Laminar cortical dynamics of visual form and motion interactions
during coherent object motion perception

by

J. Berzhanskaya*, S. Grossberg +, E. Mingolla#

Department of Cognitive and Neural Systems

and

Center for Adaptive Systems

Boston University

677 Beacon Street

Boston, MA 02215

November, 2004


Keywords: motion perception, depth perception, perceptual grouping, motion capture, prestriate cortex.

*JB was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-01-1-0397), the National Geospatial-Intelligence Agency (NMA201-01-1-2016), the National Science Foundation (NSF BCS-0235398), and the Office of Naval Research (ONR N00014-95-1-0409 and ONR N00014-01-1-0624).

+SG was supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

#EM was supported in part by the National Geospatial-Intelligence Agency (NMA201-01-1-2016), the National Science Foundation (NSF BCS-02-35398 and NSF SBE-0354378), and the Office of Naval Research (ONR N00014-01-1-0624).
Abstract

How do visual form and motion processes cooperate to compute object motion when each process separately is insufficient? A 3D FORMOTION model specifies how 3D boundary representations, which separate figures from backgrounds within cortical area V2, capture motion signals at the appropriate depths in MT; how motion signals in MT disambiguate boundaries in V2 via MT-to-V1-to-V2 feedback; how sparse feature tracking signals are amplified; and how a spatially anisotropic motion grouping process propagates across perceptual space via MT-MST feedback to integrate feature-tracking and ambiguous motion signals to determine a global object motion percept. Simulated data include: the degree of motion coherence of rotating shapes observed through apertures, the coherent vs. element motion percepts separated in depth during the chopsticks illusion, and the rigid vs. non-rigid appearance of rotating ellipses.
Introduction. Visual motion perception requires the solution of the two complementary problems of motion integration and motion segmentation. The former process joins nearby motion signals into a single percept of object motion, while the latter keeps motion signals separate as belonging to different objects. These problems become particularly acute when an object moves behind multiple occluders. Then the various object parts are segmented by the occluders, but the visual system can often integrate these parts into a percept of coherent object motion. Studying conditions under which the visual system can and cannot accomplish correct segmentation and integration provides important cues to the processes that are used by the visual system to create object motion percepts during normal viewing conditions.

The present article further develops a 3D FORMOTION model, components of which were introduced by Baloch and Grossberg (1997), Chey, Grossberg, and Mingolla (1997, 1998), Francis and Grossberg (1996b), and Grossberg, Mingolla, and Viswanathan (2001). The model explains some challenging percepts during which small changes in object or contextual cues can dramatically change motion percepts from integration to segmentation. As the model’s name suggests, it proposes how form and motion processes interact to form coherent percepts of object motion in depth. The present work focuses on the following form-motion (or formotion) binding issues: How do form-based 3D figure-ground separation mechanisms in cortical area V2 interact with directionally selective motion grouping mechanisms in cortical areas MT and MST to preferentially bind together some motion signals more easily than others? In cases where form-based figure-ground separation mechanisms are insufficient, how do motion and attentional cues from cortical area MT facilitate figure-ground separation within cortical area V2 via MT-to-V1-to-V2 feedback? Finally, how does the global organization of the motion direction field in areas MT and MST influence whether the percept of an object’s form looks rigid or deformable through time?

The model goes beyond earlier motion models both by introducing novel formotion binding mechanisms and by proposing how laminar cortical circuits realize these mechanisms. These circuits embody explicit predictions about the functional roles that are played by the corresponding cells in the brain. The model extends to the motion system a program of developing laminar models of cortical circuits that has already progressed substantially towards explaining perceptual and brain data about 3D form perception (Grossberg, 1999; Grossberg, Mingolla, and Ross, 1997; Grossberg and Raizada, 2000; Grossberg and Seitz, 2003; Grossberg and Swaminathan, 2004; Grossberg and Williamson, 2001; Grossberg and Yazdanbakhsh, 2004; Raizada and Grossberg, 2003).

The model proposes solutions to several basic problems of motion perception, including the aperture problem. 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; i.e., the normal component of motion. The aperture problem is faced by any localized neural motion sensor, such as a neuron in the early visual pathway, which responds to a local contour moving through an aperture-like receptive field. In contrast, a moving dot, line end or corner provides unambiguous information about an object’s true motion direction (Shimojo, Silverman and Nakayama, 1989). The model proposes how such moving visual features activate cells in the brain that compute feature-tracking signals which can disambiguate an object’s true direction of motion.

A key issue concerns the assignment of motion to an object boundary when motion integration interpolates two contiguous parts of a scene, since not all line ends signal motion of
an object correctly. In the example in Figure 1, motion of the left line end corresponds to the real motion of the line. The right line end is formed by the boundary between the line and a stationary occluder, and its motion provides little information about the motion of the line. This issue has been in the vision literature for a long time; e.g., see Bregman (1981) and Kanizsa (1979). Nakayama, Shimojo, and Silverman (1989) have suggested classification of terminators as *intrinsic* and *extrinsic*: An intrinsic terminator belongs to the moving object; an extrinsic one belongs to the occluder.

**Figure 1**
Motion of intrinsic terminators is taken into account in computing the motion direction of an object, while motion of extrinsic terminators is generally ignored (Shimojo et al., 1989; Duncan, Albright and Stoner, 2000). The FACADE model (Grossberg, 1994, 1997; Kelly and Grossberg, 2000) of 3D form vision and figure-ground separation proposed how boundaries in 2D images are assigned to different objects in different depth planes, thus treating terminators as either extrinsic or intrinsic ones. A precursor of the present model (Grossberg, Mingolla, and Viswananathan, 2001) proposed how FACADE figure-ground separation in cortical area V2, combined with formotion interactions from area V2 to MT, enable intrinsic terminators to create strong motion signals on a moving object, while extrinsic terminators create weak ones. These simulations assumed that figure-ground separation had already occurred within the form system, and used depth-separated boundaries from V2 as inputs to the motion system. The present model starts with motion signals in V1, where the separation in depth has not yet occurred, and predicts how V2-to-MT boundary signals can capture V1-to-MT motion signals at the correct depths, how MT-to-V1 feedback signals can bias boundary formation in V1 and V2, and how motion signals adapt in V1. The motion capture mechanism clarifies why we tend to perceive motion of visible objects and background features, but not of the intervening empty spaces between them. This may not seem to be a serious problem if we just consider the motion signals of which we are consciously aware. However, when one considers how motion signals can have an influence on visible features across empty space, as during induced motion, without causing visible motion within the intervening space that is devoid of visible features, one readily sees that it is a phenomenon that requires explanation. Motion capture in MT using depth-separated form boundaries from V2 is, we believe, a part of the explanation. Adaptation of motion signals in V1 leads to attenuation of motion signals generated by a static extrinsic terminator. Moving intrinsic terminators, on the other hand, generate strong motion signals. As local motion signal direction and strength are computed, a motion integration process in MT-MST decides the winning motion direction in the case of a single moving line, as in Figure 1.

**Figure 2**
What happens if multiple moving objects overlap? Experiments on plaids and random dot motion have demonstrated at least two possible perceptual outcomes (Ferrera and Wilson, 1987, 1990; Kim and Wilson, 1993; Snowden et al., 1991; Stoner and Albright, 1998; Stoner, Albright, and Ramachandran, 1990; Trueswell and Hayhoe, 1993) (Figure 2). First, a display can separate into two depth planes, forming a transparent motion percept, where two dot-filled planes or two gratings slide one over another. Second, if the directions of motions are compatible, then displays can produce a percept of a coherent motion of a unified pattern, and no separation in depth occurs.

**Figure 3**
While such separation can happen purely in the motion system, occluder information from the form system can modulate the calculation of motion signals (Stoner and Albright, 1996, 1998).
For example, the present article models the motion percepts that are generated by a chopsticks display (Figure 3) (Anstis, 1990). The bars in this display undergo translational motion, and may be thought of as a simplified plaid motion display. When the chopsticks move horizontally, their intersection moves vertically. In the case of visible occluders (Figure 3A), the intersection motion prevails and vertical motion of a single X-shaped object is perceived. In the case where the chopstick ends are visible (Figure 3B) — that is, the occluder is invisible — the percept is of two chopsticks moving in opposite horizontal directions and separated in depth. This depth separation cannot happen based only on the boundaries of the X-shaped form, since the boundaries near the middle of the X do not complete either bar explicitly. The 3D FORMOTION model proposes how signals from the motion to the form stream via MT-to-V1 feedback can initiate the process whereby these ambiguous boundaries can be completed and separated in depth.

Often the shape of a moving object is more complex than that of a line, and can affect the outcome of motion integration. The present article models data of Lorenceau and Alais (2001), who studied different shapes moving in a circular-parallel motion behind occluders (Figure 4). Observers had to determine the direction of motion, clockwise or counterclockwise. The percent of correct responses depended on the type of shape, and on the visibility of the occluders. In the case of a diamond (Figure 4A), a single, coherent, circular motion of a partially occluded rectangular frame was easy to perceive across the apertures. In the case of an arrow (Figure 4C), two objects with parallel sides were seen to generate out-of-phase vertical motion signals in adjacent apertures.

**Figure 4**
Local motion signals were identical in both displays, and only their spatial arrangement differed. Alais and Lorenceau suggested that certain shapes (such as arrows) “veto” motion integration across the display, while others (such as diamond) allow it. The 3D FORMOTION model explains the data without using a “veto” process. The model proposes that the motion grouping process uses anisotropic direction-sensitive receptive fields that preferentially integrate motion signals within a given direction across gaps produced by the occluders. The explanation of Figures 4D-F follows in a similar way, with the additional factor that the ends of the bars possess intrinsic terminators that can strongly influence the perceived motion direction of the individual bars.

Another example of where percepts of rotational motion involve motion grouping is the “gelatinous ellipses” display (Vallortigara et al., 1988, Weiss and Adelson, 2000). When the “thin” (high aspect ratio) and the “thick” (low aspect ratio) ellipses rotate around their centers, the perception of their shapes is strikingly different. The thin ellipse is perceived as a rigid rotating form, whereas the thick one is perceived as deforming non-rigidly through time (Figure 5). Here, the differences in 2D geometry result in differences of the spatiotemporal distribution of motion direction signals that are grouped together through time. When these motion signals are consistent with the coherent motion of a single object, then the motion grouping generates a percept of a rigid rotation. When the motion field decomposes after grouping into multiple parts, with motion trajectories incompatible with a rigid form, a non-rigid percept is obtained. Motion-to-form projections can once again help to explain these distinct outcomes. The ability of nearby “satellites” to convert the non-rigid percept into a rigid one can also be explained by motion grouping. Weiss and Adelson (2000) have proposed that such a percept can also be explained via a global optimization process. Motion grouping provides a biologically plausible alternative proposal.
In summary, all of the data considered here illustrate how the brain may use both form and motion information, and their interaction, to derive a global percept of object motion. Form and motion processes, such as those in V2/V4 and MT/MST, go on in distinct What ventral and Where dorsal cortical processing streams. Related modeling work has proposed that key mechanisms within the What and Where streams obey computationally complementary laws (Grossberg, 2000): The ability of each process to successfully compute some properties prevents it from computing other, complementary, properties. The present article clarifies some of the interactions between form and motion processes that enable them to overcome their complementary deficiencies and to thereby compute more informative representations of unambiguous object motion.

