A neural model of how the brain computes heading from optic flow in realistic scenes

by
N. Andrew Browning, Stephen Grossberg, and Ennio Mingolla
Department of Cognitive and Neural Systems
Center for Adaptive Systems
and
Center of Excellence for Learning in Education, Science and Technology
Boston University
677 Beacon Street
Boston, MA 02215

Cognitive Psychology, in press

CAS/CNS-TR-2008-006

Submitted: September, 2008

Accepted: June 20, 2009

Corresponding Author:
Stephen Grossberg
phone: 617-353-7858/7857
fax: 617-353-7755
email: steve@bu.edu

Abstract

Visually-based navigation is a key competence during spatial cognition. Animals avoid obstacles and approach goals in novel cluttered environments using optic flow to compute heading with respect to the environment. Most navigation models try either explain data, or to demonstrate navigational competence in real-world environments without regard to behavioral and neural substrates. The current article develops a model that does both. The ViSTARS neural model describes interactions among neurons in the primate magnocellular pathway, including V1, MT^{+}, and MST*d*. Model outputs are quantitatively similar to human heading data in response to complex natural scenes. The model estimates heading to within 1.5° in random dot or photo-realistically rendered scenes, and within 3° in video streams from driving in real-world environments. Simulated rotations of less than 1 degree per second do not affect heading estimates, but faster simulated rotation rates do, as in humans. The model is part of a larger navigational system that identifies and tracks objects while navigating in cluttered environments.

Keywords: navigation, optic flow, heading, motion, visual cortex, V1, MT, MST, neural model

Introduction

Cognition takes place in a world of moving animals and humans. This article develops a neural model that clarifies how humans and other mammals navigate in complex natural environments. In particular, humans and animals can avoid obstacles and approach goals in novel cluttered environments using visual information, notably optic flow, to compute heading, or direction of travel. Self-motion, goal position, and obstacle layout are quickly assessed at accuracies necessary for successful traversal towards a goal.

Most models try to either clarify data about navigation, or to demonstrate navigational competence without explaining behavioral data and neural mechanisms of navigation. The ViSTARS neural model was developed to do both. ViSTARS describes functionally characterized interactions among neurons in several visual areas of the primate magnocellular parthway, including regions from retina through cortical areas V1, MT^+, and MSTd. Model outputs quantitatively match estimates of self-motion, or heading, in response to complex natural scenes, with a heading accuracy within 1.5° in random dot or photo-realistically rendered scenes, and within 3° in video streams from driving in real-world environments. Simulated rotations of less than 1° per second do not affect model performance, but faster simulated rotation rates deteriorate performance, as also occurs in humans.

The model that is developed in this article is part of a larger navigational system. The current article describes model mechanisms that are capable of computing accurate estimates of heading in response to complex natural scenes. A companion article (Browning, Grossberg, & Mingolla, 2009) builds upon this foundation to develop an expanded model that is capable of steering to goals around obstacles in response to visual inputs, and of simulating psychophysical data of Fajen & Warren (2004) collected when humans carry out similar tasks. The model name, ViSTARS (Visual Steering, Tracking, And Route Selection), is used because the model builds upon the STARS navigation model of Elder, Grossberg, & Mingolla (2005, 2009). The STARS model did not have a vision front end that could directly process visual scenes. Rather, it used the more abstract representation of scene geometry of Longuet-Higgins & Prazdny (1980). A major accomplishment of the current model is to provide a front end that can directly process visual imagery as the model navigates and computes accurate estimates of heading that are used to steer around visually perceived obstacles towards goals; hence the model name ViSTARS. The front end that processes visual imagery adapts the 3D FORMOTION model, a neural network model which has been developed to explain and predict a wide range of perceptual and neurobiological data about human and primate visual motion perception (Baloch & Grossberg, 1997; Baloch, Grossberg, Mingolla, & Nogueira, 1999; Berzhanskaya, Grossberg, & Mingolla, 2007; Grossberg, Mingolla, and Viswanathan, 2001; Grossberg & Rudd, 1992) A diagram of model processing stages is shown in Figure 1.1. Full model equations, parameters and implementation details are given in the Appendix.

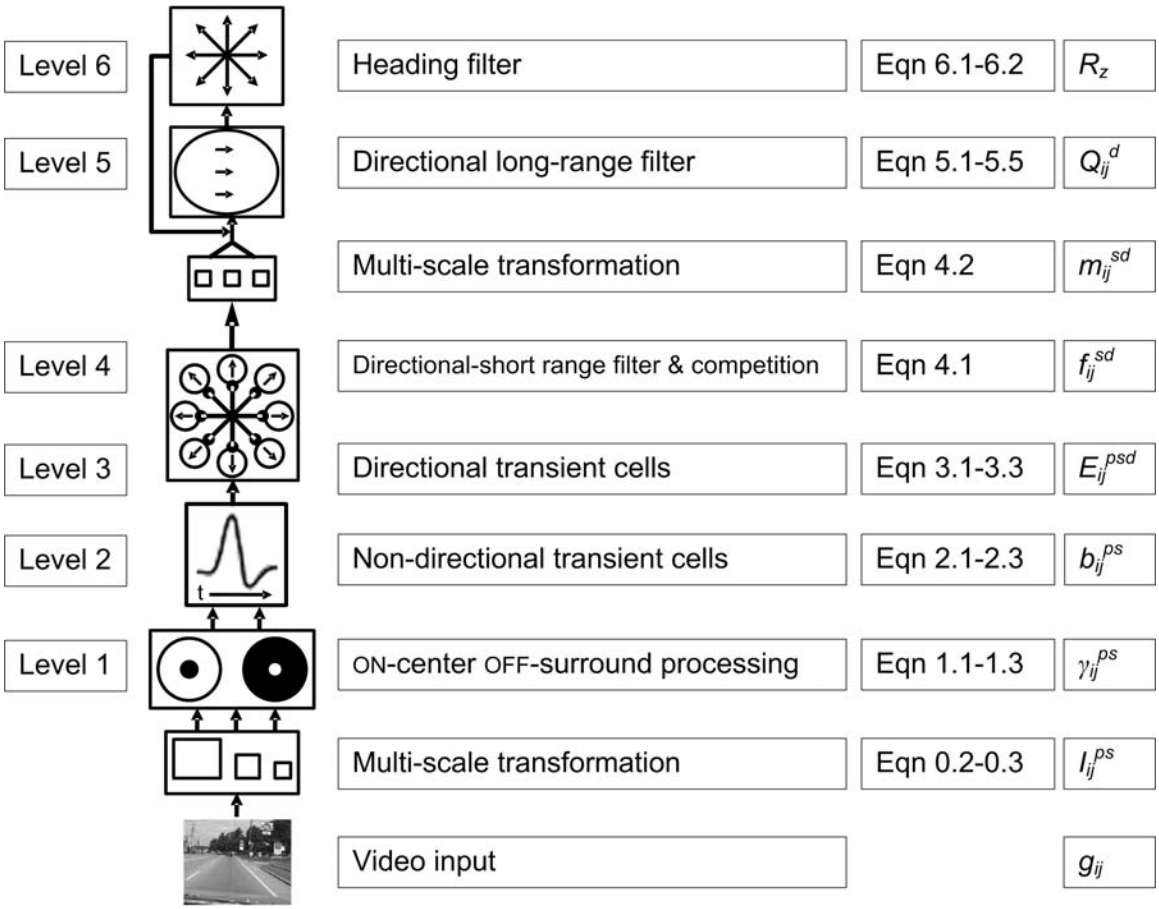

Figure 1.1. Model diagram, video input is resized to facilitate multi-scale processing and then passed to level 1: ON-center OFF-surround processing by a shunting network. Level 2 comprises non-directional transient cells that register temporal changes in the input stream. Level 3 computes local directions of motion using directional transient cells. Inter-directional competition in level 4 normalizes this activity and thereby enhances feature tracking signals. Outputs of level 4 are resized to a single scale. Level 5 sums across speed, and uses a directional long range filter to produce a more global estimate of motion direction. Heading filter cells in level 6 produce a distributed heading estimate. The maximally active cell in this distribution is the heading. Modulatory feedback from level 6 to level 5 refines the global motion estimate in level 5 and sharpens the distribution in level 6.

Inputs and flows. Visual input to mammalian retina is processed continuously to create an optical flow (Gibson, 1950). *Video* input, on the other hand, is frame-based, each frame carries aggregated information extracted from light over some finite time period. Our model is based on biological neurons that operate in continuous time, but in order to make simulation of the model tractable, we provide input in the form of a frame-based video stream. No stage of the model specifically performs frame-based processing; the differential equations determining model behavior have no "knowledge" of frame progression in the video stream. Numerical integration approximates continuous time.

Optic flow is defined as information carried by light that streams in time as a result of environmental structure and an animal's path through the environment (Gibson, 1950). *Retinal flow* is the flow of information over the retina, which additionally includes eye and head rotation information. This distinction between optic and retinal flow is not consistently applied (Warren, 1998). Optic flow is, by definition, a continuous flow of information through time, but is usually represented, and often defined, as a vector field describing instantaneous motion in the environment. Gibson noted that heading can be determined from optic flow (or the instantaneous vector field), but not necessarily from retinal flow, by finding the focus of expansion (FoE). Heading in this case is defined as the direction that the observer is traveling through the environment. Translation through a rigid environment produces a radial motion pattern with all motion vectors emanating from a single origin, the FoE (Gibson, 1950). Mathematical analysis of retinal flow patterns demonstrates that motion vectors resulting from rotations of the eye are not affected by depth, but motion vectors resulting from translation are (Longuet-Higgins & Prazdny, 1980). As a result, in flow patterns that occur when the eye is rotating, the FoE is *not* aligned with current heading. Therefore, in the retinal flow field, the focus of expansion accurately defines heading *only* when there are no components in the flow field that are due to eye rotations.

Cutting et al. (1992) estimated that humans require heading estimates accurate within 1-3 degrees to successfully navigate around cluttered environments. Humans have been shown capable of determining heading from optic flow in random dot displays to within 1-2 degrees (Warren & Hannon, 1988). When real eye movements are made, accuracy is not affected (Royden, Crowell & Banks, 1994; Warren and Hannon, 1990). When eye rotations are simulated in computer-generated psychophysics stimuli, the resulting displays are almost indistinguishable from those produced by a curvilinear path through the environment. Heading in this case is ambiguous (Warren & Hannon, 1990). Human heading accuracy is maintained in the presence of simulated eye rotations of up to 1 degree per second at translation speeds of around 2 meters per second (Banks, Ehrlich, Backus, & Crowell, 1996; Royden, Banks, & Crowell, 1992; Royden, Crowell, & Banks, 1994; Warren & Hannon, 1990). At these speeds, humans cannot distinguish between translation alone and translation with simulated eye rotation (Royden & Vaina, 2004). For faster simulated eye rotations, heading accuracy decreases (Banks et al., 1996; Royden et al., 1992; Royden et al., 1994; van den Berg, 1993; van den Berg & Brenner, 1994). There is no clear consensus as to the precise nature of the error increase and studies show large inter-observer differences (Banks et al., 1996; van den Berg & Brenner, 1994). The ratio of rotation rate to translation rate defines the absolute amount of rotation that can be tolerated in the flow field before heading errors accrue (Li, Sweet, & Stone, 2006).

The use of heading for navigation has been contested by Rushton et al. (1998) and Wilkie & Wann (2003, 2006) who have claimed that goal position information in relation to self is sufficient to explain human steering data. Warren et al. (2001) have shown that humans can make use of both strategies and suggest that, in featureless environments where heading is hard to estimate, egocentric goal position information is used, but in richer environments, heading is used.

Model types. There are three main classes of biological models of heading: *differential motion, decomposition,* and *template* models. Most models show heading

accuracy within 1-2 degrees, matching human data, and tend to be robust to noise in retinal flow. *Differential motion* models (Hildreth, 1992; Rieger & Lawton, 1985; Royden, 1997, 2002; Royden & Hildreth, 1996) and *decomposition* models (Heeger & Jepson, 1992; Lappe & Rauschecker, 1993, 1994) remove eye rotations from retinal flow before estimating heading.

Differential motion models remove the effects of rotation by looking at differences between local motion estimates. As noted above, translational motion results in flow patterns that are dependent on depth, rotational motion results in flow patterns that are not. Therefore, as long as there are multiple depths present in the environment, differential motion models can remove the constant motion due to rotations and provide accurate estimates of the translational motion. Differential motion operators can be related to ON-center OFF-surround cells in MT^- (Royden, 1997, 2002).

Decomposition models deconstruct optic flow into translational and rotational components using analytical techniques (Heeger & Jepson, 1992) based on a mathematical analysis of retinal flow (Longuet-Higgins & Prazdny, 1980). Lappe & Rauschecker (1993) suggested a way to characterize MT and MST*d* as performing a decomposition of optical flow: a feedforward neural network was constructed whereby optical flow was represented in layer 1 (MT). Weights between layer 1 and layer 2 (MST*d*) were analytically computed by finding a least squares solution to the decomposition of motion into translational and rotational components (Longuet-Higgins & Prazdny, 1980). Layer 2 cells were configured, via the weight matrices, to respond to specific combinations of translational and rotational motion. The resulting weight matrices produced cells that, under certain conditions, have response properties similar to MST*d* cells (Lappe & Rauschecker, 1993, 1994, 1995a, 1995b). A major drawback of the Lappe and Rauschecker (1993) model is that it requires a non-biologically plausible teaching signal and weight update method to train the weights. The resulting weight matrices are analogous to the use of motion templates, since layer 2 of the trained network performs a pattern match between the input pattern and the patterns represented in the weight matrices.

Template models (Beintema & van den Berg, 1998; Perrone & Stone, 1994) assess heading from retinal flow and utilize various methods to mitigate the effects of simulated and real eye rotations. The templates used in these models have been demonstrated to be learnable in both supervised and unsupervised networks (Cameron, Grossberg, & Guenther, 1998; Hatsopoulos & Warren, 1991; Zemel & Sejnowski, 1998). Gain fields have been shown to efficiently remove the effects of eye rotations in template models, allowing them to accurately detect heading while the eye is moving (Beintema & van den Berg, 1998; Elder et al., 2008). Template models are consistent with the known neurophysiology of MT^+/MST*d* (Perrone & Stone, 1998).

It is generally assumed that vector-based motion representations of retinal flow exist in V1 and are highly accurate with respect to motion in the environment. However, modeling work suggests that the generation of highly accurate representations in complex environments is challenging (Baloch, Grossberg, Mingolla, & Nogueira, 1999; Bayerl & Neumann, 2004; Chey, Grossberg, & Mingolla, 1998, 1997; Grossberg et al., 1999; Grossberg et al., 2001; Simoncelli & Heeger, 1998). There are inherent ambiguities in retinal flow due to *the aperture problem* (Marr & Ullman, 1981; Wallach, 1935; Wuerger et al., 1996), and accurate optical flow is, in general, not computable (Fermüller &

Aloimonos, 1995). As a result, there is a degree of uncertainty in all retinal flow estimations. Indeed, for navigation in complex cluttered environments, the term *optic snow* has been coined to describe the highly complex, and difficult to accurately estimate, motion patterns that can occur (Langer & Mann, 2003; Mann & Langer, 2002).

The ViSTARS model attempts to describe the key processes of primate motion processing stream, from retinal input through to heading estimate, focusing on the behavioral utility of representations at each stage. We claim, based on the data of Born & Tootell (1992), that computationally *complementary* processing streams in MT and MST have developed for object tracking and navigation, respectively, with the latter being specialized for heading estimation (Grossberg, 2000). ViSTARS produces global motion estimates that determine heading within 1-3 degrees and are highly robust to noise in the input stream. When combined with appropriate refinements to the Steering, Tracking and Route Selection (STARS) model (Elder et al., 2005, 2008), ViSTARS provides a visual front-end for reactive steering and navigation (Browning, Grossberg, & Mingolla, 2009).

Heading in ViSTARS is represented as an activity distribution across MST*d* cells. Such a distributed representation of heading has been described in primates and humans (Beardsley & Vaina, 2001; Page & Duffy, 1999). The model is computationally efficient, demonstrating that a small number of *adaptive flow filters*, or templates, distributed across visual space can produce high levels of accuracy with a wide range of inputs. The model was tested with random dot stimuli, computer-generated terrain, and video taken from a car while driving. In each case, the model produced accurate heading estimates and was consistent with human behavior, human fMRI, and monkey neurophysiology data.

What and Where processing in the ViSTARS model. The ViSTARS model embodies a number of key brain processes to compute heading from optic flow. This section reviews key processe that are represented in the model. A major organizational principles is that biological visual systems consist of functionally distinct processing pathways, or streams. The What, or ventral, cortical processing stream is devoted to object perception and recognition, whereas the Where, or dorsal, cortical processing stream is devoted to spatial localization and action in a wide range of species, including primates (Goodale & Milner, 1992; Mishkin, Ungerleider, & Macko, 1983; Schneider, 1967). The What pathway derives major input from the parvocellular (P) pathway starting in retina with midget ganglion cells. The P pathway is primarily concerned with fine form processing. The complementary Where pathway is concerned with spatial localization and, at the levels of visual processing described in this document, derives major input from the magnocellular (M) pathway. The magnocellular pathway is primarily concerned with motion processing starting in retina with Parasol cells (Rodieck, Binmoeller, & Dineen, 1985). Primate P and M pathways are in many ways similar to cat X and Y pathways, but the cells that make up these pathways do not always exhibit the same behaviors (Kaplan & Shapley, 1982).

M pathway retinal ganglion cells respond to changes in the visual input. They respond with a burst of activation when presented with a step input, they have large receptive fields relative to P pathway cells, do not seem concerned with color processing, and have high contrast sensitivity relative to P pathway cells (Benardete & Kaplan, 1999; Kaplan & Benardete, 2001) .

Primary visual cortex, V1 (striate cortex, area 17). The primate M pathway in V1 codes speed and direction of motion. M pathway retinal cells project to the two ventral layers of lateral geniculate nucleus, LGN (layers 1 and 2), and then to V1 layer 4Cα followed by layer 4B (Callaway, 2005, Livingstone, 1998). V1 layer 4B contains cells which are directionally selective, and respond more vigorously to motion in a *preferred* direction (Livingstone, 1998).

Directionally selective cells respond to objects moving in a particular direction within a range of speeds. Barlow and Levick (1965) noted that the behavior of rabbit retina directional cells was consistent with *nulling inhibition*, and not with forward excitation. The direction opposite to the *preferred* direction was defined as the *null* direction. They characterized the directional cells as time-delayed "and not" gates such that cell activation in a particular spatial position is *vetoed* by activity in a cell shifted one position in the *preferred* direction at an earlier time. Inhibition travels in the *null* direction to produce directional selectivity in the *preferred* direction.

