NEURAL DYNAMICS OF MOTION INTEGRATION AND SEGMENTATION WITHIN AND ACROSS APERTURES

Stephen Grossberg, Ennio Mingolla and Lavanya Viswanathan

Department of Cognitive and Neural Systems
and
Center for Adaptive Systems
Boston University
677 Beacon Street, Boston, MA 02215

January 2000

Technical Report CAS/CNS-2000-004

Correspondence should be addressed to:

Professor Stephen Grossberg
Department of Cognitive and Neural Systems
Boston University
677 Beacon Street, Boston, MA 02215
email: steve@cns.bu.edu
fax: 617-353-7755

Running Title: Motion Integration and Segmentation

Keywords:
motion integration, motion segmentation, motion capture, aperture problem, feature tracking, MT, MST, neural network

1. Authorship in alphabetical order. SG, EM and LV were supported in part by the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-95-1-0409). SG was also supported in part by the National Science Foundation (NSF IRI-97-20333), and the Office of Naval Research (ONR N00014-95-1-0657). LV was also supported in part by the National Science Foundation (NSF IRI-94-01659), and the Office of Naval Research (ONR N00014-92-J-1309 and ONR N00014-95-1-0657).
Abstract

A neural model is developed of how motion integration and segmentation processes both within and across apertures compute global motion percepts. Figure-ground properties, such as occlusion, influence which motion signals determine the percept. For visible apertures, a line’s extrinsic terminators do not specify true line motion. For invisible apertures, a line’s intrinsic terminators create veridical feature tracking signals. Sparse feature tracking signals can be amplified by directional filtering and competition, then integrated with ambiguous motion signals from line interiors, to determine the global percept. Filtered motion signals activate directional grouping and priming cells, which compete across space to select a winning direction, then feed back to boost consistent long-range filter activities and suppress inconsistent activities. Feedback can also attentionally prime a movement direction. This feedback process is predicted to occur between cortical areas MT and MST. Computer simulations include the barberpole illusion, motion capture, the spotted barberpole, the triple barberpole, the occluded translating square illusion, motion transparency and the chopsticks illusion.
1. Introduction

Visual motion perception requires the solution of the two complementary problems of motion integration and of motion segmentation. The former joins nearby motion signals into a single object, while the latter keeps them separate as belonging to different objects. Wallach (1935; translated by Wuerger, Shapley & Rubin, 1996) first showed that the motion of a featureless line seen behind a circular aperture is perceptually ambiguous: for any real direction of motion, the perceived direction is perpendicular to the orientation of the line, called the normal component of motion. This phenomenon was later called the aperture problem by Marr & Ullman (1981). The aperture problem is faced by any localized neural motion sensor, such as a neuron in the early visual pathway, which responds to a moving local contour through an aperture-like receptive field. Only when the contour within an aperture contains features, such as line terminators, object corners, or high contrast blobs or dots, can a local motion detector accurately measure the direction and velocity of motion.

To solve the twin problems of motion integration and segmentation, the visual system needs to use the relatively few unambiguous motion signals arising from image features to veto and constrain the more numerous ambiguous signals from contour interiors. In addition, the visual system uses contextual interactions to compute a consistent motion direction and velocity when the scene is devoid of any unambiguous motion signals. This paper develops a neural network model that demonstrates how a single hierarchically organized processing stream may be used to explain important data on motion integration and segmentation.

1.1 Plaids: Feature Tracking and Ambiguous Line Interiors

The motion of a grating of parallel lines seen moving behind a circular aperture is ambiguous. However, when two such gratings are superimposed to form a plaid, the perceived motion is not ambiguous. Plaids have therefore been extensively used to study motion perception. Three major mechanisms for the perceived motion of coherent plaids have been presented in the literature.

\[ \text{IOC} = \frac{V_x + V_y}{\sqrt{2}} \]

\[ V_{\text{avg}} = \frac{V_x V_y}{V_x^2 + V_y^2} \]

**FIGURE 1.** Type II plaids: Vector average vs. intersection of constraints (IOC). Dashed lines are the constraint lines for the plaid components. The gray arrows represent the perceived directions of the plaid components. For these two components, the vector average direction of motion is different from the IOC direction.

1. **Vector average.** The vector average solution is one in which the velocity of the plaid appears to be the vector average of the normal components of the plaid’s constituent gratings (Fig. 1).
2. **Intersection of constraints.** A *constraint line*, first defined by Adelson & Movshon (1982), is the locus in velocity space of all possible positions of the leading edge of a bar or line after some time interval $\Delta t$. The constraint line for a featureless bar, or a grating of parallel featureless bars, that is moving behind a circular aperture is parallel to the bar. The authors suggested that the perceived motion of a plaid pattern was defined by the velocity vector of the intersection in velocity space of the constraint lines of the plaid components. They named this the *intersection of constraints (IOC)* solution to the plaid problem. The IOC solution is the mathematically correct solution to the motion perception problem and, hence, is always veridical. However, as noted below, it does not always predict human motion perception even for coherent plaids.

3. **Feature tracking.** When two 1D gratings are superimposed, they form intersections which act as features whose motion can be reliably tracked. Other features are line endings and object corners. A third possible solution to the problem of plaid motion perception is that the visual system may be tracking features instead of computing a vector average or an IOC solution. At intersections or object corners, the IOC solution and the trajectory of the feature are always identical. However, in some non-plaid displays described below, the feature tracking solution differs from the IOC solution.

No consensus exists in the literature about which of these mechanisms best explains motion perception phenomena. Vector averaging tends to uniformize motion signals over discontinuities and is an efficient technique for noise suppression, especially when the feature points themselves provide ambiguous information as in the case of features formed by occlusion. However, Adelson & Movshon (1982) showed that often observers do not see motion in the direction predicted by the vector average of component motion. Ferrera & Wilson (1990, 1991) tested this rigorously by classifying plaids into Type 1 plaids, for whom the IOC solution lies inside the arc formed by the motion vectors normal to the two components, and Type 2 plaids, for whom this is not true (Fig. 1). By definition, the vector average solution always lies inside this arc. They found that, in some cases, the motion of Type 2 plaids is biased away from the IOC solution. Similarly, Rubin & Hochstein (1993) showed that moving lines can sometimes be seen to move in the vector average direction rather than the IOC direction. Further, Mingolla, Todd & Norman (1992), using multiple aperture displays, showed that, in the absence of feature information, perceived global motion was biased toward the vector average solution. However, when features were visible within apertures, the correct motion direction was perceived. Clearly, the IOC solution does not always predict what the visual system sees.

These data suggest that feature tracking signals as well as the normals to component orientations play a major role in the perceived direction of motion. Lorenceau & Shiffrar (1992) show that motion grouping across apertures becomes impossible in the presence of feature tracking signals as these signals invariably capture the motion of the lines that they belong to. In the absence of feature tracking signals, ambiguous signals from line interiors are free to propagate and combine with similar signals from nearby apertures to select a global direction of motion. Consistent with these data, the present model analyzes how both signals from line interiors and feature tracking signals may be used to determine the perceived direction of motion. Being unambiguous, feature tracking signals, when present, have the power to veto ambiguous signals from line interiors. Features such as line endings may thus decide the perceived direction of motion of the line to which they belong. When such signals are absent due to the figure-ground characteristics of the scene, ambiguous signals from line interiors may propagate across space and combine with signals from
nearby apertures to select a global direction of motion. Thus, in the absence of feature tracking signals, the model can select the vector average solution.

1.2 Intrinsic vs. Extrinsic Terminators

Not all line terminators are capable of generating feature tracking signals. When a line is occluded by a surface, it is usually perceived as extending behind that surface. The visible boundary between the line and the surface, therefore, belongs not to the line but to the occluding surface. Nakayama, Shimojo & Silverman (1989) first proposed the classification of line terminators into intrinsic and extrinsic terminators (Fig. 2). The motion of an extrinsic line terminator tells us little about the motion of the line. Such motion can at best inform us about the shape of the occluder. However, in most cases, the motion of an intrinsic line terminator signals the veridical motion of the line. As we shall soon see, the visual system treats the motion signals of intrinsic terminators as veridical signals if their motion is consistent. This makes it possible to fool the visual system by making the occluder invisible, such as by coloring it the same color as the background. In this case, the line terminators may be treated as intrinsic, but their motion is still not the veridical motion of the line. The preferential treatment displayed by the visual system of motion signals from intrinsic terminators over those from extrinsic terminators is incorporated into our model through figure-ground processes that detect occlusion events in a scene and assign edge ownership at these locations to near and far depth planes.

**FIGURE 2.** Extrinsic vs. intrinsic terminators: the boundary that is caused due to the occlusion of the gray line by the black bar is an extrinsic terminator of the line. This boundary belongs to the occluder rather than the occluded object. The unoccluded terminator of the gray line is called an intrinsic terminator because it belongs to the line itself.

Chey, Grossberg & Mingolla (1997, 1998) developed a neural model of biological motion perception, called the *Motion Boundary Contour System* (or *Motion BCS*), which used such ideas to explain several phenomena on motion grouping and speed perception. This earlier model simulated data on how speed perception and discrimination is affected by stimulus contrast and duration, dot density and spatial frequency. It also provided an explanation for the barber pole illusion, the conditions under which moving plaids cohere, and how contrast affects their perceived speed and direction. Our model both simplifies and extends this model to account for a larger set of representative data on motion grouping in 3D space, both within a single aperture and across several apertures. The next section describes in detail the design principles underlying the construction of
the model as well as the computations carried out at each stage and their functional significance. A simple simulation of a single moving line is used to demonstrate how each stage of the model functions, before other more complex data are simulated.

2. Formotion BCS Model

![Network diagram](image)

**FIGURE 3.** Network diagram. See text for details.
Fig. 3 is a macrocircuit showing the flow of information through the model. We now describe the functional significance of each stage of the model.