The 3D FORMOTION model stages and interactions between streams are shown in Fig. 6. The form stream in V1 includes oriented contrast-sensitive boundary detectors (simple cells), oriented contrast-pooling boundary detectors (complex cells), spatial and orientational competition (hypercomplex cells) stage and boundary completion (bipole cells) stage. These stages belong to the BCS, or Boundary Contour System, of the FACADE model. For simplicity, FACADE separation of boundaries in depth based on filling-in at V2 is not simulated. It is approximated by formation of 1-unit wide idealized V2 boundaries at appropriate depths as soon as separation criteria that emulate the corresponding FACADE mechanisms are reached at the V2 bipole cells. V2 boundaries project back to V1, supporting boundary completion at the appropriate depths.

The motion stream in V1 includes direction-sensitive cells, short-range anisotropic filters that are elongated in the direction of motion to accumulate local motion signals in their preferred direction, and a spatial competition stage that enhances motion signals from unambiguous moving features while suppressing ambiguous motion signals. Motion signals are then projected to MT. The MT-MST stages include: selection of motion signals by V2 boundaries, integration of motion signals through the large anisotropic receptive fields that are elongated in the direction of motion, and MT-MST interactions that allow only one motion direction to win at each spatial location. MST-to-MT feedback can also carry an attentional enhancement or attentional priming signal. MT feedback to V1 biases boundary separation in depth in the form stream when boundary depth order is ambiguous. For example, in the display of chopsticks without occluders, orthogonal boundaries compete with each other. MT feedback breaks this competition symmetry and pulls one boundary into the near depth.

Each model layer consists of a 60x60 matrix with multiple cells that code for different properties such as line orientation or motion direction at each position. A detailed model description is provided after simulations are presented and in Appendix A.

Analysis and Simulation of Psychophysical Experiments
This section is devoted to a detailed analysis and simulations of three important kinds of psychophysical displays: shapes moving behind occluders (Lorenceau-Alais, 2001), chopsticks (after Anstis, 1990) and rotating ellipses (Weiss and Adelson, 2000).

Movement behind occluders. Lorenceau and Alais (2001) created displays in which circular-parallel motion was visible through the two vertically oriented apertures, but the corners of the shapes remained hidden (Figure 4). Therefore, observers had to rely on motion integration across space to determine motion direction. The success of the motion integration process depended on the type of shape and on the contrast of the occluders. The diamond displays
resulted in a higher percentage of correct responses than the cross and arrow displays, and displays with visible occluders were easier than those with invisible ones. For example, a diamond (Figure 4A) rotating behind visible occluders created a percept of a single rotating shape. In contrast, a rotating arrow (Figures 4C and 4F) produced a percept of two disconnected shapes separately moving in their respective apertures. This disconnection was strong even in the case of visible occluders (Figure 4C) and more pronounced in the case of invisible occluders (Figure 4F).

Figure 7
Schematic representations of the motion grouping signals generated by the displays of a diamond and an arrow with visible occluders are shown in Figure 7. At the corresponding time points (for example, Figures 7A and 7B, \( t = n \)), each display has a combination of the same set of local motion signals. Perceptual dissimilarities are caused by the difference in relative positioning of those motion signals through time. Both the diamond and the arrow are visible through the apertures as four linear boundary segments. Each segment produces two types of motion signals: ambiguous signals (due to the aperture problem) from line interiors and unambiguous signals from terminators, shown in the Figure 7 inset. For the visible occluder cases, the terminator signals are extrinsic and weak. Ambiguous motion signals of the same direction from parallel segments can then combine across space using the model’s anisotropic motion grouping filters to produce the perceived object motion.

For example, in the diamond display in Figure 7A, two line segments with synchronous motion in a given direction are located in different apertures. The large anisotropic motion grouping cells that prefer this motion direction can thus integrate the diagonal motion signals across the apertures. At time \( t = n \), two segments activate the diagonal motion cells, while only one segment activates vertical or horizontal ones. The MT-MST motion grouping network therefore prefers the diagonal signals from the line interiors to the weaker vertical or horizontal groupings. Cells activated the most would be those over the center of the rotating shape. First, the cells with a 45º direction preference will be activated \( (t = n) \), then 135º cells \( (t = n + m) \), 225º, 315º, and then back to the beginning of the cycle. Simulation results are shown in Figure 8. This sequence of motion signals is consistent with the circular motion in a counter-clockwise direction, leading to a coherent percept of a rotating diamond.

Figure 8
For the arrow display in Figure 7B, vertical components of the ambiguous signals from the line interiors and vertical extrinsic signals from the line ends activate vertically oriented anisotropic long-range filter cells. Diagonal ambiguous motion signals from neighboring parallel shape segments can only weakly group together within one aperture, and so lose the directional competition that determines the winning direction. As a result, a vertical direction of motion will integrate within a certain time frame in a given aperture \( (t = n) \), but this vertical direction will develop at a later time in the other aperture \( (t = n + m) \). The result is a seesaw up-and-down translational motion that is inconsistent with rotation. In addition, such out-of-phase timing of motion signals will prevent motion integration across the two apertures. Another way of saying this is that asynchronous motions of similar directions produce a segmentation signal, thus preventing a percept of a single rotating object. Analysis of motion signals in the invisible occluder displays (Figures 4D-4F) is similar to the analysis above. Because line terminators are intrinsic, they will produce stronger vertical signals and aid the vertical motion grouping. Simulations of motion segmentation for the case of arrow with invisible occluders (Figure 4F) are shown on Figure 9.
An intermediate image configuration, such as the diamond with invisible occluders in Figure 4D, creates strong vertical feature tracking signals within each aperture that can better compete with the strong diagonal ambiguous motion grouping across apertures. The percept is thus determined by competition between two motion directions and results in a larger number of “incorrect answers” than does the percept in the visible occluder case of Figure 4A. In the case of an arrow with visible occluders in Figure 4C, the vertical signals will be weak because they are extrinsic, whereas in Figure 4F they are strong because they are intrinsic. Thus, translation will overwhelm rotation less in Figure 4C than in Figure 4F, and the number of correct responses about arrow rotation will be higher there. All of these model properties are consistent with the data of Lorenceau and Alais (2001).

Figure 9

**Chopsticks with visible and invisible occluders.** Two configurations of the chopsticks display, with visible and invisible occluders (Figures 3A and 3B), were simulated. In the case of visible occluders, chopsticks are perceived moving coherently in a vertical direction. In the case of invisible occluders, the percept is of two horizontally moving objects, one moving in front of the other. These two displays differ only at the chopsticks’ ends. The difference in motion percept here can be explained by the difference in the relative strength of unambiguous feature-tracking motion signals of the intersection and either strong (intrinsic) or weak (extrinsic) motion of the chopsticks’ ends. Aperture-ambiguous motion signals at the line interiors do not play a significant role in this percept.

Figure 10

Independent of the visibility of occluders, in the static image, the two chopsticks are perceived as one X-shaped pattern. However, in the moving image, chopsticks with invisible occluders separate in depth and are perceived as sliding one above another. Simulations of the chopstick display in the invisible occluder case are shown in Figures 10 and 11. Figure 10 shows how, in the motion system, opposite direction signals from two chopsticks separate in depth. The sequence of motion computations leading to this percept starts with strong horizontal motion direction signals from the intrinsic terminators at the chopsticks’ ends. These feature-tracking signals are amplified by anisotropic short-range motion filters of V1 that accumulate evidence in a given motion direction as the chopstick moves along, and are integrated by the long-range filters of MT. Attentional priming biases motion signal at one chopstick end (top-left) in the near depth. Competition within the MT-MST circuit includes asymmetric inhibition from the near depth (D1) to the far depth (D2) (“asymmetry between near and far”). This interaction results in the primed motion direction winning in D1 and another motion direction winning in D2.

Figure 11

Initially, the bipole cells of orthogonal diagonal orientation preferences in the form system compete with each other, but are unable to complete over the gap formed by the chopsticks’ intersection (Figure 11A). The bias that allows one chopstick to win the competition can be provided by an attentional input in the form system, by feedback from the motion system, or by introducing some inequality in the chopsticks’ properties (e.g., by making one thicker). In the current simulations, depth-selective feedback from MT modulates complex cells of the corresponding depths in V1. This feedback equals the sum of the motion signals in a given depth, and is not orientation or direction selective. Motion signals in MT are spatially restricted to one chopstick in each depth and, through the feedback, enhance boundary signals for this chopstick more than for the other. Due to this motion bias, bipole boundaries of the corresponding
chopstick complete in D1, thus pushing the second chopstick boundary in the further depth (Figure 11B).

In the case of visible occluders, the chopsticks’ ends are extrinsic terminators and do not create strong motion signals, but the vertical motion of the chopstick intersection is unambiguous and strong. The result of motion integration and competition is a coherent, vertical motion signal at the far depth, D2 (Figure 12B). This signal does not provide a segmentation bias in the feedback to the form system. The chopstick boundaries at this depth form an outline of an “X” shape (Figure 15) moving up and down, and none of the competing bipole boundaries are able to win. The form system output at the near depth, D1, consists of two static horizontal boundaries of the occluders (Figure 12A). The model predicts that these depth-separated boundaries in V2 capture motion signals in the corresponding depth representations of MT via V2-to-MT projections with excitatory centers and inhibitory surrounds. Bottom-up motion signals along the horizontal occluder boundaries consist mainly of the motion of extrinsic terminators, and are weakened by adaptation at the input layers of V1 (transient cells in Figure 6). Furthermore, surround inhibition produced by the same boundaries suppresses motion signals from interior parts of the display. This combination of narrow excitatory projections from V2 to MT with wide inhibitory surrounds results in no significant motion signal in the MT representation of the near depth, D1 (Figure 12A). On the other hand, selection by “X”-shaped boundaries in D2 picks up a strong bottom-up signal from the chopsticks’ intersection and an ambiguous signal from line interiors, resulting in a global vertical motion percept in the far depth, D2.

Figure 12

Gelatinous ellipses. The perception of rigidity of rotating ellipses depends on their shape (Figure 5). The 3D FORMOTION model suggests that the processes determining rigidity of the boundary are similar to those determining the percept of coherent vs. incoherent motion, as well as the percept of a single object vs. assignment of neighboring boundaries to different objects, possibly at different depths. In the non-rigid case (thick ellipse), analysis of local motion signals shows that local motion signals perpendicular to the ellipse boundary may prevail. As in the case of incoherent Lorenceau-Alais displays (arrow), each segment of the ellipse boundary moves in the manner inconsistent with a single (object) motion in the display (Figure 13A).

