (1998). Junctions: Detection, classification, and reconstruction. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **20**, 687-698.

QIAN, N., ANDERSEN, R. A., & ADELSON, E. H. (1994). Transparent motion perception as detection of unbalanced motion signals: III. Modeling. *Journal of Neuroscience*, **14**, 7381-7392.

REICHARDT, W. (1961). Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. In W. A. Rosenblith (Ed.), *Sensory communications* (pp. 303-317). Cambridge, MA: MIT Press.

SHIMOJO, S., SILVERMAN, G. H., & NAKAYAMA, K. (1989). Occlusion and the solution to the aperture problem. *Vision Research*, **29**, 619-626.

SIMONCELLI, E. P. (1993). *Distributed representation and analysis of visual motion*. Unpublished doctoral thesis, Massachusetts Institute of Technology, Cambridge.

SIMONCELLI, E. P., & HEEGER, D. J. (1998). A model of neuronal responses in visual area mt. *Vision Research*, **38**, 743-761.

SINGH, A. (1991). *Optical flow computation*. Los Alamitos, CA: IEEE Press.

STONER, G. R., & ALBRIGHT, T. D. (1993). Image segmentation cues in motion processing: Implications for modularity in vision. *Journal of Cognitive Neuroscience*, **5**, 129-149.

STONER, G. R., & ALBRIGHT, T. D. (1996). The interpretation of visual motion: Evidence for surface segmentation mechanisms. *Vision Research*, **36**, 1291-1310.

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

TERMAN, D., & WANG, D. L. (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D*, **81**, 148-176.

TRUESWELL, J. C., & HAYHOE, M. H. (1993). Surface segmentation mechanisms and motion perception. *Vision Research*, **33**, 313-328.

USHER, M., & DONNELLY, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature*, **394**, 179-182.

VAN DOORN, A. J., & KOENDERINK, J. J. (1982). Spatial properties of the visual detectability of moving spatial white noise. *Experimental Brain Research*, **45**, 189-195.

VAN DOORN, A. J., & KOENDERINK, J. J. (1984). Spatiotemporal integration in the detection of coherent motion. *Vision Research*, **24**, 47-53.

VAN SANTEN, J. P. H., & SPERLING, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America A*, **2**, 300-321.

WALLACH, H. (1935). Über visuell wahrgenommene bewegungsrichtung. *Psychologische Forschung*, **20**, 325-380.

WANG, D. L., & TERMAN, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks*, **6**, 283-286.

WANG, D. L., & TERMAN, D. (1997). Image segmentation based on oscillatory correlation. *Neural Computation*, **9**, 805-836 (For errata, see *Neural Computation*, **9**, 1623-1626, 1997)

WANG, J. Y. A., & ADELSON, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing*, **3**, 635-638.

WANG, R. (1997). A network model of motion processing in area mt of primates. *Journal of Computational Neuroscience*, **4**, 287-308.

WATAMANIUK, S. N. J., & DUCHON, A. (1992). The human visual system averages speed information. *Vision Research*, **32**, 931-941.

Watamaniuk, S. N. J., & Sekuler, R. (1992). Temporal and spatial integration in dynamic random dot stimuli. *Vision Research*, **32**, 2341-2347.

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: The integration of direction information. *Vision Research*, **29**, 47-59.

Weber, J., & Malik, J. (1997). Rigid body segmentation and shape description from dense optical flow under weak perspective. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **19**, 139-143.

Weiss, Y. (1998). *Bayesian motion estimation and segmentation.* Unpublished doctoral thesis, Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge.

Weiss, Y., & Adelson, E. H. (1996). A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 321-326). Los Alamitos, CA: IEEE Computer Society Press.

Weiss, Y., & Adelson, E. H. (1997). Vector averaging as Bayesian IOC. *Investigative Ophthalmology & Visual Science*, **38**, S936.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, **5**, 598-604.

Williams, D., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research*, **24**, 55-62.

Wilson, H. R., Ferrera, V. P., & Yo, C. (1992). A psychophysically motivated model for two-dimensional motion perception. *Visual Neuroscience*, **9**, 79-97.

Wilson, H. R., & Kim, J. (1994). A model for motion coherence and transparency. *Visual Neuroscience*, **11**, 1205-1220.

Yo, C., & Wilson, H. R. (1992). Perceived direction of moving 2-dimensional patterns depends on duration, contrast and eccentricity. *Vision Research*, **32**, 135-147.

APPENDIX

We demonstrate the performance of our model with synthetic image sequences and visual stimuli used in psychophysical studies. All the image sequences have three frames with a size of 256×256, and we show only the middle frame as the input scene. We choose $N_M = N_S = 5 \times 5$, which is also the initial size of N_B. The thresholds for M_s, M_a, and the maximum edge length for the block size are $\theta_s = 1.0$, $\theta_a = 0.1$, and $e_m = 15$, respectively. The maximum displacement and the threshold set are $R = 10$ and $(\theta_{M,1}, \theta_{M,2}, \theta_B) = (20, 10, 10)$, respectively, for all the scenes. The standard deviation of 2-D Gaussian in Equation 4, σ, is selected to have the maximum displacement, R, within its halfbandwidth—namely, $\sigma \geq R/\sqrt{[2ln(2)]}$, resulting in $\sigma \cong 8.5$. The size of T- and X-junction templates in the integration stage is 7×7. The template set for T-junctions includes eight different orientations for a T-configuration. There are three different T-configurations in which the vertical part of the T makes three different angles—that is, 45º, 90º, 135º—with the roof of the T. These conditions are tuned to detect the configurations in which the boundary of the occluded surface makes these three different angles with the boundary of the occluding surface. In the detection of X-junctions, note that each of the four regions could be the area where the overlapping surface is. Similar to T-junctions, we allow three different angles between the two oblique lines forming the shape *X*—that is, 45º, 90º, and 135º. Thus, we obtain 12 different configurations. Noting that a particular configuration detected by 1 of the 12 detectors could also occur when this configuration is rotated around the common corner of the four regions, we obtained three additional versions for each of the 12 configurations, resulting in 48 templates in the X-junction set.

In the simulation of LEGION networks, an algorithmic version (D. L. Wang & Terman, 1997) is employed for computational efficiency, for which only the following parameters are needed: local coupling neighborhood, $N = 3 \times 3$, and its contribution, $W = 1.0$; the local similarity threshold, $\theta = 2.0$; the potential neighborhood, $N_p = 7 \times 7$, and its contribution, $W_p = 0.4$; the relevant threshold, $\theta_p = 36.75$; and finally, the contribution of global inhibitor, $W_z = 1.0$. We adjust $\theta_{M,1}$ and $\theta_{M,2}$, the thresholds for the initial and the final segmentations in the motion network, respectively, and θ_B, that of the luminance network. We require that $\theta_{M,1}$ and θ_B initially do not group locations with significantly different motion and luminance. Since refined estimates allow for a sharper distinction, the selection is relatively easier.

Finally, for better visualization, estimates in needle diagrams are displayed after a spatial subsampling by a factor of 10 in all results.

(Manuscript received June 19, 2000;
accepted for publication August 26, 2001.)