2.1 Level 1: Input After Preprocessing by FACADE

One sign of occlusion in a 2D picture is a T-junction (Fig. 4). The black bar in Fig. 4(A) forms a T-junction with the gray bar. The top of the T belongs to the occluding black bar while the stem belongs to the occluded gray bar. When no T-junctions are present in the image, such as in Fig. 4(B), it is harder to see depth in the image due to occlusion. Since extrinsic terminators are generated due to occlusion events in a scene, T-junctions are one important way of distinguishing between extrinsic and intrinsic object contours in an image. Clearly, any 3D motion system capable of using feature tracking signals to compute a global motion percept must be able to recognize occlusion events caused by T-junctions. The present model achieves this functionality by using the output of a static form processing system called the FACADE model (Grossberg, 1994, 1997; Grossberg & Kelly, 1999; Grossberg & McLoughlin, 1997; Grossberg & Pessoa, 1998; Kelly & Grossberg, 1998) as the input to the motion system via a form-motion interaction (Baloch & Grossberg, 1997; Francis & Grossberg, 1996). The form system is proposed to occur in the cortical stream that passes through the interblobs of V2, and the motion system is proposed to occur in the cortical stream that passes through MT (see DeYoe & Van Essen, 1988, for a review). The form-motion interaction is proposed to include signals from V2 to MT.

FACADE is a neural model of 3D figure-ground separation that explains how the visual system can see occluded and occluding objects and, hence, form a 3D representation from a 2D pictorial input. FACADE detects T-junctions in a picture without the explicit use of T-junction detectors, but through a neural circuit that includes oriented bipole cells (Grossberg & Mingolla, 1985), similar to the V2 cells reported in vivo by von der Heydt, Peterhans & Baumgartner (1984). In a bipole cell, if the oriented inputs to each of the two oriented branches of its horizontal receptive field are simultaneously sufficiently large, then the cell can fire an output signal (Fig. 5). This ensures that the cell fires beyond an oriented contrast such as a line-end only if there is evidence for a linkage with another similarly oriented contrast, such as a second collinear line-end.

At a T-junction, horizontal bipole cells get cooperative support from both sides of their receptive field from the top of the T, while vertical bipole cells only get activation on one side of their receptive field from the stem of the T. As a result, horizontal bipole cells are more strongly activated that vertical bipole cells and win a spatial competition for activation. This cooperative-competitive interaction leads to the detachment of the vertical edge of the T at the location where it joins the horizontal edge, creating an end-gap in the vertical boundary (Fig. 6). Grossberg, Mingolla & Ross (1997) and Grossberg & Raizada (2000) have shown how the bipole cell property can be implemented between collinear coaxial pyramidal cells in layer 2/3 of visual cortex via long-range excitatory horizontal connections and short-range inhibitory connections that are mediated by interneurons.

Disparity-sensitive competition between multiple spatial scales that obey a size-disparity correlation results in the top of the T being to assigned to a nearer depth -- that is, to the occluding object -- while the stem of the T is assigned to a farther depth -- that is, to the occluded object. FACADE has provided explanations for a variety of figure-ground percepts, including the Bregman-Kanizsa amodal completion illusion, Kanizsa stratification, Munker-White assimilation, the Benary cross, the Kanizsa checkerboard, and Fechner’s paradox.
FIGURE 4. T-junctions signalling occlusion. In the 2D image (A), the black bar appears to occlude the gray bar. When the black bar is colored white, and thus made invisible, as in (B), it is harder to perceive the gray regions as belonging to the same object.

FIGURE 5. Bipole Cells (adapted from Grossberg, 1997). Horizontally-tuned hypercomplex cells feed their signals into each of the two lobes of a horizontally-tuned bipole cell. When the activity in both lobes is above threshold, the cells fires its output down to the horizontally-tuned hypercomplex cell in the middle. At a T-junction, horizontal bipole cells get cooperative support from both sides of their receptive field from the top of the T.

FIGURE 6. T-junction sensitivity of bipole cells. Cooperative-competition generates end-gaps. Long-range cooperation is effected by bipole cells (+ regions) while short-range competition is achieved through hypercomplex cells (- regions) (adapted from Grossberg, 1997).
These FACADE mechanisms generate the boundary representations shown in Fig. 7 at the farther depth for a partially occluded line and an unoccluded line. Note that when the occluders are invisible, the occluded line does not appear to be occluded any more. These boundary representations, computed at each frame of a motion sequence, serve as the inputs to our model. It is important to note, however, that any other system capable of detecting T-junctions in an image and assigning a depth ordering to the components of the T could also provide the inputs to the current model.

2.2 Level 2: Transient Cells

A directionally selective neuron is one that fires vigorously when a stimulus is moved through its receptive field in one direction (called the preferred direction), while motion in the reverse direction (termed the null direction) evokes little response. The second stage of the model comprises undirectional transient cells, directional interneurons and directional transient cells. Undirectional transient cells are cells that respond to image transients such as luminance increments and decrements. They are analogous to the Y cells of the retina (Enroth-Cugell & Robson, 1966; Hochstein & Shapley, 1976a, b).

The connectivity between the three different cell types in Level 2 of the model incorporates three main design principles that are consistent with the available data on directional selectivity in the retina and visual cortex: (a) directional selectivity is the result of asymmetric inhibition along the preferred direction of the cell, (b) inhibition in the null direction is spatially offset from excitation, and (c) inhibition arrives before, and hence vetoes, excitation in the null direction. Fig. 8 shows schematically how asymmetrical directional inhibition works in a 1D simulation of a two-frame motion sequence. When the input arrives at the leftmost transient cell in frame 1, all interneurons at that location, both leftward-tuned and rightward-tuned, are activated. The rightward-tuned interneuron at this location, in its turn, inhibits the leftward-tuned interneuron and directional cell one unit to the right of the current location. When the input reaches the new location in frame 2, the leftward-tuned cells, having already been inhibited, can no longer be activated. Only the rightward-tuned cells are activated, consistent with motion from left to right. Further, mutual inhibition between the interneurons ensures that a directional transient cell response is relatively uniform across a wide range of speeds. This allows the directional transient cells to respond equally well to slow and fast speeds. Directional transient cell outputs for a 2D simulation of a single moving line are shown in Fig. 9(A). The signals are ambiguous as several motion directions are activated and the aperture problem is clearly visible.
FIGURE 8. Schematic diagram of a 1D implementation of the transient cell network showing the first two frames of the motion sequence. Thick circles represent active undirectional transient cells while thin circles are inactive undirectional transient cells. Ovals containing arrows represent directionally-selective neurons. Unfilled ovals represent active cells, cross-filled ovals are inhibited cells and gray-filled ovals depict inactive cells. Excitatory and inhibitory connections are labelled by ‘+’ and ‘-’ signs respectively.
FIGURE 9. Model activities for a 2D simulation of a moving tilted line. (A) Directional transient cells. (B) Thresholded short-range filter cells. (C) Competition network cells. (D) MT cells. (E) MST cells: model output. The gray region in each diagram represents the position of the input at the current frame. The inset diagram in (A) enlarges the activities of cells at one x-y location. The dot represents the center of the x-y pixel. Since all simulations in this paper use eight directions, there are eight cells, each with a different directional tuning at every spatial location. At the location shown, three of the eight cells, those tuned to east, south-east and south directions, are active. This is depicted through velocity vectors oriented along the preferred directions of each cell. The length of each vector is proportional to the activity of the corresponding cell. This convention is used for all the model outputs in the paper.

2.3 Level 3: Short-range Filter

Although known to occur in vivo, the veto mechanism described in the previous section exhibits two computational uncertainties in a 2D simulation. First, the extremely short spatial range over which it operates results in the creation of spurious signals near line endings as can be seen in Fig. 9(A). Second, vetoing eliminates the wrong (or null) direction, but does not actively select the correct direction. It is especially important to suppress spurious signals and boost the correct motion direction at line endings because the unambiguous signals from these features must be made strong enough to track the correct motion direction and to overcome the much more numer-
ous ambiguous signals from line interiors. In Level 3 of the model, the directional transient cell signals are space- and time-averaged to create these feature tracking signals. A short-range filter cell accumulates evidence from directional transient cells of similar directional preference within a spatially anisotropic region that is oriented along the preferred direction of the cell (Fig. 3). This simple computation is sufficient to build up feature tracking signals at unoccluded line endings, object corners and other featural regions in the scene. Note that, to compute the motion of features, it is not necessary for the network to first identify form discontinuities that may constitute features and match their positions from frame to frame. We thus avoid the feature correspondence problem to which correlational models (Reichardt, 1961; van Santen & Sperling, 1985) are prone; namely, how to match features in one frame with those in another frame.

Another key concept in the short-range filter is the introduction of multiple spatial scales. Each scale responds preferentially to a specific speed, and larger scales respond better to faster speeds than do smaller scales. This is achieved by thresholding the outputs of the short-range filter by a self-similar threshold; that is, a threshold that increases with filter size. Short-range filter outputs for a single moving line are shown in Fig. 9(B). We can now see relatively unambiguous feature tracking signals at line endings while all points in the interior of the line still exhibit the aperture problem.

2.4 Level 4: Spatial Competition and Opponent Direction Inhibition

Spatial competition among cells of the same spatial scale and that prefer the same motion direction further boosts the amplitude of feature tracking signals relative to that of ambiguous signals. This happens without making the signals from line interiors so small that they will be unable to group across apertures in the absence of feature tracking signals. Spatial competition also results in speed tuning curves for each scale; see Chey et al. (1997, 1998).

This stage of the model also uses opponent direction inhibition, or inhibition between cells tuned to opposite directions. This ensures that at a single spatial location, cells tuned to opposite directions of motion cannot be simultaneously active. Outputs of the competition stage for a moving line are shown in Fig. 9(C).

2.5 Level 5: Long-range Directional Grouping and Attentional Priming

Level 5 of the model consists of two groups of cells: the long-range filter activates model MT cells which in turn activate model MST cells. The long-range filter pools signals over larger spatial areas, opposite contrast polarities, and multiple orientations. However, the pooling is restricted to cells of the same scale that are tuned to the same direction. The MT and MST cells together comprise the grouping and priming network. MST cells implement a winner-take-all competition across directions. The winning direction is then fed back down to MT through a top-down matching and priming pathway (Fig. 3). This kind of attentional priming was proposed by Carpenter & Grossberg (1987) as part of Adaptive Resonance Theory (ART); see Grossberg (1999) for an interpretation of how it is realized by identified cells in the visual cortex. Cells tuned to the winning direction in MST have an excitatory influence on MT cells tuned to the same direction. However, they also nonspecifically inhibit all directionally tuned cells in MT. For the winning direction, the excitation cancels the inhibition. But for all other directions, having lost the competition in MST and not receiving excitation from MST to MT, there is net inhibition in MT. This attentional modulation of MT by MST leads to net suppression of all directions other than the
winning direction. The activities of MT and MST cells in the model for a single tilted line moving to the right are shown in Fig. 9(D, E).