Figure 13

In the rigid ellipse case (thin ellipse) the dominant motion signal is consistent with a single object rotation that is tangential to the boundary at the points of the highest curvature (Figure 13B). The resulting motion percept in the ellipse displays is determined by the competition among ambiguous local signals integrated through large MT receptive fields. This hypothesis is supported by the “satellite effect” (Weiss and Adelson, 2000): dots moving outside of the ellipse can bias the perception of rigidity. If dots, which provide unambiguous motion signals, move along circular trajectories, then the ellipse, even a thick one, is perceived as rigidly rotating (Figure 14A). If dots oscillate in the direction orthogonal to the contour, the ellipse, even a thin one, is perceived as deforming (Figure 14B).

Weiss and Adelson (2000) reported that the capture of an ambiguous ellipse motion by unambiguously moving satellites happens even if both lie in different depth planes (as defined by disparity). Moreover, in the case of two pairs of satellites, the closer one in depth captures ambiguous ellipse motion and determines the global percept. These data can be explained by the depth-selectivity of V2 _ MT projections (Bradley and Andersen, 1998). For example, the maximum capture signal will be at the depth of the satellites, and the strength of the capture signal will decrease with the difference in depth between the satellites and the ambiguous motion
signals. The ambiguous motion signals that are closest to the depth of the satellites will thus be captured more easily within their depth plane. The outcome of the competition of two sets of satellites will be determined by the one with the stronger motion signal in the ellipse depth plane. These effects are not simulated in the present article, but they are clearly implied by the 3D FORMOTION model.

Figure 14

3D FORMOTION Model
The 3D FORMOTION model (Figures 6 and 15) builds upon two previous models: FACADE and Formotion.

Figure 15

Motion Processing System. The motion processing part of the model consists of six stages that represent cell dynamics homologous to LGN, V1, MT, and MST (Figure 15, right).

Level 1: Input from LGN. A precursor of the present model (Grossberg et al., 2001) used FACADE output from V2 as the input to the Motion system. In the 3D FORMOTION model, the boundary input (Level 1) is not depth-specific. Rather, the 2-cell wide boundary input models the signals that come from LGN into V1 (Xu, Bonds and Casagrande, 2002). This boundary is represented in both ON and OFF channels. After a few stages of V1 processing, the motion signal then goes on to MT and MST. The 3D figure-ground separated boundary inputs in the current model come from V2 to MT and select bottom-up inputs from V1 in a depth-specific way. This biologically more realistic input scheme proposes how the visual system overcomes a significant challenge, since the input itself does not separate the occluder boundaries from the moving boundaries into different depth planes. The present model proposes how a combination of habituative (Appendix equations (A4)-(A6)) and depth capture (Appendix equation (A11)) mechanisms accomplishes the required depth segregation of motion signals.

These mechanisms are proposed to also have several other roles in motion processing. In particular, habituative mechanisms are part of the preprocessing whereby motion cues trigger the activation of transient cells; see below. Because the occluder boundaries are static, at least relative to the continuously moving chopsticks, their signals become much weaker over time. As a result, when the chopsticks (Figure 3) move along the fixed locations of the static occluders, they generate much weaker motion signals than the same chopsticks moving without occluders. This habituative property helps to explain why visible occluders generate weaker motion signals at all depth planes. It does not, however, separate intrinsic from extrinsic boundaries. The motion capture mechanism does this by using depth-separated occluder and occluding boundary signals from V2 to MT. For simplification, after the BCS completes contours in corresponding depths (Appendix equations (A23)-(A24)), these signals are approximated by 1-pixel wide, depth-separated boundaries. The model shows how these boundaries can capture only the appropriate motion signals onto their respective depth planes in MT (Figure 12).

3D FORMOTION uses both ON and OFF input cells. For example, when a bright chopstick moves to the right on a dark background (Figure 3, polarities are reversed for illustration purposes), ON cells respond to its leading edge, but OFF cells respond to its trailing edge. Likewise, when the chopstick reverses direction and starts to move to the left, its leading edge now activates ON cells and its trailing edge OFF cells. By differentially activating ON and OFF cells in different parts of this motion cycle, these cells have more time to recover from habituation, so that the system remains more sensitive to repetitive motion signals. Model ON and OFF responses are thus relevant to the role played by habituative mechanisms in generating transient cell responses and in weakening the boundaries of occluders.
**Level 2: Transient cells.** The second stage of the motion processing system (Figures 6 and 15) consists of non-directional transient cells, directional interneurons and directional transient cells. The non-directional transient cells respond briefly to a change in the image luminance, irrespective of the direction of movement (Appendix equations (A4)-(A6)). Such cells respond well to moving boundaries and poorly to the static occluder because of the habituation, or adaptation that creates the transient response. Adaptation occurs not only in V1 (Abbott, Sen, Varela and Nelson, 1997; Carandini and Ferster, 1997; Chance, Nelson and Abbott, 1998; Varela, Sen, Gibson, Fost, Abbott and Nelson, 1997) but also at earlier stages in the visual system, such as at the Y cells of the cat retina (Enroth-Cuggell and Robson, 1966; Hochstein and Shapley, 1976a, 1976b). For simplicity, here we consider only the V1 input stage. As noted above, after the transient cells adapt in response to a static boundary, then boundary segments that belong to a static occluder (extrinsic terminators, Figure 12A) produce weaker signals than those that belong to a continuously moving object. In the invisible occluder display (Figure 12B), the horizontal motion signals at the chopstick ends will be strong, and thus influence the final outcome.

A directionally selective neuron fires vigorously when a stimulus is moved through its receptive field in one direction (called the preferred direction), while motion in the reverse direction (called the null direction) evokes little response (Barlow and Levick, 1965). Mechanisms of direction selectivity include asymmetric inhibition along the preferred cell direction, notably an inhibitory veto of null-direction signals (Appendix equations (A7)-(A8)), as in Grossberg et al. (2001).

**Level 3: Short-range filter.** A key step in solving the aperture problem is to strengthen unambiguous feature tracking signals relative to ambiguous motion signals. Feature tracking signals are often generated by a relatively small number of moving features in a scene, yet can have a very large effect on motion perception. One process that strengthens feature tracking signals relative to ambiguous aperture signals is the short-range spatial filter (Figure 15). Cells in this filter accumulate evidence from directional transient cells of similar directional preference within a spatially anisotropic region that is oriented along the preferred direction of the cell. This computation selectively strengthens the responses of short-range filter cells to feature-tracking signals at unoccluded line endings, object corners, and other scenic features (Appendix equation (A9)). The use of a short-range spatial filter followed by competition at Level 4 avoids the feature correspondence problem that various other models need to solve (Reichardt, 1961; van Santen and Sperling, 1985).

**Level 4: Spatial competition and opponent direction competition.** Two kinds of competition further enhance the relative advantage of feature tracking signals (Figures 6 and 15, Appendix equation (A10)). These competing cells are proposed to occur in layer 4B of V1 (Figure 15, bottom-right). Spatial competition among cells of the same spatial scale that prefer the same motion direction boosts the amplitude of feature-tracking signals relative to those of ambiguous signals. Feature tracking signals are contrast-enhanced by such competition because they are often found at motion discontinuities, and thus get less inhibition than ambiguous motion signals that lie within an object’s interior. Opponent-direction competition also occurs at this processing stage, similar to the V1 cells described by Rust, Majaj, Simoncelli and Movshon (2002) both in exhibiting an opponent direction mechanism, and in having the correct spatial scale for such interactions. The activity pattern at this model stage is consistent with data of Pack, Gartland and Born (2004). First, in their experiments, V1 cells demonstrate an apparent suppression along visible occluders. A similar suppression occurs in the model due to the
adaptation of transient inputs to static boundaries. Second, cells in the middle of a grating (influenced only by ambiguous signals) respond more weakly than cells at the edge of the grating (influenced by intrinsic terminators). This effect is explained in the model by spatial competition between motion signals, which results in attenuation of signals from line interiors.

Level 5: Long-range filter and formotion capture. Motion signals from model layer 4B of V1 input to model area MT. Area MT also receives a projection from V2 (Anderson and Martin, 2002; Rockland, 1995) that carries depth-specific figure-ground-separated boundary signals. These V2 form boundaries capture the motion signals (formotion capture) by selectively assigning to different depths the motion signals coming into MT from layer 4B of V1 (Appendix equation (A11)). When the dynamically formed V2 boundary signals satisfy an appropriate criterion (Appendix equations (A23) - (A24)), they are projected to MT as 1-pixel wide depth-separated boundaries. This approximation eliminates the need to do a complete FACADE model simulation.

Formotion capture, or selection of motion signals in depth by corresponding boundaries, is proposed to occur via a narrow excitatory on-center, broad inhibitory off-surround projection from V2 to layer 4 of MT. For example, in response to the chopsticks display with visible occluders (Figures 3A), the motion capture mechanism for depth D1 selects motion signals at its positions in D1, which lie along the visible occluder boundaries, and suppresses motion signals at other locations in depth D1. The resulting activation in D1 will be weak, due to the habituated bottom-up input from V1 along the selected occluder boundary positions (Figure 12A). The V2 boundary signals that correspond to the moving boundaries capture strong motion signals at depth D2 (Figure 12B).

A similar type of inter-stream gating signal is proposed to play a key role in explaining challenging data about stereopsis, 3D surface perception, and figure-ground separation (Cao and Grossberg, 2004; Fang and Grossberg, 2004 Grossberg, 1994, 1997; Grossberg and Yazdanbakhsh, 2004). This gating signal is proposed to operate within the form system, namely from the thin stripes to the pale stripes of V2, and allows 3D surface feedback to modulate the strength of 3D boundaries that control visible 3D form percepts. Thus it seems that several different types of gating occur across the parallel visual processing streams at the V2 and MT processing levels.

The boundary-gated signals from layer 4 of MT are proposed to input to the upper layers of MT (Figure 15, top-right), where they activate directionally-selective, spatially anisotropic filters via long-range horizontal connections (Appendix equation (A12)). In this long-range filter, motion signals coding the same directional preference are pooled from object contours with multiple orientations and opposite contrast polarities. This pooling process creates a true directional cell response (Chey, et al., 1997; Grossberg et al., 2001; Grossberg and Rudd, 1989, 1992). Earlier versions of the long-range filter used a spatially isotropic kernel, for simplicity. In order to explain the types of data analyzed in this paper, we propose that the long-range filter accumulates evidence of a given motion direction using a kernel that is elongated in the direction of that motion, much as in the case of the short-range filter. This hypothesis is consistent with data showing that approximately 30% of the cells in MT show a preferred direction of motion that is aligned with the main axis of their receptive fields (Xiao, Raiguel, Marcar and Orban, 1997).