Directional cells in cat and primate V1 are nonlinear, and space-time plots of these cells support a nulling inhibition mechanism (DeAngelis, Ohzawa, & Freeman, 1995; Livingstone, 1998). Livingstone (1998) characterized primate V1 directional cells as having nulling inhibition combined with an asymmetric excitatory dendritic tree. Asymmetric dendritic trees are found in human and primate directional cells (Elston & Rosa, 1997; Livingstone, 1998) but asymmetry alone is not sufficient to explain directional selectivity (Anderson, Binzegger, Kahana, Martin, & Segev, 1999). Subsequent rabbit retina data showed that starburst interneurons provide spatially asymmetric inhibition that comes from the null side (Fried, Münch, & Werblin, 2002, 2005). In previous modeling work (Chey, Grossberg, & Mingolla, 1997, 1998), interneurons with properties of the subsequently reported starburst cells were predicted to enable directional cells to maintain their sensitivity to a wide range of speeds. This prediction has not yet been tested.

Middle temporal area (MT). The Where stream in monkeys projects from V1 to extrastriate cortical area MT (middle temporal, V5) and then to area MST (medial superior temporal). In humans, V1 projects to an area known as human MT complex (hMT, V5+). The two species show reasonably high levels of similarity with respect to motion processing, although humans have more brain areas involved in motion processing (Orban et al., 2003). For a full review of the structure and function of primate MT, see Born and Bradley (2005). Only the most relevant details are included here.

Primate MT has two main populations of directional cells. MT^+ consists of cells with large additive receptive fields and projects primarily to dorsal MST (MST*d*). MT^- consists of cells with ON-center OFF-surround receptive fields and projects primarily to ventral MST (MST*v*) (Born & Tootell, 1992). MT^+/MST*d* is implicated in navigation based upon optic flow, notably motion-derived heading estimation (Duffy, 1998; Duffy & Wurtz, 1995, 1997). MT^-/MST*v* is implicated in motion-based object segmentation and tracking (Duffy, 1998). These two substreams of the Where stream have been shown to have computationally complementary properties (Grossberg, 2000).

The present article models aspects of the MT^+/MST*d* processing stream. Our companion paper models MT^-/MST*v* dynamics (Browning, Grossberg & Mingolla, 2009) and how the two complementary MT^+/MST*d* and MT^-/MST*v* streams cooperate to enable navigation towards goals around obstacles. MT cells have much larger receptive

fields than V1, but like V1 cells respond to direction and speed of motion (Born & Bradley, 2005). Macaque MT cells have been shown to represent an aperture-resolved motion signal. MT cells initially respond to the perpendicular direction of bar orientation but after a period of 100-200ms respond to the true direction of motion (Pack & Born, 2001). MT is the first primate brain area demonstrated to have an aperture-resolved motion signal. These data confirm predictions of the 3D FORMOTION model which the ViSTAR model incorporates (Chey, Grossberg, & Mingolla, 1997) in which model cells that integrate over space and time initially respond to the direction orthogonal to orientation but gradually respond to the true direction of motion.

Dorsal medial superior temporal area, MSTd. MST*d* cells respond to patterns that occur during self-motion through the environment (Duffy, 1998; Grossberg, Mingolla, & Pack, 1999; Stone & Perrone, 1994, 1997a) . The human homolog of monkey MST is also implicated in human heading detection tasks (Beardsley & Vaina, 2001). There are MST*d* cells tuned to planar, radial, and spiral motion (Duffy & Wurtz, 1995; Graziano, Andersen, & Snowden, 1994; Saito et al., 1986). Radial motion occurs when an observer travels along a forward or backward path through the environment, whereas planar motion occurs due to a sideward path through the environment or due to rotation of the head or eyes. Heading appears to be represented as a distributed population code in both primate MST*d* and human hMT+ (Beardsley & Vaina, 2001; Page & Duffy, 1999).

Methods

The ViSTARS model

Before describing each of the model stages in detail, we provide a brief overview: Input to the model is initially processed by a multiple-scale-sensitive shunting ON-center OFF-surround network that enhances spatial discontinuities and normalizes the intensity of the input stream (Level 1 in Figure 1.1). Spatial discontinuities often correspond to object boundaries or spatial features in the environment. In this article, when we refer to objects, object boundaries, or spatial features, we mean spatial discontinuities in the intensity image. Transient cells in the model retina respond to the leading and trailing boundaries of moving objects (Levels 2 and 3 in Figure 1.1). When the observer is moving, transient cells additionally respond to the boundaries of stationary objects. Directional cells in the model's cortical area V1 accumulates directional evidence for local motion within a given range of directions and speeds (Level 4 in Figure 1.1). These estimates are ambiguous due to noise in the input stream and the aperture problem (see below). Directional competition reduces ambiguities in the local motion estimates (Level 4 in Figure 1.1). Spatiotemporal integration through a directional long-range filter, sharpened by competition, in MT^+ produces more global motion estimates (Level 5 in Figure 1.1). An additional stage of bottom-up filtering of these estimates generates heading estimates in MST*d* (Level 6 in Figure 1.1). Model cortical area MT^+ and MST*d* form a recurrent processing loop: The heading estimates feed back to MT^+, where they choose motion vectors consistent with the current heading and suppress motion vectors that are not.

Model Levels 1 and 2: Input normalization and non-directional transients. Magnocellular pathway retinal ganglion cells in ViSTARS consist of a shunting ON-center-OFF-surround network (Appendix equations 1.1-1.3) that feeds into a *non-directional transient cell level* (Appendix equations 2.1-2.3). The ON-center OFF-

surround network splits the input into ON and OFF channels, enhances spatial discontinuities, and normalizes the intensity of the input; cf. Baloch, Grossberg, Mingolla, & Nogueira, 1999; Chelian & Carpenter, 2005. Spatial discontinuities in intensity generally correspond to spatial features such as object boundaries and corners. Transient cells activate in response to the leading and trailing boundaries of moving objects, as modeled in Baloch et al., (1999), Berzhanskaya, Grossberg, & Mingolla, (2007), and Grossberg, Mingolla, & Viswanathan (2001). When the observer is moving, stationary boundaries also produce a response. Figure 1.2 illustrates the output of the *transient cells* in response to a variety of stimuli.

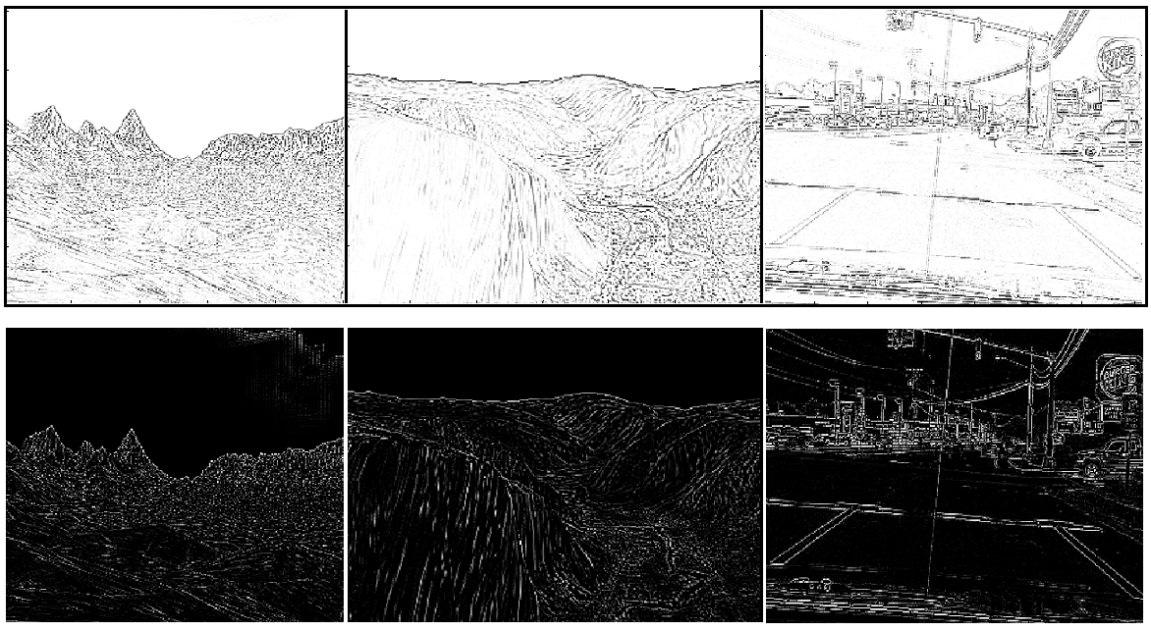

Figure 1.2. Output of the *non-directional transient cells (model level 2).* Left, OpenGL stimulus; middle, Yosemite sequence; right, video taken from car while driving. Top, ON-channel. Bottom, OFF-channel. Descriptions and pictorial representations of all input sequences are given in the Methods section.

Model level 3: Directional transient cells. Level 3 of the model corresponds to primate V1. Model *directional transient cells* (Appendix equations 3.1-3.3) respond to changes in intensity that move in a *preferred* direction. Eight directions are represented at 45 degree increments. These cells are variants of those introduced in Chey et al. (1997) and incorporate offset nulling inhibition via inhibitory interneurons to produce a local motion estimate that is sensitive to a wide range of speeds. The model is processed at 3 spatial scales. These spatial scales respond selectively to different ranges of speed. The output of Level 3 is a local motion estimate represented by 8 directions and 3 speeds across ON and OFF channels.

Motion estimates in Level 3 are limited by the neural *aperture problem* and thus are unable to accurately determine motion direction (Marr & Ullman, 1981; Wallach, 1935; Wuerger et al., 1996). Within a small aperture, such as the receptive field of a neuron, the direction of motion of a line, or object, is inherently ambiguous. Motion within the aperture is perceived in the direction perpendicular to the orientation of the

bar, irrespective of the true direction of motion. Three main approaches have been suggested to solve the aperture problem and determine true direction of motion:

(1) *Population average:* motion vectors within a local area are pooled to determine an accurate motion estimate (Horn & Schunck, 1981).

(2) *Feature tracking:* spatial features, such as line ends and corners, in the visual scene do not suffer from the aperture problem. Tracking these features therefore can give an accurate motion estimate (Lucas & Kanade, 1981; Mingolla, Todd, & Norman, 1992; Tomasi & Kanade, 1991).

(3) *Intersection of constraints (IOC):* perceived motion of a plaid pattern is in the direction defined by the velocity vector of the intersection in velocity space of the constraint lines of the plaid component motion (Adelson & Movshon, 1982). None of these approaches alone can explain all relevant data.

Based on prior modeling work, we utilize a combination of *feature enhancement* in V1 to facilitate *feature tracking* in MT along with *population averaging* in MT via long-range spatio-temporal integration (Chey et al., 1997). The aperture problem is usually considered in relation to determining accurate object motion. However, it also affects heading. For example, if one imagines an environment consisting of only nearly horizontal lines, forward camera motion in this environment will produce (due to the aperture problem) a relatively large amount of upward and downward motion energy and a relatively small amount of leftward and rightward energy. Thus, to accurately determine the heading in this environment, some process needs to ensure that the global motion estimate is not entirely dominated by the upward and downward motion signals. Such a process need not explicitly solve the aperture problem. However, it must enable the system as a whole to overcome the aperture problem and compute a reasonably accurate estimate of global motion direction. For example, Fermüller and Aliomonos (1995) outlined a method by which heading can be estimated based on the sign of 2D motion vectors across the visual field. This method does not explicitly solve the aperture problem to resolve local motion, but the effects of the aperture problem are mitigated such that they do not adversely affect heading judgments.

Model level 4: Directional short-range filter and competition. Level 4 of the model combines the ON and OFF cell streams and implements a directional short-range filter and competition (Appendix equation 4.1). The short-range filter sums directional evidence across multiple scales. This is followed by competition within a shunting, or membrane equation, on-center off-surround network. In such a network, competition within a dimension results in divisive normalization across that dimension (Grossberg, 1973). Level 4 is based on a shunting equation that incorporates competition across direction and thus normalizes cell population activity across direction. Cross-directional normalization enhances the least ambiguous motion signals and suppresses the most ambiguous signals (Bayerl & Neumann, 2004; Chey et al., 1997). The least ambiguous regions, such as line ends or corners, correspond to spatial features in the input, or feature tracking signals. The enhancement and suppression of features does not totally eliminate ambiguities. Rather, it increases the magnitude of activation at less ambiguous locations and decreases the magnitude of activation at more ambiguous locations. Level 4 of the model produces a feature-enhanced local motion estimate represented by 8 directions and 3 speeds. Figure 1.3 illustrates the output of level 4 of the model for various stimuli.

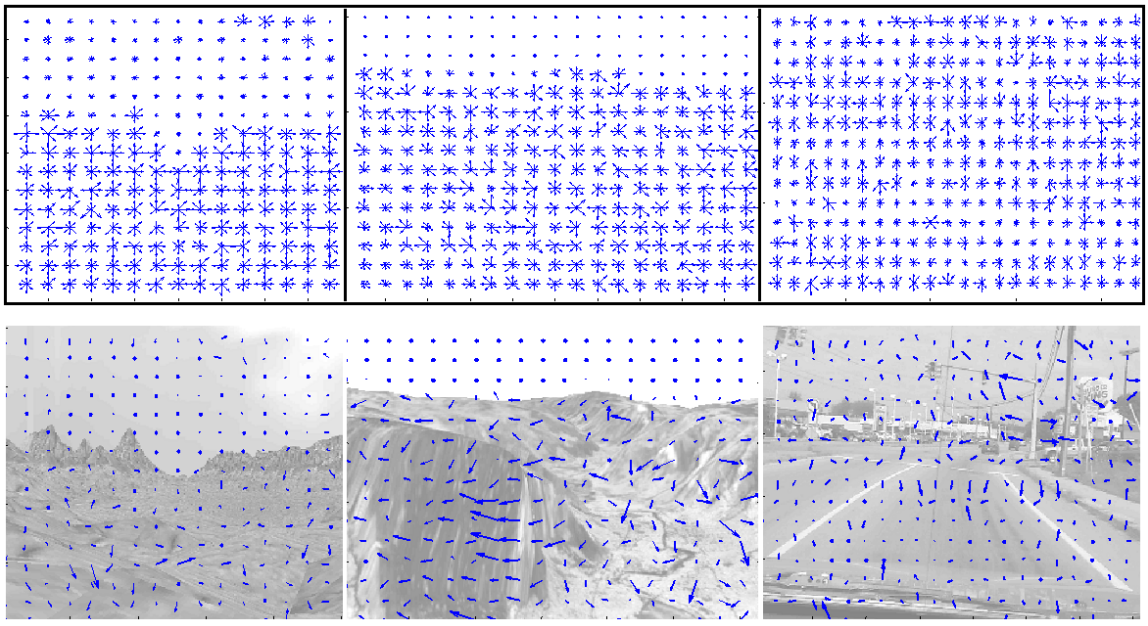

Figure 1.3. *Feature-enhanced directional* output of level 4. Left, OpenGL stimulus moving leftward; middle, Yosemite sequence moving rightward; right, video taken from car driving rightward. Top: vector representation of activation, at each position. All 8 directions are shown. Arrow length is a measure of speed. Bottom: vector population average overlaid on frame of stimulus. When computing the population average, opposing estimates within the same position cancel each other. Therefore arrow length denotes confidence in the direction estimate. Vectors are shown at every 16th vertical and horizontal position for clarity.

Model Level 5: Directional long-range filter. Level 5 of ViSTARS corresponds to MT$^+$ (Appendix equations 5.1-5.3). Input from V1 (equation 4.2) is summed across speed and then spatially integrated through a directional long-range-filter (equation 5.2). Summing across speed allows the model to perform a heading estimate on the basis of the direction patterns, irrespective of how fast the observer is moving. Representation of motion at slower speeds is weighted more heavily in the sum to compensate for the relative spatial scarcity of motion signals at those scales. The long-range filter, L, is a Gaussian filter elongated in the *preferred* direction of the cell: it accumulates evidence for motion in a particular direction across a region of space.

Bottom-up input from the long-range filter is modulated by feedback from level 6 (model MST*d*). This feedback enhances inputs that support the current heading estimate in MST*d* (Appendix equation 6.2). Model MT$^+$ is implemented as a shunting network (Appendix equation 5.1) whereby directional evidence is accumulated over space and time. Level 5 implements a *choice network* (Grossberg, 1973), also known as a winner-take-all-network. A choice network consists of balanced self-excitation and mutual inhibition through a nonlinear feedback signal function that results in a winner-take-all computation. The lateral connectivity in primate MT is unknown, but competitive interactions within MT are supported by data demonstrating that MT cell responses can be suppressed under some transparent motion conditions (Snowden, Treue, Erickson, &

Andersen, 1991). Models have shown that competition between directions in MT (Royden, 2002; Berzhanskaya et al., 2007; Grossberg et al., 2001; Chey et al., 1997) generates more accurate motion estimates.

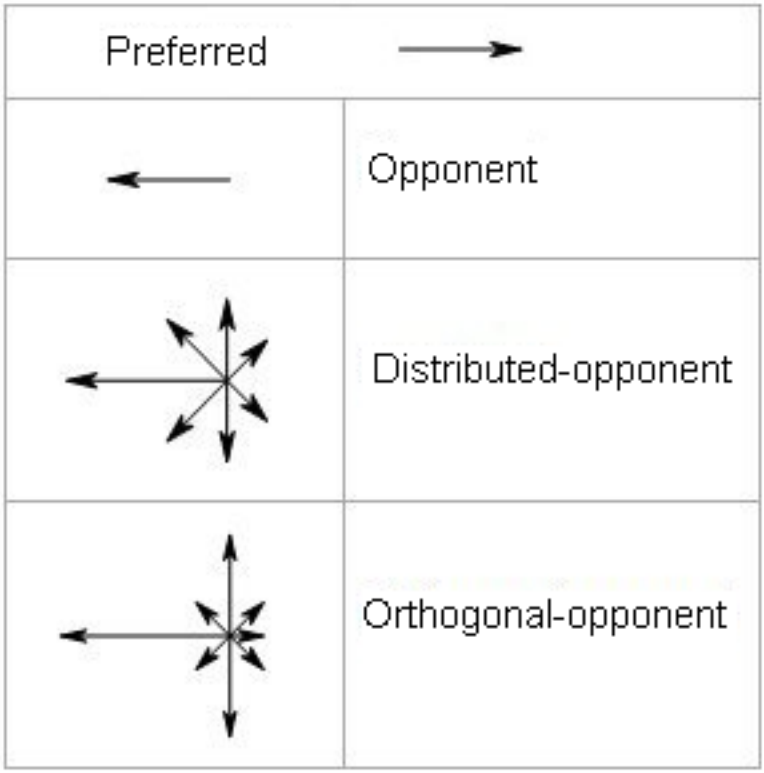

Figure 1.4. MT^+ lateral inhibition weighting functions, v_{dD}, for a cell with a preferred direction to the right. Opponent, distributed-opponent, and orthogonal-opponent competitive weighting examples are depicted from top to bottom.

We implemented 4 types of lateral connectivity (Appendix equations 5.4-5.7): *no-competition, opponent, distributed-opponent, and orthogonal-opponent* competition, represented pictorially by Figure 1.4. *No-competition* acts as a control. *Opponent* directional mechanisms are implicated throughout primate motion processing (Heeger, Boynton, Demb, Seidemann, & Newsome, 1999). *Distributed-opponent* competition implements inhibition proportional to the angular distance between directions. *Orthogonal-opponent* competition implements strong inhibition from the orthogonal and opponent directions, but only weak inhibition from other directions. It was found through parameter search to produce good results. Output from, and feedback within, MT^+ is thresholded, half-wave rectified, and squared (Appendix equation 5.3).