The model is called the Formotion BCS Model because a form-motion interaction primes the mechanisms of a Motion BCS model with the figure-ground separated boundaries of the FACADE model.

3. Model Computer Simulations

In this section, we describe several classical and recent motion percepts and how the model mechanisms can be used to explain and simulate them.

3.1 Classic Barber Pole

Due to the aperture problem, the motion of a line seen behind a circular aperture is inherently ambiguous. The same is true for a grating of equally-spaced parallel lines moving coherently. However, Wallach (1935) showed that if such a grating were viewed behind an invisible rectangular aperture, then the grating appears to move in the direction of the longer edge of the aperture.

FIGURE 10. Moving grating illusions. The left column shows the physical stimulus presented to observers and the right column depicts their percept. (A,B) Classic barber pole illusion. (C,D) Motion capture. (E,F) Spotted barber pole illusion.

Due to the aperture problem, the motion of a line seen behind a circular aperture is inherently ambiguous. The same is true for a grating of equally-spaced parallel lines moving coherently. However, Wallach (1935) showed that if such a grating were viewed behind an invisible rectangular aperture, then the grating appears to move in the direction of the longer edge of the aperture.
Thus, for a horizontal aperture, such as the one shown in Fig. 10(A), the grating appears to move horizontally from left to right, as in Fig. 10(B).

Line terminators play a major role in explaining this illusion by acting as features with unambiguous motion signals (Hildreth, 1984; Nakayama & Silverman, 1988a, b). As in the tilted line simulation, our model uses line terminators to generate feature tracking signals. In the short-range filter stage (Level 3), line terminators generate feature tracking signals that gain in strength through spatial competition (Level 4). In a horizontal rectangular aperture, there are more line terminators along the horizontal direction than along the vertical direction (Fig. 10). Hence, in this example, there are more feature tracking signals signalling rightward motion than downward motion. Rightward motion signals therefore prevail over downward motion signals in the winner-take-all interdirectional competition in the long-range directional grouping and priming MT-MST network. Top-down priming of the winning motion direction, here rightward motion, from MST to MT suppresses all losing directions across MT. The model hereby explains how, in the presence of multiple feature tracking signals (here, grating terminators) signalling motion in different directions, interdirectional and spatial competition ensure that the direction favored by the majority of features determines the global motion percept of the barber pole pattern (Fig. 11(A)).

### 3.2 Motion Capture

The barber pole illusion demonstrates how the motion of a line is determined by the unambiguous signals formed at its terminators. Are motion signals restricted to propagate only from unambiguous motion regions to ambiguous motion regions within the same object or can they also propagate from unambiguous motion regions of an object to nearby ambiguous motion regions of other objects? Ramachandran & Inada (1985) addressed this question with a motion sequence in which random dots were superimposed on a classic barber pole pattern such that the dots on any one frame of the sequence were completely uncorrelated with the dots on the subsequent frame. Despite the noisiness of the motion signals of the dots from frame to frame in this display, subjects remarked that the dots appeared to move in the same direction as the barber pole grating (Fig. 10(C,D)). The motion of the dots appeared to have been captured by the motion of the grating. The authors named this phenomenon motion capture.

Our model explains motion capture as follows (Fig. 11(B)): since the dots are not stationary but flickering, they activate the transient cells in Level 2. However, due to the noisy and inconsistent motion of the dots in consecutive frames, no feature tracking signals are generated for the dots in the short-range filter. The ambiguous noisy motion signals lose the competition in the MT-MST loop. The winning barber pole motion direction inhibits the inconsistent motion directions of the dots, so that these now appear to move with the grating.

### 3.3 Spotted Barber Pole

Another interesting twist on the classic barber pole stimulus, the spotted barber pole (Shiffrar, Li & Lorenceau, 1995), involves the superposition of random dots on a barber pole grating as in motion capture. However, unlike in motion capture, the dots move coherently downwards. Far from seeing two separate overlapping motion fields, one for the barber pole grating and the other for the dots, observers see the grating move downwards with the dots. Thus, the motion of the dots now captures the perceived motion of the grating (Fig. 10(E,F)).
At first, this phenomenon may seem to be difficult to explain. One may expect that, as in the classic barber pole, for each line of the grating, the unambiguous motion of its terminators would...
determine its perceived motion. Since the stimulus still contains more lines with rightward moving terminators than downward moving terminators, it would seem that the grating should still appear to move rightward rather than downward. However, as we saw in the case of motion capture, unambiguous motion signals need not be restricted to propagate only within a single object. These signals can also influence the perceived motion of spatially adjacent regions. This is achieved in our model by using long-range filter kernels that are large enough to overlap several feature tracking signals from spatially contiguous regions. In the spotted barber pole, the superimposed dots also generate strong feature tracking signals signalling downward motion. When these downward signals combine with those produced by the few downward moving grating terminators, they outnumber the rightward signals formed by the remaining grating terminators. Downward energy predominates over rightward energy in the MT-MST loop and wins the interdirectional competition. As a result, the model successfully predicts that both the grating and the dots would appear to move downward (Fig. 11(C)).

3.4 Line Capture

The previous simulations have demonstrated the importance of line terminators in determining the perceived direction of motion in a moving sequence. However, all terminators are not created equal. While intrinsic terminators appear to belong to the line, extrinsic terminators, which are artifacts of occlusion, do not. The following simulations, which are related to the motion capture stimuli of Ramachandran & Inada (1985), predict how the visual system assigns differing degrees of importance to intrinsic and extrinsic terminators to determine the global direction of motion in a scene.

**FIGURE 12.** Line capture stimuli: Percept and model input from FACADE. Small arrows near line terminators depict the actual motion of the terminators. Larger gray arrows represent the perceived motion of the lines. (A,B) Single line translating behind visible rectangular occluders. (C,D) Line behind visible occluders with flanking unoccluded rightward moving lines.
FIGURE 13. Model MST output for line capture. (A) Partially occluded line. (B) Horizontal line capture.
3.4.1 Partially Occluded Line

When a line’s terminators are occluded such that they become extrinsic, their motion signals are ambiguous. However, in the absence of any other disambiguating motion signals in the scene, the visual system is forced to accept the motion of these terminators as the most likely candidate for the motion of the line (Fig. 12(A)). In the visual system and the model, extrinsic terminators can produce feature tracking signals but these are weaker than those produced by intrinsic terminators. They play a role in determining the global percept only when intrinsic features are lacking (Fig. 13(A)).

3.4.2 Horizontal Line Capture

When the same partially occluded line is presented with flanking unoccluded lines, the perceived motion of the ambiguous line is captured by the unambiguous motion of the flanking lines (Fig. 12(C)). The terminators of the unoccluded lines, being intrinsic, generate strong feature tracking signals in the short-range filter stage of the model. These are strong enough to capture not only the motion of the line that they belong to but also that of nearby ambiguous regions such as the partially occluded line which only has extrinsic terminators (Fig. 13(B)).

3.5 Triple Barber Pole

**FIGURE 14.** Triple Barber Pole. Thin black arrows represent the possible physical motions of the barber pole patterns. Thick gray arrows represent the perceived motion of the gratings.
FIGURE 15. Model MST output for the triple barber pole illusion. (A) Visible occluders, i.e., extrinsic horizontal line terminators. (B) Invisible occluders, i.e., intrinsic horizontal line terminators.

Shimojo, Silverman & Nakayama (1989) further explored the relative strength of the feature tracking signals produced at intrinsic and extrinsic line terminators. They combined three barber pole patterns, as shown in Fig. 14, and found that when the occluding bars are visible, i.e., when
the horizontal barber pole terminators are perceived to be extrinsic, observers saw a single downward-moving vertical barber pole behind the occluding bars. However, when the occluding bars are invisible, i.e., when the barber pole terminators are intrinsic, the percept was that of three rightward-moving horizontal barber pole patterns. Tommasi & Vallortigara (1999) performed a similar experiment in which they emphasized the importance of figure-ground segregation on the final motion percept.

It is easy to see why the three barber pole gratings appears to move rightward when the occluders are invisible: in each grating, rightward moving terminators outnumber downward moving terminators. Although this is still the case when the occluders are made visible, the rightward moving line endings, being extrinsic, can only produce very weak feature tracking signals while the downward moving endings, being intrinsic, continue to produce strong feature tracking signals. Downward activities, although fewer, are larger than the more numerous, but weaker, rightward activities. Downward motion wins the MT-MST competition and determines the percept (Fig. 15).

3.6 Translating Square seen behind Multiple Apertures

All the phenomena described so far have only involved the integration of motion signals into a global percept. We now describe data in which the nature of terminators is solely responsible for whether motion integration or segmentation takes place. This set of stimuli was developed by Lorenceau & Shiffrar (1992) while studying the effect of the shape and color of apertures on the ability of human subjects to group local motion signals into a global percept. These results are specially significant because they demonstrate the importance of features in determining the final percept. Since the physical motion in each of the three cases described below is identical and the only parameters being varied are the luminance and shape of the occluders, a solution computed on the basis of the intersection of constraints (IOC) model (Adelson & Movshon, 1982) would predict the same percept for each case. The visual percept, however, varies widely from case to case and depends entirely on the strength of the feature tracking signals generated in each case.

3.6.1 Visible Rectangular Occluders

Suppose that a square translates behind four visible rectangular occluders (Fig. 16(A)) such that the corners of the square (potential features) are never visible during the motion sequence. Observers are then able to amodally complete the corners of the square and see it consistently translating southwest (Fig. 16(B)). For computational simplicity, we can, without loss of generality, consider just the top and right sides of the square (Fig. 16(C)). When the occluders are visible, the extrinsic line terminators generate weak feature tracking signals that are unable to block the spread of ambiguous signals from line interiors across apertures. The southwest direction gets activated from both apertures while the other directions only get support from one of the two apertures (Fig. 17(A)). This is because the ambiguous motion positions activate a range of motion directions, including oblique directions, in addition to the direction perpendicular to the moving edge. The southwest direction hereby wins the interdirectional competition in MST. Top-down priming from MST to MT boosts the southwest motion signals while suppressing all others (Fig. 17(A)). Thus, in the model computer simulation, both lines appear to move in the same direction (Fig. 18(A)). Motion integration of local motion signals occurs.
Figure 16. Square translating behind rectangular occluders. (A,B,C) Visible occluders. Dark gray dashed lines represent the corners of the square that are never visible during the translatory motion of the square. (D,E,F) Invisible occluders. Light gray dashed lines depict the invisible corners of the square; dashed rectangular outlines represent the invisible occluders that define the edges of the apertures.

Figure 17. Schematic of how model mechanisms explain the translating square illusion. (A) when occluders are visible, motion integration across apertures takes place. (B) when occluders are invisible, motion segmentation occurs.

3.6.2 Invisible Rectangular Occluders

This display is identical to the previous one except that the occluders are made invisible by making them the same color as the background (Fig. 16(D)). This small change drastically affects the percept. Now, observers are no longer able to tell that the lines they see belong to a single object, a square, that is translating southwest. The lines appear to move independently of one another (Fig. 16(E)). Again, for simplicity, we consider only the top and right sides of the square (Fig. 16(F)). When the occluders are invisible, the intrinsic line terminators produce strong feature tracking signals. For each line, the feature tracking signals of its terminators veto the ambiguous...
signals from its interior. Each line appears to move in the direction computed by its terminators. The intrinsic terminators thus effectively block the grouping of signals from line interiors across apertures (Fig. 17(B)). Motion segmentation occurs when the intrinsic terminators move consistently in directions that are incongruent with the global direction of physical motion. Models outputs are shown in Fig. 18(B).

![FIGURE 18. Model MST output for the translating square behind multiple apertures. (A) Visible rectangular occluders. (B) Invisible rectangular occluders. (C) Invisible jagged occluders.](image)

The role of inhibition between motion signals from line endings and line interiors was emphasized by Giersch & Lorenceau (1999). They boosted inhibition through the use of lorazepam, a substance that facilitates the fixation of inhibitory neurotransmitter GABA on GABA_A receptors. This selectively affected performance in the invisible rectangular occluders case but not in the vis-
ible rectangular occluders case. Enhanced inhibition did not affect motion integration when the
occluders were visible, but it boosted motion segmentation when the occluders were invisible.
This is consistent with our model’s prediction.

3.6.3 Invisible Jagged Occluders

Lorenceau & Shiffrar (1992) showed that if the occluders are invisible as before but jagged
instead of rectangular, then observers are once again able to group the motion of individual lines
into the percept of a global translating square (Fig. 19). Clearly, intrinsic terminators do not
always generate feature tracking signals that are strong enough to block motion grouping across
apertures. The jagged edges of the occluders cause the motion of the line terminators to change
direction constantly and, thus, be very noisy. As a result, the short-range filter is unable to accu-
mulate enough evidence for motion along any particular direction at line endings. Therefore,
strong feature tracking signals are not produced at line endings. Signals from line interiors can
again freely group across apertures (Fig. 18(C)). In summary, for features such as line endings and
dots to produce reliable feature tracking signals, they must be intrinsic and must generate suffi-
cient evidence for consistent motion in a particular direction.

![Figure 19: Square translating behind invisible jagged apertures: Model input and predicted output.](image)
3.7 Motion Transparency

Motion transparency is the phenomenon by which the visual system perceives transparency in a display purely as a result of motion cues. A typical display consists of two fields of random dots superimposed on each other. When the direction of motion of the two fields is different, the visual system perceives one field of dots to be closer than the other. The motion dissimilarity between the two fields is alone responsible for their depth segregation (Fig. 20).

FIGURE 20. Motion transparency. Note that, in this figure, shading has been used solely to identify the two fields. In the actual display, the two fields are identical in all respects except their motion.

While opponent direction inhibition in MT is useful to reduce noisy local motion signals, it can also have the undesirable effect of suppressing neuron responses under transparent conditions and rendering the visual system blind to transparent motion. For example, Snowden, Treue, Erickson & Andersen (1991) showed that the response of an MT cell to the motion of random dots in the cell’s preferred direction is strongly reduced when a second, transparent dot pattern moves in the opposite direction. Recanzone, Wurtz & Schwartz (1997) demonstrated that this result extended to cells in MST and can also be observed when discrete objects are substituted for whole-field motions. However, Bradley, Qian & Andersen (1995) and Qian & Andersen (1994) showed that, since opponent direction inhibition occurs mainly between motion signals with similar disparities, the disparity-selectivity of MT neurons can be used effectively to extract information about transparency due to motion cues. Our model explains how the use of multiple spatial scales, with each scale being sensitive to a particular range of depths according to the size-disparity correlation, achieves this functionality.

Just as FACADE (Grossberg, 1994) uses multiple scales for depth sensitivity and the Motion BCS (Chey et al., 1997) uses multiple scales for speed sensitivity, the Formotion BCS model uses multiple scales for motion segmentation in depth. The transparent motion percept is bistable and attention determines which of the two fields is seen in front of the other. We implement this in the model by randomly selecting one of the two active directions of motion, say rightward motion, within a given scale, say scale 1, and inside a foveal region and attentionally enhancing the MST signals for that direction. The attentional enhancement acts as a gain control mechanism that adds a DC value to all cells tuned to rightward motion within the attentional locus (O’Craven, Rosen, Kwong, Treisman & Savoy, 1997; Treue & Martinez Trujillo, 1999; Treue & Maunsell, 1996, 1999). Consistent with these data, the enhancement does not change the tuning curves of the cells and only increases their activity. The attentional gain is applied only within the selected direction and scale and inside the attentional locus. In our simulation, the locus of attention is at the center.
of the display and covers 6.25% of the total display area. The boost to rightward motion signals in scale 1 makes this direction win the interdirectional competition in scale 1. Interscale inhibition from the near scale, scale 1, to the far scale, scale 2, within direction and at each spatial location suppresses rightward motion in scale 2. Leftward motion signals in scale 2 are now disinhibited and can win the interdirectional competition in this scale. Two different motion directions become active at two different depths (Fig. 21). Thus, by the use of two scales representing two different depths, the model explains how a 2D input sequence can lead to the perceptual segregation in depth of two surfaces based solely on motion cues.

![Frame no. 15; Scale 1](A)

![Frame no. 15; Scale 2](B)

**FIGURE 21.** Model MST output for motion transparency. (A) Scale 1. (B) Scale 2.

### 3.8 Chopsticks Illusion: Coherent and Incoherent Plaids

The chopsticks illusion (Anstis, 1990) is similar, but not identical, to the plaids stimulus (Fig. 22). Two overlapping lines of the same luminance move in opposite directions. When the lines are viewed behind visible occluders, they appear to move together as a welded unit in the downward direction. When the occluders are made invisible, the lines no longer cohere but appear to slide one on top of the other. The first case is similar to coherently moving plaids while the second resembles the percept of incoherently moving plaids. The chopsticks display contains two kinds of feature: the line terminators of each line and the intersection of the two lines. Of the line terminators, two move leftward while the other two move rightward. The intersection of the two lines moves downward. All these features have unambiguous motion signals.

#### 3.8.1 Visible Occluders

Clearly, when the line terminators are made extrinsic by making the occluding bars visible, their motion signals are given less importance by the visual system. The feature tracking signals due to the intersection of the two lines are stronger than those due to the extrinsic line terminators. The downward moving signals at the intersection win the competition in the MT-MST loop and propagate outward to capture the motion of the lines. Both lines appear to move downward as a single coherent unit (Fig. 23(A)).
3.8.2 Invisible Occluders

The percept of incoherency involves the interplay of more complicated mechanisms. We argue that this percept cannot be explained by considering the motion system alone, but requires a form-motion interaction of the form and motion systems. In this view, incoherency is the combination of two percepts that occur simultaneously: (a) the perceived inconsistency of the motion velocities of the two lines, and (b) perceptual form transparency with one line perceived as being superimposed in front of the other. The two percepts are interlinked and can each cause the other. For instance, Stoner, Albright & Ramachandran (1990) showed that form transparency cues at the intersections of two plaids can lead to perceptual incoherency of the plaids. This is an example of a form-to-motion interaction. However, Lindsey & Todd (1996) argued that form transparency is necessary but not sufficient for the perception of motion incoherency in plaids; that is, form transparency cues in a plaid pattern do not guarantee the perception of motion incoherency. They showed that incoherency may arise from prolonged viewing, and suggested that motion adaptation may also play a role. In displays such as the chopsticks illusion, where there are no form cues that robustly lead to perceptual transparency in a static version of the illusion, motion cues can themselves lead to the percept of depth segregation of the two lines. This is an instance of a motion-to-form interaction. Models that have attempted to simulate incoherent plaids without using a form-to-motion interaction (Chey et al., 1997; Liden & Pack, 1999) have failed to produce the perceived motion signals at the plaid intersections.
In the chopsticks illusion, when the line terminators are intrinsic, their motion signals are at least as strong as those due to the intersection of the two lines. The different motion signals arising from line terminators leads to the depth segregation of the two lines. When this happens, the feature arising from the intersection of the two lines no longer perceptually exists. This is consistent with the data of Bressan, Ganis & Vallortigara (1993) and Vallortigara & Bressan (1991). To understand how the visual system sees this stimulus, it is necessary to consider our model as part of a broader framework of models that perform figure-ground segmentation within the form system and implement interactions both from the form to the motion system and from the motion to the form system.

Fig. 24 shows the neural pathways and connections that we predict are involved in providing a complete explanation of the incoherent chopsticks illusion. A complete simulation of this circuit is beyond the scope of the present article. The input motion sequence appears at V1 after retinal and LGN processing. Figure-ground processing between V1 and V2, such as in FACADE (Grossberg, 1994, 1997; Grossberg & Kelly, 1999), detects occlusion events in the form of T-junctions and assigns depth ordering to object boundaries at the site of an occlusion. This stage, labelled as 1 in Fig. 24, represents one input stage (Level 1) of the Formotion BCS model. Form-to-motion interaction from V2 to MT, similar to the processes described by Baloch & Grossberg (1997) and Francis & Grossberg (1996), extracts motion signals from the form inputs and computes feature tracking signals at the intrinsic line terminators of the chopsticks as well as at their intersection. This stage, labelled as 2 in Fig. 24, represents the middle stages (Levels 2-4) of the Formotion BCS model.

The grouping and priming MT-MST loop, labelled as 3 in Fig. 24 and corresponding to Level 5 of the Formotion BCS model, detects the lack of a clear directional winner due to the conflicting motion signals from the line terminators. In the MT-MST feedback loop, these conflicting signals propagate from the line terminators to the intersection. At the intersection, top-down attention in MST randomly or volitionally selects one of the two chopsticks. Attentional enhancement of the motion signals along the form boundaries of the attended chopstick occurs. This top-down atten-
tional priming effect from MST to MT can then propagate to V1 via top-down MT-to-V1 signals, labelled 4 in Fig. 24. This attentional effect is predicted to amplify the boundaries that are formed at the attended chopstick. This advantage is sufficient to activate figure-ground separation of the boundaries corresponding to the two chopsticks (Grossberg, 1994, 1997).

As a result, the form system modally completes the boundaries of the favored chopstick and amodally completes the boundaries of the “far” chopstick behind the near one. The attentional bias can hereby propagate in an MST-MT-V1-V2-MT loop. The motion-to-form interaction from MT-to-V1 along pathway 4 in Fig. 24 is predicted to act like a top-down ART-like attentional prime (Grossberg, 1999; Grossberg & Raizada, 2000). This proposal is supported by neurophysiological data showing that feedback connections from MT-to-V1 are important in the differentiation of figure from ground (Hupe, James, Payne, Lomber, Girard & Bullier, 1998). These feedback connections facilitate responses to moving objects in the center and inhibit responses in the surround of V1 cells, as also occurs with an ART attentional prime.

Once figure-ground separation is initiated, another loop through the model MT-MST interactions is required to determine the perceived motion directions of the lines at each depth. A simulation of this second loop is shown in Fig. 23(B,C). A more complete simulation of the process described above would require simulation of the entire system depicted in Fig. 24.

4. Discussion

We have presented a hierarchical neural network model, called the Formotion BCS, that successfully performs the conflicting tasks of the integration and segmentation of motion cues into a unified global percept. Interconnections between neurons in the model are consistent with, and functionally clarify, currently known data on the connectivity between cortical areas devoted to visual motion processing such as the retina, primary visual cortex (V1), prestriate visual cortex (V2), middle temporal cortex (MT) and the medial superior temporal cortical area (MST). The model extracts feature tracking signals from a 2D motion sequence without explicit feature detection or feature matching. The model combines unambiguous motion signals from features with ambiguous signals from the interiors of regions that arise from the aperture problem. The two types of signals are computed by the same mechanisms. Competition between motion signals from feature tracking regions and other parts of the scene determines the final 3D percept. Simulations show how a range of challenging perceptual phenomena can be explained by a single unified theory of motion perception.

4.1 The Motion Boundary Contour System

The Motion Boundary Contour System, which has been further developed in this paper, was originally proposed by Grossberg & Rudd (1989, 1992). They simulated data on short-range and long-range apparent motion, including beta, gamma, delta, reverse and split motion and Ternus motion (Grossberg & Rudd, 1989, 1992). Grossberg (1991, 1998) extended this model to explain how a moving target can be tracked when it is intermittently occluded by intervening objects. Grossberg & Mingolla (1993) extended the model to suggest a solution to the global aperture problem; i.e., how a coherent motion signal is imparted to all regions of a moving figure, not just to regions at which unambiguous motion signals exist.
Baloch & Grossberg (1997) and Francis & Grossberg (1996) further extended the model to explain the interaction of the form and motion systems. They simulated Korte’s laws, the line motion illusion, motion induction and transformational apparent motion. Baloch, Grossberg, Mingolla & Nogueira (1999) developed this model even further to simulate first- and second-order motion stimuli, including the reversal of perceived motion direction with distance from the stimulus (gamma display), and data about directional judgments as a function of relative spatial phase or spatial and temporal frequency.

Chey et al. (1998) added multiscale dynamics to the model to explain the size-speed correlation and simulate data on how visual speed perception and discrimination are affected by stimulus contrast, duration, dot density and spatial frequency. Chey et al. (1997) presented a model of motion integration that simulates the conditions under which components of moving stimuli cohere into a global direction of motion, as in barberpole and Type I and Type II plaids. The model also simulates the temporal dynamics of how unambiguous feature tracking signals from line terminators spread to and conquer the ambiguous signals from line interiors. This paper extends the model further to perform motion integration as well as motion segmentation by combining figure-ground mechanisms and formotion interactions with motion mechanisms showing how to combine the unambiguous motion signals of image features with ambiguous signals arising as a result of the aperture problem.

4.2 Neurophysiological evidence

4.2.1 Level 2: Transient Cells

Directionally sensitive cells, similar to those in Level 2 of the model, have been found both in the retina of rabbit (Barlow, Hill & Levick, 1964) and in simple and complex cells in V1 (Hubel & Wiesel, 1968), as well as in later stages in the visual processing stream. Barlow & Levick (1965) first suggested that directional sensitivity in ganglion cells of the rabbit retina is mainly a result of the lateral spread of inhibition in an asymmetric fashion, so that it blocks excitation which subsequently arrives on one side of it, but not on the other. This forward inhibition has a certain rise time and decay and serves to veto cell responses to the null direction. This approach argues against the Reichardt (1961) hypothesis that directional selectivity is achieved by the cross-correlation of a signal with delayed excitation from one side.

Since the pioneering work of Barlow & Levick (1965), there has been growing evidence in favor of their theory. Several pharmacological studies conducted in the retinae and primary visual areas (V1) of rabbits, cats and monkeys (Ariel & Daw, 1982; Sato, Katsuyama, Tamura, Hata & Tsumoto, 1995; Sillito, 1975, 1977; Wyatt & Daw, 1976) conclude that antagonists to the inhibitory neurotransmitter gamma-aminobutyric acid (GABA) abolish or greatly reduce directional selectivity. Ariel & Daw (1982) further observed that when a potentiator for the excitatory neurotransmitter acetylcholine (ACh) is used, substantial excitation is produced which overcomes or outlasts the null direction GABA inhibition within the receptive field. The spatial extent of GABA inhibition is asymmetric to and larger than the spatial extent of ACh excitation.

Other physiological studies (Emerson, Citron, Vaughn & Klein, 1987; Emerson & Coleman, 1981; Emerson & Gerstein, 1977; Ganz, 1984; Ganz & Felder, 1984) compared responses to single static flashes at various receptive field locations in either the preferred or the null direction with responses to sequence pairs of static flashes at those same locations. They found that the
response to a single bar was smaller when it was preceded by a stimulus from the null side. Hammond & Kim (1994) and Innocenti & Fiore (1974) systematically mapped excitatory and suppressive receptive fields and found that their profiles were spatially offset, especially along the preferred direction such that, for stimuli moving in the non-preferred direction, the inhibition lay ahead of the excitation. Further, Ganz & Felder (1984), Goodwin, Henry & Bishop (1975a, b) and Heggelund (1984) argued against Hubel & Wiesel’s (1959, 1962) hypothesis that directional selectivity can be explained on the basis of a simple linear combination of the responses from adjacent ON and OFF regions of the neuron. Several of these neurophysiological studies (Barlow & Levick, 1965; Emerson et al., 1987; Emerson & Gerstein, 1977; Ganz, 1984; Ganz & Felder, 1984) agree about the existence of direction-selective subunits distributed across the receptive field and contributing their inputs to a directionally selective neuron.

However, another theory for directional selectivity exists (Dean & Tolhurst, 1986; DeAngelis, Ohzawa & Freeman, 1993a, b; Jagadeesh, Wheat & Ferster, 1993; Jagadeesh, Wheat, Kontsevich, Tyler & Ferster, 1997; McLean & Palmer, 1989; McLean, Raab & Palmer, 1994; Movshon, Thompson & Tolhurst, 1978; Reid, Soodak & Shapley, 1987, 1991). This is referred to as spatiotemporal inseparability (Adelson & Bergen, 1985). According to this hypothesis, differences in excitatory response timing across the receptive field causes directional sensitivity. A stimulus moving in the preferred direction would activate faster and faster responses which summate optimally if the stimulus speed matches the shift in response time course.

In a recent study on alert fixating macaque monkeys, Livingstone (1998) suggested that delayed asymmetric inhibition may contribute to the shifting excitatory response time course. Her data suggest that asymmetric forward inhibition is the major determinant for directionality in V1 cells. She shows how the morphology and connectivity of Meynert cells, that are large, direction-selective, MT-projecting cells in layer 6 of V1, can be used to explain the role of inhibition in direction-selectivity. A Meynert cell has asymmetrical basal dendrites extending in one direction within layer 6. It receives excitatory inputs from its distal dendrites and relatively denser inhibitory inputs from the synapses formed by inhibitory interneurons with its cell body. This structure ensures that the cell receives excitatory and inhibitory inputs from different regions of the visual field. Besides, due to dendritic conduction delays, excitatory inputs from distal dendritic tips would arrive at the cell body later than the inhibitory inputs from interneurons. These simple properties enable the cell to use asymmetric inhibition to achieve directional selectivity.

4.2.2 Level 4: Spatial Competition and Opponent Direction Inhibition

Several neurophysiological studies confirm that the opponent direction inhibition used in Level 4 of the model exists in MT but has not been found in V1 (Bradley et al., 1995; Heeger, Boynton, Demb, Seidemann & Newsome, 1999; Qian & Andersen, 1994; Recanzone et al., 1997; Snowden et al., 1991).

4.2.3 Level 5: Long-range Directional Grouping and Attentional Priming

Several studies have shown that cells in the middle temporal cortical area (MT) are directionally selective (Albright, 1984; Maunsell & van Essen, 1983a; Zeki, 1974a, b). They respond more strongly to moving stimuli, irrespective of direction of contrast, than to static stimuli suggesting that they are sensitive to motion cues but not to form or color cues. They have broad directional tuning curves and narrow orientational tuning curves. Directional tuning of MT neurons is unaffected by the featural properties, first-order or second-order form cues, generating the motion sig-
nal (Albright, 1992). Psychophysical evidence using heterogeneous-cue plaids (Stoner & Albright, 1992) shows that motion signals are integrated irrespective of whether they were produced by first- or second-order form cues. It is suggested that this integration of local motion signals to form a global percept occurs in MT. The discovery of two types of MT neuron, those that respond to component motion and those that respond to pattern motion of plaids (Movshon, Adelson, Gizzi & Newsome, 1985; Rodman & Albright, 1989) confirms the hypothesis that MT is the first cortical area in the visual processing stream where integration of motion cues occurs.

Outputs from MT feed into the medial superior temporal (MST) area (Desimone & Ungerleider, 1986; Maunsell & van Essen, 1983b). Cells in this area are known to be directionally selective and have large receptive fields. The dorsal part of MST, called MST_d, responds selectively to expansion, contraction, and clockwise or counterclockwise rotation (Saito, Yukie, Tanaka, Hikosaka, Fukada & Iwai, 1986) and favors movements of a wide textured field, probably caused by movements of the observer, over those of objects in the visual field (Duffy & Wurtz, 1991a, b; Komatsu & Wurtz, 1988; Orban, Lagae, Verri, Raiguel, Xiao, Maes & Torre, 1992; Tanaka & Saito, 1989). Grossberg, Mingolla & Pack (1999) have modeled how this part of MST may be involved in navigation based on optic flow stimuli. The ventral part of MST, termed MST_v, prefers object movements to whole-field movements. Cells in this region represent the relative motion of an object with respect to its background. This is the sort of motion processing that we have used in our model of MT-MST directional selection and attentional priming.

Such attentional modulation of motion signals has been shown to occur in vivo (O’Craven et al., 1997; Treue & Martinez Trujillo, 1999; Treue & Maunsell, 1996, 1999). Treue & Maunsell (1996, 1999) demonstrated a strong modulatory influence of attention on motion processing in the directionally selective cells of MT and MST in macaque monkeys. Using functional magnetic resonance imaging (fMRI) in human subjects, O’Craven et al. (1997) found greater activation in MT-MST in the presence of voluntary attention. Further, attention acts as a nonspecific gain control mechanism that enhances responses within the locus of attention without narrowing the direction-tuning curves (Treue & Martinez Trujillo, 1999).

4.3 Comparison with other motion models

Several theories of motion perception have been proposed in the literature. However, most of these offer explanations for either motion integration or motion segmentation but not both. It is relatively easy to explain either of these tasks independent of the other. However, since they have contradictory yet complementary goals, it is more difficult to present a theory that can solve both problems simultaneously through the same set of mechanisms. We describe models that have attempted to solve these problems and compare them to our approach. A summary of this analysis is presented in Table 1.

One model of motion integration which is also a popular theory for the perceived motion direction of coherent plaids is the intersection of constraints (IOC) model (Adelson & Movshon, 1982). This theory predicts that observers would always see the veridical motion of a coherent plaid pattern. However, a growing body of data suggests that this is not the case (Bowns, 1996; Bressan et al., 1993; Cox & Derrington, 1994; Derrington & Ukkonen, 1999; Ferrera & Wilson, 1990, 1991; Rubin & Hochstein, 1993; Vallortigara & Bressan, 1991). The role of features such as dots, line terminators, object corners and plaid intersections in determining the global direction of motion has been emphasized in both plaid displays (Alais, Burke & Wenderoth, 1996; Alais, van der
Given that the motion signals from features plays an important role in the final percept, we are still faced with the problem of how to compute this motion. Correlational models (Lappin & Bell, 1972; Reichardt, 1961; van Santen & Sperling, 1985) suggest that this is done by a pair of receptors separated by some physical distance such that the delayed output of one receptor is multiplied by the output of the other receptor. This matching of corresponding points in succeeding frames can be done at two levels. Feature matching models (Reichardt, 1961; van Santen & Sperling, 1985) detect salient features and match corresponding features to compute image velocities. Global matching models (Lappin & Bell, 1972) perform template matches over larger regions of

<table>
<thead>
<tr>
<th>Paper</th>
<th>Type of model</th>
<th>Type of data simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelson &amp; Bergen (1985)</td>
<td>spatiotemporal energy</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Adelson &amp; Movshon (1982)</td>
<td>intersection of constraints (IOC)</td>
<td>motion integration: coherent plaids</td>
</tr>
<tr>
<td>Del Viva and Morrone (1998)</td>
<td>feature tracking</td>
<td>motion integration and segmentation</td>
</tr>
<tr>
<td>Fennema &amp; Thompson (1979)</td>
<td>gradient</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Hildreth (1984)</td>
<td>regularization / smoothing</td>
<td>motion integration</td>
</tr>
<tr>
<td>Horn &amp; Schunck (1981)</td>
<td>regularization / smoothing</td>
<td>motion integration: optic flow</td>
</tr>
<tr>
<td>Jasinschi, Rosenfeld &amp; Sumi (1992)</td>
<td>correlational &amp; IOC</td>
<td>motion integration and segmentation</td>
</tr>
<tr>
<td>Jin &amp; Srinivasan (1990)</td>
<td>gradient</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Johnston, Mcowan &amp; Buxton (1992)</td>
<td>gradient</td>
<td>first- and second-order motion</td>
</tr>
<tr>
<td>Lappin &amp; Bell (1972)</td>
<td>correlational</td>
<td>apparent motion</td>
</tr>
<tr>
<td>Liden &amp; Pack (1999)</td>
<td>feature tracking</td>
<td>motion integration and segmentation</td>
</tr>
<tr>
<td>Loffler &amp; Orbach (1999)</td>
<td>feature tracking</td>
<td>motion integration: coherent plaids</td>
</tr>
<tr>
<td>Marr &amp; Ullman (1981)</td>
<td>gradient</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Nowlan &amp; Sejnowski (1994)</td>
<td>spatiotemporal energy</td>
<td>motion segmentation: transparency</td>
</tr>
<tr>
<td>Poggio, Torre &amp; Koch (1985)</td>
<td>regularization / smoothing</td>
<td>motion integration: barber-pole</td>
</tr>
<tr>
<td>Qian, Andersen &amp; Adelson (1994)</td>
<td>subtractive and divisive inhibition</td>
<td>motion segmentation: transparency</td>
</tr>
<tr>
<td>Reichardt (1961)</td>
<td>correlational</td>
<td>low-level vision</td>
</tr>
<tr>
<td>Sachtlter &amp; Zaidi (1995)</td>
<td>center-surround shearing</td>
<td>motion segmentation</td>
</tr>
<tr>
<td>van Santen &amp; Sperling (1985)</td>
<td>correlational</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Watson &amp; Ahumada (1985)</td>
<td>spatiotemporal energy</td>
<td>directional and speed sensitivity</td>
</tr>
<tr>
<td>Yuille &amp; Grzywacz (1988)</td>
<td>regularization / smoothing</td>
<td>motion integration: motion capture</td>
</tr>
<tr>
<td>Zemel &amp; Sejnowski (1994)</td>
<td>adaptive learning neural network</td>
<td>motion segmentation</td>
</tr>
</tbody>
</table>
space by sliding images in subsequent frames to obtain optimal matches. Both kinds of correlational model are susceptible to the correspondence problem; namely, how to establish correspondences across successive frames, especially when the similarity of objects in the images suggests that more than one kind of correspondence is possible (Anstis, 1980). Clearly, velocity estimates in the scene depend crucially on which correspondence is chosen. We therefore need a method of computing the motion of features without explicitly detecting and matching features.

Spatiotemporal motion energy models (Adelson & Bergen, 1985; Watson & Ahumada, 1985) are similar to correlational models in that they recover speed and direction estimates from spatiotemporal information in the scene. However, to do this, they use linear filters whose Fourier transforms are oriented in space-time. Velocity sensitivity is achieved through orientation sensitivity in space-time. Motion energy models are formally equivalent to elaborated Reichardt detectors in that they compute identical outputs for any given input (van Santen & Sperling, 1985). Emerson, Bergen & Adelson (1992) presented neurophysiological evidence that the responses of directionally selective complex cells in the cat’s striate cortex are consistent neither with correlational models (Reichardt, 1961; van Santen & Sperling, 1985) nor with an opponent combination of motion energy models (Adelson & Bergen, 1985; Watson & Ahumada, 1985). Moreover, like correlational models, motion energy models are incapable of detecting motion in second-order motion stimuli; that is, stimuli that consist of a moving spatial modulation of the contrast of a non-moving pattern.

Gradient models (Fennema & Thompson, 1979; Jin & Srinivasan, 1990; Marr & Ullman, 1981) compute velocity by using local spatial and temporal derivatives of the image’s spatiotemporal luminance profile. Speed sensitivity is coded by the magnitudes of the gradients. Since derivatives are computed at single spatial locations, gradient schemes successfully bypass the correspondence problem. However, they succumb to the aperture problem since the expression used to compute velocity in the case of moving 1D bars is ill-conditioned. In an attempt to solve this problem, Johnston and colleagues (Johnston & Clifford, 1995; Johnston, McOwan & Benton, 1999; Johnston, McOwan & Buxton, 1992) proposed a model that combines a gradient scheme with the IOC procedure to detect first- and second-order motion in the presence or absence of static noise. The resulting multi-channel gradient model is capable of reliably detecting the motion of a grating superimposed on a static random binary noise pattern. The model is consistent with the data of Lu & Sperling (1995) whose experiments using contrast-modulated noise patterns found no evidence for feature tracking in first- and second-order motion detection. However, when contrast-modulated sine-wave gratings are substituted for contrast-modulated noise patterns, second-order motion detection is disrupted by the superimposition of a pedestal, thus suggesting that the motion of contrast envelopes is detected by a mechanism that tracks features (Derrington & Ukkonen, 1999). Moreover, although the multi-channel gradient model is well-conditioned for velocity coding, it would fail in the same way as IOC in explaining data on Type II plaids.

Regularization theories (Hildreth, 1984; Horn & Schunck, 1981; Koch, Wang & Mathur, 1989; Poggio, Torre & Koch, 1985; Yuille & Grzywacz, 1988) minimize a cost function by applying a smoothness constraint to the velocity field. They make the assumption that real-world objects have smooth surfaces, whose projected velocity field is usually smooth. Such techniques are robust to noise and are good for motion integration but can perform motion segmentation only by explicitly detecting discontinuities in the motion field, such as when the spatial gradient of the velocity field between two neighboring points is larger than some threshold. Further, the iterative minimization of the cost functional is computationally expensive, subject to getting trapped in local minima for non-quadratic functionals, and difficult to interpret biologically.
Marshall (1990) and Wang (1997) presented adaptive neural networks in which weights and connections between neurons are modified during an iterative training phase in which motions of various directions and speeds are presented. However, it remains to be seen whether the perception of motion illusions such as those presented in this paper is the result of adaptive learning.

Other models primarily address the problem of motion segmentation (Nowlan & Sejnowski, 1994; Qian, Andersen & Adelson, 1994; Sachter & Zaidi, 1995; Zemel & Sejnowski, 1994). They detect local motion discontinuities and use these to segment the scene. They fail to integrate motion signals across discontinuities that arise from noise in the stimulus.

Computational models of feature tracking have traditionally faced two problems: I. What constitutes a feature? How should features be detected in a scene? Definitions of features have typically been vague. Dots, line terminators, object corners and plaid intersections are examples of easily detectable features. However, corners of objects formed by subjective contours can also constitute features and these are considerably harder to detect. II. Even if features can be reliably detected in a scene, how should features in one frame of a motion sequence be matched to features in the next frame? This is the correspondence problem discussed earlier.

Jasinschi, Rosenfeld & Sumi (1992) proposed a model that combines a feature matching scheme similar to that of correlational models with IOC to explain motion transparency and coherence. The model uses a velocity histogram that combines votes from the velocities of features such are corners and line endings (computed by template matching) with those from the intersections of all possible constraint lines due to the motion of image contours. The model succeeds in explaining motion transparency; namely, how two velocities can be perceived at the same spatial location, as well as the bistability of motion transparency and coherence in plaid displays. However, the use of global correlational matching as well as IOC makes the model susceptible to the drawbacks of both types of scheme.

Del Viva and Morrone (1998) detect features by computing peaks of spatial local energy functions and compute feature velocities using a spatiotemporal motion energy scheme. Such a technique would fail to detect features formed by subjective contours. Loffler & Orbach (1999) presented a model of motion integration in coherent plaids which uses two parallel pathways (Fourier and non-Fourier) to perform feature tracking without the explicit use of feature detectors such as end-stopped cells. However, there is psychophysical evidence against the existence of two pathways (Bowns, 1996; Cox & Derrington, 1994). Moreover, none of the models described so far can explain how the intrinsic-extrinsic classification of features influences the global motion percept. For instance, intrinsic line terminators have unambiguous motion signals while the motion of extrinsic terminators is discounted by the visual system; while the former can block motion grouping across apertures, the latter fail (Lorenceau & Shiffrar, 1992).

Liden & Pack (1999) proposed a neural network model of motion integration and segmentation that consists of two separate but interacting systems of cells, one specialized for integration and the other for segmentation. The model takes into account the relative strengths of intrinsic and extrinsic features by hypothesizing that local motion signals near T-junctions signalling occlusion are masked. In this way, the motion signals generated by extrinsic features are excluded from computations of global motion while those of intrinsic features are preserved. This mechanism predicts the existence of form-to-motion interaction whereby form cues such as T-junctions inhibit motion signals at nearby locations. Further, the nature of the interaction between the inte-
gration and segmentation networks precludes the possibility of two motion velocities being active at the same spatial location. Therefore, the model cannot explain motion transparency.

Our model suggests that a single system is capable of performing the dual tasks of motion integration and segmentation. The model performs neither feature detection nor feature matching, thus circumventing both the problems faced by most feature tracking models. Nevertheless, we can reliably compute feature tracking signals. For a motion signal at a given spatial location to be attributed to the motion of a feature, it is sufficient that the signal be consistent and have few competitors both across direction at the same spatial location and across space from similar directions. Model dynamics then ensure that these signals are made strong enough to dominate the final percept. Our model differs from that of Liden & Pack (1999) in that only form cues are inhibited at T-junctions leaving motion cues intact. The use of multiple spatial scales makes it possible for distinct motion velocities to be active at the same spatial location but at different scales, thus allowing an explanation of depth segregation due to motion transparency.

5. Appendix: Model Equations

We first describe the symbols and notations used in the network equations. Each cell activity is denoted by a variable whose letter indicates the cell type. Subscripts indicate the spatial position of the cell. Superscripts indicate the directional tuning and scale of the cell. For example, \( F_{ij}^{ds} \) indicates the activity of a thresholded short-range filter cell at spatial location \((i,j)\), directional preference \(d\) and scale \(s\). The notation \([w]^+ = \max(w, 0)\) stands for half-wave rectification. Similarly, \([w-t]^+\) denotes rectification with threshold at \(t\). The outputs of every level of the model are rectified before being fed into the next level. The notation \(|S|\) indicates the size of the set \(S\).

Some equations involve interactions between opponent directions. We compute the direction \(D^O\) exactly opposite to the direction \(d\) as follows:

\[
D^O = \left( d + \frac{ND}{2} \right) \mod(ND)
\]

where \(ND\) is the total number of discrete directions used in the simulation and \(\mod\) is the modulo operator.

All the simulations shown in this paper use 8 directions, so \(ND = 8\). Only the motion transparency and chopsticks simulations use 2 scales while all others use a single scale. These two simulations are different from the others in that they require interscale competition. Other than this difference, all simulations used the same set of parameters. Only the inputs were varied from one simulation to the next.

5.1 Level 1: Input

The input consists of a series of static frames each of which represents a time slice of a motion sequence. As mentioned in Section 2.1, the boundary representations at the farther depth, com-
puted by FACADE at each frame of the sequence, serve as the inputs, \( I_{ij} \), to the Formotion BCS Model. Input dimensions for each simulation are listed in Table 2.

**TABLE 2.** Input dimensions for all simulations.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Display Width (in pixels)</th>
<th>Display Height (in pixels)</th>
<th>No. of frames in the motion sequence</th>
<th>Other input specific parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic Barber Pole</td>
<td>60</td>
<td>30</td>
<td>15</td>
<td>No. of horizontal terminators = 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 2</td>
</tr>
<tr>
<td>Motion Capture</td>
<td>60</td>
<td>30</td>
<td>15</td>
<td>No. of horizontal terminators = 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of dots = 4</td>
</tr>
<tr>
<td>Spotted Barber Pole</td>
<td>60</td>
<td>30</td>
<td>15</td>
<td>No. of horizontal terminators = 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of dots = 4</td>
</tr>
<tr>
<td>Line Capture</td>
<td>71</td>
<td>71</td>
<td>10</td>
<td>None</td>
</tr>
<tr>
<td>Triple Barber Pole</td>
<td>60</td>
<td>90</td>
<td>15</td>
<td>No. of horizontal terminators = 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 6</td>
</tr>
<tr>
<td>Translating Square:</td>
<td></td>
<td></td>
<td></td>
<td>None</td>
</tr>
<tr>
<td>Visible Rectangular Occluders</td>
<td>33</td>
<td>33</td>
<td>15</td>
<td>No. of horizontal terminators = 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 15</td>
</tr>
<tr>
<td>Invisible Rectangular Occluders</td>
<td>33</td>
<td>33</td>
<td>15</td>
<td>No. of horizontal terminators = 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 15</td>
</tr>
<tr>
<td>Invisible Jagged Occluders</td>
<td>37</td>
<td>37</td>
<td>15</td>
<td>No. of horizontal terminators = 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No. of vertical terminators = 15</td>
</tr>
<tr>
<td>Motion Transparency</td>
<td>20</td>
<td>20</td>
<td>15</td>
<td>No. of dots = 20</td>
</tr>
<tr>
<td>Chopsticks</td>
<td>57</td>
<td>35</td>
<td>15</td>
<td>None</td>
</tr>
</tbody>
</table>

5.2 Level 2: Transient cell network

Undirectional transient cell activities, \( b_{ij} \), are computed by:

\[
b_{ij} = x_{ij} z_{ij}, \tag{A2}
\]

where simple cell activities, \( x_{ij} \), perform leaky integration of their inputs as follows:

\[
\frac{dx_{ij}}{dt} = 10[-x_{ij} + (2-x_{ij})I_{ij}] \tag{A3}
\]

and \( z_{ij} \) are habituative transmitter gates defined by:

\[
\frac{dz_{ij}}{dt} = 0.03[1-z_{ij} - 100x_{ij}z_{ij}] \tag{A4}
\]
The constants outside the brackets in (A3) and (A4) depict the rates of change of the simple cell activities, \( x_{ij} \), and the transmitter gates, \( z_{ij} \), respectively. In (A3), the constant 2 represents the maximum value that the simple cell activities can reach. The term \( 1 - z_{ij} \) in (A4) signifies that the transmitter gates \( z_{ij} \) can reach a maximum value of 1. The term \(-100x_{ij}z_{ij}\) in this equation says that the transmitter habituates in proportion to the strength of the signal passing through the gate with 100 being the constant of proportionality. Thus, the transmitter gates \( z_{ij} \) accumulate at a constant rate to a finite maximum value and habituate, or are inactivated, at a rate proportional to the strength \( x_{ij} \) of the signal. The undirectional transient cell responses, \( b_{ij} \), computed in Equation (A2) are the gated signals of (A4). These cell activities correspond to the lowest layer of cells depicted in Fig. 8.

Directional interneurons, \( c_{ij}^{d} \), perform a time-average of undirectional transient cell activities:

\[
\frac{dc_{ij}^{d}}{dt} = -c_{ij}^{d} + b_{ij} - 10[c_{XY}^{Do}]^+ \tag{A5}
\]

Each of these cells has a preferred directional tuning and helps to achieve directional selectivity. Each cell receives excitatory input, \( b_{ij} \), from the undirectional transient cell at the same spatial location, and inhibitory input, \( c_{XY}^{D_o} \), from the directional interneuron that is tuned to the opposite direction \( D^O \) and at a location \( (X, Y) \) that is spatially offset from location \((i, j)\) by one unit along the preferred direction of the cell, \( d \). For example, a directional interneuron tuned to leftward motion at location \((i, j)\) receives inhibitory input from the directional interneuron located one unit to its left and tuned to rightward motion (see Fig. 8). The inhibition that a cell receives is stronger than the excitation that it receives; cf., coefficient 10 in (A5).

The dynamics of directional transient cell activities, \( e_{ij}^{d} \), are similar to those of directional interneurons. These cells receive excitatory input from undirectional transient cells, \( b_{ij} \), and inhibitory input from directional interneurons, \( c_{XY}^{D_o} \):

\[
\frac{de_{ij}^{d}}{dt} = 10(-e_{ij}^{d} + b_{ij} - 10[c_{XY}^{D_o}]^+) \tag{A6}
\]

In Equations (A5) and (A6), direction \( D^O \) is the direction opposite to direction \( d \) and is computed by (A1). The output of Level 2 is rectified before being sent to Level 3: \( E_{ij}^{d} = [e_{ij}^{d}]^+ \).

Equations (A5) and (A6) implement a vetoing mechanism through spatially asymmetric inhibition. The need for inhibitory directional interneurons is not only biologically motivated, as discussed in Sections 2.2 and 4.2.1, but is also functionally essential. A veto mechanism based solely on inhibitory connections between neighbouring transient cells is insufficient because vetoed transient cells are incapable of further vetoing their neighbours. This problem is solved by introducing
inhibitory interneurons that are capable of maintaining their activities independently of the transient cells that they veto. Besides, interneurons can operate over a time scale different from that of the transient cells. Vetoing can thus be performed robustly at a variety of speeds. Mutual inhibition between interneurons is necessary to construct transient cells that respond preferentially to a range of directions of motion and whose response is essentially invariant with input speed and to preserve the speed tuning of the short-range filters at higher stimulus speeds.

5.3 Level 3: Short-range filter network

The short-range filter cell activities, \( f_{ij}^{ds} \), perform space- and time-averaging of directional transient cell responses. Each activity, \( f_{ij}^{ds} \), receives excitatory input from directional transient cells tuned to the same direction and within a Gaussian receptive field, \( G_{ijXY}^{ds} \), that is oriented along the preferred direction, \( d \), of the cell. The scale, \( s \), of each cell determines the size of its receptive field:

\[
\frac{df_{ij}^{ds}}{dt} = 4 \left[ -f_{ij}^{ds} + \sum_{(X,Y)} E_{XY}^{d} G_{ijXY}^{ds} \right].
\]  

(A7)

The Gaussian kernel, \( G_{ijXY}^{ds} \), is defined by the following equation for upward and downward directions of motion:

\[
G_{ijXY}^{Ds} = \exp \left[ -\frac{1}{2} \left( \frac{(X-i)^2}{\sigma_{Gx}} + \frac{(Y-j)^2}{\sigma_{Gy}} \right) \right],
\]  

(A8)

where \( D = 1 \) or \( 5 \), \( \sigma_{Gx} = 0.5 \) and \( \sigma_{Gy} = s + 0.5 \). The kernels for the remaining motion directions are obtained by rotating the kernel in (A8) and aligning it parallel to the current motion direction. Thresholded short-range filter cell outputs, \( F_{ij}^{ds} \), result from a self-similar threshold being applied to the short-range filter activities, \( f_{ij}^{ds} \). In other words, they threshold short-range filter activities by an amount that increases linearly with filter size. This ensures that each scale is activated preferentially by a different speed range that increases with scale size:

\[
F_{ij}^{ds} = \left[ f_{ij}^{ds} - \frac{s^2}{4} \right]^+
\]  

(A9)

5.4 Level 4: Competition network

Competition network cell activities, \( h_{ij}^{ds} \), implement both spatial competition within each direction and opponent directional inhibition within each scale. Shunting automatically gain controls cell responses to the inputs:
\[
\frac{dh_{ij}^{ds}}{dt} = 20 \left[ -h_{ij}^{ds} + (1 - h_{ij}^{ds}) \left( \sum_{(X,Y)} F_{XY} J_{ijXY}^d \right) - 10(\sigma_{ij} + 0.1) \left( \sum_{(X,Y)} F_{XY} K_{ijXY}^d \right) - 50h_{ij}^{ds} F_{ij}^o \right]
\]  
(A10)

Direction \(D^o\) is again the direction opposite to direction \(d\) and is calculated in (A1). The excitatory and inhibitory Gaussian kernels, \(J_{ijXY}^d\) and \(K_{ijXY}^d\), are defined by the following equations for upward motion, that is, direction \(d = 1\):

\[
J_{ijXY}^1 = \left( \frac{1}{2\pi \sigma_{J_x} \sigma_{J_y}} \right) \exp \left[ -\frac{1}{2} \left( \frac{(X-i)^2}{\sigma_{J_x}^2} + \frac{(Y-j)^2}{\sigma_{J_y}^2} \right) \right]
\]  
(A11)

and

\[
K_{ijXY}^1 = \left( \frac{1}{2\pi \sigma_{K_x} \sigma_{K_y}} \right) \exp \left[ -\frac{1}{2} \left( \frac{(X-i)^2}{\sigma_{K_x}^2} + \frac{(Y-(j-1))^2}{\sigma_{K_y}^2} \right) \right].
\]  
(A12)

The excitatory kernel, \(J_{ijXY}^1\), is spatially anisotropic with \(\sigma_{J_x} = 0.5\) and \(\sigma_{J_y} = 2.5\). On the other hand, the inhibitory kernel, \(K_{ijXY}^1\), is spatially isotropic with \(\sigma_{K_x} = \sigma_{K_y} = 4\), but it is offset from the spatial location of the cell, \((i, j)\), by one unit along the direction opposite to the preferred direction of the cell, that is, by one unit in the downward direction. In other words, the inhibition that a cell receives spatially lags behind the excitation that it sees, along the preferred direction of the cell. As with (A8), the kernels for the remaining motion directions are computed by rotating the kernels in (A11) and (A12) and aligning them parallel to the current motion direction. For instance, for the simulations in this paper all of which use 8 directions of motion, the kernels for north-east motion, that is, direction \(d = 2\), are obtained by rotating the kernels (A11) and (A12) clockwise by 45°. The activity of Level 4 is rectified before being output to Level 5: \(H_{ij}^{ds} = [h_{ij}^{ds}]^+\).

### 5.5 Level 5: Long-range Directional Grouping and Attentional Priming

The long-range filter sums over large spatial extents:

\[
N_{ij}^{ds} = \sum_{(X,Y)} (H_{XY}^{ds})^2 L_{ijXY}.
\]  
(A13)

In (A13), \(L_{ijXY}\) is an isotropic Gaussian kernel centered at position \((i, j)\) and defined by

\[
L_{ijXY} = \exp \left[ -\frac{1}{2} \left( \frac{(X-i)^2}{\sigma_{L_x}^2} + \frac{(Y-j)^2}{\sigma_{L_y}^2} \right) \right].
\]  
(A14)
where $\sigma_{Lx} = \sigma_{Ly} = 20$. Each model MT cell activity, $m_{ij}^d$, receives bottom-up excitation from the long-range filter and top-down inhibition from model MST cells, $n_{ij}^{Ds}$, tuned to all directions $D$ other than the preferred direction $d$ of the cell:

$$\frac{dm_{ij}^{ds}}{dt} = \left( -m_{ij}^{ds} + (1 - m_{ij}^{ds})N_{ij}^{ds} - (1 + m_{ij}^{ds}) \sum_{D \neq d} [n_{ij}^{Ds}]^+ \right). \quad (A15)$$

The output $M_{ij}^{ds} = [m_{ij}^{ds}]^+$.

### 5.5.1 Case I: Without Interscale Competition

Except for the cases of motion transparency and the chopsticks illusion, all simulations shown in this paper used only one scale and did not require interscale competition. The simplified equation for model MST cells used for these simulations is:

$$\frac{dn_{ij}^{ds}}{dt} = \left( -n_{ij}^{ds} + (1 - n_{ij}^{ds})M_{ij}^{ds} - 5 \sum_{D \neq d} [n_{ij}^{Ds}]^+ \right). \quad (A16)$$

By (A16), each model MST cell activity, $n_{ij}^{ds}$, receives excitation, $M_{ij}^{ds}$, from model MT cells and lateral inhibition from model MST cells tuned to all directions $D$ other than the preferred direction $d$ of the cell. This competition between model MST cells helps to choose a winning direction which then boosts activities in model MT cells that are tuned to the same direction, via Equation (A15).

### 5.5.2 Case II: With Interscale Competition: Motion Transparency & Chopsticks

The motion transparency and chopsticks simulations use two scales that compete with each other. In addition to the competition in Equation (A16), the equation for model MST cell activities, $n_{ij}^{ds}$, includes asymmetric inhibition from smaller to larger scales:

$$\frac{dn_{ij}^{ds}}{dt} = \left( -n_{ij}^{ds} + (1 - n_{ij}^{ds})(m_{ij}^{ds} + A_{ij}) - 0.01 \sum_{(X, Y)D \neq d} Z^{Dd} [n_{ij}^{Ds}]^+ P_{ijXY} - \sum_{S < s} n_{ij}^{ds} \right). \quad (A17)$$

In (A17), $P_{ijXY}$ is an isotropic Gaussian kernel defined by

$$P_{ijXY} = \exp\left[ -\frac{1}{2} \left( \frac{(X - i)^2}{\sigma_{P_x}} + \frac{(Y - j)^2}{\sigma_{P_y}} \right) \right], \quad (A18)$$

where $\sigma_{P_x} = \sigma_{P_y} = 20$. $Z^{Dd}$ is a kernel that ensures that inhibition between opponent directions is greater than that between any other two directions:
A_{ij}^{ds} is attentional enhancement that is specific to both direction and scale and directed to a given region of space. No attentional enhancement was used for the chopsticks simulation. For the motion transparency simulation, attention was directed to a particular direction, say $D^A$, and a specific scale, say $S$, within a given rectangular region of space centered at the center of the display $(C_X, C_Y)$, and with half-width $R_X = 5$ and half-height $R_Y = 5$. Direction $D^A$ is the direction for which the total activity in the long-range filter in the rectangular region is maximum. We assume that attention is always allocated to the closest depth; i.e., the smallest scale, so, $S = 1$:

$$A_{ij}^{ds} = \begin{cases} 
0.01 & |i - C_X| \leq R_X, |j - C_Y| \leq R_Y, d = D^A, s = S \\
0 & \text{otherwise}
\end{cases}$$

(A20)
6. References