The predicted long-range filter cells in layer 2/3 of MT are proposed to play a role in binding together 3D directional information that is homologous to the orientationally selective, coaxial and collinear accumulation of evidence within layer 2/3 of the pale stripes of cortical
area V2 for the purpose of 3D perceptual grouping of form (Grossberg 1999; Grossberg and Raizada, 2000. This anisotropic long-range motion filter allows motion signals to be selectively integrated across occluders with variable degrees of success in response to the various shapes in the Lorenceau-Alais displays of Figure 4.

**Level 6: Directional grouping.** The model processing stages up to now do not fully solve the aperture problem. Although they can amplify feature tracking signals and assign motion signals to the correct depths, they cannot yet explain how feature tracking signals can propagate across space to select consistent motion directions from ambiguous motion directions, without distorting their speed estimates, and at the same time suppress inconsistent motion directions. They also cannot explain how motion integration can compute a vector average of ambiguous motion signals across space to determine the perceived motion direction when feature tracking signals are not present at that depth. The final stage of the model accomplishes this goal by using a motion grouping network (Appendix equations (A13)-(A14)), interpreted to occur in ventral MST (MSTv). This motion grouping network is predicted to determine the coherent motion direction of discrete moving objects. The motion grouping network works as follows: Cells that code the same direction in MT — and also possibly similar directions, but this possibility is not explored herein — send convergent inputs to cells in MSTv via the motion grouping network. Within MSTv, directional competition at each position determines a winning motion direction. This winning directional cell then feeds back to its source cells in MT. This feedback supports the activity of MT cells that code the winning direction, while suppressing the activities of cells that code all other directions. This motion grouping network enables feature tracking signals to select similar directions at nearby ambiguous motion positions, while suppressing other directions there. On the next cycle of the feedback process, these newly unambiguous motion directions select consistent MSTv grouping cells at positions near them. The grouping process propagates across space as the feedback signals cycle through time between MT and MSTv. Chey et al. (1997) and Grossberg et al. (2001) first used this process to simulate data showing how the model solves the aperture problem.

**Form Boundary System.**

**Perceptual Grouping and Figure-Ground Separation of 3D Form.** The FACADE boundary completion system, called the Boundary Contour System, or BCS (Figures 6 and 15, left), predicts how boundaries of occluding surfaces are separated from occluded surfaces in depth, including the separation of extrinsic vs. intrinsic boundaries (Grossberg, 1994, 1997; Grossberg and Yazdanbakhsh, 2004; Kelly and Grossberg, 2000), within the pale stripes of V2. One cue of occlusion in a 2D image is a T-junction. The black bar in Figure 16A forms a T-junction with the gray bar (Figure 16B). The top of the T belongs to the occluding black bar, while the stem belongs to the occluded gray bar. Bipole long-range grouping (Figure 16C) strengthens the horizontal boundary, while short-range competition weakens the vertical boundary (Figure 16D). This end gap in the boundary initiates the process of separating occluding and occluded boundaries. In other words, basic properties of perceptual grouping are predicted to initiate the separation of figures from their backgrounds, without the use of explicit T-junction operators. In order to simplify the simulations, the model does not include all the stages of boundary and surface interaction that complete figure-ground separation. That these mechanisms work has been demonstrated elsewhere (Fang and Grossberg, 2004; Grossberg and Cao, 2004, Grossberg and Yazdanbakhsh, 2004; Kelly and Grossberg, 2000). Instead, as soon as T-junctions have been detected by the model dynamical equations, boundaries are
algorithmically separated in depth. A further computational simplification in our simulations is that result of such separation is represented by 1-cell wide boundaries; corresponding to the demonstrated ability of bipolar cells to form sharp boundaries. For example, at a T-junction, a long horizontal boundary will be represented in Depth 1 and a short vertical boundary in Depth 2. These idealized boundaries are positioned at the same locations as input boundaries. The effect of motion on boundary position shifts is explored in simulations of flash-lag and flash-drag effects by Berzhanskaya, Grossberg and Mingolla (2004). V2 boundaries are used to provide both V2-to-MT motion selection signal (Appendix equation (A11)) and V2-to-V1 depth-biasing feedback (Appendix equation (A17)) (Figure 15, top-left). While V2-to-V1 feedback is orientation-specific, V2-to-MT projection sums boundary signals over all orientations, just as motion signals do at MT (Albright, 1984).

**Motion Modulation of Figure-Ground Separation.** Form cues are not always available to initiate figure-ground separation. Motion cues can initiate figure-ground separation even when form cues are not available. One such route in the model is via feedback projections from MT to V1 (Figures 6 and 15, Appendix equation (A17)), which have been reported both anatomically and electrophysiologically (Bullier, 2001; Jones, Grieve, Wang and Sillito, 2001; Movshon and Newsome, 1996). Above it was shown how this MT-to-V1 feedback can strengthen the boundary signals of one chopstick enough to trigger FACADE mechanisms of figure-ground separation, even when the enhanced motion signals from this chopstick may be the only cue for depth separation in the form system (Figure 3B). A certain direction of motion can, for example, be enhanced within the MT/MST complex at a given depth by an attentional prime (Treue and Maunsell, 1999), which can then propagate to V1 via MT to help separate opposite motions in depth (Grossberg et al., 2001, Bradley, Chang and Andersen, 1998), as well as to separate orthogonal boundaries.

**Discussion**

**Previous models of motion integration and segmentation.** A number of motion models have dealt with mechanisms of directional selectivity, motion integration and segmentation. For a review, see Grossberg et al. (2001). Only few of them have addressed the issues of extrinsic vs. intrinsic terminators, and the effect of this dichotomy on the motion processing. Lidén and Pack (1999) proposed that T-junctions, which indicate occlusion in 2D images of 3D scenes, can suppress motion signals in their vicinity. Their model does not, however, explain how occluding and occluded objects are separated in depth, or how varying the relative contrasts at the X- and T –junctions can cause totally different outcomes, as explained in Grossberg and Yazdanbaksh (2004). Wilson, Ferrera and Yo (1992) proposed that these are parallel Fourier and non-Fourier channels in motion processing. However psychophysical data do not support the existence of these pathways (Bowns, 1996; Cox and Derrington, 1994).

Authors of the three sets of data simulated in this article proposed explanations for their respective data that differ from those offered by the 3D FORMOTION model. For example, Lorenceau and Alais (2001) suggested that some shapes rotating behind occluders produce weak rotational motion percepts because of a “veto” imposed on motion integration. Only the “bad” shapes, those that cannot form a closed contour, would veto motion integration. Mechanisms and cortical locations of the “veto” process were not specified. In contrast, the 3D FORMOTION model suggests that anisotropic receptive fields integrate motion across apertures as a part of the process generating a coherent object motion percept. Some MT cells have elongated receptive fields (Xiao et al., 1997) that can be formed by long-range anisotropic projections (Schmidt, Goebel, Lowell, Singer, 1997; Sincich and Blasdel, 2001) in the upper laminae of MT (Malach,
Schriman, Harel, Tootell and Malonek, 1997). The 3D FORMOTION model thus explains differences in motion percepts using known cortical mechanisms.

Several prior models compute motion signals for gratings and plaids. However, none of them can explain in detail the different percepts for the chopstick illusion, which can be considered as a limiting case of plaid consisting of just two bars: the visible occluder case produces a coherent vertical motion, while the invisible occluder case results in motion separation in depth. Typically, alternative motion models concentrate on motion mechanisms and do not explain how 3D figure-ground separation mechanisms form extrinsic and intrinsic terminators, and how these terminators affect global motion computations. Grossberg et al. (2001) provided a partial explanation of how local motion signals in the ambiguous positions can be overwhelmed by the propagation of the strong feature-tracking signals from the chopsticks’ ends. The 3D FORMOTION model uses the same propagation of feature-tracking signals together with the new form-motion interactions to more fully explain all aspects of the chopsticks illusion.

Previous models of the ellipse illusion have either accounted for the differences between rigid and nonrigid cases, but not for the effect of satellites (Hildreth, 1983), or for the effect of satellites but not of background motion (Grzywacz and Yuille, 1991). Multiple depth layers in combination with a smoothness constraint helped Weiss et al. (2000) to explain a rigidity percept as a function of the aspect ratio, the effect of satellites and the effect of a background motion. That work, however, did not suggest a neural implementation. Our model suggests specific mechanisms: depth-specific boundary selection of motion, together with motion integration and motion transparency mechanisms, allows it to address all variations of the ellipse display.

A number of more recent models of vision employ Bayesian techniques. One that is particularly relevant to this work is that of Weiss, Simoncelli and Adelson (2002). A traditional intersection of constraints approach is enhanced by introducing an individual’s decision uncertainty and priors into the process of motion computation. The 3D FORMOTION model can be viewed as the brain’s way of using normalized patterns of form and motion activities as “real-time probabilities” that work together to contextually overcome uncertainty.

Mechanisms for Interaction of Form and Motion streams. One of the most important components of the 3D FORMOTION model is interaction between form and motion processing. Form and motion processing streams in the visual cortex are traditionally considered as separate from each other (Mishkin, Ungerleider and Macko, 1983). Separation starts at the retinal level. Lesion data seem to support the separation idea: Lesions of the parvocellular, or P-pathway, do not affect performance in pure motion tasks; lesions of the magnocellular, or M-pathway, do not affect color or fine spatial frequency sensitivity (Schiller and Logothetis, 1990). However, when more complicated scenes are considered, independence of the two pathways is questionable, as in the Lorenceau-Alais, chopsticks, and gelatinous ellipse displays.

Motion signals can change based on the occlusion information present in the display. For example, the difference between the motion of extrinsic and intrinsic terminators explains chopstick displays and some of the Lorenceau-Alais displays. Previously, it was suggested that FACADE figure-ground separation would provide a basis for such a distinction. However, separation of boundaries in depth does not happen until V2, or at least the upper layers of V1. Here we suggest that some difference between extrinsic and intrinsic terminators can already be detected in the input layers of V1, and it is established in part by adaptation to static boundaries. Electrophysiological recordings of V1 cell activity in response to a similar display, the diagonal grating with and without horizontal occluders (Pack et al., 2004), can be interpreted as a support
to the adaptation hypothesis. These data are also consistent with properties of the model feedback from the V2 figure-ground separation mechanisms to the V1 motion stream. Because these authors did not study temporal dynamics of the suppression along occluders, or vary other parameters affecting depth order of the grating and horizontal occluders, based on their data it is hard to distinguish between feedforward and feedback mechanisms.