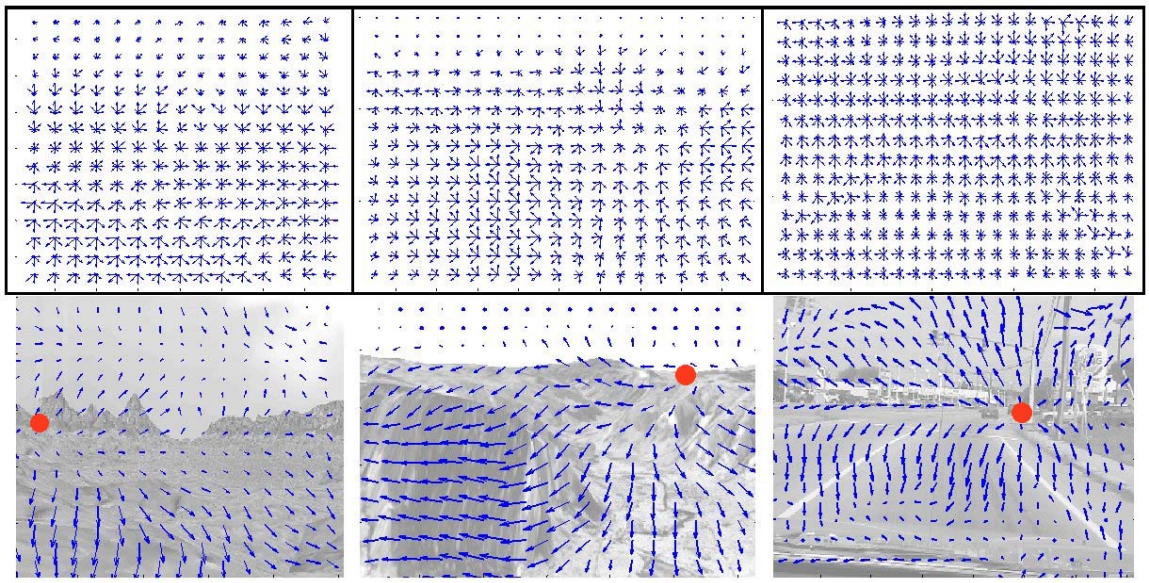

Figure 1.5. *Global motion estimate in MT+ cell* output of level 5. Left, OpenGL stimulus with true heading (red circle) to left of center; middle, Yosemite sequence with true heading to right of center; right, video taken from car while driving around a rightward bend. Top: vector representation of activation, at each position. All 8 directions are shown. Bottom: population average overlaid on frame of stimulus. Arrow length denotes confidence in the direction estimate, as described in Figure 1.3. Vectors are shown at every 16th vertical and horizontal position for clarity.

Model MT+ global motion outputs are shown in Figure 1.5. The global structure of motion in the environment is hereby extracted by MT+ from the short-term noisy local motion estimates represented in V1, as shown in Figure 1.3. Ambiguities in the motion estimates are vastly reduced and the global motion pattern broadly matches what one would expect from motion towards each of the stimulus headings.

Model level 6: Adaptive Flow Filter. Level 6 of the model corresponds to MST*d* (Appendix equations 6.1 and 6.2). Model MST*d* computes heading through a bottom-up heading filter of the global motion estimate in MT+. We call this filter an *adaptive flow filter.* Flow filters represent the idealized motion pattern, or template, that occurs due to translation in a rigid environment towards the point in space represented by its target MSTd cell. Cameron et al. (1998) demonstrated how this type of flow filter can be learned using a self-organizing feature map (SOFM) (Cameron et al., 1998; Elder et al., 2008). As the self-organizing system is exposed to global motion patterns, such as translational motion patterns, it will learn to distinguish between different translational headings. In ViSTARS, the flow filters are fixed, but they can, in principle, be learned. The motion vector at each position is normalized since only direction of motion, and not speed, is represented in the filter. This assumption is imposed for computational simplicity. We do not claim that MSTd cells have no speed tuning, but show that such tuning is not necessary to demonstrate accurate heading estimates.

Input from MT+ is filtered by all of the adaptive flow filters to provide a distributed representation of heading. Flow filters are created for two rows of headings,

one at 1/2 of the screen height and one at 5/8 of the screen height. The vertical position of each row was chosen arbitrarily based on likely horizon positions. As discussed above, the FoE of a motion pattern is situated on the point in the image towards which the observer is moving. For forward translation, therefore, the FoE tends to be near the horizon. If one were headed towards the ground, the FoE would be positioned on the ground. Humans often look at the ground while they walk. However, they are rarely heading in that direction. Thus, the FoE is still likely to be near the horizon in this case, even though it could be outside the field of view of the observer. The human heading data that we replicate in this article are based on forward translations with the horizon visible and situated between ½ and 5/8 of the screen height.

Concentrating templates on likely horizon positions reduces the number of cells required for accurate heading estimation. Each cell row consists of headings spaced at every third pixel, resulting in between 21 and 30 MST*d* cells per row, depending on the input resolution of the stimulus. This further reduces the number of cells required for accurate heading estimation but makes it unlikely that any single filter will exactly match the motion pattern in MT^+. Despite the sparse distribution of cells, the activation pattern across MST*d* cells provides an accurate measure of heading.

The field of view of our model is between 35 and 45 degrees, depending on the input stimulus. Primate MST*d* pattern cells typically span >50 degrees of the visual field (Duffy & Wurtz, 1997). We therefore use filters that cover the full 35-45 degree visual field of our inputs. Temporal integration by the shunting equation produces a temporally stable heading estimate (Appendix equation 6.1), resulting in smooth transitions in the MST*d* population distribution as an observer's heading changes. Model MST*d* implements a recurrent *contrast-enhancing* network (Grossberg, 1973) that selects a small number of the most active cells through balanced self-excitation and broad lateral inhibition with feedback via a sigmoid signal function (Appendix equation 6.2). Output from MST*d* is a distribution across these heading cells, whose activities scale with how well the corresponding filter matches the global motion estimate in MT^+. Representative outputs are shown in Figure 1.6.

The two rows of heading cells typically respond with the row closest to the horizon having higher overall activation and the peak of that distribution providing the most accurate heading estimate. For navigation, ViSTARS uses a distributed representation of heading. However, for analysis of heading accuracy, the horizontal pixel position of the maximally active MST*d* cell is selected to be the estimated heading. The multiple-scale specialization of model levels 1-3 combined with the new model heading computation levels 4-6 provide a demonstration of the utility of neural models on natural image sequences. ViSTARS is the first model to utilize any of these model components for natural image stream processing. ViSTARS is also the first model to demonstrate how these model components integrate to produce heading estimates in a similar manner to humans. A companion article demonstrates how ViSTARS utilizes these processing stages for reactive obstacle avoidance and goal approach (Browning et al., 2009).

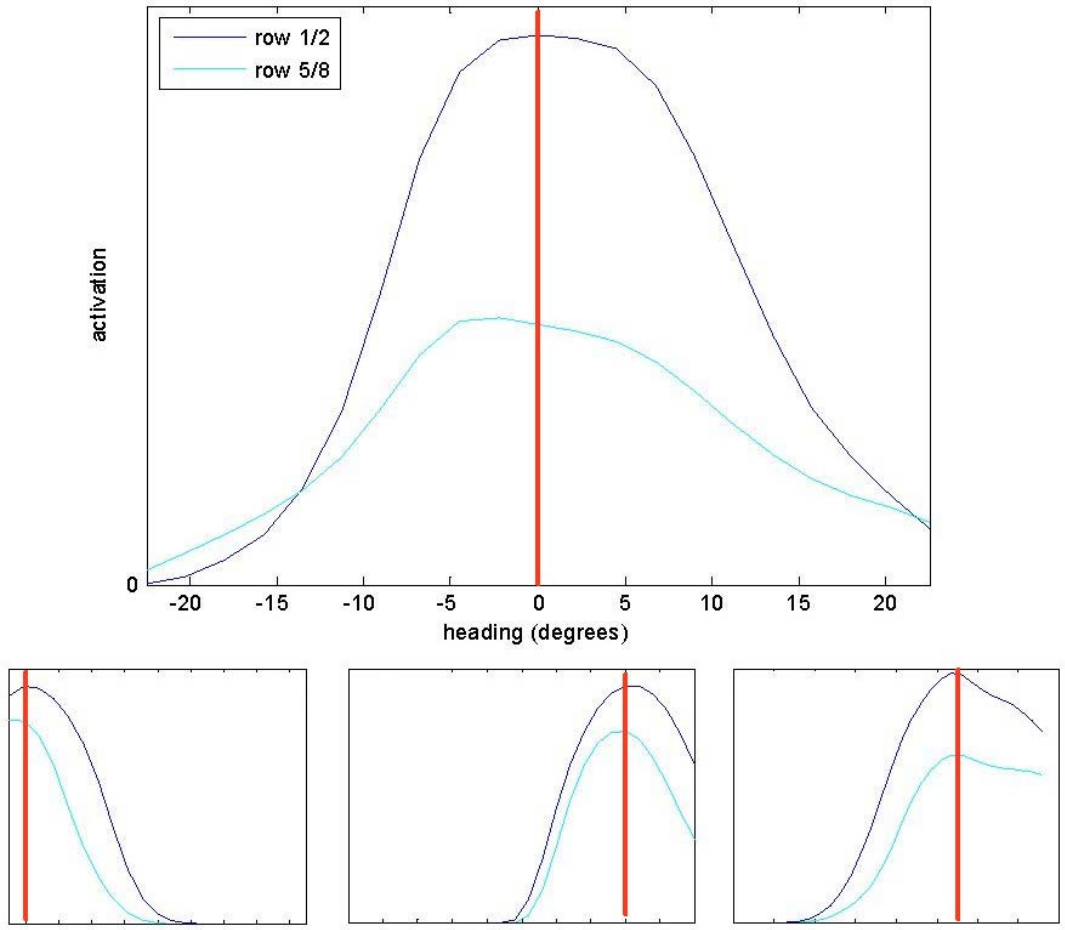

Figure 1.6. H*eading estimate in MSTd cell* output from level 6. MST*d* cells form 2 rows, one row at 1/2, and one row at 5/8, of the vertical resolution of the input. The horizontal position of the heading is given by the vertical line. Top: response to OpenGL stimulus with an observer heading at 0 degrees. Bottom left: response to OpenGL stimulus with an observer heading at -19 degrees. Bottom center: response to Yosemite sequence with observer heading at 16 degrees (assuming a 50 degree field of view). Bottom right: response to video taken in a car driving round a rightward bend (precise ground truth not available; see Results).

Implementation

The ViSTARS model is defined by a system of differential equations. It was tested using psychophysical stimuli, computer-generated animations, and video taken while driving. Simulations were performed in MATLAB R14 (MathWorks, 2005) on a dual 2Ghz AMD Opteron (AMD, 2003) workstation with 8Gb of RAM running Microsoft Windows XP x64 (Microsoft, 2003). Input to the model was a frame-based input stream, either computer-generated image sequences or video from a camera, as described below. Euler's method was used to numerically integrate the solution to the equations with a time-step of 0.1. Each frame of video input was presented for 10 time steps and then the next frame of input was presented on the subsequent time step. We

define the frame rate of the input at 15 frames per second. The time course of the model non-directional transient cells shows that a burst response occurs after about 75ms of simulated time, consistent with data describing M-pathway retinal ganglion cells (Benardete & Kaplan, 1999; Kaplan & Benardete, 2001). The equations were not integrated to equilibrium. Rather, the activations of model cells ebb and flow with changes in the input. Whether or not a particular cell in a particular spatial position reaches an equilibrium state is dependent on the magnitude of changes in the input stream at that spatial position. A mathematical model definition and implementation discussion is given in the Appendix.

In order to improve the computational efficiency of the model we used input resizing to represent different spatial scales, rather than explicitly coding the model with multiple receptive field sizes. This resizing allows the model to use a single set of parameters for all processing scales. This method is known in the computational literature as hierarchical processing and is common in optic flow algorithms (Beauchemin & Barron, 1995). The video input is stored at 3 resolutions: scale 1 processes at full resolution and responds to motion at slow speeds (~ 1 pixel per frame). Scale 2 processes at half of the original resolution and corresponds to medium speeds (~ 2 pixels per frame). Scale 3 processes at quarter of the original resolution and corresponds to motion at faster speeds (~ 4 pixels per frame). Model retina and V1 process at all 3 scales.

Output from all speeds in V1 is resized to a quarter of the original resolution. For example, if the input video has a resolution of 256 x 256, scale 1 has resolution 256 x 256, scale 2 has resolution 128 x 128, and scale 3 has resolution 64 x 64. Output from V1 is resized to 64 x 64 for all scales. See Appendix equations 0.2 and 4.2 for more details.

Random dot stimuli were created in MATLAB based on those described in Royden et al. (1994). Stimuli were created for a *ground plane, 3D dot cloud, 2m frontal plane* and *8m frontal plane*. Stimuli were created such that a dot had a value of 1 and the background had a value of 0, and the resolution of the stimuli was 256 x 256 pixels. The horizontal visual angle was defined as 30°. Observer height was 1.6m. For ground plane and 3D dot cloud stimuli, depth was limited at 37.3m, dot density was set at 0.6 dots per square meter, and translation speed was fixed at 1.9 meters per second. For frontal plane stimuli, the plane started at 2 or 8 meters with 625 dots visible, and the translation rate was 0.5 meters per second. Dots that left the visual space of the screen were not replaced. Translational stimuli were created for *ground plane*, *3D dot cloud,* and *2m frontal plane* with headings at ± 10, ± 5, and 0 degrees, resulting in 15 translational random dot stimuli. Simulated rotations for ± 10, ± 5, ± 2.5, and ± 1 degree(s) per second were added to a 0 degree translational heading stimulus in all 4 stimuli groups, resulting in 36 simulated rotation random dot stimuli. Projection of the 3D scene on to an image plane was performed in accordance with Longuet-Higgins and Prazdny (1980). Figure 2.1 illustrates the ground plane and 3D dot cloud stimuli.

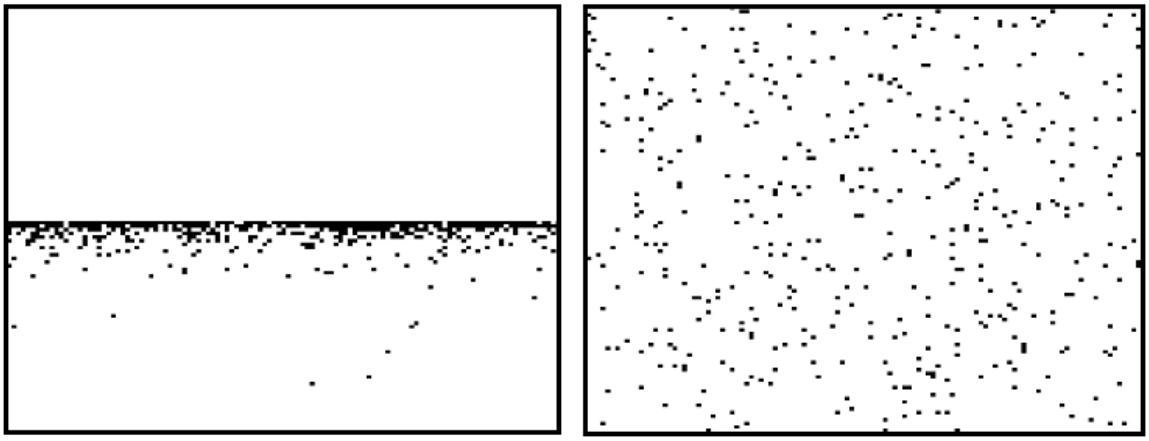

Figure 2.1. Left: frame from ground plane random dot stimuli. Right: frame from 3D dot cloud random dot stimuli. Dots are enlarged and contrast-inverted for display.

The *Yosemite sequence* was obtained from Michael Black's website (Black, 2007) and was pre-processed according to instructions in that document. The pixel intensity values were scaled between 0 and 1. The image sequence was originally created by Lynn Quam at SRI, aerial imagery of the Yosemite Valley was mapped onto a depth reading of the valley, and a simulated fly-through was created. There are 14 frames of input with a resolution of 316 x 252 pixels. The Yosemite sequence was used to assess model robustness to noise. Gaussian noise was added to the input frames using the *randn* MATLAB function. Noise was added at each pixel position for each frame of the input, and the resulting value was clipped at 0 and 1. The signal to noise ratio (SNR) was calculated by summing the absolute differences between the noisy and the original inputs and dividing the sum of the original input by this difference. A total of 9 different magnitudes of noise were added, and each test was repeated for 10 trials with different randomly generated numbers. The mean output of the 10 trials, for each of the 9 noise magnitudes, was used to report results. A frame of the Yosemite sequence is shown in Figure 2.2. As noted in the Introduction, computing optic flow from natural image sequences is highly challenging (Langer & Mann, 2003; Mann & Langer, 2002). The addition of noise to the Yosemite sequence, which contains only translational motion, produces a test of the model's robustness to noise in the pixel intensities in the input.

OpenGL stimuli were generated in OpenGL by taking a depth map of a fictional terrain and mapping small textures onto it. Textures were taken from NASA Mars photographs. The specific textures chosen at any point in the terrain were based on the height of that piece of terrain. The resulting terrain is homogeneous but realistic looking, Figure 2.2 illustrates the terrain. Homogeneous terrains are challenging for optic flow algorithms since they provide few trackable features to disambiguate the flow. The environment was rendered with a 45° horizontal visual angle at a resolution of 256 x 256 pixels. Pixel values were grayscale, scaled between 0 and 1. 12 stimuli were created, 5 slow and 7 fast, each stimulus consisting of 14 frames. Units of measurement are arbitrary in OpenGL: the slow stimuli had a translation speed of 1 unit per second and the fast stimuli had a translation speed of 10 units per second. Slow stimuli were generated

for headings ± 19, ± 9 and 0 degrees. Fast stimuli were generated for headings ± 20, ± 10, ± 6 and 0 degrees.