The projection from V1 to MT is unlikely to carry depth-selective signals (Movshon and Newsome, 1996). How, then, is the motion of two overlapping objects separated in MT? Palanca and DeAngelis (2003) have shown that cells in MT have disparity tuning even in the absence of motion. V2 cells appear to participate in figure-ground separation (Bakin, Nakayama and Gilbert, 2000, Zhou, Friedman and von der Heydt, 2000). The 3D FORMOTION model predicts that the V2 thick stripe projection to MT can carry occlusion information necessary to resolve the motion of different surfaces in depth. Such an excitatory-center inhibitory-surround projection of depth-separated boundaries from V2 to the motion stream can also help to explain the absence of motion in the near depth of chopsticks (or any other) display with visible occluders. Occluder boundaries represented in the near depth plane would select relatively weak “extrinsic” motion signals along them and suppress motion signals anywhere else at that depth. This mechanism predicts that a proportion of cells in MT representing closer depths will be suppressed when occluder boundaries are presented. While neuronal recordings where either disparity-defined (Duncan et al., 2000) or contrast-defined (Pack et al., 2004) occluders were presented do not offer such evidence, protocols used in these studies did not include a control case of motion presented without occluders. Because only motion-sensitive cells are usually selected for recordings, the cell populations that are suppressed by form boundaries would be easy to overlook.

Other Form-Motion interaction phenomena can be explained by feedback projections between cortical areas. Different motion signals coexisting in the image can create a motion-defined boundary (separation in 2D plane) or two motion planes (separation in depth). This suggests that projections from the motion system go to the form boundary/surface processing system. Such a projection from MT to V1 was used in the model to explain the perceived separation of chopsticks in depth in the invisible occluder case. Studies of the function of the MT-to-V1 projection (Movshon, and Newsome, 1996, Jones et al., 2001) used either microstimulation or microinjection techniques in the context of the simple local motion displays. The effect of the feedback projection was often excitatory, sometimes inhibitory, but its overall function was not clear. We predict that it is realized by a modulatory on-center off-surround network, much like in the MST-to-MT feedback pathway, and within the form processing stream (Grossberg, 1999; Raizada and Grossberg, 2003).

Projections from the motion to the form stream can also distort boundaries of objects under certain conditions, as in the case of gelatinous ellipses. In this article we show only the result of computations in the motion stream: a tangential motion in the case of the rigid ellipse, and radial motion in the case of non-rigid ellipse. Tangential and radial biases are consistent with rotation or deformation, respectively. Further experiments are needed to test if the MT-to-V1 projection can shift perceived boundary positions in the manner consistent with deformation in the case of non-rigid ellipse.

One important difference between the form and motion systems is the difference in timing. In particular, the timing of boundary completion is sometimes slow because it may involve feedback and competition between different depth planes. There are also latency differences between parvocellular and magnocellular streams. The motion signal to MT is very
quick with a latency of 40 ms, compared to more than 50 ms in orientation-selective simple and complex cells in V1 (Bullier, 2001; Bair, Cavanaugh, Smith and Movshon, 2002). While adaptation mechanisms resulting in the intrinsic/extrinsic terminator distinction are feedforward and quick, boundary selection mechanisms require an additional stage of cortical processing and are slower. On the other hand, motion signals, even in a simple moving line display, suffer from the aperture problem. In the visible occluder case of the chopsticks display, the 3D FORMOTION model predicts that, initially, the direction of motion in both depth representations of MT corresponds to an ambiguous motion signal, and that the correct motion signal develops through time. Boundary suppression through the V2-to-MT projection is predicted to inhibit the motion signal in the near depth plane, concurrently with the development of the correct motion signal in the farther depth plane. This effect could be noticeable in the depth-modulated barberpole illusion, as in Duncan et al. (2000). These authors have not, however, done the analysis of timing of motion-sensitive cells relative to the boundary onset. Pack, Berezovskii and Born (2001) did demonstrate a switch from ambiguous to veridical direction of motion over a period of 50-70ms as detected in the response of certain MT cells to a modified barberpole illusion. The effect of suppression of motion in the corresponding depth remains to be shown.

The 3D FORMOTION model explanations are consistent with those of many other motion data by earlier versions of the model (Baloch and Grossberg, 1998; Chey et al., 1997; Francis and Grossberg, 1996b; Grossberg et al., 2001). The same mechanisms can be also applied to illusory boundaries from motion (Anderson and Barth, 2000), aperture discontinuity (Palmer and Kellman, 2001), flash lag and flash-drag effects (Nijhawan, 1994; Whitney and Cavanagh, 2000), and motion induction/motion capture effects (Murakami 1999). Some of these issues are addressed in a follow-up of the current model (Berzhanskaya et al., 2004).
Appendix: 3D FORMOTION equations, parameters and implementation

All stages of the model (except simple cells in the form system (Equation (A15))) were numerically integrated using a 4th order Runge-Kutta method with a fixed integration step. The activity of simple cells was computed at equilibrium. Each layer, including the input, was represented by a 60x60 matrix for each combination of attributes used at the given layer. For example, for the motion system, if there were 2 spatial scales, there were (2x8) cells sensitive to different combinations of scale and direction at each point of the matrix. For the scale-sensitive form system cells, there were (2x4) scale and orientation cells at each point in the image.

I. Motion system

All motion sequences are given to the network as series of static 2D frames representing black-and-white image snapshots at the consecutive moments of time. In both form and motion systems, inputs are not separated in depth; i.e., both occluder and occluded objects exist in the same image plane. Activities at each layer \( y_i \) are results of computation in a dynamical system, where the rate of activity change is proportional to some function \( f \) of this layer’s activities, inputs \( I \) and, sometimes, feedback \( F \). Dynamics can be described in a general form as:

\[
\frac{dy_n}{dt} = A_n f(y_n, I, F),
\]

where \( A_n \) scales how fast \( y_n \) changes. High values of \( A_n \) result in fast dynamics, while low values of \( A_n \) result in slow dynamics. Outputs of all stages are rectified: \( Y_n = \left[ y_n \right] = \max(y_n, 0) \).

All model equations are membrane equations:

\[
C_m \frac{dV}{dt} = -[V - E_{\text{excit}}]g_{\text{excit}} - [V - E_{\text{inhib}}]g_{\text{inhib}} - [V - E_{\text{leak}}]g_{\text{leak}}
\]

(A2)

In this equation, \( g_{\text{excit}} \) and \( g_{\text{inhib}} \) represent the total inputs from excitatory and inhibitory neurons synapsing on the cell; \( g_{\text{leak}} \) is a leakage conductance. Parameters \( E_{\text{excit}}, E_{\text{inhib}}, \) and \( E_{\text{leak}} \) are reversal potentials for excitatory, inhibitory and leakage conductances, respectively. All conductances contribute to the divisive normalization of the membrane potential, \( V \), as shown by equilibrium solution for \( V \):

\[
V = \frac{(E_{\text{excit}}g_{\text{excit}} + E_{\text{inhib}}g_{\text{inhib}} + E_{\text{leak}}g_{\text{leak}})}{(g_{\text{excit}} + g_{\text{inhib}} + g_{\text{leak}})}
\]

(A3)

(Grossberg, 1973, 1980; Grossberg and Raizada, 2000). Reversal potentials in the following simulations were (for simplicity) set to \( E_{\text{excit}} = 1, E_{\text{inhib}} = -1, \) and \( E_{\text{leak}} = 0 \) (unless noted otherwise). When the reversal potential of the inhibitory channel, \( E_{\text{inhib}}, \) is close to the resting potential, the inhibitory effect is pure “shunting”; i.e., decreasing effect of excitation only through an increased membrane conductance. It balances excitatory inputs and prevents network activities from saturating. In the equations where saturation effects are not possible (for example (A9)), the shunting term was not used.

Depending on a layer’s functionality, activities at each position \((i,j)\) are represented as \( x_{ij}^p \), where \( p \in \{1,2\} \) indicates whether the cell (population) belongs to an ON or OFF stream; or as \( x_{ij}^d \), where \( d \in \{1,..,8\} \) indicates directional preference within a single spatial scale; or else as \( x_{ij}^{ds} \), where \( d \in \{1,..,8\} \) indicates motion directional preference, and \( s \in \{1,2\} \) indicates spatial scale.

**Level 1: Input.** Motion processing starts from the input layer of V1 \((4C_a)\). Previous models (Baloch et al., 1997) analyzed how LGN ON and OFF streams interact to create
boundaries from a 2D image. They demonstrated that in static images ON-center /OFF-surround and OFF-center/ON-surround mechanisms would create thin boundaries on the edges of the object. In moving images boundaries at the leading edge of the bright bar are represented mainly by ON stream, while boundaries at the trailing edge by mainly OFF stream. Based on the results of Baloch et al. (1997) a simplified input \( I_{ij}^p \) to the visual cortex was represented by 2-cell wide boundaries in two separated ON and OFF channels. This simplification was motivated by the fact that we used simple black-and-white images. The boundary on the leading edge of the object was represented by the ON channel, the boundary on the trailing edge by the OFF channel. No interaction between ON and OFF channels was simulated.

**Level 2: Transient cells.** At the first stage of V1, non-directional transient cell activities \( b_{ij} \) are computed as a sum of ON (\( p = 1 \)) and OFF (\( p = 2 \)) channels:

\[
b_{ij} = \sum_p x_{ij}^p z_{ij}^p , \tag{A4}
\]

where input cell activities, \( x_{ij}^p \), perform leaky integration on their inputs \( I_{ij}^p \):

\[
\frac{dx_{ij}^p}{dt} = A_1 (-B_1 x_{ij}^p + (C_1 - x_{ij}^p) z_{ij}^p) . \tag{A5}
\]

Non-zero activation \( x_{ij}^p \) results in slow adaptation of a habituative transmitter gate \( z_{ij}^p \):

\[
\frac{dz_{ij}^p}{dt} = A_2 \left( -z_{ij}^p - K_2 x_{ij}^p z_{ij}^p \right) . \tag{A6}
\]

(Abbott et al., 1997; Grossberg, 1980). In (A5), \( A_1 B_1 x_{ij}^p \) is the rate of passive decay and \( C_1 \) is a maximum activity \( x_{ij}^p \) can reach. For non-zero inputs \( I_{ij}^p \), \( x_{ij}^p \) approaches \( C_1 \) with a rate proportional to \( (C_1 - x_{ij}^p) \). When a nonzero input \( x_{ij}^p \) is presented, \( z_{ij}^p \) adapts with the rate of \( A_2 K_2 x_{ij}^p \) in (A6). When the input returns to 0, \( z_{ij}^p \) recovers to 1 at the rate \( A_2 \). The parameters used in Level 2 simulations are: \( A_1 = 10, B_1 = 3, C_1 = 1, A_2 = 0.01, \) and \( K_2 = 20 \).

Input activity \( x_{ij}^p \) combined with transmitter gate \( z_{ij}^p \) results in transient non-directional cell activities \( b_{ij} \) that model activity of the non-directionally selective cells in layers 4C of with circular receptive fields (Livingstone and Hubel, 1984). ON and OFF inputs summate at this stage. For visual inputs with a short dwell time, such as moving boundaries, activities \( b_{ij} \) respond well. A static input, on the other hand, produces only a weak response after an initial presentation period, because of the habituation (Muller, Metha, Krauskopf, and Lennie, 2001).