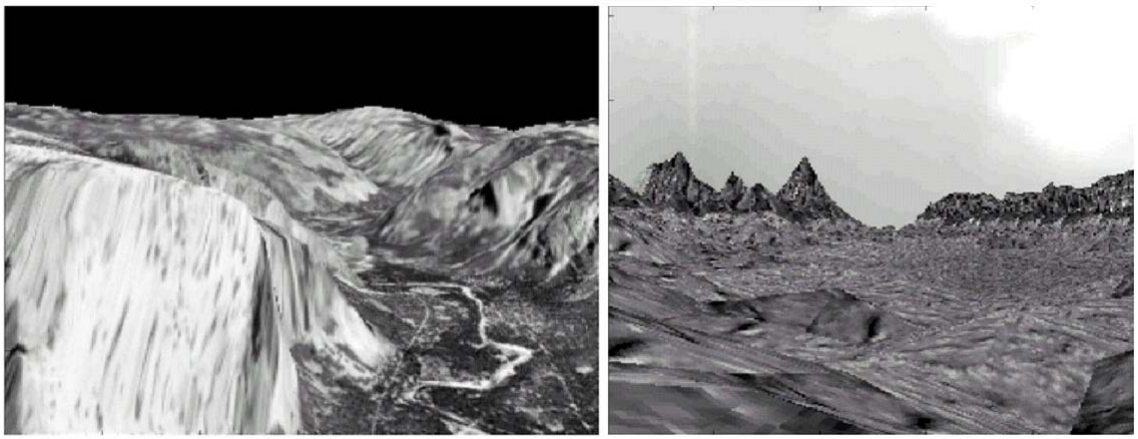

Figure 2.2. Left: frame from Yosemite Sequence. Right: frame from OpenGL stimuli

Driving video was taken over a number of sessions using a SONY DCR-TRV70 (Sony, 2003) digital video camera. The camera was mounted on a tripod and loosely fixed between the front seats of a sports utility vehicle with the camera pointing straight forward through the windscreen. The camera was visually aligned to straight ahead using the camera's view screen. The horizontal visual angle of the camera was calculated using Sony's published specifications at 37.2 degrees. Video was converted to 360 x 240 pixels at 15 frames per second, and converted to grayscale between 0 and 1. Twenty-five one-second video clips were generated under various conditions. Ten clips were generated from driving in the rain, fifteen clips were generated in dry conditions. Video was taken on rural roads, suburban main roads, and highways at speeds up to 65 miles per hour. No efforts were taken to ensure that the camera was in precisely the same position for each session and the camera moved during each session. Figure 2.3 illustrates frames from three driving videos. All stimuli are available for download from http://cns.bu.edu/vislab/projects/buk/.

Figure 2.3. Frames from driving video clips: rural, highway and suburban sequences, respectively.

As noted above, although heading is defined in the model in terms of an activity distribution, for analysis we define heading as the pixel position of the maximally active

heading cell. Heading error is initially calculated as the distance in pixels between the model heading and the true heading. For comparison with human data, we report error in terms of visual angle, in degrees. Pixel error is converted to angle error by multiplying by the ratio of *horizontal visual angle* to *horizontal resolution.* True heading is defined in the OpenGL and random dot stimuli by an angle; this is converted to a true pixel heading by multiplying angular heading by the ratio of *horizontal resolution* to *visual angle*. True heading for the Yosemite sequence was calculated by finding the pixel position of the minimum of motion in the ground truth. The ground truth motion is supplied as part of the Yosemite sequence. True heading was estimated for the driving video clips based on the type of maneuver. See the results section for more details.

There are a large number of parameters (39) in the model (Appendix; Table 5). This is due to the fact that each stage of the model simulates a network of cells in primate magnocellular pathway. These parameters typically govern basic cell properties, such as the temporal decay rate, the upper and lower bounds on activation, or the balance between excitatory and inhibitory inputs. The same parameters were used in all simulations (Table 5). Note that the two normalization constants M_6 and N_7 are functions of the resolution of the input stimulus and are therefore different for different input sources.

Parameters at each stage of the model were set to simulate cells found in neurophysiological data. Parameters for levels 1 and 2 of the model, corresponding to primate retina, were configured such that they produce a burst response after around 75ms in order to match the primate data (Benardete & Kaplan, 1999; Kaplan & Benardete, 2001). Parameters for levels 3 and 4 of the model, corresponding to primate V1, were configured such that they produced a direction estimate in response to moving objects. The directional response exhibited the aperture problem, as it does in primate V1 (Livingstone, 1999). Parameters for level 5 were chosen such that, over time, model MT represents an aperture-resolved motion signal, as it does in primate (Pack & Born, 2001), as predicted by Chey, Grossberg, & Mingolla (1997). Model levels 1-5 were parameterized using binary stimuli containing moving bars. Parameters for model level 6, corresponding to primate MSTd, were chosen so that level 6 responded optimally to various translational motion patterns. Model level 6 was parameterized by feeding motion patterns in to model level 5 rather than by processing visual stimuli. All parameters were chosen by hand and the model behavior is reasonably robust to parameter changes around the chosen values. As with all dynamical neurophysiological models, large changes in parameter values can significantly affect the behavior of the model by upsetting the balance that enables all the mechanisms to achieve their functional properties.

Parameter F_1, in Appendix equation 1.2, was then tuned to optimize overall model performance. This parameter, F_1, governs the inhibitory surround of the ON-center OFF-surround retinal ganglion cell network. Manipulation of this parameter determines the contrast normalization characteristics of the network and thus what parts of the input are salient enough for further processing. The same value of F_1 was used in all simulations, including the driving video, despite the fact that each of the input sources had quite different characteristics. For example, the pixel resolution, visual angle, and distribution of pixel intensity in the image differ across each stimulus set. Parameter F_1 was chosen to minimize average heading error across the five slow OpenGL stimuli,

translational random dot patterns, and the Yosemite sequence. It should be noted that the goal of this work was not to demonstrate that the model, when configured for one set of stimuli, will operate well on another set of stimuli, as it would be in a machine learning exercise. We tuned the parameter F_1 across most of the tested artificial translational stimuli in order to demonstrate that emergent properties due to interactions among the model's neurophysiological mechanisms can quantitatively explain key behavioral competencies of primates and humans.

Results

Accurate heading estimates. The model was tested using the 12 OpenGL stimuli, the Yosemite sequence, and 15 random dot stimuli. The 15 random dot stimuli were *ground plane, random dot cloud,* and *2m frontal plane,* each with translational headings at ± 10, ± 5, and 0 degrees. When processing with *distributed-opponent* competition (Appendix equation 5.6) in MT^+ (Appendix equation 5.1), the model predicts the nearest pixel to the analytically computed true heading for the Yosemite sequence. On the random dot stimuli, it achieves a mean error of 1.2 degrees, and for the OpenGL stimuli, a mean error of 1.4 degrees is obtained. The maximum error is 3.83 degrees. For comparison with human data, results for translational stimuli are shown in Table 1.

	No Competition	*Opponent*	*Distributed-Opponent*	*Orthogonal-Opponent*
OpenGL	4.7	0.9	1.4	1.1
Yosemite	1.5	1.5	0	1.5
Random dots	1.4	1.4	1.2	1.4
Mean	**2.9**	**1.2**	**1.3**	**1.3**

Table 1. Mean error, in degrees, for 12 OpenGL, Yosemite and 15 Random dot stimuli with no rotations, processed by the model using different competitive conditions in MT^+.

The results are broadly the same for *opponent, distributed-opponent* and *orthogonal-opponent* competition in MT^+. The *no competition* case performs well with random dots and Yosemite, but does not perform well with the OpenGL stimuli. With the exception of the *no competition* case, these results match data showing that humans are capable of accurately determining heading to within 1-2 degrees accuracy when there is no rotation in the stimulus (Warren & Hannon, 1990). The results are also within the 1-3 degree range required for obstacle avoidance behaviors (Cutting, Springer, Braren, & Johnson, 1992). It should be noted that by modifying the inhibitory gain parameter F_1 (Appendix equation 1.2) it is possible to improve results on some stimuli. For example, the best obtained performance on the OpenGL stimuli was with *orthogonal-opponent* competition (Appendix equation 5.7) and parameter F_1 equal to 10, resulting in a mean error of 0.7 degrees. However, this parameter value did not produce as good results on the Yosemite sequence indicating, perhaps, a need for parameter tuning to the specific input source.

Fast rotations impair performance. The model was presented with 36 random dot stimuli containing simulated rotations in addition to translation. There were 9 stimuli for each of four groups: *ground plane, 3D dot cloud, 2m frontal plane,* and *8m frontal plane.*

As noted above, the 9 stimuli were defined with the following rotation rates: ±10, ±5 ±2.5, ±1, 0 degrees per second. In accord with human data, for all random dot stimuli, as rotation rates increase, the heading error increases (Banks et al., 1996; van den Berg, 1993; Royden et al., 1992; Royden et al., 1994; Warren & Hannon, 1990). As stated in the introduction, there are a number of inter-observer differences and inter-study differences, therefore only a qualitative analysis is given below. Results for *distributed-opponent* competition for all simulated rotation stimuli are shown in Figure 3.1.

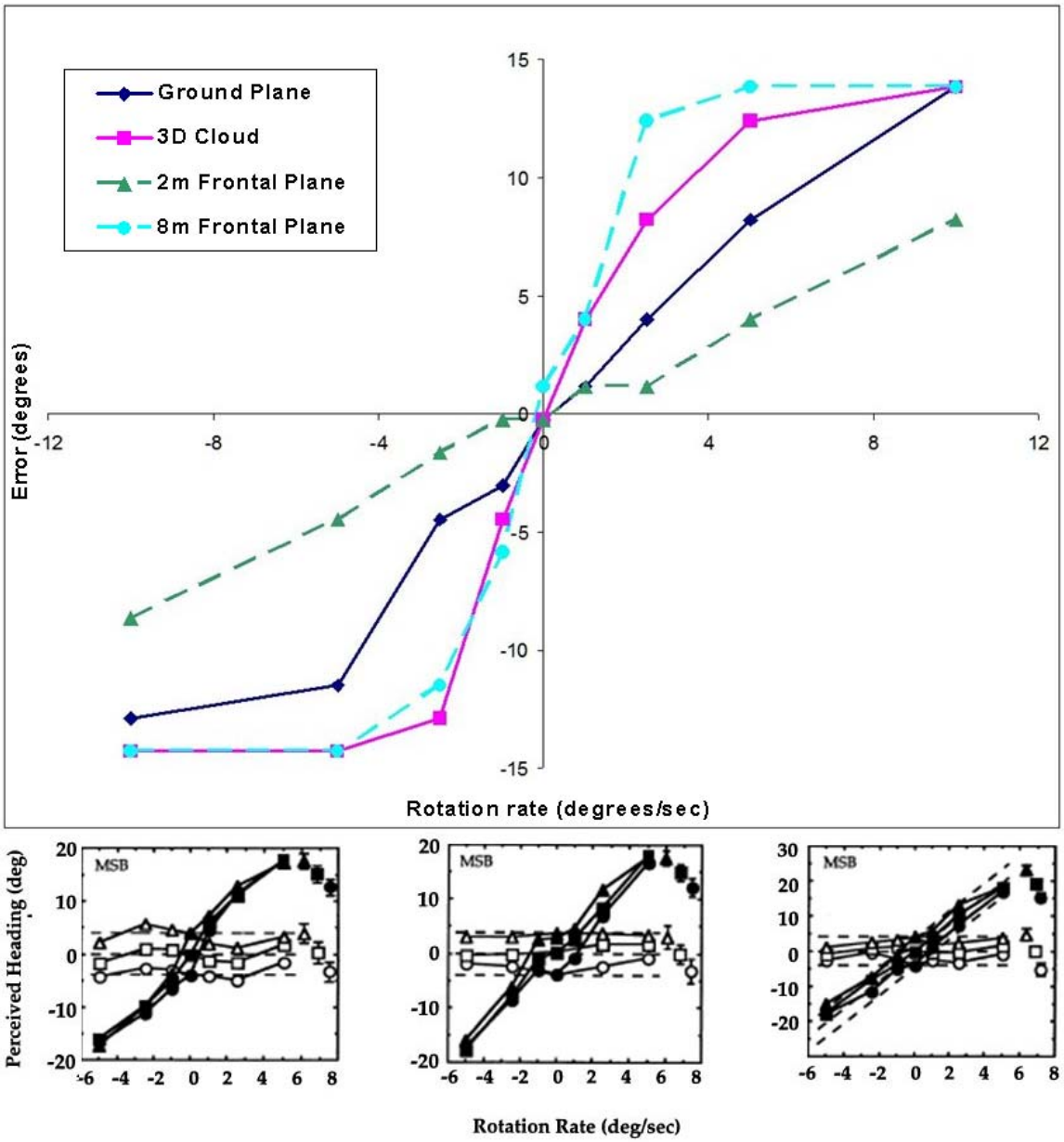

Figure 3.1. Top: Mean error by rotation rate for *distributed-opponent* competition in model MT^+ across four stimuli sets: ground plane, 3D dot cloud, 2m frontal plane and 8m frontal plane. In all cases, error increases as rotation rates increase. Rotation rates of 1 degree per second produce errors in the same

range as when no rotation is present. Bottom: Human heading data from Royden et al. (1994) for subject MSB. Black symbol lines show performance during simulated rotation, for 3 different headings: -4°, 0° and 4°, as depicted by triangle, square and circle, respectively. White symbol lines show performance in the presence of real eye rotations for the same 3 headings. Left: ground plane. Center: 3D dot cloud. Right: frontal plane.

The *no competition* network performed badly on the ground plane, producing large errors even for small rotations. With the exception of the *no competition* case, all results show the same pattern. Inspection of the *distributed-opponent competition* results, shown in Figure 3.1, show that the ground plane, 3D dot cloud, and 8m frontal plane all produce errors which increase rapidly as the rotation rate increases. We based all rotation simulations on a zero degree heading with a 30 degree field of view; therefore the maximum possible error is ±15°. Maximum error occurred for rotation rates ≥ 5° per second in both the 8m frontal plane and 3D dot cloud. Ground plane performance deteriorates slightly less rapidly due to the limits of resolution in the input stream. Rotation has a larger effect, relative to translation, on far depths (Longuet-Higgins & Prazdny, 1980). In our ground plane stimuli, dots at far depths tend to overlap to produce a horizon line; see Figure 2.1. This better performance is therefore an artifact of the low resolution, 256 x 256 pixels, of the input, in relation to the dot density, rather than any intrinsic model preference for ground planes. Note that humans viewed stimuli with the same dot densities, visual angle, etc., but at a higher resolution. In order to make computation tractable for a dynamical model of this complexity, we were unable to process stimuli with the same resolution as humans.

For the 2m frontal plane, ViSTARS is able to determine heading within a 5 degree error during rotations of up to 5 degrees per second. As noted above, rotations affect far depths more than close depths. The 2m depth plane is only moderately affected by rotations in the range tested and, as such, the focus of expansion is only moderately shifted. These simulated rotation results, for ground plane, 3D dot cloud, and frontal planes, are a good qualitative match for human psychophysical data (Banks et al., 1996; van den Berg, 1993; Royden et al., 1992; Royden et al., 1994; Warren & Hannon, 1990).

These model heading estimates are consistent with finding a *shifted* FoE in the global motion estimate, and thus an estimate of curvilinear heading rather than instantaneous heading, even though the model is implemented with only radial heading templates. These results indicate that ViSTARS is capable of finding the focus of expansion in motion patterns that also contain rotation information.

Robust in the presence of noise. Gaussian noise was added to the Yosemite sequence, as defined in the Methods section. When using *no competition,* the system could not consistently determine heading in the presence of this noise. Results for 9 signal to noise ratios (SNR, as defined in the Methods section) for *opponent*, *distributed-opponent* and *orthogonal-opponent* competition are shown in Figure 3.2. *Opponent* competition is capable of handling inputs with an SNR greater than 3, after which error increases rapidly. *Distributed-opponent* and *orthogonal-opponent* competition are better able to cope with noise, producing only small errors for SNRs greater than 1.5. For SNRs less than 1.5, errors increase rapidly.

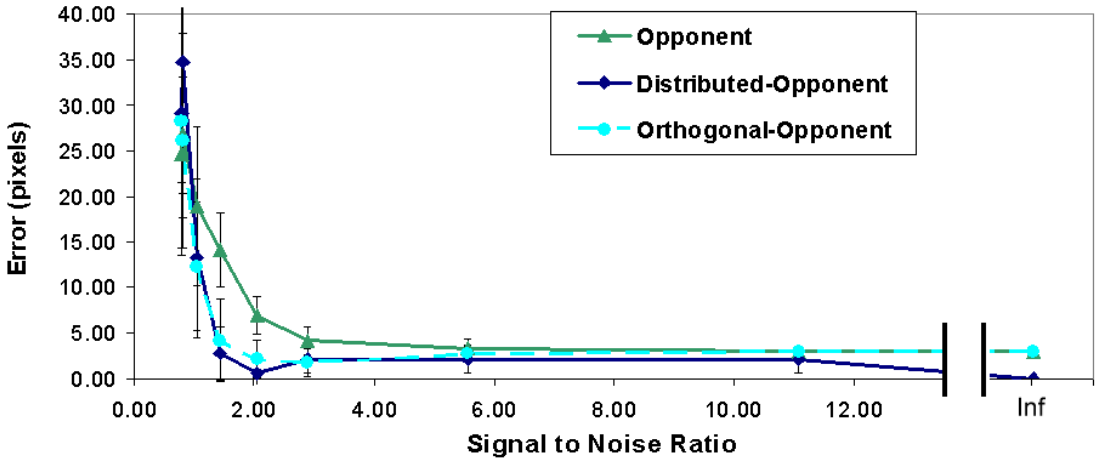

Figure 3.2. Error in pixels vs. Signal to Noise Ratio (SNR) for model processing Yosemite sequence with Gaussian noise added. *Opponent*, *distributed-opponent,* and *orthogonal-opponent* competition in MT^+ are shown. Note that the ground truth for the Yosemite sequence is given in pixels, and therefore error is measured in pixels. Mean values are shown. Error bars depict the standard deviation over 10 trials. Noise was randomly generated for each trial.

In order to specifically investigate the robustness of the MT^+/MST*d* processing loop (Appendix equations 5.1 – 6.2), a second set of simulations was performed. A motion pattern was created for self-motion towards a zero degree heading. This motion pattern was corrupted with Gaussian noise. This noisy motion pattern was input directly to MT^+ (Appendix equation 5.1). Under these conditions, all competition scenarios in MT^+ were able to handle high levels of noise in the input. There was zero error with SNRs greater than 0.26, but for SNRs less than 0.26, error increased rapidly. By an SNR of 0.25, error was in the range 15-20 pixels. Adding Gaussian noise to a vector-based motion pattern results in an equal number of corruptions in every direction. In our scheme, noise survives for only a single frame of the input, after which a new noise signal is generated. The long-range directional filter in MT^+ (Appendix equation 5.2) performs spatiotemporal integration of motion vectors, thereby cancelling out random noise until such point as the noise completely overwhelms the signal. Systematic, as opposed to random noise, could be addressed by a retuning of templates since it should be somewhat predictable by definition; see the Discussion section. The combination of the long-range filter and competition between directions in MT^+ with feedback from MSTd allows the model to produce accurate heading and motion estimates even in the presence of high levels of noise.

Good performance on driving video. The 25 video clips were split into 3 categories, left, straight and right. This classification split was performed on the basis of the maneuver being undertaken: turns, bends and lane changes were considered rightward or leftward. All other maneuvers were considered straight ahead. The video clips contain examples of maneuvers at speeds between 15 mph and 65 mph. It is important to restate that at higher translational speeds, higher rotation rates are required to affect the position

of the FoE (Longuet-Higgins & Prazdny, 1980). The model processed the video clips and the results were compared with this 3 class grouping.

No competition in MT^+ was unable to reliably produce a heading estimate when presented with these video clips; all other forms of competition produced similar results.

The horizontal pixel resolution of MT^+ for the driving clips was 90 pixels. For the 3 class problem: any MST*d* cell corresponding to the left-most 36 pixels was considered leftward vehicle-motion, any MST*d* cell in the central 18 pixels was considered straight vehicle-motion, and any cell in the right-most 36 pixels was considered rightward vehicle-motion. The model achieved a correct classification rate of 96% (vs. chance at 33%). The misclassified clip was a rightward lane change which the model classified as straight ahead. Given this good performance, we proceeded with a 7 class task. In the 7 class task, video clips were classified into more detailed maneuver groupings. Maneuvers were considered *turns* when the car turned at a junction, *bends* when the car turned a bend in the road, and *lane changes* when the car changed from one lane to another on a multi-lane road. The 7 classes were defined as follows, with corresponding pixel positions shown in parenthesis: left-turn (1-4), left bend (5-20), left lane change (21-36), straight ahead (37-54), right lane change (55-70), right bend (71-86), right turn (87-90).

These class-pixel correspondences were chosen based on the authors' judgement of the source driving video. While this is a merely qualitative method, it provides a conservative way to quantify the performance of the system as a proof of concept. There is little human data on the accuracy of heading estimates in response to driving video. If one were to use a human study to assess a class-pixel correspondence, one would also have to deal with large discrepancies between respondents (Banks et al., 1996; van den Berg, 1993; van den Berg & Brenner, 1994; Ehrlich, Beck, Crowell, Freeman, & Banks, 1998; Kaiser, Perrone, Stone, Banks, & Crowell, 1993; Royden et al., 1992; Royden et al., 1994; Warren & Hannon, 1990).

On average, model heading estimates were within 3 degrees of the center of the class to which they were assigned. Figure 3.3 shows results for *opponent*, *distributed-opponent,* and *orthogonal-opponent* competition in MT^+. Accuracy was not dependent on the type of maneuver or the speed of the vehicle. Table 2 is a confusion matrix showing results for *distributed-opponent competition*.

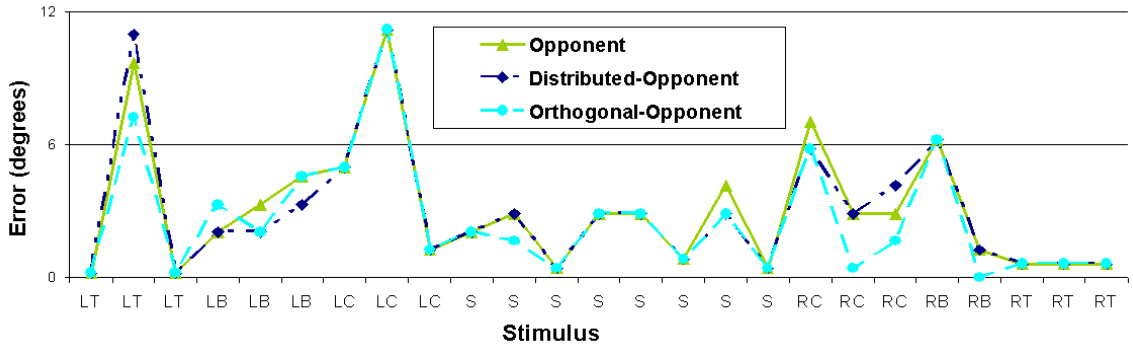

Figure 3.3. Error in degrees for the 25 driving video clips, the 7 class grouping of each clip is shown on the x-axis. The 7 classifications are: left turn (LT), left bend (LB), lane change to the left (LC), straight (S), lane change to the right

(RC), right bend (RB), and right turn (RT). See text for a more detailed description of the classifications. O*pponent, distributed-opponent* and *orthogonal-opponent* results are shown.

Predicted / *Actual*	*Turn left*	*Left bend*	*Left lane change*	*Straight*	*Right lane change*	*Right bend*	*Turn right*
Turn left	**2**	0	1	0	0	0	0
Left bend	1	**2**	0	0	0	0	0
Left lane change	0	2	**1**	0	0	0	0
Straight	0	0	0	**8**	0	0	0
Right lane change	0	0	0	1	**2**	0	0
Right bend	0	0	0	0	1	**1**	0
Turn right	0	0	0	0	0	0	**3**

Table 2. Confusion matrix showing classification of video-clips using distributed-opponent competition. 19 out of 25 clips were correctly classified, 5 were classified to be in a neighboring class. Only 1 clip was classified neither correctly nor in the neighboring class, in this case the direction (left) was correctly determined.

In the 7 class task, the model achieved a correct classification rate of 19 out of 25, or 76% (vs. chance at 14%). Of the 6 misclassifications, 2 were within 5 degrees of the center of the correct classification, 2 were between 5 and 7 degrees from the center of the correct classification. 2 were over 10 degrees away from the correct classification. With distributed-opponent competition, all but one video clip is classified within either the correct class or an immediately adjacent class. Lane change maneuvers accounted for half of the misclassifications. The misclassifications were equally split between over and under estimates. These results, obtained using the same parameter set as all other stimuli sets, indicate that the model has strong potential for further development on live video streams.

Discussion

As demonstrated above, the ViSTARS model builds on prior work to produce an integrated neural model capable of processing natural image sequences and estimating heading in a human-like fashion. Motion estimates from the lower levels of the model (levels 1-3) are highly affected by the aperture problem and other sources of noise. We have demonstrated that heading estimates, and to a lesser extent global optic flow estimation, in ViSTARS are highly robust to this noise. This noise tolerance was demonstrated to be due to the combination of spatiotemporal averaging, in the long-range filter and competitive interactions within MT^+. A predicted feedback path from MSTd to MT^+ further refines the estimates by selecting motion signals that are congruent with the current heading estimate. The dynamics of this recurrent processing loop cause bottom-up motion signals from level 4 to be refined over time: signals that persist in the receptive field of the long-range filter accumulate evidence and suppress the spurious

sporadic signals caused by noise. This MT^+- MSTd feedback loop for the computation of heading is computationally homologous to the MT^--MSTv feedback loop that has been used to explain how the brain solves the aperture problem for computing object motion direction (Berzhanskaya, Grossberg, & Mingolla, 2007; Browning, Grossberg, & Mingolla, 2009; Grossberg, Mingolla, & Viswanathan, 2001). These two parallel circuits MT^+-MSTd and MT^--MSTv are part of the complementary organization of the visual cortex for carrying out optic-flow navigation and object tracking, respectively (Grossberg, 2000).

The ViSTARS model does not fit neatly into any of the standard categories for the computation of optic flow. However, it may useful to describe our model in the language used for alternative computational methods. For a review of optic flow computation methods, see Beauchemin and Barron (1995). Here we will compare our model to other biologically-inspired models of optic flow and heading estimation.

Our measure of optic flow *quality* is dependent on the behavioral consequences of the representation. Model global motion estimates are considered accurate if they generate accurate heading estimates. We do not deny the importance of constraints such as occlusion, object motion, and transparency for other tasks. For example the 3D FORMOTION model (Berzhanskaya et al., 2007; Grossberg et al., 2001) explains object motion, occlusion, and transparency percepts using motion representations. Our results suggest that detailed computation of occlusion and transparency may be unnecessary for the accurate calculation of heading.

Reliability and confidence in the model are measured by how *peaky* the heading estimate is; see Figure 1.6. Peakyness was not calculated in this analysis. In the case of a Gaussian distribution, peakyness is synonymous with its standard deviation. More generally, it is a measure of the deviation in the distribution from the maximum peak of the distribution.

A clear winner in MST*d* requires that there be a high degree of confidence in the optic flow representation, or more specifically a low degree of directional ambiguity at each spatial position represented in MT^+.

As discussed in the Introduction, ViSTARS is a template model, and a neural network model, as well as a *temporal refinement* model (Black & Anandan, 1991; Fleet & Langley, 1995; Singh, 1990), the quality of whose representation is refined over time. Model retinal stages perform a *spatial and temporal differentiation* of the input that produces a representation of moving contours. This use of spatial and temporal differentiation is different from *gradient* and *differentiation* methods. These stages accomplish feature tracking, since the temporal differentiation is applied after the spatial differentiation and not directly on the image.

The directional transient cell level (Figure 1.1, Level 3) followed by the directional short-range filter and competition (Figure 1.1. Level 4) can be considered a form of correlation-based matching with penalty term. A more mechanistically correct description of these processing stages is as a spatially-short-range filter that begins to accumulate directional evidence which selectively enhances informative feature tracking signals while attenuating ambiguous motion direction signals. The dynamics of the model as a whole helps to explain a wide range of psychophysical percepts, including, but not limited to: aperture problem, chopsticks illusion, rigidity of rotating ellipses, motion coherence, barberpole illusion, spotted barberpole illusion, triple barberpole

illusion, occluded translating square illusion, motion transparency, gamma motion, beta motion, and second-order motion illusions (Baloch et al., 1999; Berzhanskaya et al., 2007; Chey et al., 1997; Grossberg et al., 2001).

The directional long-range Gaussian filter in level 5 (Figure 1.1) of model MT^+ spatiotemporally groups and smoothes the local motion estimates of the model which, when coupled with feedback from the heading estimate in level 6 of model MST*d* provides a coherent global motion estimate that is consistent with the estimated heading. Thus, the optic flow representation in model MT^+ is not an estimate of instantaneous motion. Rather, it is an estimate of the flow of motion over time.

The elaborated Reichardt detector (ERD, van Santen & Sperling, 1985) and energy models (Adelson & Bergen, 1985) produce local motion estimates and are a natural alternative to our retina-V1 processing stages. Both the ERD and energy models fit various data, have physiological interpretations, and are consistent with space-time plots of direction-selective V1 cells in cat (DeAngelis et al., 1995). The ViSTARS, ERD, and energy models all start with spatial differentiation followed by temporal differentiation. The ERD and energy models are equivalent to each other under some circumstances (Adelson & Bergen, 1985). However, our directional transient cell network differs from the ERD or energy models. ViSTARS builds upon the Barlow and Levick (1965) model of directional transient cells in rabbit and shows how to refine their model to retain sensitivity to multiple speeds (Grossberg et al., 2001). It hereby clarifies how the motion system can respond to stationary transients, and how such transients can elicit motion percepts; c.f. Grossberg & Rudd, 1992. Finally its two cell layers of have direct homologs with physiologically observed neurons and inhibitory interneurons that are involved in direction selectivity *in vivo* (Fried et al., 2002; Livingstone, 1998).

The model of Bayerl and Neumann (2004) produces accurate optic flow estimates from natural image sequences. Their model implements a recurrent loop between MT and V1 to explain how the aperture problem is resolved in MT (Pack and Born, 2001). There are some parallels between the Bayerl and Neumann model and ours, but there are also significant differences: Their model used elaborated Reichardt detectors (ERD; van Santen & Sperling, 1985) to produce local motion estimates. Their model also solves differential equations, describing V1 and MT cell activity, at equilibrium and then applies an iterative computation to explain the time-course of aperture resolution. We use a dynamical system to explain the time-course of all cells, including how the aperture problem is overcome in MT, which predicted the data of Pack and Born (2001) (Chey et al., 1997). Our implementation is agnostic as to whether or not an equilibrium state is achieved. Numerical integration techniques attempt to broadly mimic the temporal dynamics that occur *in vivo*. Whether or not a cell in any given position and stage of the model reaches an equilibrium state depends on the spatiotemporal properties of the stimulus.

Both models utilize cross-directional normalization in the motion stream as a form of spatial feature enhancement in V1, as has been used in prior models (Chey et al., 1997), and produce a global motion estimate in MT. The Bayerl and Neumann model focuses on the production of an accurate instantaneous vector field representation of optic flow, whereas our model produces a temporally evolving global motion estimate for the purpose of determining heading. As with the ERD and energy models discussed above, we believe that our model offers a more complete description of the neural circuitry and

dynamics of primate magnocellular pathway, and its ability to control navigation based on optic flow.

Calow et al (2005) presented a biologically motivated model of heading estimation which is capable of detecting heading from video to within 1 degree accuracy. Their model uses space-variant filtering; central regions are processed in more detail than peripheral regions. This is analogous to log-polar mapping and cortical magnification in primate visual systems, which produce a space-variant representation of the world (Schwartz, 1977). The Calow et al. (2005) algorithm was tested on only two video sequences. In both cases, the translational heading was straight ahead, the roads were mostly empty, and there were relatively few vehicles in the scenes. How the model performs on more complex driving sequences remains unclear. Performance of their space-variant filter model was slightly better than for a model using the same sized filter for all regions of space. There are two possible reasons for this: central region motion vectors are more accurate and should be weighted more heavily, or large filters produce beneficial spatio-temporal smoothing. Wagner, Polimeni & Schwartz (2005) claimed that peripheral motion vectors provide the most robust representation of ego-motion, thereby suggesting that the use of large filters and/or peripheral motion could be more important than the weighting of central regions.

Computing a log-polar map is computationally expensive and reduces the information available when processing video streams generated by a standard camera. At the resolutions processed in this article (256x256, 320x240), the information loss from a log-polar transform is significant. Given our desire to demonstrate the technological capabilities of the model, and computational constraints when implementing such a large dynamical system, we decided to forgo a log-polar mapping. Our results indicate that a log-polar mapping is not needed to fit the data considered herein. However using a log-polar mapping may allow the model to more completely match human steering data (Browning, Grossberg, & Mingolla, 2008a, 2008b; Elder et al., 2008; Mingolla, Browning, & Grossberg, 2008). Data from van den Berg and Brenner (1993, 1994) show humans can accurately determine heading from ground plane stimuli with simulated rotation rates up to 5 degrees per second if the fixation point is on the ground plane. Grossberg et al (1999) demonstrated that spiral MST*d* cells described in log-polar coordinates can replicate these data. Future work will therefore extend the model by incorporating space-variant filters to embody the log-polar cortical mapping.

As noted in the Introduction, there are three main classes of biological models for heading estimation: *differential motion, template,* and *decomposition* models. *Differential motion* models (Rieger & Lawton, 1985; Hildreth, 1992; Royden, 1997, 2002; Royden & Hildreth, 1996) and *decomposition* models (Heeger & Jepson, 1992; Lappe & Rauschecker, 1994, 1993) take great pains to remove eye rotations from retinal flow before estimating heading. However, as also noted in the Introduction, humans are affected by rotations in the optic flow field if they are not due to a real eye movement. Human heading estimations decrease in accuracy in proportion to simulated rotations added to the flow field (Banks et al., 1996; van den Berg, 1993; van den Berg & Brenner, 1994; Ehrlich, Beck, Crowell, Freeman, & Banks, 1998; Kaiser, Perrone, Stone, Banks, & Crowell, 1993; Royden et al., 1992; Royden et al., 1994; Warren & Hannon, 1990). Indeed, most studies note that naïve observers tend to report curvilinear heading rather than "accurate" instantaneous heading. Our model is a kind of *template* model (Perrone

& Stone, 1994), albeit one that uses biologically plausible adaptive flow filters, and fits the above behavioral data well with simulated rotation of the optic flow field. Eye movement sensitive gain fields have been shown to efficiently deal with *real* eye rotations in template models (Beintema & van den Berg, 1998; Elder et al., 2008). Consistent with such a mechanism, some primate MST*d* cells have responses that are modified during pursuit eye movements (Page & Duffy, 1999). Although these data are consistent with a template model with gain fields, they are hard to explain using models that produce a rotation-free optic flow representation before inputting to MST, such as *decomposition* or *differential* models.

There are, however, complicating factors suggesting that humans do not completely ignore the issue of rotations in the optic flow field. A number of researchers have demonstrated the importance of the instructions given to naïve observers. During simulated rotation studies, when instructed to report instantaneous heading, observer accuracy is higher than when given neutral instructions (Stone & Perrone, 1997b; Li & Warren, 2000, 2004). These data suggest that using ocular motor information to remove effects of rotation in MT/MST may not be sufficient, on its own, to explain human behavioral data. Li and Warren (2004) took this analysis further and demonstrated that environmental structure can allow observers to make more accurate instantaneous heading estimations in the presence of simulated rotations. Our model, as it is currently specified, does not explain these data. A wider range of templates in MSTd, including some that contain rotational components, such as planar or spiral motion cells, may provide enough information to explain these data. However, it is important to note that higher-level cognitive processing, notably task-selective focusing of spatial attention (c.f., Fazl, Grossberg, and Mingolla, 2008), may also be involved. We did not implement planar or spiral templates because radial cells sufficed to explain key human psychophysics data concerning heading estimation. See Elder et al. (2008) for how this can be done using log-polar preprocessing. If planar or spiral flow filters are added, the model will continue to function due to the normalization constants (M_6, N_7) in Appendix equations 5.1 and 6.1.

Differential motion models can be related to ON-center OFF-surround motion processing cells in MT^- (Royden, 1997, 2002). As noted above, MT^-/MST*v* processing is concerned with object motion and tracking, whereas MT^+/MST*d* processing is concerned with navigation, including the computation of optic-flow based heading (Duffy & Wurtz, 1995; Graziano et al., 1994; Saito et al., 1986). Cutting and Wang (2000) suggested a strategy by which MT^- (differential motion) cells could be used to determine heading. This strategy matches heading against the differential motion representation. It applies ON-center OFF-surround motion filters to analytically computed optic flow and then determines heading from this differential motion estimation using MST*d*-like operators. The Cutting and Wang model (2000) highlights a difference between approaches that take single cell properties and link them arbitrarily to produce behavioral functions, versus system models that attempt to understand the organizational principles and functional architectures that underlie demonstrated brain dynamics. In the present case, it misses the key fact that MT^+/MST*d* and MT^-/MST*v* compute *complementary* properties (Grossberg, 2000). Cutting and Wang (2000) do not attempt to explain the sub-division of processing between MT^+/MST*d* and MT^-/MST*v*, and their model is difficult to reconcile with these sub-divisions.

Differential motion models offer the best explanation of human data showing misperceptions of heading in the presence of large moving objects (Royden & Hildreth, 1996, Royden, 2002; Warren, 1998). Moving objects only significantly affect heading perception when they cover a large part of the visual field, and they affect heading differently depending on their direction of motion. The effects of these objects cannot be explained by simply pooling their motion vectors with those due to translation, nor by removing their motion vectors from the heading estimate (Royden & Hildreth, 1996; Warren, 1998; Warren & Saunders, 1995). In this case, it may be that object-based MT^- signals dominate perception. Whether or how these MT^- signals are incorporated into MST*d* heading estimations, or whether some other brain area additionally performs heading estimation on the basis of this information is unknown. Pack et al. (2001) developed a model of object tracking and predictive pursuit, on which the ViSTARS model builds, which incorporates feedback interactions between MST*v* and MST*d* to maintain accurate pursuit of objects during tracking. The Pack et al. (2001) framework should be extensible to explain the influence of large objects on heading perception by allowing the interactions between MSTv and MSTd to affect heading estimates.