The next two cell layers provide a directional selectivity mechanism that can retain its sensitivity in response to variable speed inputs (Chey et al., 1997). As noted above, index \( d \) denotes the directional preference of a given cell. First, directional interneuron activities \( c_{ij}^d \) integrate transient cell inputs \( b_{ij} \):

\[
\frac{dc_{ij}^d}{dt} = A_3 \left( -B_3 c_{ij}^d + C_3 b_{ij} - K_3 \int \frac{dK_{xy}^D}{dx} \right) . \tag{A7}
\]

A directional inhibitory interneuron \( c_{ij}^d \) receives excitatory input from a transient non-directional cell activity \( b_{ij} \) at the same position, and suppression from directional interneuron \( c_{xy}^D \) of
opposite direction preference $D$ at the position $(X,Y)$ offset by 1 cell in the direction $d$. For example, for the direction of motion $45^\circ$, $X = i+1$, $Y = j+1$, and $D = 135^\circ$.

Activity $c_{ij}^d$ increases proportionally to input $b_{ij}$ with coefficient $A_3C_3$ and passively decays to zero with rate $A_3B_3c_{ij}^d$. The strength of opponent inhibition is $K_3c_{ij}^D$. Inhibition is stronger than excitation and “vetoes” a directional signal if the stimulus arrives from the null direction. Thus, a bar arriving from the left into the rightward directional interneuron receptive field would activate it well; while a bar arriving from the right would inhibit it even if activation $b_{ij}$ is non-zero. The parameters are: $A_3 = 5, B_3 = 2, C_3 = 0.5,$ and $K_3 = 20$.

Directional transient cell activities $e_{ij}^d$ at the next level combine transient input $b_{ij}$ with inhibitory interneuron activity $c_{ij}^d$. Their dynamics are similar to those of $c_{ij}^d$:

$$\frac{de_{ij}^d}{dt} = A_4 \left( B_4e_{ij}^d + C_4b_{ij} - K_4c_{ij}^D \right)$$

(A8)

Activity $e_{ij}^d$ increases proportionally to transient input $b_{ij}$, passively decays with the fixed rate, and is inhibited by an inhibitory interneuron tuned to the opponent direction. The parameters are: $A_4 = 30, B_4 = 1, C_4 = 0.5,$ and $K_4 = 20$. Computation at Level 2 results in multiple directions activated in response to a moving line, which is consistent with the ambiguity caused by the aperture problem due to the limited size of V1 receptive fields.

**Level 3: Short-range motion filter.** Short-range anisotropic filter activities, $f_{ij}^{ds}$, accumulate motion in each direction $d$:

$$\frac{df_{ij}^{ds}}{dt} = A_5 \left( -f_{ij}^{ds} + \sum_{XY} E_{ij}^{ds} G_{ijXY}^{ds} \right).$$

(A9)

Here $E_{ij}^{ds}$ is the rectified output of $e_{ij}^{ds}$ from Level 2, and $G_{ijXY}^{ds}$ is a Gaussian receptive field that depends on both direction $d$ and scale $s$:

$$G_{ijXY}^{ds} = G \exp \left( -0.5 \left( \frac{X-i}{\sigma_x^s} \right)^2 + \left( \frac{Y-j}{\sigma_y^s} \right)^2 \right).$$

(A9a)

Scale $s$ determines a receptive field size, and therefore the extent of spatiotemporal integration of lower-level motion signals. Larger receptive fields respond selectively to larger speeds, smaller receptive fields to smaller speeds; cf., Chey et al. (1998). While in our simulations speed did not vary much, in more motion-rich environments speed-depth correlations can help to assign an approximate depth order to the moving objects. The kernel $G_{ijXY}^{ds}$ is elongated in the direction of motion. For a horizontal motion direction, the kernel has $\sigma_x^s = 1.5, \sigma_y^s = 0.5$ for $s = 1, \sigma_x^s = 2.5, \sigma_y^s = 0.5$ for $s = 2; G = 0.15$. Kernels for other directions are derived by a rotation which aligns the major kernel axis with the preferred direction of motion. Output of the short-range filter is thresholded and rectified, $F_{ij}^{ds} = \left[ f_{ij}^{ds} - \theta_s \right]$, with threshold $\theta_1 = 0.04, \theta_2 = 0.08$. Self-similar scale-specific thresholds provide different speed sensitivity for two spatial scales. If thresholds for two scales were the same, the larger scale would be always activated more strongly. With the larger threshold it prefers larger speeds. The full simulation of speed sensitivity was performed in a similar system by Chey et al. (1997). The value of constant $A_5 = 50$. 

19
Level 4: Spatial competition and opponent direction inhibition. The next cell layer activities, \( h_{ij}^{ds} \), combine spatial competition within one motion direction across the area determined by the kernel \( K_{ijXY}^{ds} \) with inhibition from opponent direction cells \( F_{ij}^{Ds} \) in the same spatial position. A membrane, or shunting, equation combines these effects:

\[
\frac{dh_{ij}^{ds}}{dt} = A_0 \left( -h_{ij}^{ds} + (1 - h_{ij}^{ds}) \sum_{x} F_{x}^{ds} J_{ijXY}^{ds} - 0.1 + h_{ij}^{ds} \left\{ C_0 \sum_{x} F_{x}^{ds} K_{ijXY}^{ds} + D_0 F_{ij}^{Ds} \right\} \right). \tag{A10}
\]

Rectified activities, \( F_{ij}^{ds} \), from Level 3 define the spatial competition through the excitatory Gaussian kernel \( J_{ijXY}^{ds} \), which is spatially anisotropic with \( \sigma_x = 2.5 \) and \( \sigma_y = 0.5 \) (for horizontal motion):

\[
J_{ijXY}^{ds} = \frac{J}{2\pi \sigma_x \sigma_y} \exp \left\{ -0.5 \left( \frac{(X - i)^2}{\sigma_x^2} + \frac{(Y - j)^2}{\sigma_y^2} \right) \right\}, \tag{A10a}
\]

and the inhibitory kernel \( K_{ijXY}^{ds} \), which is isotropic with \( \sigma = 4 \):

\[
K_{ijXY}^{ds} = \frac{K}{2\pi \sigma^2} \exp \left\{ -0.5 \left( \frac{(X - i)^2 + (Y - j)^2}{\sigma^2} \right) \right\}. \tag{A10b}
\]

The center of inhibitory kernel \( K_{ijXY}^{ds} \) is offset from the \((i,j)\) position by one cell in the direction opposite to the cell preferred direction \( d \). This arrangement results in inhibition trailing excitation. The strength of spatial competition is determined by parameter \( C_6 \), and that of opponent inhibition by \( D_6 \). Parameters are: \( A_0 = 50 \), \( C_6 = 5 \), \( D_6 = 100 \), \( J = 2 \), and \( K = 2 \). \( D_6 \) is opposite to \( d \).

It usually takes few frames of motion to accumulate and accurately compute motion signals through the Level 2-4 mechanisms (Equations (A4)-(A10)). However, a motion span (maximal displacement in one direction) of the Lorenceau-Alais displays is small. The radius of rotation and the motion span there are limited by the geometry of the input; in particular, corners of the shape that provide unambiguous motion signals are not visible. To accumulate enough information for the motion mechanisms to adequately sample the moving stimulus, one may increase the size of the network by supersampling and scale the motion sequence correspondingly. For example, a 3-pixel sequence of motion in one direction becomes a 9-pixel sequence (scaling by a factor of three). In order to keep the simulation times reasonable, this scaling was done only up to Level 4 (see Figure 15, layers 4C- 4B, and Equations (A4)-(A10)). Furthermore, due to memory restrictions, displays were computed piece-wise: four segments of each shape were processed by a 60 x 60 network each. Output activities at Level 4 were then subsampled by a factor of 3, in order to compensate for the previous supersampling, and combined into one 60 x 60 display at Level 5 (Equation (A11)). A supersampled 9-pixel motion sequence thus becomes a subsampled 3-pixel sequence, thereby returning to the original cellular dimensions, but motion signals are more thoroughly processed due to the finer scale at the input levels. Piece-wise simplification was possible because four segments of an individual Lorenceau-Alais shape are separated in space and do not interact with each other at the spatial scale of Level 2-4 computations. When interactions between segments become essential (Level 5 and later), activities are combined. Computations for Lorenceau-Alais used the same parameters as for other displays.
**Level 5: Formotion capture and long-range filter.** Rectified motion output signals, $H_{ij}^{ds}$, from V1 (model Level 4) are selected by form boundary signals, $\bar{Z}_{ij}^s$, from V2 in the input layers 4 and 6 of MT. The activities, $q_{ij}^{ds}$, of these MT cells combine motion and boundary signals via a membrane equation:

\[
\frac{dq_{ij}^{ds}}{dt} = A_7 \left( -q_{ij}^{ds} + (I - q_{ij}^{ds}) H_{ij}^{ds}(K_e + K_z \bar{Z}_{ij}^s) - K_b \left( + q_{ij}^{ds} \sum_{XY} \bar{Z}_{ijXY}^s I_{ijXY}^{ds} \right) \right). \tag{A11}
\]

In (A11), an input from the V1 motion stream $K_e H_{ij}^{ds}$ is positively modulated by boundaries $K_z \bar{Z}_{ij}^s$ in the excitatory term of the equation (A11). In addition, boundaries inhibit unmatched motion signals via term $\sum_{XY} \bar{Z}_{ijXY}^s I_{ijXY}^{ds}$. This modulatory on-center off-surround network allows boundaries to select motion signals at their positions and corresponding depths. Parameter $K_e$ determines the strength of feedforward inputs $H_{ij}^{ds}$, and $K_z$ the strength of modulation by V2 boundaries. The V2 boundary projection to MT is stronger than the bottom-up motion projection; that is, $K_e << K_z$. The strength of the inhibitory effect of V2 boundaries $\bar{Z}_{ij}^s$ is proportional to the coefficient $K_b$, and its spatial reach is determined by inhibitory Gaussian kernel $I_{ijXY}^{ds}$:

\[
I_{ijXY}^{ds} = \frac{I}{2\pi\sigma^2} \exp \left( -0.5 \left( \frac{(X - i)^2 + (Y - j)^2}{\sigma^2} \right) \right). \tag{A11a}
\]

When no boundary is provided and $\bar{Z}_{ij}^s$ is 0 everywhere (for example, the parvocellular stream is inactivated), motion signals can still activate MT via the term $K_e H_{ij}^{ds}$ in (A11). In this case, no inhibition is present as well. In the presence of boundary input, motion signals at the boundary positions are strong, whereas those outside of the boundary position are suppressed. Activity $\bar{Z}_{ij}^s$ in (A11) codes a 1-cell wide boundary simulating output of V2. It simplifies boundaries $\bar{Z}_{ij}^s$ separated in depth by the form system (Equation (A23)), positioned at the locations of input boundaries $I_{ij}$ (Equation (A4)). Parameter $s = 1$ corresponds to the near depth, $s = 2$, to the far depth. The parameters are: $A_7 = 100$, $K_e = 1$, $K_z = 10$, $K_b = 0.12$, $I = 0.1$, and $\sigma = 6$.

Next, MT cell activities, $m_{ij}^{ds}$, in layer 2/3 receive MT signals, $N_{ij}^{ds}$, from layer 4 via a long-range filter and top-down matching signals, $T_{ij}^{ds}$, from MST:

\[
\frac{dm_{ij}^{ds}}{dt} = A_8 \left( -m_{ij}^{ds} + (I - m_{ij}^{ds}) N_{ij}^{ds} - D_s \left( I + m_{ij}^{ds} \sum_{c_{XY}} w_{c_{XY}} T_{c_{XY}}^{ds} P_{c_{XY}}^{ds} \right) \right). \tag{A12}
\]

To compute the long-range filter inputs, $N_{ij}^{ds}$, the MT input activities, $q_{ij}^{ds}$, are rectified ($Q_{ij}^{ds}$), squared, and passed through an anisotropic filter $L_{ijXY}^{ds}$, thresholded, and rectified again:

\[
N_{ij}^{ds} = \left[ \sum_{XY} (Q_{ijXY}^{ds} L_{ijXY}^{ds} - \theta_n)^+ \right]. \tag{A13}
\]

In (A12), $L_{ijXY}^{ds}$ is an anisotropic Gaussian kernel.
\[ L_{ds}^{xy} = \frac{L}{2\pi \sigma_x \sigma_y} \exp \left( -0.5 \left( \frac{X-x}{\sigma_x} \right)^2 + \left( \frac{Y-y}{\sigma_y} \right)^2 \right) \]  

(A13a)

that is elongated in the direction of preferred motion. For the horizontal motion direction, for example, \( \sigma_x = 10, \sigma_y = 4 \). \( L = 20, \) and \( \theta = 0.03 \). Kernels for other directions are derived by rotation, for example to compute 45° kernel, the horizontal kernel is rotated 45° counterclockwise.

This long-range anisotropic Gaussian filter accumulates motion in its preferred direction over time and space. The anatomical basis for such integration can be provided by long-range horizontal projections in layers 2/3 of MT. The squaring operation gives higher preference to larger signals, which leads to winner-take-all dynamics in competitive recurrent networks (Grossberg, 1973, 1988).

The strength of MST feedback is determined by coefficient \( D_\delta \). Its spatial extent is determined by the isotropic kernel \( P_{ds}^{xy} \):

\[ P_{ds}^{xy} = \frac{1}{2\pi \sigma^2} \exp \left( -0.5 \left( \frac{(X-x)^2 + (Y-y)^2}{\sigma^2} \right) \right) \]  

(A13b)

with \( \sigma = 8 \). The suppression is from all directions except \( d \). The inhibitory weight \( w^{de} \) between given direction \( d \) and another direction \( e \) is given by:

\[ w^{de} = \begin{cases} 
0, & e = d \\
1, & e \neq d, e \neq D \\
2, & e = D
\end{cases} \]  

(A13c)

where \( D \) is the direction opposite to \( d \). Because excitatory input \( N_{ij}^{ds} \) is from the preferred direction, this asymmetric suppression effectively amplifies \( d \) and suppresses other motion directions. Motion from unambiguous feature-tracking signals propagates to ambiguous motion positions through the large kernel \( P_{ds}^{xy} \). As in the case of the V2-to-MT and MT-to-V1 projections, MST-to-MT feedback is defined by a modulatory on-center off-surround network. The parameters are: \( A_s = 200 \) and \( D_s = 5 \).

**Level 6: Directional grouping and suppression in depth.** The MT-MST circuit acts in a winner-take-all mode, selecting a single direction of motion at each point. MST activity \( T_{ij}^{ds} \) is described by

\[ \frac{dT_{ij}^{ds}}{dt} = A_s \left[ -T_{ij}^{ds} + (1 - T_{ij}^{ds}) M_{ij}^{ds} \left[ (1 + O_{ij}^{ds}) - D_d \sum_{D,XY} w^{ddT_{XY}^{ds} P_{ds}^{xy}} - C_s \sum_{s \neq S} T_{ij}^{ds} \right] \right] \]  

(A14)

where \( M_{ij}^{ds} \) is the rectified MT output. Inhibition is provided by recurrent connections within MST. Its strength is determined by coefficient \( D_\nu \), and its spatial extent by the kernel \( P_{ds}^{xy} \). The weighting coefficient \( w^{de} \) and surround suppression kernel \( P_{ds}^{xy} \) are the same as in equation (A13a). MST also includes direction-specific suppression from the near depth (D1, \( s=1 \)) to the far depth (D2, \( s=2 \)), which is important for motion transparency simulations. If the motion in the direction \( d \) wins in D1, the same direction will be suppressed in D2. This allows the model to avoid a single motion direction being represented in both depths. In the case of transparent
motion, suppression of one direction in a given depth would allow the other direction to win. The parameters are: \(A^f = 400\), \(C^f = 1\), and \(D^f = 10\).

MST can be modulated by attention via term \(O_{ij}^{ds}\) in equation (A14). If attention is attracted by features in the near depth plane, this would help one motion direction to win in the near depth. Attention was used only in chopsticks simulations with invisible occluders to break the symmetry between competing motion signals from two chopsticks moving in opposite directions. Attention was applied as a single Gaussian “spot” in the near depth \((s=1)\) and rightward direction \((d=5)\):

\[
O_{ij}^{ds} = A \exp \left( -0.5 \left( \frac{(x_i - j)^2 + (y_i - j)^2}{\sigma^2} \right) \right). \tag{A14a}
\]

In (A14a), \(\sigma = 2\), \(A = 0.05\), and \((x_i, y_i)\) are the spatial coordinates of the top-left moving chopstick end. This bias is similar to the one used in the case of transparent motion in Grossberg et al. (2001) and allows a single motion signal to win in D1.

II. Form system
A reduced version of the FACADE model was implemented as the form system in order to keep simulations manageable. Only part of the Boundary Contour System (BCS) was simulated. Binocular inputs were not considered. Filling-in was not simulated. After completion of an object boundaries, it was assumed that filling-in processes and boundary competition in a full FACADE implementation would complete separation of occluded and occluding objects in different depths. For more complete descriptions of FACADE mechanisms, see Grossberg and Howe (2003), Grossberg and Swaminathan (2004), and Kelly and Grossberg (2000).

Depending on a layer’s functionality, activities at each position \((i,j)\) are represented as \(x_{ij}^p\), where \(p\) indicates either whether a cell population belongs to ON or OFF streams, or whether it is an odd/even filter; or as \(x_{ij}^r\), where \(r \in \{1,2,3,4\}\) indicates orientational preference, and \(s \in \{1,2\}\) indicates the spatial scale or depth plane.

Level 1: Input. Input to V1, \(X_{ij}^p\), corresponds to the input processing by LGN through circular center-surround receptive fields. As in the motion system, a simplified input \(X_{ij}^{ON/OFF}\) was represented by 1-cell wide boundaries in two distinct ON and OFF channels. This simplification was motivated by our use of simple black-and-white images. No interaction between ON and OFF channels was simulated.

Level 2: Simple cells. Model simple cells respond to oriented contrasts in the image in a polarity-sensitive manner. Simple cell activities for orientation \(d\) and spatial scale \(s\), \(S_{ij}^{rs}\), are computed by convolution of V1 input \(X_{ij}^{ON/OFF}\) with even-symmetric and odd-symmetric oriented filters \(s_{ijpq}^{rs,even}\) and \(s_{ijpq}^{rs,odd}\), respectively:

\[
S_{ij}^{rs,ON/OFF,even} = \sum_{pq} s_{ijpq}^{rs,even} X_{pq}^{ON/OFF} - \sum_{pq} s_{ijpq}^{rs,even} X_{pq}^{OFF/ON}, \tag{A15a}
\]

\[
S_{ij}^{rs,ON/OFF,odd} = \sum_{pq} s_{ijpq}^{rs,odd} X_{pq}^{ON/OFF} - \sum_{pq} s_{ijpq}^{rs,odd} X_{pq}^{OFF/ON}. \tag{A15b}
\]
where odd and even filters are given by

\[ s_{ijpq}^{rs, odd} = \exp \left[ -0.5 \left( \frac{(i-p)^2}{\sigma_x^2} + \frac{(j-q)^2}{\sigma_y^2} \right) \right] - \exp \left[ -0.5 \left( \frac{(i-p-h)^2}{\sigma_x^2} + \frac{(j-q)^2}{\sigma_y^2} \right) \right] \]  

(A15c)

and

\[ s_{ijpq}^{rs, even} = k_1 \exp \left[ -0.5 \left( \frac{(i-p)^2}{\sigma_x^2} + \frac{(j-q)^2}{\sigma_y^2} \right) \right] - k_2 \exp \left[ -0.5 \left( \frac{(i-p-h)^2}{\sigma_x^2} + \frac{(j-q)^2}{\sigma_y^2} \right) \right] - k_2 \exp \left[ -0.5 \left( \frac{(i-p+h)^2}{\sigma_x^2} + \frac{(j-q)^2}{\sigma_y^2} \right) \right]. \]  

(A15d)

Four orientations \((r)\) and two scales \((s)\) were used. For the vertically oriented filter the parameters are: \(\sigma_x^{r1} = \sigma_z^{r2} = 0.75\), \(\sigma_x^{r1} = 1\), \(\sigma_y^{r2} = 2.5\); \(h = 1\), \(k_f = 1.6\), and \(k_2 = 0.8\).

**Level 3: Complex cells.** Complex cells pool inputs from simple cells of the same orientation and opposite polarity. Complex cell activities \(c_{ij}^{rs}\) combine odd and even, and ON and OFF, inputs from simple cells:

\[ G_{ij}^{rs} = \left| S_{ij}^{rs, even, ON} - S_{ij}^{rs, even, OFF} \right| + \left| S_{ij}^{rs, odd, ON} - S_{ij}^{rs, odd, OFF} \right|. \]  

(A16)

The activity of the complex cells is computed as:

\[ \frac{dc_{ij}^{rs}}{dt} = -A_2 \left( c_{ij}^{rs} + B_2 (1 - c_{ij}^{rs}) G_{ij}^{rs} (1 + k_{ex} M_{ij}^s + k_{v2} \tilde{z}_{ij}^s) - D_2 (1 + c_{ij}^{rs}) \sum_{s} c_{ij}^{rs} \right). \]  

(A17)

The last term introduces competition between boundaries of the same orientation \(r\), at each point \((i,j)\), across scales \(S\) and \(s\), thus allowing a given orientation boundary to be represented in only one depth.