A major criticism of template models is that a huge number of templates may be required to fully explain motion patterns that include arbitrary combinations of 3D structure and eye rotation data at any speed (Lappe, Bremmer, & van den Berg, 1999). However, human data indicate that heading may be estimated from a simplified 2D representation of space (Li et al., 2006) and humans do not, in general, deal well with rotations in the flow field unless accompanied by real eye rotations (although see discussion above). Templates that correspond to different speeds are not necessary for heading detection (Grossberg et al., 1999) since direction of motion, not rate, defines heading. The ViSTARS model uses a relatively small number of templates ($\leq$60) which are able to match human behavioral performance on heading detection tasks and a variety of input stimuli. These templates are distributed across two rows of the input. The rows do not correspond exactly with the vertical positions of the focus of expansion in the input stimuli, but are sufficient for accurate heading detection. We do not suggest that two rows of templates are sufficient for all input types. For example, humans can vary the position of the horizon across the full vertical range, whereas a vehicle-mounted camera will likely have a more limited range. We do, however, claim that it is possible to accurately estimate heading with a vastly reduced distribution of heading-sensitive cells, and that the density of the cells should be based on the distributions of likely heading positions in the environment. A simplified representation of space and a small number of templates are thus sufficient to explain key human data, in addition to being computationally efficient.

A number of approaches have been proposed for calculating heading for autonomous robotics. Most use feature tracking, sometimes with stereo vision, in concert with a global positioning system (GPS) and inertial sensor-based data (Agrawal, Konolige, & Bolles, 2007; Alenyà, Martinez, & Torras, 2004; Olson, et al., 2003). Autonomous robots are assessed in terms of their ability to navigate from some start point to a goal while avoiding obstacles. This use of behavioral measures of the quality of representation is consistent with our own approach. While these robotic methods for heading estimation can be effective, they are not intended to, and do not, model the biological mechanisms of visual processing. They do demonstrate that integrated

information from multiple sensor types has real engineering value, further highlighting the importance of large-scale integrated systems models encompassing many sensory modalities. However, it is unclear how well these algorithms will work with loss of one or more input modality, for example without the GPS signal or with a single camera. Biological systems, by their very nature, have multiple redundant processes and, due to the noisy nature of neurons, tend to use noise-tolerant distributed representations. Our model provides proof of concept that low-resolution, monocular video processing can be sufficient for reasonably accurate heading detection, and hereby makes a case for developing robotic navigation applications from biological models.

When ViSTARS processed driving videos, it was on average accurate to within 3 degrees, with 84% of video clips classified within 5 degrees. We assigned video clips to classes on the basis of how the car maneuvered on the road. While not arbitrary, this classification task is an imperfect method for assessing model performance, since cars have an instantaneous heading which is in the direction that the car is facing. Any system that attempts to determine car heading is therefore trying to assess *curvilinear* heading. It is possible with GPS and/or inertial measurement devices to ascertain ground truth to high precision for moving vehicles. However, a behavioral measure of performance may be more informative. For navigation, heading estimates need to be accurate *relative* to *object position* in *visual coordinates*. Obstacle avoidance and reactive navigation require a positional representation of heading, goal and obstacles in the same frame of reference. GPS and/or inertial measurements can accurately determine heading in a *global* frame of reference. To be useful for obstacle avoidance, this global heading must be mapped to a visual frame of reference, or goal and obstacle positions must be mapped to a global frame of reference. By detecting heading in the visual frame of reference, as it is in the current model, the task of comparing heading with obstacle and goal position is simplified. No additional mappings are required, and heading, goal and obstacle positions can be directly compared. Thus, although GPS and inertial motion sensors may be useful for calibration, it seems likely that a predominantly visual solution for detecting heading and object position will allow for the use of much simpler steering decision systems, as in the Steering, Tracking and Route Selection (STARS) model (Elder et al., 2008).

The current model is designed to feed into the STARS model (Elder et al., 2008) of obstacle avoidance and goal approach. The STARS model demonstrates how interactions between visual representations of object position and heading result in obstacle avoidance and steering strategies. The STARS model estimates heading, goal, and obstacles as distributed activation patterns. By summing the distributed positional representations, the STARS model steers around obstacles towards a goal in a manner similar to humans performing the same task. Our representation of heading in MST*d* provides a distributed positional representation that naturally feeds into STARS. A companion article develops model MT^-/MST*v* for the purposes of object segmentation and localization, and uses a distributed positional representation for integration with the STARS model (Browning et al., in 2008b).

Our results show that model MT^+ is better able to resolve a global motion estimate when there is some form of competitive dynamics in MT^+. This finding supports prior modeling research demonstrating that inhibitory connectivity is beneficial for heading estimation (Beardsley & Vaina, 2001; Pack, et al., 2001; Royden, 2002).

Utilizing *no-competition* (Appendix equation 5.4) in MT^+ (Appendix equation 5.1), produced a system that was less able to cope with noise and was unable to produce heading estimates for the driving video clips. Our model thus supports the claim that MT^+ requires some form of competition, specifically to ensure stability and robustness under a wide range of possible inputs. The nature of this competition *in vivo* is unclear, although *opponent* competition has been previously postulated (Born & Bradley, 2005). We tested 3 forms of competition. Each behaved in a similar manner. These results suggest that random dot stimuli, which are common for optic flow and monkey studies, are unsuitable for probing the nature of competition in MT^+. The signals in such stimuli are sparse and easily segmented, providing little challenge for our model and, by extension, for the human or primate visual system. The more advanced virtual reality (VR) displays utilized by more recent heading studies (Fajen & Warren, 2003; Li & Warren, 2000, 2004; Li et al., 2006; Loomis & Beall, 1998; Tarr & Warren, 2002; Warren & Fajen, 2004) offer a more useful probe, but again this may depend on the actual content of the virtual reality stimulus. In our work, we demonstrate that the Yosemite sequence produces relatively unambiguous motion estimates and, as a result, the *no-competition* scenario in MT^+ is able to determine heading as accurately as *opponent* and *orthogonal-opponent* scenarios. In contrast, the computer-generated terrain that we created provided a much tougher challenge for the *no-competition* case, with an average error of 4.7 degrees. All other competition scenarios produced an average error of less than 1.5 degrees on the same stimuli. Our OpenGL terrains were specifically created to have a relatively homogeneous visual appearance to make accurate assessment of motion difficult. This indicates that the importance of competition in MT^+ appears to be related to the difficulty of accurately resolving local motion in the input sequence.

Model robustness was demonstrated using inputs containing high levels of Gaussian noise. Appropriate types of spatiotemporal integration throughout the model eliminate the effects of this noise. Spatiotemporal integration will not help in the case of systematic noise in the input stream. If the noise produces systematic bias in motion estimates, then the templates can be adaptively recalibrated to account for the bias. It should be possible to deal with context-dependent systematic motion bias using gain fields in much the same way as the effect of eye movements can be mitigated (Beintema & van den Berg, 1998; Elder et al., 2008).

In summary, the ViSTARS models the primate motion pathway to show how a relatively sparse heading map across space and a reasonably accurate global motion estimate are sufficient for human-like performance on heading tasks. The model is robust to high levels of noise in the input stream and is capable of performing on real-world video as well as psychophysical stimuli. The model may also be useful in technological applications, such as navigational control of a mobile robot to carry out vision-based tasks.

Acknowledgements

NAB, SG and EM were supported in part by CELEST, an NSF Science of Learning Center (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624). SG and EM were supported in part by the SyNAPSE program of DARPA (HR0011-09-3-0001, HR0011-09-C-0011). NAB and EM were supported in part by National Science Foundation (NSF BCS-0235398). EM was also supported in part by the National Geospatial Intelligence Agency (NMA201-01-1-2016).

Appendix. Model equations, parameters, and implementation

All stages of the model are defined by differential equations and are numerically integrated using Euler's method with a time-step of 0.1. Each frame of video input is presented for 10 time steps and then the next frame of input is presented on the subsequent time step. We define the frame rate of the input at 15 frames per second. This calibrates each integration time step at roughly 7 ms in simulated time. Computation time for the same time step in MATLAB is roughly 2.5 seconds on a dual 2Ghz AMD Opteron (AMD, 2003) based workstation with 8Gb of RAM running Microsoft Windows XP x64 (Microsoft, 2003). Figure 1.1 describes the functional stages of the model with respect to their equation numbers and variable labels.

Stages of the model are designed to elucidate the processes by which biological neurons perform their calculations. Each differential equation specifies the activation state of individual neurons or populations of neurons. Model cells are typically controlled by shunting, or membrane, equations (Grossberg, 1968, 1973; Hodgkin, 1964; Sperling, 1970) that perform a leaky integration of inputs. Equation (0.1) defines a shunting equation wherein x represents cell activity in response to excitatory inputs E and inhibitory inputs I:

$$\frac{dx}{dt} = -Ax + (B - x)E - (C + x)I \,. \tag{0.1}$$

In Equation (0.1), parameter A determines the *decay rate* of the cell; B determines the upper bound, or excitatory saturation point, of x; E is the excitatory input; C determines the lower bound, or inhibitory saturation point, of x; and I is the inhibitory input.

Signal functions define how cell activity generates an output signal. Common signal functions include half-wave rectification, squaring and sigmoid functions. Half-wave rectification is denoted by $[x]^+ = \max(x,0)$. The output may be interpreted as the firing rate of a neuron.

Index	Meaning
ij	Spatial position
p	ON / OFF channel
s	Scale (speed)
d	Direction
z	Heading cell index
XY	Dummy index for spatial position
D	Dummy index for direction
ε	Dummy index for heading cell

Table A.1. Indices used in model equations.

In the equations that follow: lowercase letters correspond to variables, whereas uppercase letters correspond to output signals. For example, r corresponds to an activity

representing MSTd, whereas R corresponds to output signal from MSTd. Subscript indices correspond to spatial position. Superscript indices correspond to non-spatial dimension values, such as speed or direction. Uppercase indices correspond to the use of a dummy index. Parameters are labeled as uppercase letters with numerical subscripts. When equations have been previously published, variables and indices have been labeled consistently wherever possible to make cross-referencing easier. In cases where following these conventions makes an equation ambiguous or confusing, Greek letters are used. Variable labels and related descriptions are listed in Table A.1, index labels and definitions are listed in Table A.2. Parameters and their values are listed in Table A.3.

Eqn	Variable	Dimensions	Function
	g	ij	Grayscale pixel value of video input, scaled between 0 and 1.
0.2 0.3	I	ij, p, s	Resized video input pixel values at multiple scales with ON and OFF channels.
1.1 1.3	a y	ij, p, s	ON-center-OFF-surround processing of video input, finds spatial features for tracking, output (y) through sigmoid signal function.
2.1 2.2 2.3	x z b	ij, p, s	Sustained response neuron. Depleting neuro-transmitter. Half-wave rectified, gated output responds to changes in the input stream.
3.1 3.2 3.3	c e E	ij, p, s, d	Directional interneuron. Directionally selective transient cell. Local motion estimate with speed and direction.
4.1 4.2	f m	ij, s, d ij, s, d	Cross directional normalization enhances the least ambiguous motion directions producing motion based feature enhancement. Resizing of feature enhanced local motion estimate (f) for comparison across speeds.
5.1 5.3	q Q	ij, d	Long-range filter produces global motion direction estimate from local motion estimate (m). Modulatory feedback from heading estimate (R) and directional competition refines global estimate. Output (Q) through square function.
6.1 6.2	r R	z	Performs template pattern match on global motion estimate (Q) to determine heading. Competition across heading cells normalizes activity. Output (R) through sigmoid function.

Table A.2. Model variables and descriptions.

Input (g). Input, g_{ij}, is converted to grayscale and scaled between 0 and 1. OpenGL and random dot stimuli resolutions were 256 x 256 pixels, the Yosemite stimulus resolution was 316 x 252, and the driving video stimuli set resolution was 360 x 240. Driving video sequences consisted of 15 frames. All other sequences consisted of 14 frames.

Eqn	Parameter	Value	Function
1.1	A_1	0.001	Decay rate
	B_1	1	Upper bound of activation
	C_1	2	Excitation scale factor
	D_1	0.25	Lower bound of activation
1.2	F_1	10.225	Inhibition scale factor
	σ_1	1	Inhibitory kernel variance
1.3	G_1	0.031623	Sigmoid function scale factor
	φ_1	0.1	Firing threshold
2.1	A_2	10	Scale factor
	B_2	1	Decay rate
	C_2	2	Upper bound of activation
2.2	D_2	0.01	Scale factor
	K_2	20	Inhibition scale factor
3.1	A_3	1	Scale factor
	B_3	1	Decay rate
	C_3	1	Excitation scale factor
	K_3	2	Inhibition scale factor
3.2	A_4	10	Scale factor
	B_4	1	Decay rate
	C_4	1	Excitation scale factor
	K_4	2	Inhibition scale factor
4.1	A_5	0.1	Decay rate
	B_5	1	Upper bound of activation
	C_5	0.01	Lower bound of activation
5.1	A_6	0.5	Decay rate
	B_6	1	Upper bound of activation
	C_6	0.5	Modulatory MSTd feedback scale factor
	D_6	0.5	Excitation scale factor
	M_6	42, 52, 60	The number of heading cells in MSTd (OpenGL, Yosemite, and Driving videos).
5.3	θ_6	0.2	Firing threshold
5.2	L_6	2	Long range filter scale factor
	σ_x	3	Long range filter horizontal variance (horizontal direction preference)
	σ_y	2	Long range filter vertical variance (horizontal direction preference)
6.1	A_7	0.5	Decay rate
	B_7	1	Upper bound of activation
	C_7	4	Scale factor for feedforward excitation
	D_7	0.25	Scale factor for self-excitation
	E_7	0.25	Scale factor for inhibition from other heading cells
	N_7	4103, 4984, 5407	Template normalization factor (OpenGL, Yosemite, and Driving videos).
6.2	G_7	0.1	Sigmoid function scale factor
	θ_7	0.2	Firing threshold

Table A.3. Model parameters, values, and descriptions.

Multi-scale transformation (I). Rather than implement each stage of the model multiple times with multiple receptive field sizes, we resized the video input and used the same receptive field size for each input scale. This allows one set of parameters to process any number of scales at some cost of aliasing, as described below. Resizing is also computationally more efficient, with larger scales being processed at a lower resolution. We implemented three scales using a pixel averaging procedure. Scale 1 is

defined at the resolution of the input, scale 2 computes the mean value of groups of 4 pixels (2 x 2), and scale 3 computes the mean value of groups of 16 pixels (4 x 4). Figure A.1 and Equation (0.2) demonstrate this procedure.

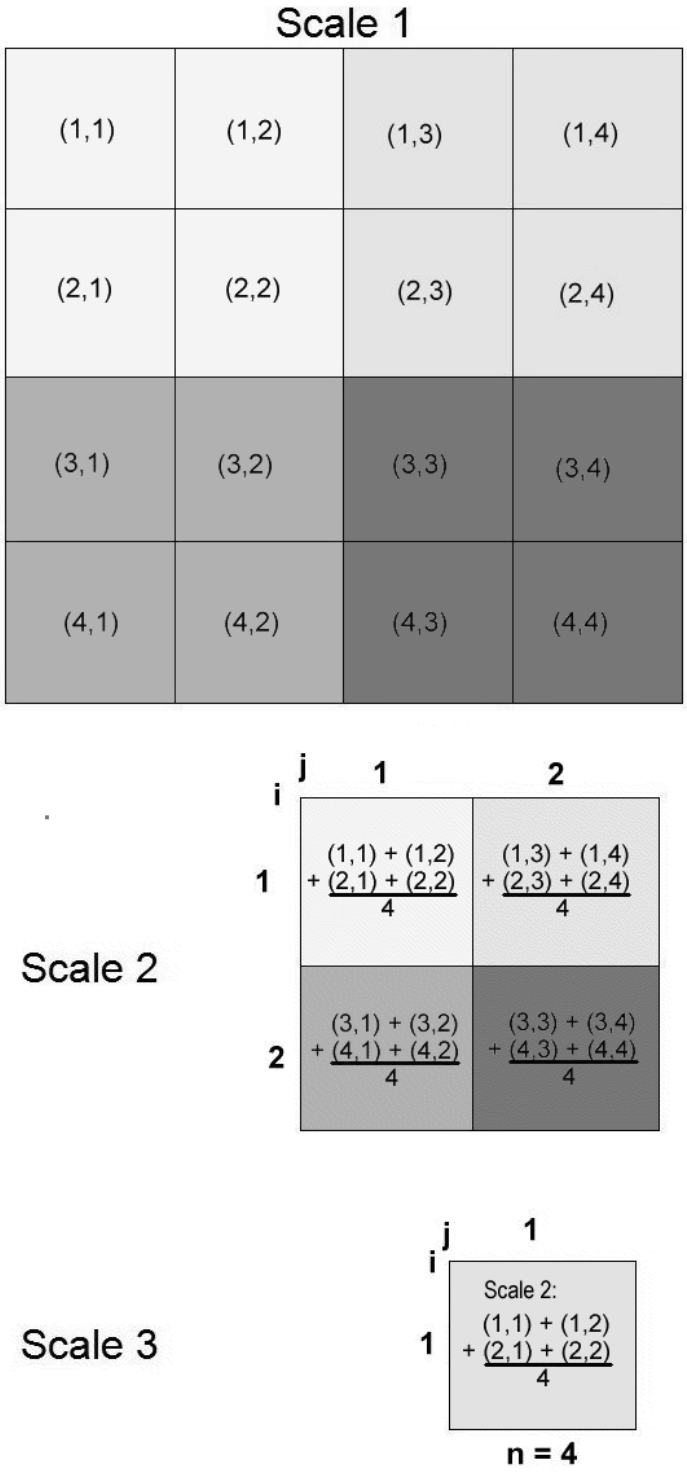

Figure A.1. Resizing algorithm, groups of 4 pixels (2 x 2) at scale 1 are averaged to give the value of a single pixel at scale 2, groups of 16 pixels (4 x 4) at scale 1

are averaged to give the value of a single pixel at scale 3. Grayscale corresponds to magnitude of the filter: dark denotes large values, light small values.

In practice, all resizing algorithms introduce some form of aliasing. In the present case, the algorithm has no overlap between regions that are grouped together. As a result, if an object with a size of 1 pixel exists on an odd-numbered column in the input image and it moves 1 pixel to the right, to an even-numbered column, this movement will *not* be visible at scale 2 (if the input consists of just these input frames). However, if the 1 pixel object exists on an even-numbered column in the input image and it moves 1 pixel to the right, the movement *will* be visible at scale 2. This aliasing effect is minimal since scale 2 is looking for movements in the range of 2 pixels per frame over some temporal window. Since the object described above moves at 1 pixel per frame and alternates between visible and not visible, it will produce a weak signal in scale 2. At scale 1, both odd and even pixels produce the same response. The object described above would therefore produce a strong signal in scale 1, the actual speed at which it is moving.

Ogden, Adelson, Bergen, & Burt (1985) presented pyramid schemes as a method of representing an image at multiple scales. Pyramid schemes use a Gaussian filter to produce overlapping regions, which are then decimated to reduce size. This method implicitly weights the chosen pixel more than its neighbors, and can therefore introduce a complex aliasing effect. In theory, when Gaussian filters are chosen such that the spatial frequency that they are tuned to exactly matches that of the associated image scale, this aliasing can be eliminated. In practice, it is not possible to exactly match a Gaussian filter with a pixel based image. The Gaussian filter is computationally more expensive than the algorithm we use, and the effects of the Gaussian filter combined with decimation on motion vectors are unclear.

We therefore chose the simpler model for clarity, computational convenience, and to enable the same resizing algorithm to be used, and its limitations understood, in higher-level motion representations.

The grayscale intensity values of the resized input stream are defined as the ON-*channel* (equation 0.2), and the complement of its activity is defined as the OFF-*channel* (equation 0.3); cf., the use of complement coding with an OFF-channel in Chelian & Carpenter (2005). Equations (0.2) and (0.3) define I_{ij}^{ps}, the multi-scale input, indexed by spatial position (i, j), ON/OFF channel (p = 1,2), and scale (s = 1,2,3):

$$I_{ij}^{1s} = \frac{1}{n_s^2} \sum_{X=(i-1)n+1, Y=(j-1)n+1}^{ni,nj} g_{XY} \quad \text{(ON-channel)} \tag{0.2}$$

and

$$I_{ij}^{2s} = 1 - I_{ij}^{1s} \quad \text{(OFF-channel)} \tag{0.3}$$

In equation 0.2, g_{XY} is the input intensity, i = 1, 2, …, $\frac{i_{\max}}{n}$ and j = 1, 2, …, $\frac{j_{\max}}{n}$, i_{max} = horizontal resolution, and j_{max} = vertical resolution of the input, and $n_s = 2^{s-1}$.

Level 1: ON-center OFF-surround network (γ). The first level of model processing is a shunting *ON*-center *OFF*-surround network (Grossberg, 1973), which normalizes network activity while enhancing areas of high spatial discontinuity, such as image edges,

and corners. The *ON*-center is a single pixel. The *OFF*-surround is inversely weighted by distance from the center using a Gaussian kernel:

$$\frac{da_{ij}^{ps}}{dt} = -A_1 a_{ij}^{ps} + (B_1 - a_{ij}^{ps}) C_1 I_{ij}^{ps} - (D_1 + a_{ij}^{ps}) \sum_{XY} F_{ijXY} I_{XY}^{ps} \tag{1.1}$$

In equation (1.1) a_{ij}^{ps} is the cell activity at position *(i, j)*, channel (*p)*, and scale (*s)*. Parameter A_1 is the decay rate, B_1 is excitatory saturation potential, C_1 is the input gain, and D_1 is the inhibitory saturation potential. In our simulations, $A_1 = 0.001$, $B_1 = 1$, $C_1 = 2$, and $D_1 = 0.25$. I_{ij}^{ps} is the input from equations (0.2) and (0.3). F_{ijXY} is a Gaussian inhibitory surround kernel, truncated to a 7x7 filter (see Figure A.2, equation 1.2):

$$F_{ijXY} = \frac{F_1}{2\pi\sigma_1} \exp\left(-\frac{(X-i)^2 + (Y-j)^2}{\sigma_1^2} \right) \tag{1.2}$$

where F_1 scales the inhibitory kernel gain, and σ_1 is the inhibitory kernel variance. In our simulations, $F_1 = 10.225$, and $\sigma_1 = 1$. The output signal γ_{ij}^{ps} is a sigmoid function of activity a_{ij}^{ps}:

$$\gamma_{ij}^{ps} = \frac{\left(\left[a_{ij}^{ps} - \varphi_1 \right]^+ \right)^2}{G_1^2 + \left(\left[a_{ij}^{ps} - \varphi_1 \right]^+ \right)^2} \tag{1.3}$$

In equation (1.3), parameter G_1 defines the value at which the output signal attains one-half of its maximum value, and term φ_1 is the firing threshold. In our simulations $G_1^2 = 0.001$, and $\varphi_1 = 0.1$.

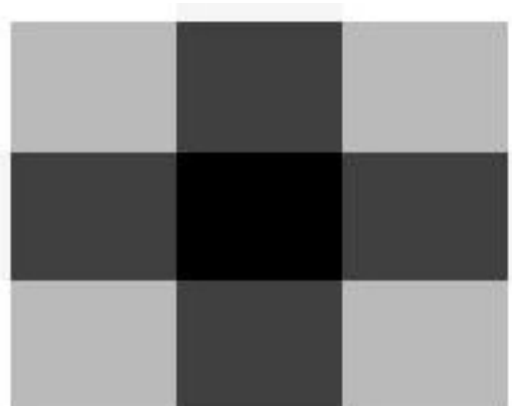

Figure A.2. Inhibitory surround kernel, defined in equation 1.2.

Level 2: Non-directional transient cells (b). Non-directional transient cells respond to changes in the input stream. The non-directional transient cell activities b_{ij}^{ps} are computed as follows:

$$b_{ij}^{ps} = \left[x_{ij}^{ps} z_{ij}^{ps} \right]^+, \tag{2.1}$$

where input cell activities, x_{ij}^{ps}, perform leaky integration on their inputs γ_{ij}^{ps} (equation 1.3):

$$\frac{dx_{ij}^{ps}}{dt} = A_2\left(-B_2 x_{ij}^{ps} + (C_2 - x_{ij}^{ps})\gamma_{ij}^{ps}\right) \tag{2.2}$$

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

$$\frac{dz_{ij}^{ps}}{dt} = D_2\left(1 - z_{ij}^{ps} - K_2 x_{ij}^{ps} z_{ij}^{ps}\right) \tag{2.3}$$

(Grossberg, 1980). In equation (2.1), parameter A_2 determines how fast the cell responds, B_2 scales the passive decay rate, and C_2 is excitatory saturation point. For non-zero inputs, x_{ij}^{ps} approaches C_2 at a rate proportional to $(C_2 - x_{ij}^{ps})$. In our simulations, $A_2 = 10$, $B_2 = 1$, and $C_2 = 2$. In equation (2.3), parameter D_2 determines how fast the cell responds, and K_2 scales the habituation (or transmitter depletion) rate which is also proportional to x_{ij}^{ps}. When x_{ij}^{ps} is zero, activity at z_{ij}^{ps} recovers to 1 at rate D_2. In our simulations, $D_2 = 0.01$, and $K_2 = 20$.

Input activity x_{ij}^{ps} combined with transmitter gate z_{ij}^{ps} results in transient non-directional cell activities b_{ij}^{ps}. For visual inputs with a short dwell time, such as moving boundaries, activities b_{ij}^{ps} respond well. A static input on the other hand, produces only a weak response after the initial presentation period.

Level 3: Directionally selective transient cells (E). This model level provides a directional selectivity mechanism that can retain its sensitivity in response to variable speed inputs (Chey et al., 1997). Eight directions were implemented at 45 degree increments. First, directional interneuron activities, c_{ij}^{psd}, integrate transient cell inputs, b_{ij}^{ps}:

$$\frac{dc_{ij}^{psd}}{dt} = A_3\left(-B_3 c_{ij}^{psd} + C_3 b_{ij}^{ps} - K_3\left[c_{XY}^{psD}\right]^+\right) \tag{3.1}$$

In equation (3.1), a directional inhibitory interneuron, c_{ij}^{psd}, receives excitatory input from transient non-directional cell activity, b_{ij}^{ps}, and suppression from directional interneuron, c_{XY}^{psD}, of opposite direction preference, D, at position (X, Y) offset by 1 cell in direction d. For example if d = 45°, X = i+1, Y = j+1, D = 135°.

Activity c_{ij}^{psd} increases proportionally to input b_{ij}^{ps} with coefficient A_3C_3 and decays to zero with rate $A_3B_3\,c_{ij}^{psd}$. The strength of opponent inhibition is $K_3\left[c_{XY}^{psD}\right]^+$. Inhibition is stronger than excitation and "vetoes" a direction signal if the stimulus arrives from the null direction. In our simulations, $A_3 = 1$, $B_3 = 1$, $C_3 = 1$, and $K_3 = 2$.

Directional transient cell activities, e_{ij}^{psd}, combine transient input b_{ij}^{ps}, with inhibitory interneuron activity, c_{ij}^{psd}. Their dynamics are similar to those of c_{ij}^{psd}:

$$\frac{de_{ij}^{psd}}{dt} = A_4\left(-B_4 e_{ij}^{psd} + C_4 b_{ij}^{ps} - K_4\left[c_{XY}^{psD}\right]^+\right). \tag{3.2}$$

Activity e_{ij}^{psd} increases proportionally to transient input b_{ij}^{ps}, passively decays with the fixed rate and is inhibited by an inhibitory interneuron tuned to the opposite direction. In our simulations, $A_4 = 10$, $B_4 = 1$, $C_4 = 1$, and $K_4 = 2$.

The output of the directional transient cell network is the half-wave rectified activity of e_{ij}^{psd}:

$$E_{ij}^{psd} = \left[e_{ij}^{psd} \right]^{+} \tag{3.3}$$

Computation at level 3 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 4: Directional short-range filter and competition (f). Due to the neural aperture problem, outputs from the directional transient cell network (equation 3.3) do not unambiguously signal the direction of object motion (Marr & Ullman, 1981; Wallach, 1935; Wuerger et al., 1996). The directional short-range filter sums over multiple channels p in equation (3.3) to accumulate directional evidence over a short spatial scale, and to thereby begin to enhance unambiguous feature tracking signals. This filter inputs to a cross-directional normalizing competitive network which enhances the least ambiguous motion directional signals regions and suppresses the most ambiguous regions, thus further strengthening feature tracking signals (Bayerl & Neumann, 2004; Berzhanskaya et al., 2007; Chey et al., 1997), and thereby helping to reduce the effects of the aperture problem (Lucas & Kanade, 1981; Mingolla, Todd & Norman, 1992). Equation (4.1) combines across ON and OFF channel directional transient cell inputs, and competitively normalizes across direction:

$$\frac{df_{ij}^{sd}}{dt} = -A_5 f_{ij}^{sd} + (B_5 - f_{ij}^{sd}) \sum_{p} E_{ij}^{psd} - (C_5 + f_{ij}^{sd}) \sum_{D \neq d} \sum_{p} E_{ij}^{psD} \tag{4.1}$$

In equation (4.1), activity, f_{ij}^{sd}, integrates excitatory input from the directional transients across channels (p) at the same position (i, j), scale (s) and directional preference (d), and is suppressed by directional transients at the same scale and position, in both channels, with directional preferences D ≠ d. Parameter A_5 is the passive decay rate, B_5 is the excitatory saturation potential, and C_5 is the inhibitory saturation potential. In our simulations, $A_5 = 0.1$, $B_5 = 1$, and $C_5 = 0.01$.

For efficient computation across scales in subsequent model levels, the output of level 4 is resized so that all scales are represented at the lowest pixel resolution, which is that of the highest scale (s = 3). Variable m_{ij}^{sd} computes the mean activity across groups of cells with the same scale and directional selectivity:

$$m_{ij}^{sd} = \frac{1}{n_s^2} \sum_{X=(i-1)n+1, Y=(j-1)n+1}^{ni, nj} f_{XY}^{sd}, \tag{4.2}$$

where $n_s = 2^{(3-s)}$, s = 1, 2, 3; i = 1, 2, …, $\frac{i_{\max}}{4}$, and j = 1, 2, …, $\frac{j_{\max}}{4}$, where i_{max} is the horizontal resolution, and j_{max} is the vertical resolution of input g_{ij}.

Level 5: Directional long-range filter (Q). Motion estimates from level 4 are integrated across scale and filtered by a directional long-range filter to produce a more globally-sensitive direction estimate in activities q_{ij}^{d}:

$$\frac{dq_{ij}^{d}}{dt} = -A_6 q_{ij}^{d} + \left(B_6 - q_{ij}^{d}\right)\left(\left(\sum_{XY} L_{ijXY}^{d}\left(\sum_{s} n_s m_{XY}^{sd}\right)\right)\left(1 + \frac{C_6}{M_6}\sum_{z} R_z w_{ijz}^{d}\right) + D_6 Q_{ij}^{d}\right) - q_{ij}^{d}\sum_{D} v_{dD} Q_{ij}^{D} \tag{5.1}$$

In equation (5.1), excitatory input signals m_{ij}^{sd} from equation (4.2) are integrated across scale (s), weighted by $n_s = 2^{(3-s)}$ and filtered by a directional long-range filter kernel, L_{ijXY}^{d} (see equation 5.2), and modulated by feedback that is proportional to heading cell activity, $\sum_{z} R_z w_{ijz}^{d}$, where R_z is the heading cell output from equation (6.2), z indexes heading, and w_{ijz}^{d} defines the translational motion pattern that occurs towards heading z, at spatial position (i, j) and directional selectivity (d). Recurrent connections within equation (5.1) implement a winner-take-all, or choice, network (Grossberg, 1973) via self-excitation and lateral inhibition across direction from cells in the same position. The inhibitory strength is governed by the kernel v_{dD} (equations 5.4-5.7). Parameter A_6 defines the passive decay rate, B_6 is the excitatory saturation point, C_6 scales heading feedback, M_6 is the number of heading cells implemented (see Table 5), and D_6 governs self-excitatory gain. In our simulations, $A_6 = 0.5$, $B_6 = 1$, $C_6 = 0.5$, and $D_6 = 0.5$.

The directional long-range filter, L_{ijXY}^{d}, is an anisotropic Gaussian elongated along the filter's direction of selectivity (see Figure A.4):

$$L_{ijXY}^{d} = \frac{L_6}{2\pi\sigma_x\sigma_y}\exp\left(-0.25\left(\left(\frac{X-i}{\sigma_x}\right)^2 + \left(\frac{Y-j}{\sigma_y}\right)^2\right)\right), \tag{5.2}$$

where L_6 is the long-range filter gain, σ_x is the horizontal variance, and σ_y is the vertical variance. Values less than 0.005 were truncated. In our simulations, $L_6 = 2$, and for horizontal filters, $\sigma_x = 3$, and $\sigma_y = 2$.

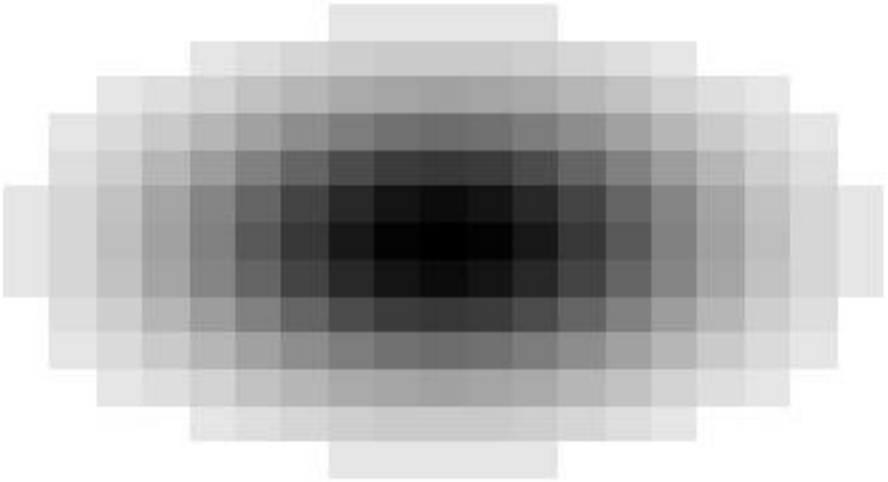

Figure A.3. Long-range filter for horizontal orientations, defined in equation 5.2.

Output from level 5 is half-wave rectified and squared.

$$Q_{ij}^{d} = \left(\left[q_{ij}^{d} - \theta_6 \right]^{+} \right)^2, \tag{5.3}$$

where $\theta_6 = 0.2$ is the firing threshold.

The four lateral inhibition weighting functions v_{dD} are defined as follows (see Figure 1.4):

no-competition

$$v_{dD} = \{0 \quad \text{for all } D \tag{5.4}$$

opponent

$$v_{dD} = \begin{cases} 5 & D = d \pm 180° \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

distributed-opponent

$$v_{dD} = \begin{cases} 0 & D = d \\ 0.5 & D = d \pm 45° \\ 1 & D = d \pm 90° \\ 1 & D = d \pm 135° \\ 10 & D = d \pm 180° \end{cases} \tag{5.6}$$

orthogonal-opponent

$$v_{dD} = \begin{cases} 1 & D = d \pm 90° \\ 10 & D = d \pm 180° \\ 0.25 & \text{otherwise} \end{cases} \tag{5.7}$$

Level 6: Heading filter (R). Adaptive flow filters, or templates, w_{ijz}^{d}, were generated to represent the 2D translational motion vectors produced when moving towards a specific heading. The flow filters were normalized such that in each position the flow filter represented only direction and not speed. As noted in the text, this is consistent with filters learned using a self-organizing map (Cameron et al., 1998; Elder et al., 2008). Two rows of flow filters were created corresponding to two rows of heading cells, one at 1/2 the height of the input and one at 5/8 of the height of the input. Within each row, heading cells were spaced at every third pixel, starting on pixel 2. Heading cell activity, r_z, results from matching the flow filters, w_{ijz}^{d}, against inputs Q_{ij}^{d} from level 5 (equation 5.3):

$$\frac{dr_z}{dt} = -A_7 r_z + \left(B_7 - r_z \right) \left(\frac{C_7}{N_7} \sum_{d} \sum_{ij} w_{ijz}^{d} Q_{ij}^{d} + D_7 R_z \right) - r_z \left(E_7 \sum_{\mathcal{E} \neq z} R_{\mathcal{E}} \right), \tag{6.1}$$

for a particular heading (z), summed across spatial positions (i, j) and directional selectivity (d). The pattern match is weighted by $\frac{C_7}{N_7}$, where N_7 is the energy of the flow

filter; see Table 5. Self-excitation and mutual inhibition via a sigmoid feedback signal R_z (equation 6.2) produce a contrast-enhancing network (Grossberg, 1973). Parameter A_7 defines the passive decay rate, and B_7 is the excitatory saturation point. In our simulations, $A_7 = 0.5$, $B_7 = 1$, $C_7 = 4$, $D_7 = 0.25$, and $E_7 = 0.25$.

The sigmoid signal R_z, is defined by:

$$R_z = \frac{\left(\left[r_z - \theta_7\right]^+\right)^2}{G_7^2 + \left(\left[r_z - \theta_7\right]^+\right)^2}, \tag{6.2}$$

where parameter $G_7 = 0.1$ defines the value at which the output signal attains one-half of its maximum value, and $\theta_7 = 0.2$ is a firing threshold. For simplicity the feedback and output signals R_z are the same.

The heading is the position $\tilde{Z}$ of the maximally active MSTd cell:

$$\tilde{Z} = \arg\max_z \left(R_z\right) \tag{6.3}$$

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Optical Society of America, Journal, A: Optics and Image Science*, *2*, 284-299.

Adelson, E. H., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature*, *300*(5892), 523-525.

Agrawal, M., Konolige, K., & Bolles, R. C. (2007). Localization and Mapping for Autonomous Navigation in Outdoor Terrains: A Stereo Vision Approach. In *Eighth IEEE Workshop on Applications of Computer Vision (WACV'07), 2007*.

Alenyà, G., Martinez, E., & Torras, C. (2004). Fusing visual and inertial sensing to recover robot egomotion. *Journal of Robotic Systems*, *21*(1), 23-32.

AMD. (2003). AMD. *Sunnyvale, CA 94088*.

Anderson, J. C., Binzegger, T., Kahana, O., Martin, K. A. C., & Segev, I. (1999). Dendritic asymmetry cannot account for directional responses of neurons in visual cortex. *Nature Neuroscience*, *2*, 820-824.

Baloch, A. A., & Grossberg, S. (1997). A neural model of high-level motion processing: Line motion and formotion dynamics. *Vision Research*, *37*(21), 3037-3059.

Baloch, A. A., Grossberg, S., Mingolla, E., & Nogueira, C. A. M. (1999). Neural model of first-order and second-order motion perception and magnocellular dynamics. *Journal of the Optical Society of America A*, *16*(5), 953-978.

Banks, M. S., Ehrlich, S. M., Backus, B. T., & Crowell, J. A. (1996). Estimating heading during real and simulated eye movements. *Vision research*, *36*(3), 431-43.

Barlow, H., & Levick, W. (1965). The mechanism of directionally selective units in rabbit's retina. *J Physiol*, *178*(3), 477-504.

Bayerl, P., & Neumann, H. (2004). Disambiguating Visual Motion Through Contextual Feedback Modulation. *Neural Computation*, *16*(10), 2041-2066.

Beardsley, S. A., & Vaina, L. M. (2001). A laterally interconnected neural architecture in MST accounts for psychophysical discrimination of complex motion patterns. *Journal of Computational Neuroscience*, *10*(3), 255-280.

Beauchemin, S. S., & Barron, J. L. (1995). The computation of optical flow. *ACM Comput. Surv*, *27*(3), 433-466.

Beintema, J. A., & van den Berg, A. V. (1998). Heading detection using motion templates and eye velocity gain fields. *Vision Research*, *38*(14), 2155-2179.

Benardete, E., & Kaplan, E. (1999). The dynamics of primate M retinal ganglion cells. *Visual Neuroscience*, *16*(02), 355-368.

van den Berg, A. V. (1993). Perception of heading. *Nature*, *365*(6446), 497-498.

van den Berg, A. V., & Brenner, E. (1994). Humans combine the optic flow with static depth cues for robust perception of heading. *Vision Res*, *34*(16), 2153-67.

Berzhanskaya, J., Grossberg, S., & Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision*, *20*(4), 337-395.

Black, M. J. (2007). Yosemite Sequence FAQs. *Michael Black personal web pages*. Retrieved October 1, 2007, from http://www.cs.brown.edu/~black/.

Black, M. J., & Anandan, P. (1991). Robust dynamic motion estimation over time. *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, 296-302.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annu Rev Neurosci*, *28*, 157-189.

Born, R. T., & Tootell, R. B. H. (1992). Segregation of global and local motion processing in primate middle temporal visual area. *Nature*, *357*(6378), 497-499.

Browning, N. A., Grossberg, S., & Mingolla, E. (2008). Visually guided navigation and steering: motion based object segmentation and heading estimation in primates. . In *Twelfth International Conference on Cognitive and Neural Systems. May 2008, 67.* Boston, MA.

Browning, N. A., Grossberg, S., & Mingolla, E. (2009). ViSTARS: Cortical dynamics of navigation and steering in natural scenes: Motion-based object segmentation, heading, and obstacle avoidance. *Neural Networks, in press.*

Callaway, E. M. (2005). Structure and function of parallel pathways in the primate early visual system. *J Physiol*, *566*(1), 13-19.

Calow, D., Krüger, N., Wörgötter, F., & Lappe, M. (2005). Biologically motivated space-variant filtering for robust optic flow processing. *Network: Computation in Neural Systems*, *16*(4), 323-340.

Cameron, S., Grossberg, S., & Guenther, F. H. (1998). A self-organizing neural network architecture for navigation using optic flow. *Neural Comput*, *10*(2), 313-52.

Chelian, S., & Carpenter, G.A. (2005). DISCOV: A neural model of colour vision, withapplications to image processing and classification. In *Proceedings of AIC05: 10th congress of the international colour association*, Granada, Spain, May.

Chey, J., Grossberg, S., & Mingolla, E. (1997). Neural dynamics of motion grouping: From aperture ambiguity to object speed and direction. *Journal of the Optical Society of America*, *14*(10), 2570-2594.

Chey, J., Grossberg, S., & Mingolla, E. (1998). Neural dynamics of motion processing and speed discrimination. *Vision Research*, *38*(18), 2769-2786.

Cutting, J. E., Springer, K., Braren, P. A., & Johnson, S. H. (1992). Wayfinding on foot from information in retinal, not optical, flow. *Journal of Experimental Psychology: General*, *121*(1), 41-72.

Cutting, J. E., & Wang, R. F. (2000). Heading judgments in minimal environments: the value of a heuristic when invariants are rare. *Perception & psychophysics*, *62*(6), 1146-59.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends Neurosci*, *18*(10), 451–458.

Duffy, C. J. (1998). MST neurons respond to optic flow and translational movement. *Journal of Neurophysiology*, *80*(4), 1816-1827.

Duffy, C. J., & Wurtz, R. H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience*, *15*(7), 5192-5208.

Duffy, C. J., & Wurtz, R. H. (1997). Planar directional contributions to optic flow responses in MST neurons. *Journal of Neurophysiology*, *77*(2), 782-796.

Ehrlich, S. M., Beck, D. M., Crowell, J. A., Freeman, T. C. A., & Banks, M. S. (1998). Depth information and perceived self-motion during simulated gaze rotations. *Vision Research*, *38*(20), 3129-3145.

Elder, D. M., Grossberg, S., & Mingolla, E. (2005). A neural model of visually-guided steering, obstacle avoidance, and route selection. *Soc for Neurosci, Washington DC, Abstract Viewer and Itinerary Planner CD-ROM, Prog*.

Elder, D. M., Grossberg, S., & Mingolla, E. (2009). A neural model of visually guided steering, obstacle avoidance, and route selection. *Journal of Experimental Psychology: Human Peception & Performance,* in press.

Elston, G. N., & Rosa, M. G. (1997). Morphological variation of layer III pyramidal neurones in the occipitotemporal pathway of the macaque monkey visual cortex. *Cerebral Cortex*, *8*, 278-294.

Fajen, B. R., & Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. *J Exp Psychol Hum Percept Perform*, *29*(2), 343-62.

Fajen, B. R., & Warren, W. H. (2004). Visual guidance of intercepting a moving target on foot. *Perception*, *33*(6), 689-715.

Fazl, A., Grossberg, S., & Mingolla, E. (2008). View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology (/in press/)*

Fermüller, C., & Aloimonos, Y. (1995). Direct perception of three-dimensional motion from patterns of visual motion. *Science (New York, N.Y.)*, *270*(5244), 1973-6.

Fleet, D. J., & Langley, K. (1995). Recursive filters for optical flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *17*(1), 61-67.

Fried, S. I., Münch, T. A., & Werblin, F. S. (2002). Mechanisms and circuitry underlying directional selectivity in the retina. *Nature*, *420*(6914), 411-4.

Fried, S. I., Münch, T. A., & Werblin, F. S. (2005). Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina. *Neuron*, *46*(1), 117-27.

Gibson, J. J. (1950). *The Perception of the Visual World*. Houghton Mifflin Boston.

Goodale, M. A., & Milner, D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience, 15*, 10–25.

Graziano, M. S., Andersen, R. A., & Snowden, R. J. (1994). Tuning of MST neurons to spiral motions. *J. Neurosci.*, *14*(1), 54-67.

Grossberg, S. (1968). Some physiological and biochemical consequences of psychological postulates. *Proceedings of the National Academy of Sciences,* 59, 368-372.

Grossberg, S. (1972). A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences*, *15*, 253-285.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverbarating neural networks. *Studies in Applied Mathematics*, *52*, 213-257.

Grossberg, S. (1980). Intracellular mechanisms of adaptation and self-regulation in self-organizing networks: The role of chemical transducers. *Bulletin of Mathematical Biology*, *42*(3), 365-396.

Grossberg, S. (2000). The complementary brain: unifying brain dynamics and modularity. *Trends in Cognitive Sciences*, *4*(6), 233-246.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychol Rev*, *92*(2), 173-211.

Grossberg, S., Mingolla, E., & Pack, C. C. (1999). A neural model of motion processing and visual navigation by cortical area MST. *Cereb. Cortex*, *9*(8), 878-895.

Grossberg, S., Mingolla, E., & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research*, *41*(19), 2521-2553.

Grossberg, S., & Rudd, M. E. (1992). Cortical dynamics of visual motion perception: short-range and long-range apparent motion. *Psychol Rev*, *99*(1), 78-121.
Grossberg, S., & Todorović, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: a unified model of classical and recent phenomena. *Perception & psychophysics*, *43*(3), 241-77.
Hatsopoulos, N. G., & Warren, W. H. (1991). Visual navigation with a neural network. *Neural Networks*, *4*(3), 303-317.
Heeger, D. J., & Jepson, A. D. (1992). Subspace methods for recovering rigid motion I: Algorithm and implementation. *International Journal of Computer Vision*, *7*(2), 95-117.
Heeger, D. J., Boynton, G. M., Demb, J. B., Seidemann, E., & Newsome, W. T. (1999). Motion opponency in visual cortex. *J. Neurosci.*, *19*(16), 7162-7174.
Hildreth, E. C. (1992). Recovering heading for visually-guided navigation. *Vision Research*, *32*(6), 1177-1192.
Hodgkin, A. L. (1964). The Conduction of the Nerve Impulse, C C. *Thomas, Springfield, Illinois*.
Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1-3), 185-203.
Kaiser, M., Perrone, J. A., Stone, L. S., Banks, M. S., & Crowell, J. A. (1993). Extracting heading and temporal range from optic flow: Human performance issues. *Proceedings of the Workshop on Augmented Visual Display(AVID) Research p 379-396(SEE N 94-25490 06-54)*.
Kaplan, E., & Shapley, R. M. (1982). X and Y cells in the lateral geniculate nucleus of macaque monkeys. *J Physiol*, *330*(1), 125-143.
Kaplan, E., & Benardete, E. (2001). The dynamics of primate retinal ganglion cells. *Prog Brain Res*, *134*, 17-34.
Langer, M. S., & Mann, R. (2003). Optical snow. *International Journal of Computer Vision*, *55*(1), 55-71.
Lappe, M., Bremmer, F., & van den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences*, *3*(9), 329-336.
Lappe, M., & Rauschecker, J. P. (1993). A neural network for the processing of optic flow from ego-motion in man and higher mammals. *Neural computation*, *5*(3), 374-391.
Lappe, M., & Rauschecker, J. P. (1994). Heading detection from optic flow. *Nature*, *369*(6483), 712-713.
Lappe, M., & Rauschecker, J. P. (1995a). Motion anisotropies and heading detection. *Biological Cybernetics*, *72*(3), 261-277.
Lappe, M., & Rauschecker, J. P. (1995b). An illusory transformation in a model of optic flow processing. *Vision Res*, *35*(11), 1619-31.
Li, L., Sweet, B. T., & Stone, L. S. (2006). Humans can perceive heading without visual path information. *Journal of Vision*, *6*(9), 874-881.
Li, L., & Warren, W. H. (2000). Perception of heading during rotation: sufficiency of dense motion parallax and reference objects. *Vision research*, *40*(28), 3873-94.
Li, L., & Warren, W. H. (2004). Path perception during rotation: influence of instructions, depth range, and dot density. *Vision Res*, *44*(16), 1879-89.

Livingstone, M. S. (1998). Mechanisms of direction selectivity in macaque V1. *Neuron*, *20*(3), 509-526.

Longuet-Higgins, H. C., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *208*(1173), 385-397.

Loomis, J. M., & Beall, A. (1998). Visually controlled locomotion: Its dependence on optic flow, three-dimensional space perception, and cognition. *Ecological Psychology*, *10*(3-4), 271-285.

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *Proc. DARPA Image Understanding Workshop*, 121-130.

Mann, R., & Langer, M. S. (2002). Optical snow and the aperture problem. In *International Conference on Pattern Recognition, 2002*.

Marr, D., & Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *211*(1183), 151-180.

MathWorks. (2005). The MathWorks Inc. *Natick, Mass*.

Microsoft. (2003). Microsoft Corporation. *Redmond, WA 98052*.

Mingolla, E., Browning, N. A., & Grossberg, S. (2008). Neural dynamics of visually-based object segmentation and navigation in complex environments [Abstract]. *Journal of Vision*, *8*(6), 1154.

Mingolla, E., Todd, J., & Norman, J. (1992). The perception of globally coherent motion. *Vision Research*, *32*(6), 1015-1031.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends Neurosci*, *6*(10), 414-417.

Ogden, J. M., Adelson, E. H., Bergen, J. R., & Burt, P. J. (1985). Pyramid-based computer graphics. *RCA Engineer*, *30*(5), 4-15.

Olson, C. F., Matthies, L. H., Schoppers, M., & Maimone, M. W. (2003). Rover navigation using stereo ego-motion. *Robotics and Autonomous Systems*, *43*(4), 215-229.

Orban, G. A., Fize, D., Peuskens, H., Denys, K., Nelissen, K., Sunaert, S., et al. (2003). Similarities and differences in motion processing between the human and macaque brain: evidence from fMRI. *Neuropsychologia*, *41*(13), 1757-1768.

Pack, C. C., & Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, *409*(6823), 1040-2.

Pack, C. C., Grossberg, S., & Mingolla, E. (2001). A neural model of smooth pursuit control and motion perception by cortical area MST. *J. Cogn. Neurosci.*, *13*(1), 102-120.

Page, W. K., & Duffy, C. J. (1999). MST Neuronal Responses to Heading Direction During Pursuit Eye Movements. *J Neurophysiol*, *81*(2), 596-610.

Perrone, J. A., & Stone, L. S. (1994). A model of self-motion estimation within primate extrastriate visual cortex. *Vision Res*, *34*(21), 2917-38.

Perrone, J. A., & Stone, L. S. (1998). Emulating the visual receptive-field properties of MST neurons with a template model of heading estimation. *Journal of Neuroscience*, *18*(15), 5958-5975.

Rieger, J., & Lawton, D. (1985). Processing differential image motion. *Optical Society of America, Journal, A: Optics and Image Science*, *2*, 354-360.

Rodieck, R. W., Binmoeller, K. F., & Dineen, J. (1985). Parasol and midget ganglion cells of the human retina. *The Journal of Comparative Neurology*, *233*(1), 115-132.

Royden, C. S. (1997). Mathematical analysis of motion-opponent mechanisms used in the determination of heading and depth. *Journal of the Optical Society of America A*, *14*(9), 2128-2143.

Royden, C. S. (2002). Computing heading in the presence of moving objects: a model that uses motion-opponent operators. *Vision Res*, *42*(28), 3043-58.

Royden, C. S., Banks, M. S., & Crowell, J. A. (1992). The perception of heading during eye movements. *Nature*, *360*(6404), 583-585.

Royden, C. S., Crowell, J. A., & Banks, M. S. (1994). Estimating heading during eye movements. *Vision research*, *34*(23), 3197-214.

Royden, C. S., & Hildreth, E. C. (1996). Human heading judgments in the presence of moving objects. *Percept Psychophys*, *58*(6), 836-56.

Royden, C. S., & Vaina, L. M. (2004). Is precise discrimination of low level motion needed for heading discrimination. *Neuroreport*, *15*(6), 1013-7.

Rushton, S. K., Harris, J. M., Lloyd, M. R., & Wann, J. P. (1998). Guidance of locomotion on foot uses perceived target location rather than optic flow. *Current Biology*, *8*(21), 1191-1194.

Saito, H., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., & Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *J. Neurosci.*, *6*(1), 145-157.

van Santen, J. P. H., & Sperling, G. (1985). Elaborated Reichardt detectors. *J. Opt. Soc. Am. A*, *2*(2), 300-320.

Schneider, G. E. (1967). Contrasting visuomotor functions of tectum and cortex in the golden hamster. *Psychological Research*, *31*(1), 52-62.

Schwartz, E. L. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological cybernetics*, *25*(4), 181-94.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743-761.

Singh, A. (1990). An estimation-theoretic framework for image-flow computation. *Computer Vision, 1990. Proceedings, Third International Conference on*, 168-177.

Snowden, R., Treue, S., Erickson, R., & Andersen, R. (1991). The response of area MT and V1 neurons to transparent motion. *J. Neurosci.*, *11*(9), 2768-2785.

Sony. (2003). Sony Corporation of America. *New York, NY* .

Sperling, G. (1970). Model of visual adaptation and contrast detection. *Percept. Psychophys*, *8*, 143–157.

Stone, L. S., & Perrone, J. A. (1994). A role for MST neurons in heading estimation. In *RECON no. 20010116589. Society for Neuroscience, Miami Beach, FL, United States, 13-18 Nov. 1994.*

Stone, L. S., & Perrone, J. A. (1997a). Quantitative simulations of MST visual receptive field properties using a template model of heading estimation. *Soc Neurosci Abstr*, *23*, 1126.

Stone, L. S., & Perrone, J. A. (1997b). Human heading estimation during visually simulated curvilinear motion. *Vision Res*, *37*(5), 573-90.

Tarr, M. J., & Warren, W. H. (2002). Virtual reality in behavioral neuroscience and beyond. *Nature Neuroscience*, *5*(supp), 1089-1092.

Tomasi, C., & Kanade, T. (1991). Detection and tracking of point features. *School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-91–132*.

Wagner, R. E., Polimeni, J. R., & Schwartz, E., L., (2005). Gibson meet topography: the dipole structure of extra striate cortex facilitates navigation via optical flow [abstract]., *Journal of Vision,* 5(8), 895.

Wallach, H. (1935). On the visually perceived direction of motion. *Psychologische Forschung*, *20*, 325-380.

Warren, W. H. (1998). *High Level Motion Processing (ed. Watanabe, T.)*. MIT Press, Cambridge, MA.

Warren, W. H., & Fajen, B. R. (2004). From optic flow to laws of control. *Optic flow and beyond. L.M. Vaina, S.A. Beardsley, & S.K. Rushton (Eds.), Kluwer Academic Publishers, Norwell MA*, 307-337.

Warren, W. H., & Hannon, D. J. (1988). Direction of self-motion is perceived from optical flow. *Nature*, *336*(6195), 162-163.

Warren, W. H., & Hannon, D. J. (1990). Eye movements and optical flow. *Optical Society of America, Journal, A: Optics and Image Science.*, *7*, 160-169.

Warren, W. H., Kay, B. A., Zosh, W. D., Duchon, A. P., & Sahuc, S. (2001). Optic flow is used to control human walking. *Nature Neuroscience*, *4*, 213-216.

Warren, W. H., & Saunders, J. A. (1995). Perceiving heading in the presence of moving objects. *Perception*, *24*(3), 315-331.

Wilkie, R. M., & Wann, J. P. (2003). Controlling steering and judging heading: Retinal flow, visual direction and extra-retinal information. *Journal of Experimental Psychology*, *29*(2), 363-378.

Wilkie, R. M., & Wann, J. P. (2006). Judgments of path, not heading, guide locomotion. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 88–96.

Wuerger, S., Shapley, R., & Rubin, N. (1996). "On the visually perceived direction of motion" by Hans Wallach: 60 years later. *Perception*, *25*(1317), 67.

Zemel, R. S., & Sejnowski, T. J. (1998). A model for encoding multiple object motions and self-motion in area MST of primate visual cortex. *J. Neurosci.*, *18*(1), 531-547.