Both V2 and MT modulate the response of V1 complex cells to the simple cell inputs, \(G_{ij}^{rs}\), via the terms that multiply, and thus modulate, \(G_{ij}^{rs}\) in equation (A17). Thus, the model predicts that MT feedback \(M_{ij}^s\) (activity computed in (A13)) to layer 2/3 via apical dendrites in layer 1 of V1 (Callaway, 1998; Shipp and Zeki, 1989) can affect not only the motion system but also complex cells in the form system (second term in (A17)). This feedback is scale and depth specific, but not orientation or direction specific, and provides excitatory modulation only:

\[ M_{ij}^s = \sum_u M_{ij}^{ds} - \theta_{MT}. \]  

(A17b)

MT-to-V1 feedback may also modulate layer 4 or V1, but this projection would have no effect on simulation of the targeted data.

The second feedback component of the second term in (A17) is provided by a 1-cell wide approximation \(\tilde{z}_{ij}^{rs}\) of V2 boundaries \(z_{ij}^{rs}\) (see equations (A23)-A(24)). \(\tilde{z}_{ij}^{rs}\) positions correspond to the boundary positions at the input level \(I_{ij}\), and their value is 1 on a background of 0's. These boundaries provide a depth-specific bias to V1 complex cells, so if FACADE mechanisms assign a certain boundary to depth 1 at the level of V2, this boundary will be strengthened in scale 1 at the complex cell layer in V1, and will be weakened in scale 2. This mechanism helps to ensure
that a given boundary is represented in one depth only. Parameters are: $A_2 = 2$, $B_2 = 1$, $k_{ex} = 25$, $\theta_{MT} = 0.15$, and $D_2 = 10$.

**Level 4: Hypercomplex cells.** The hypercomplex cell level has both spatial, $y_{ij}^{rs}$, and orientational, $n_{ij}^{rs}$, competition stages. This level models the process of end-stopping. It combines feedforward inputs from complex cells, $C_{pq}^{rs}$, through the on-center off-surround terms $C_4$ and $E_4$, respectively, with feedback inputs from bipole cells through the on-center off-surround terms $C_7$ and $E_7$, respectively. The activity, $y_{ij}^{rs}$, at the spatial competition stage is described by:

\[
\frac{dy_{ij}^{rs}}{dt} = -A_3 \left( 0.1y_{ij}^{rs} + (1 - y_{ij}^{rs}) \left[ C_{4ij}^{rs} + C_{7ij}^{rs} \right] - (y_{ij}^{rs} + 1) \left[ E_{4ij}^{rs} + E_{7ij}^{rs} \right] \right)
\]  
(A18)

The on-center, $C_{4ij}^{rs}$, and the off-surround, $E_{4ij}^{rs}$, inputs from complex cells obey:

\[
C_{4ij}^{rs} = \frac{C}{2\pi\sigma_c^2} \sum_{pq} \exp \left[ -0.5 \left( \frac{(i-p)^2 + (j-q)^2}{\sigma_c^2} \right) \right] C_{pq}^{rs},
\]
(A19a)

and

\[
E_{4ij}^{rs} = \frac{E}{2\pi\sigma_s^2} \sum_{pq} \exp \left[ -0.5 \left( \frac{(i-p)^2 + (j-q)^2}{\sigma_s^2} \right) \right] C_{pq}^{rs},
\]
(A19b)

where $C_{pq}^{rs} = \left[ C_{pq}^{rs} \right]$ is rectified input from the complex cell in (A17). The on-center feedback from the bipole stage is provided by one-to-one projections:

\[
C_{7ij}^{rs} = k_{ex} Z_{ij}^{rs},
\]
(A20a)

where $Z_{ij}^{rs} = \left[ Z_{ij}^{rs} \right]$ is the bipole output from the corresponding orientation $r$ and scale $s$ (equation (A23)). The off-surround feedback from the bipole stage is given by:

\[
E_{7ij}^{rs} = k_{inh} \sum_r \sum_{pq} \frac{E}{2\pi\sigma_s^2} \exp \left[ -0.5 \left( \frac{(i-p)^2 + (j-q)^2}{\sigma_s^2} \right) \right] Z_{pq}^{rs}.
\]
(A20b)

Summation of inhibitory feedback over all orientations $r$ provides a spatial competition property, and suppresses hypercomplex cell activities, $y_{ij}^{rs}$, in the case of ambiguous boundary signals. This feedback is the part of competitive mechanism that breaks the stems of T-junctions from their tops during figure-ground separation. Parameters are: $A_3 = 50$, $C = E = 1$; $\sigma_c = 0.5$, $\sigma_s = 1$, $k_{ex} = 0.02$, and $k_{inh} = 0.2$.

The orientation competition activities, $n_{ij}^{rs}$, receive rectified inputs $Y_{ij}^{rs} = \left[ Y_{ij}^{rs} \right]$ from the spatial competition stage (A18):

\[
\frac{dn_{ij}^{rs}}{dt} = -A_4 \left( n_{ij}^{rs} + (1 - n_{ij}^{rs}) \sum_{d \in \{3, 5, 7, 9\}} C_{ij}^{dr} Y_{ij}^{ds} - (n_{ij}^{rs} + 1) \sum_{d \in \{3, 5, 7, 9\}} E_{ij}^{dr} Y_{ij}^{ds} \right).
\]
(A21)

In (A21), orientation signals at each point $(i,j)$ compete across orientations. Both the excitatory, $C_{ij}^{dr}$, and inhibitory, $E_{ij}^{dr}$, kernels have Gaussian profiles across orientations so that orthogonal orientation suppression is the strongest. The excitatory orientation competition kernel is:
\[ C^{dr} = \frac{c}{2\pi \sigma^2_c} \exp \left[ -0.5 \left( \frac{d - r}{\sigma_c} \right)^2 \right]. \]  

(A21a)

The inhibitory orientation competition kernel is:

\[ E^{dr} = \frac{s}{2\pi \sigma^2_s} \exp \left[ -0.5 \left( \frac{d - r}{\sigma_s} \right)^2 \right]. \]  

(A21b)

Parameters are: \( A_d = 100, \sigma_c = 0.5, \sigma_s = 0.75, c = 0.5, \) and \( s = 0.75. \)

**Level 5. Long-range bipole grouping.** Bipole cell activities, \( z_{ij}^{rs}, \) cooperatively group hypercomplex cell inputs aligned in the same orientation as the bipole orientation preference, and allow boundaries to complete over gaps:

\[ \frac{dz^{rs}_{ij}}{dt} = -A_d \left( z^{rs}_{ij} + H^{rs}_{ij} - (1 + z^{rs}_{ij})H^{rs}_{ij} \right), \quad R \perp r. \]  

(A22)

The bipole input is defined by:

\[ H^{rs}_{ij} = \left[ g(A^{rs}_{ij}) + g(B^{rs}_{ij}) + k_s N^{rs}_{ij} - \theta_1 \right]. \]  

(A23)

The two terms \( A^{rs}_{ij} \) and \( B^{rs}_{ij} \) sum collinear hypercomplex cell signals, \( N^{rs}_{ij}, \) from opposite sides of the bipole cell, where \( N^{rs}_{ij} = \left[ h^{rs}_{ij} \right] \) is the rectified output of the orientation competition stage (A21). The signal function:

\[ g(x) = \frac{|x - \theta_2|}{D_2 + |x - \theta_2|} \]  

(A23a)

and threshold \( \theta_1 \) are chosen so that both branches \( A^{rs}_{ij} \) and \( B^{rs}_{ij} \) must be sufficiently active to fire \( H^{rs}_{ij} \) in the absence of the direct bottom-up input \( N^{rs}_{ij}. \) The simplified bipole kernel includes only spatial pooling across the same orientation from both bipole branches:

\[ A^{rs}_{ij} = \sum_{pq} \exp \left[ -0.5 \left( \frac{(i - p)^2}{\sigma_x^2} + \frac{(j - q - h)^2}{\sigma_y^2} \right) \right] N^{rs}_{pq}, \]  

(A23b)

and

\[ B^{rs}_{ij} = \sum_{pq} \exp \left[ -0.5 \left( \frac{(i - p)^2}{\sigma_x^2} + \frac{(j - q + h)^2}{\sigma_y^2} \right) \right] N^{rs}_{pq}, \]  

(A23c)

where \( \sigma_x = 1, \sigma_y = 2, \) and offset \( h = 5. \) These kernels are for the vertically oriented bipole. Each kernel A (top branch) and B (bottom branch) is elongated along the orientation it is pooling to facilitate grouping of the corresponding orientation across boundary gaps. Other orientation kernels are derived by rotation. Parameters are \( D = 0.2, \theta_1 = 0.5, \theta_2 = 0.3, k_s = 1.5. \) The last term in (A22) introduces competition between orthogonal orientations. Parameter \( A_d = 50. \)

In a 2D image, 3D information such as occlusion is often represented by T-junctions. FACADE has proposed how T-junction detection and figure-ground separation occur without using explicit T-junction detectors. Simulation of the complete dynamics of boundary separation in depth would require large-scale simulations (Kelly and Grossberg, 2000; Grossberg and Yazdanbakhsh, 2004), and were not implemented. Instead, it was assumed that after one chopstick boundary wins in the near depth, D1, these already demonstrated FACADE interactions will
complete the boundary separation in depth. Algorithmically, a certain boundary orientation \( r \) won when the ratio, \( a \), of the total boundary activation with orientation \( r \) to the total competing boundary \( R \) activation exceeds a threshold \( T \):

\[
\alpha = \frac{\sum_y Z_{r y}^{x}}{\sum_y Z_{R y}^{x}} \geq T .
\]

This ratio was computed in the circular neighborhood of a given junction. The radius of the circle was 5 cells. For a T-junction, the ratio of perpendicular orientations \( r \) and \( R \) in the neighborhood of T-junction was computed. For a Y-junction, the ratio of competing orientations, \( r \) and \( R \), 45° apart was computed. After the criterion (A24) was met, further V2 processing of the V1 bipole signal \( z_{ij}^{x} \) (A23) was simplified by representing the corresponding boundaries by 1-cell wide boundary activities, \( z_{ij}^{x} \). These boundaries were positioned at the same locations as non-zero input boundary values \( I_{ij} \). When one boundary wins in the near depth, it suppresses the same orientation via V2-to-V1 feedback in the far depth at the complex cell stage level (A17), thus resulting in a given boundary being represented only in one depth. The V2-to-MT projection in (A11) was calculated as the sum of the bipole activations across all orientations \( Z_{y r}^{x} = \sum_{r} Z_{y}^{x} \), and then simplified with a 1-cell wide depth-separated boundary \( Z_{y}^{x} \).

References:


