Cortical Dynamics of Navigation and Steering in Natural Scenes:

Motion-Based Object Segmentation, Heading, and Obstacle Avoidance

by

N. Andrew Browning, Stephen Grossberg, 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

December, 2008

Neural Networks, in press

CAS/CNS Tech Report #CAS/CNS-TR-08-007

Corresponding Author: Stephen Grossberg

Department of Cognitive and Neural Systems
Boston University
677 Beacon Street
Boston, MA 02215
617-353-7858/7
617-353-7755 (fax)
steve@bu.edu

Abstract

Visually guided navigation through a cluttered natural scene is a challenging problem that animals and humans accomplish with ease. The ViSTARS neural model proposes how primates use motion information to segment objects and determine heading for purposes of goal approach and obstacle avoidance in response to video inputs from real and virtual environments. The model produces trajectories similar to those of human navigators. It does so by predicting how computationally complementary processes in cortical areas MT^-/MSTv and MT^+/MSTd compute object motion for tracking and self-motion for navigation, respectively. The model retina responds to transients in the input stream. Model V1 generates a local speed and direction estimate. This local motion estimate is ambiguous due to the neural aperture problem. Model MT^+ interacts with MSTd via an attentive feedback loop to compute accurate heading estimates in MSTd that quantitatively simulate properties of human heading estimation data. Model MT^- interacts with MSTv via an attentive feedback loop to compute accurate estimates of speed, direction and position of moving objects. This object information is combined with heading information to produce steering decisions wherein goals behave like attractors and obstacles behave like repellers. These steering decisions lead to navigational trajectories that closely match human performance.

KEYWORDS: Optic flow, navigation, MT, MST, motion segmentation, object tracking

1. Introduction

The ViSTARS (Visually-guided Steering, Tracking, Avoidance, and Route Selection) model demonstrates how the primate magnocellular pathway may generate sufficient information for reactive navigation, route selection, and target tracking tasks (Figure 1). When immersed in a realistic visual world, ViSTARS is capable of human-like steering behaviors towards goals and around obstacles in response to realistic visual scenes.

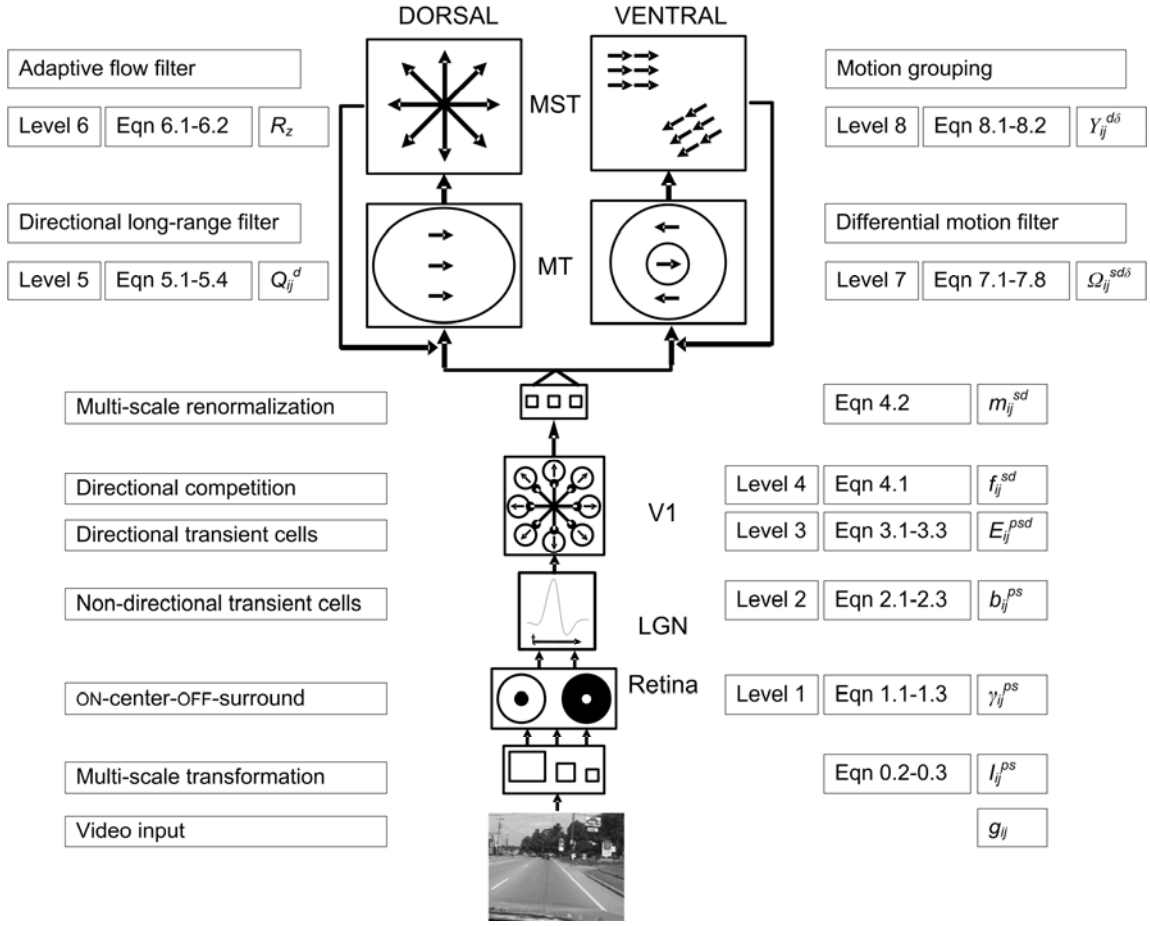

Figure 1. Model overview: pictorial representation of each processing stage coupled with functional description, model level number, corresponding equation number, and output variable labels. See text for model description.

ViSTARS is a synthesis and further development of two previous models: The STARS model of Elder et al. (2007) is capable of reactive steering towards goals and around obstacles, and accurately simulates human navigational data of Fajen and Warren (2003), among others. However, the STARS model did not directly process visual scenes. Rather, it used the equations that describe scenic geometry of Longuet-Higgins & Prazdny (1980) as model inputs. Browning, Grossberg, and Mingolla (2008b) showed how ViSTARS could build upon STARS to directly process visual data, notably virtual world animations and driving video sequences of realistic visual scenes, as well as random dot displays, to compute accurate heading, or direction of travel, estimates at human-like accuracies. To accomplish this, the motion processing front end of ViSTARS adapted a biological motion perception model, called the 3D FORMOTION model, that has been

progressively developed to explain and predict large perceptual and neurobiological data bases about motion perception (Baloch & Grossberg, 1997; Chey, Grossberg, & Mingolla, 1997; Berzhanskaya, Grossberg, & Mingolla, 2007; Grossberg, Mingolla, & Viswanathan, 2001). The current extension of ViSTARS shows, in addition, how heading estimates can be joined to STARS navigational mechanisms to achieve reactive navigation and object tracking estimates in response to realistic visual scenes.

This synthesis clarifies how the brain exploits computationally *complementary* processes for navigation and object tracking, respectively (Grossberg, 2000). As will be seen in greater detail below, the processing stream through cortical areas MT^+/MSTd is specialized for visually based navigation, whereas the parallel processing stream through cortical areas MT^-/MSTv is specialized for visual tracking of moving objects. In particular, navigating a body moving with respect to the world uses *additive* processing, whereas tracking an object moving with respect to that body uses *subtractive* processing.

By showing navigational competence in realistic settings, ViSTARS provides an example of how real-time adaptive control systems can accomplish visually-based autonomous robotic navigation. By linking identified brain regions and cell types to navigational behaviors, the model clarifies how the brain accomplishes visually-based navigation and object tracking. Previous models typically contribute to one of these goals, but not both.

ViSTARS processing levels correspond to brain regions from retina through cortical areas V1, MT, and MST. Before describing the model, a short summary of pertinent experimental data will be given.

Neurophysiology. The early primate visual system consists of two distinct pathways: the parvocellular (P) pathway is concerned with high resolution, color, static information, whereas the magnocellular (M) pathway is concerned with low resolution, monochromatic, transient information (Kandel, Schwartz, & Jessell, 2000). The parvocellular pathway processes object form and identity in the What, or ventral, cortical processing stream. The magnocellular pathway processes motion, object location, and action in the Where, or dorsal, cortical processing stream (Mishkin, Ungerleider, & Macko, 1983; Schneider, 1967). Although form processing influences motion perception and navigational tasks through form-motion interactions (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg et al., 2001; Ponce, Lomber & Born, 2008), the results reported herein focus on the magnocellular pathway.

Magnocellular retinal cells respond with a burst of activation when presented with a step input (Benardete & Kaplan, 1999; Cleland, Dubin & Levick, 1971; Enroth-Cugell & Robson, 1966; Kaplan & Benardete, 2001; Valois, Albrecht & Thorell, 1982). M pathway retinal cells project to lateral geniculate nucleus (LGN) layers 1 and 2 and then to primary visual cortex (V1) (Callaway, 2005). V1 cells are directionally selective, responding more vigorously to motion in a preferred direction at a preferred speed, and are disparity selective (Hubel & Wiesel, 1959, 1962, 1968; Livingstone & Hubel, 1987; Livingstone, 1998; Schiller, Finlay & Volman, 1976).

V1 projects to area MT (middle temporal cortex, or V5) which, in turn, projects to area MST (medial superior temporal cortex) (Albright, 1984; Born & Bradley, 2005). Cells in MT respond preferentially to motion in a particular direction within a range of speeds and depths (Albright, 1984; Born & Bradley, 2005). Macaque MT has two main sub-divisions: MT^+ consists of cells with large *additive* receptive fields that project

primarily to dorsal MST (MST*d*); MT^- consists of cells with *subtractive* (that is, ON-center OFF-surround opponent-motion) receptive fields that project primarily to ventral MST (MST*v*) (Allman, Miezin, & McGuinness, 1985; Born, 2000; Born & Tootell, 1992). The MT^+/MST*d* stream carries out visually-guided navigation, including heading estimation (Born & Tootell, 1992; Duffy, 1998; Duffy & Wurtz, 1995, 1997). The MT^-/MST*v* stream carries out object-based segmentation and tracking (Born & Tootell, 1992; Duffy, 1998). In order to realize their complementary tracking and navigation functions, ventral and dorsal MST cells have different response properties. MST*v* cells respond to relative direction of object motion across a background (Duffy, 1998; Tanaka, Sugita, Moriya, & Saito, 1993). MT^- cells respond to objects moving within a specific range of speeds, whereas MST*v* cells respond more vigorously to faster speeds (Tanaka et al., 1993). MST*d* cells respond to large motion patterns such as those that occur during self-motion through the environment (Duffy, 1998; Grossberg et al., 1999; Stone & Perrone, 1994, 1997a). Thus MST appears to integrate information across MT. The human homolog of monkey MST is also implicated in human heading detection tasks (Beardsley & Vaina, 2001). Heading appears to be represented as a population code in both primate MST*d* and human MT complex (hMT+) (Beardsley & Vaina, 2001; Page & Duffy, 1999).

It is the MT^- ON-center OFF-surround network that enables it to respond to differential motion. Its cells detect motion discontinuities at object boundaries, when an observer moves, or due to independent object motion (Grossberg et al., 1999; Longuet-Higgins & Prazdny, 1980; Nakayama & Loomis, 1974; Pack et al., 2001; Rieger & Lawton, 1985). Humans also appear to utilize disparity information to segment moving objects. When no disparity information is available, humans take longer to respond and are less accurate when discriminating moving objects (Rushton, Bradshaw, & Warren, 2007; Rushton & Warren, 2005a, 2005b; Warren & Rushton, 2007). Indeed, MT^- cells can have ON-centers that prefer one disparity and OFF-surrounds that prefer another (Bradley & Andersen, 1998; Born & Bradley, 2005). These cells can respond to both disparity and differential motion, thereby combining the segmentation abilities of each.

A line viewed through a small isotropic aperture, such as a circular receptive field of a neuron, always appears to move in the direction that is perpendicular to its orientation. This is known as the aperture problem (Marr & Ullman, 1981; Wallach, 1935; Wuerger, Shapley, & Rubin, 1996). MT^- cells initially respond to the perpendicular direction of a line or bar's orientation if the line extends beyond the cell's receptive field, but after a period of 100-200ms respond to the true direction of motion (Pack & Born, 2001). Thus MT^- computes an aperture-resolved object motion signal.

Psychophysics. Fajen and Warren (2003) demonstrated how humans steer with a smooth trajectory around obstacles towards a goal (Figure 2 panels A and B). They found that the deviation from a straight line of the chosen trajectory is dependent on the distance between the observer and the obstacle. When obstacles are at a fixed depth but the visual angle between current heading and the obstacle is varied, smaller visual angles result in wider trajectories around the obstacle. When visual angle remains fixed and depth is varied, humans steer earlier and produce wider trajectories around obstacles at closer depths. In both cases, closer obstacles result in larger deviations from a straight line trajectory.

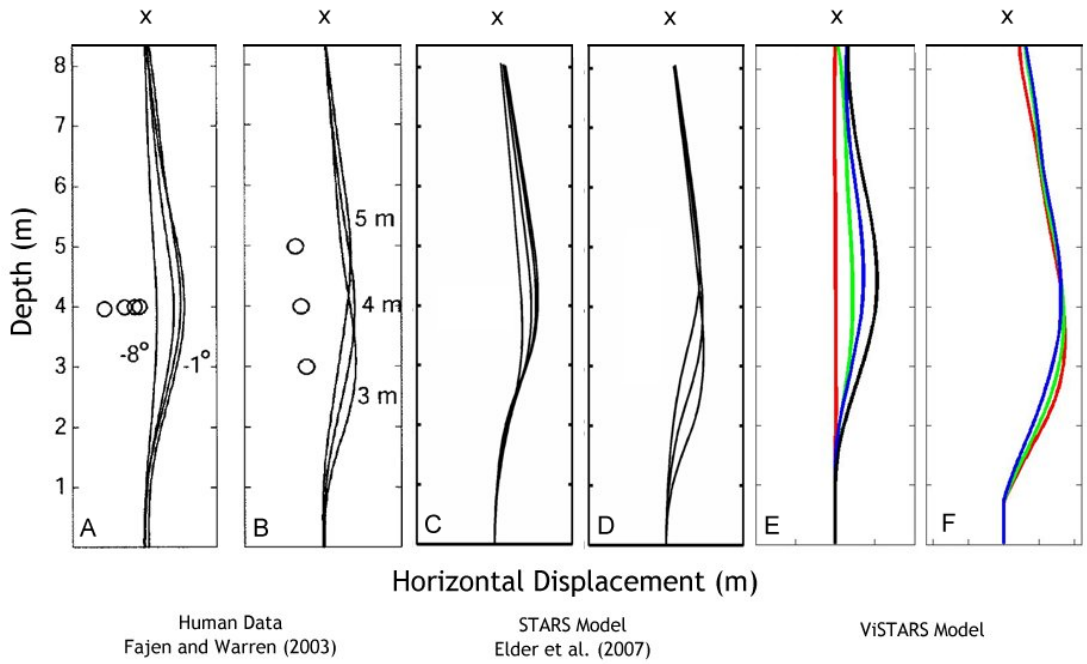

Figure 2. Mean human trajectories around obstacles (circles) towards a goal (x) are shown in panels A and B. Reproduced from Fajen and Warren (2003, Figure 10) with permission. STARS trajectories are shown in panels C and D and ViSTARS trajectories are shown in panels E and F. When the obstacles are at fixed depth but variable visual angle (panels A, C and E), humans deviate more from a straight path for smaller visual angles. When the obstacles are at fixed visual angle but variable depth (panels B, D and F), humans deviate more quickly and with larger magnitude for smaller depths.

Prior Modeling. Fajen and Warren (2003) proposed a behavioral model whereby goals are treated as attractors and obstacles as repellers. The model takes the form of a damped spring equation. Utilizing a third-person geometrical description of objects in the world and observer heading, it provided a good fit to human steering data (Fajen & Warren, 2003). In their model, object positions are compared against the current heading; if a collision with an obstacle is likely, then heading is repelled by the obstacle to produce a trajectory that avoids the obstacle. If heading is not congruent with the goal position, then the goal attracts heading to produce a trajectory that approaches the goal. The balance of attraction and repulsion defines the final trajectory.

The STARS model of Elder et al. (2007) demonstrated how the data observed by Fajen and Warren (2003) can be explained by a first-person dynamical explanation of how goal position, obstacle position, and heading may be computed from optic flow. These representations form three Gaussian activation distributions across a cortical map. Trajectories are computed by adding together these activity distributions, with the obstacle Gaussian subtracted from the sum of the other two Gaussians. The resulting *steering field* is a net distribution of activity (Figure 3) whose direction and magnitude control angular steering velocity. STARS exhibits steering behavior almost indistinguishable from the Fajen and Warren (2003) model (Figure 2, panels C and D).

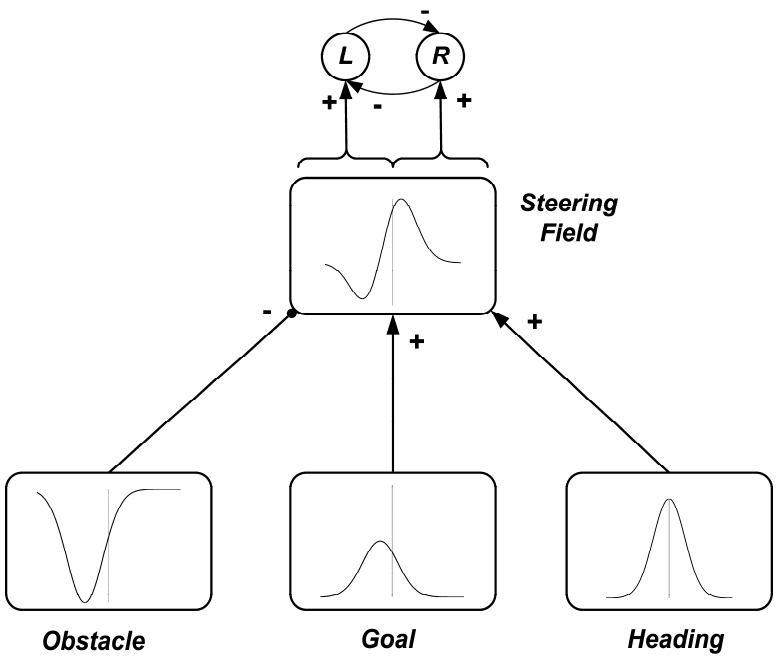

Figure 3: The STARS model combines obstacle, goal, and heading position to determine a steering trajectory. Spatial position of obstacles, goal and heading are represented as unimodal activation distributions. Summation of the distributions, with a negative sign on the obstacle distribution, results in a net distribution in the steering field. The position of the peak of this distribution determines the magnitude and direction of the steering command to the left or right. [Reproduced from Elder et al. (2007) with permission.]

As noted in the Introduction, STARS does not process visual imagery. Instead, optic flow is computed analytically from a scene geometry whose noise-free description is assumed to be provided (Longuet-Higgins & Prazdny, 1980). Optic flow in this representation is then processed by the model. This optic flow estimation is highly precise and accurate, and is dense in the sense that motion vectors are estimated for each pixel in the visual input. Optic flow is transformed into log-polar coordinates whose cortical V1 representations are comparable to those found *in vivo* (Schwartz, 1977; Wagner, Polimeni, & Schwartz, 2005). The log-polar transformation produces high levels of detail in the central, or *foveal*, region and low levels of detail in the periphery. Model MT and MST cells process this input. In accordance with a predicted V2-MT pathway (Berzhanskaya et al., 2007; Grossberg, 1991; Maunsell & Van Essen, 1983), interactions from model V2 to MT provide disparity information to the MT$^-$ cells. The output of MST is a heading estimate in MSTd, and object motion estimates in MSTv. Gain fields compensate for eye and head rotations, and map V1 retinotopic representations into body-centric coordinates in MT$^-$/MST.

In contrast with the Fajen and Warren (2003) model, STARS computes estimates of the visual angle of heading and objects, and of object depth, directly from optic flow in realistic visual imagery.

The use of heading by humans during navigation has been contested (Rushton et al., 1998; Wilkie & Wann, 2003, 2006). In some cases, goal position relative to the navigator's position is sufficient to explain human steering data. The Ruston et al. (1998) ego-centric model orients itself towards the goal and moves directly towards it. Heading

is not required. Warren et al. (2001) demonstrated that humans can make use of both strategies: in featureless environments where heading is hard to estimate, ego-centric goal position is used, but in richer environments, heading information is also used. STARS demonstrated the same behavior (Elder et al., 2007).

As noted in the Introduction, ViSTARS extends STARS to process visual imagery. Realistic images are noisier than analytically computed optic flow. Their pixel intensity values are prone to *sensor noise*, since the sensor responds directly to light, which is by nature stochastic, and *aliasing*, the sensor has fixed pixel size which may not match the size of elements in the environment (Bradski & Kaehler, 2008; Langer & Mann, 2003; Mann & Langer, 2002). Indeed, the term *optic snow* has been coined to describe motion estimates from natural image sequences (Langer & Mann, 2003; Mann & Langer, 2002). The aperture problem causes additional ambiguities in motion estimates that are derived from scene statistics at each point in the image (Marr & Ullman, 1981; Wallach, 1935; Wuerger, et al., 1996). Motion estimates also suffer from the *correspondence problem*: in a cluttered scene, objects do not have unique intensity values across space or time, so that ambiguity is introduced to motion estimates when tracking pixel features across time, especially in scenes with cast shadows, specular highlights, and other vagaries of illumination across discrete video frames (Aggarwal & Nandhakumar, 1988; Bradski & Kaehler, 2008).

The 3D FORMOTION model combines form and motion information to produce an accurate motion estimate from ambiguous inputs when either process alone is insufficient (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg et al., 2001). Model stages corresponding to cortical area V2 represent boundaries in depth. These boundary signals are, via a V2-to-MT inter-stream interaction, used to capture directional motion signals at the corresponding depth in MT. In a complementary interaction, motion signals in MT are used to disambiguate incomplete or ambiguous boundary signals in V2 via feedback to V1. This indirect feedback from MT to V2 allows motion information to determine object shape. Utilization of form processing information allows the motion processing stages to perform object feature tracking and thereby reduce ambiguity. Spatially anisotropic grouping integrates the feature-enhanced motion signals to produce a global object motion percept.

The 3D FORMOTION model has previously explained many perceptual and neurobiological data about motion perception. However, it has not previously been demonstrated capable of processing natural image streams. Additionally, the model processes a single MT/MST stream that is most consistent with MT^-/MSTv. ViSTARS includes the two complementary MT-MST streams, MT^-/MSTv for object tracking and MT^+/MSTd for optic-flow based navigation, that are found in the brain.

The ViSTARS model and its properties are described before simulations are provided of how it simulates human performance.

2. The ViSTARS Model

Implementation. The model is defined as a system of differential equations that are described in the Appendix. The model was tested using computer-generated animations, publicly available video, and video taken from a moving vehicle while driving. Simulations were performed in MATLAB R14 (MathWorks, 2005) on a dual 2Ghz AMD Opteron (AMD, 2003) based workstation with 8Gb of RAM running

Microsoft Windows XP x64 (Microsoft, 2003). Input to the model was in the form of a frame-based input stream, either computer-generated image sequences or video, as described below. Euler's method was used to numerically integrate the solution to the equations. 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 cell at a particular spatial position reaches an equilibrium state is dependent on the magnitude of changes in the input stream at that spatial position. Code samples, input videos, and demonstrations of results can be obtained from http://cns.bu.edu/vislab/objectmotion.

Inputs and preprocessing. Inputs to the model are in the form of an 8-bit grayscale image stream. The resolution and frame rate of the image stream are dependent on the source. For example, videos that we created while driving with a camera mounted in a car were processed at 15 frames per second (fps), due to constraints from the camera, post-processing, and computation time. The effective frame rate in the immersed environment, where more flexibility in generating the image sequence was afforded, was 47 fps. See Appendix Table 1 for detailed descriptions of each video source. The same model parameters were used irrespective of the source. When the frame rate was slower than the integration time step, each video frame was presented for multiple time steps. Intensity values for the grayscale pixels were scaled between 0 and 1 by dividing the 8-bit intensity value by 255. The ON-cell channel response was defined as the image intensity value, and the OFF-cell channel response computes the complementary activity (Chelian & Carpenter, 2005); namely, one minus the ON-cell channel response (Appendix, equations 0.2 and 0.3).

Three scales of inputs were created by reducing the size of the input by successive factors of 2. The first scale is the original image, the second scale reduces the height and width of the original image by a factor of 2, and the third scale reduces the height and width of the original image by a factor of 4. Size reduction is performed by taking the mean intensity value of a group of pixels as described in the Appendix. All such resizing methods introduce some aliasing, as discussed in the Appendix. The model processes 6 image streams: the ON channel at scales 1, 2 and 3, and the OFF channel at scales 1, 2, and 3. The same parameters were used for all image streams and are defined in the Appendix.

For navigation simulations in the virtual environment, coarse depth information was incorporated into the model inputs: the near depth was defined as everything in front of the goal, and the fixation depth was defined as everything at the same depth as the goal (cf., Elder et al., 2007). These stimuli did not include far depths. Thus navigation simulations process 12 image streams, six for near depth and six for fixation depth. The parameters used for both depth planes were the same. Interactions between the depths are defined at model area MT (Figure 1, levels 5 and 7).

Retina – LGN. The first stage of the model (Appendix, equations 1.1-1.3) is an ON-center OFF-surround shunting that contrast enhances and normalizes the input image. The ON-center is narrow, consisting of a single pixel. The OFF-surround is a 2-dimensional Gaussian truncated at 7x7 pixels. The output of this stage is through a thresholded sigmoid signal function, which further enhances contrast.

Level 2 of the model (Appendix, equations 2.1-2.3) consists of a non-directional transient cell network (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg et al., 2001). The transient cell network responds with a burst response at the onset of a

stimulus. This occurs when a light object appears (or a dark object disappears) in the ON stream and when a dark object appears (or a light object disappears) in the OFF stream. The time course of the model cells is configured such that the peak of the response occurs after 70-75ms, similar to M-pathway retinal cells (Benardete & Kaplan, 1999; Kaplan & Benardete, 2001).

Primary visual cortex (V1). Level 3 of the model (Appendix, equations 3.1-3.3) consists of directional transient cells (Berzhanskaya et al., 2007; Chey et al., 1997; Grossberg et al., 2001). These cells respond to motion in a preferred direction. We implemented 8 directions at 45 degree increments. The three input scales correspond to speed in the vector velocity domain: a directional cell that responds at scale 3 (one quarter of the size of the original image) is responding to motion 4 times as fast as a cell in the same position at scale 1.

The directional transient cell network was designed to allow model cells to respond to motion at a wide range of object speeds (Berzhanskaya et al., 2007; Chey et al., 1997, 1998; Grossberg et al., 2001). This is accomplished by incorporating a stage of directional inhibition via directional interneurons. Inhibition travels in the direction opposite the preferred direction of the cell and is thus called nulling inhibition. Nulling inhibition has been found *in vivo* in rabbit (Barlow & Levick, 1965; Fried, Münch, & Werblin, 2002, 2005), cat (DeAngelis, Ohzawa, & Freeman, 1995), and primate V1 (Livingstone, 1998). It thus appears to be a widespread mechanism across mammals. The recently discovered Starburst inhibitory interneurons in the rabbit have been shown to provide directional nulling inhibition and occur in a network that strikingly resembled the model directional transient cell network (Fried et al., 2002, 2005). It remains to be tested if they enable their target directional cells to retain directional selectivity in response to a wide range of speeds. Level 4 of the model (Appendix, equations 4.1-4.2) combines ON and OFF channels and uses directional competition to normalize activity across direction. Normalization across direction suppresses activity in positions where there is a high degree of directional ambiguity and enhances activity in positions where there is a low degree of directional ambiguity (Bayerl & Neumann, 2004; Chey et al., 1997, 1998). Directional normalization hereby aids in the resolution of the aperture problem.

The output of level 4 is resized such that all scales are represented at the same pixel resolution. Thus, the height and width of scale 1 are reduced by a factor of 4, and the height and width of scale 2 are reduced by a factor of 2. Size reduction is performed via the pixel averaging procedure described by Appendix equation 4.2. The responses in model V1 are calibrated both by their parameters and by the timescale of model retina responses. As noted above, model retina responds with a peak of activation after roughly 75ms. As a result, in V1 scale 1 represents speed in the range of 1 pixel every 75ms, scale 2 in the range of 2 pixels every 75ms, and scale 3 in the range of 4 pixels every 75ms.

Middle temporal area – additive cells (MT^+) and dorsal medial superior temporal area (MSTd). Level 5 (Appendix, equations 5.1-5.4), corresponds to MT^+ and implements a long-range directional filter. The filter pools information across speed to produce a global direction of motion estimate. The directional long-range filter is a 2D Gaussian elongated in the preferred direction of the cell that realizes anisotropic spatial integration (Berzhanskaya et al., 2007). Temporal integration is provided by the

dynamics of the shunting network. The elongation of the spatial filter, when combined with the temporal integration of the network, allows MT^+ to track motion across space and time. Recurrent shunting ON-center OFF-surround interactions within the level result in a choice, or winner-take-all, network (Grossberg, 1973), which reduces ambiguities introduced by the spatiotemporal integration. Feedback from the MSTd heading estimate (level 6) modulates activity to ensure that motion patterns which are consistent with the current heading estimate are enhanced.

Level 6 of the model (Appendix, equations 6.1-6.2), corresponds to MSTd and estimates heading through a bank of adaptive flow filters. The adaptive flow filters perform a template match against the global motion estimate MT^+. Competition within this stage results in contrast enhancement to ensure that only a small subset of cells is highly active at any one time. In accordance with neurophysiological data (Duffy & Wurtz, 1995, 1997), the output of MSTd is an accurate heading estimate that is highly robust to noise in the input stream.

Middle temporal area – subtractive cells (MT^-) and ventral medial superior temporal area (MSTv). Level 7 of the model (Appendix, equations 7.1-7.4), corresponds to MT^- and determines motion boundaries in the scene. Motion boundaries occur when an object moves differently from its background, either due to a large depth discontinuity as the observer navigates towards a stationary object, or due to the independent motion of the object. Directional filters detect differential motion, recurrent ON-center OFF-surround connectivity reduces directional ambiguities, and attentive matching feedback from MSTv (level 8) further reduces ambiguities in the object motion estimate. In accord with neurophysiological data (Born & Bradley, 2005) and the STARS model (Elder et al., 2007), MT^- was implemented with an ON-center and OFF-surround that respond preferentially to different depths. For example, if the ON-center prefers the near depth, then the OFF-surround prefers the fixation depth, and vice-versa. The output of MT^- is the motion boundaries in a scene.

Level 8 (Appendix, equations 8.1-8.2) corresponds to MSTv and combines object motion estimates from MT^- in a given direction across speed, such that the activity in MSTv increases with the speed of object motion. Directional competition selects the strongest directional signal. In accord with neurophysiological data (Duffy, 1998) and the STARS model (Elder et al., 2007), model MSTv activity represents the position, direction, and speed of object motion the scene.

3. Results

The STARS model (Elder et al., 2007) combines heading with goal and obstacle positions for reactive steering. ViSTARS has previously been shown capable of human-like accuracy in heading detection tasks in realistic scenes (Browning et al., 2007a, 2007b, 2007c, 2008b). The present article also demonstrates how the MT^-/MSTv stream can represent object position in response to realistic scenes at accuracies that enable the STARS steering circuitry to necessary to reactively avoid obstacles and approach goals. Speed and direction of object motion estimates were not utilized by STARS. They are, however, a major function of the MT^-/MSTv processing stream (Albright, 1984; Allman, Miezin, & McGuinness, 1985; Born, 2000; Born & Bradley, 2005; Born & Tootell, 1992; Tanaka et al., 1993) and are demonstrated in Figure 4.

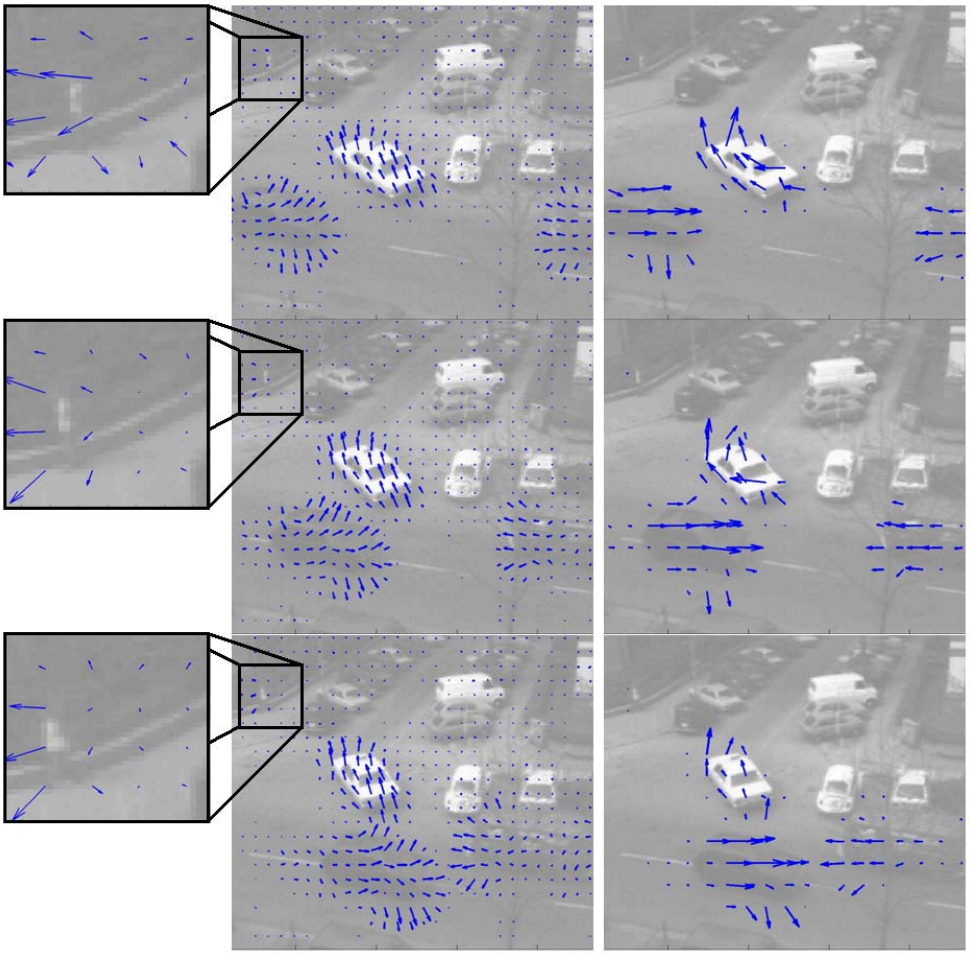

Figure 4: Model activations in MT⁻ and MSTv for frames 20, 30 and 40 of the Hamburg taxi sequence. The right panel shows MSTv outputs, the middle panel shows MT⁻ outputs, the left panel shows a blow-up of MT⁻ outputs in an area of input where a pedestrian is present. The background image is the video frame, blue arrows represent motion, arrow length corresponds to speed, and arrow direction corresponds to the estimated motion direction. The model representations of three moving cars and a moving pedestrian produce speed and direction of the object motions.

Figure 4 shows a series of frames of the Hamburg taxi sequence (obtained from http://i21www.ira.uka.de/image_sequences/#taxi, hosted by Institut fuer Algorithmen und Kognitive Systeme) with the MT⁻ and MSTv activations overlaid as motion vectors, on the left and right, respectively. Note how the white taxi is tracked by activity in both MT⁻ and MSTv with the speed and direction of motion being estimated as it turns through the junction. The two darker vehicles moving into the scene from the left and right, respectively, are tracked with the speed and direction of motion estimated. A pedestrian in the top left is also tracked and his speed and direction of motion estimated, as shown in the box on the left. The cars are estimated to be traveling at much the same speed as each other, whereas the speed of the pedestrian is estimated to be much slower than that of the cars. The Hamburg taxi sequence thus demonstrates model competence with a stationary camera.

The model also performs well with a moving camera. Figure 5 illustrates MSTv activation when processing an image sequence where a car is driving on a highway and

another car is overtaking it on the left hand side. The motion of the overtaking car on the left is estimated, as is the relative motion of any stationary objects that have a significant depth difference between them and the background. No other objects produce motion estimates. In particular, the cars in the distance move at roughly the same speed as the camera.

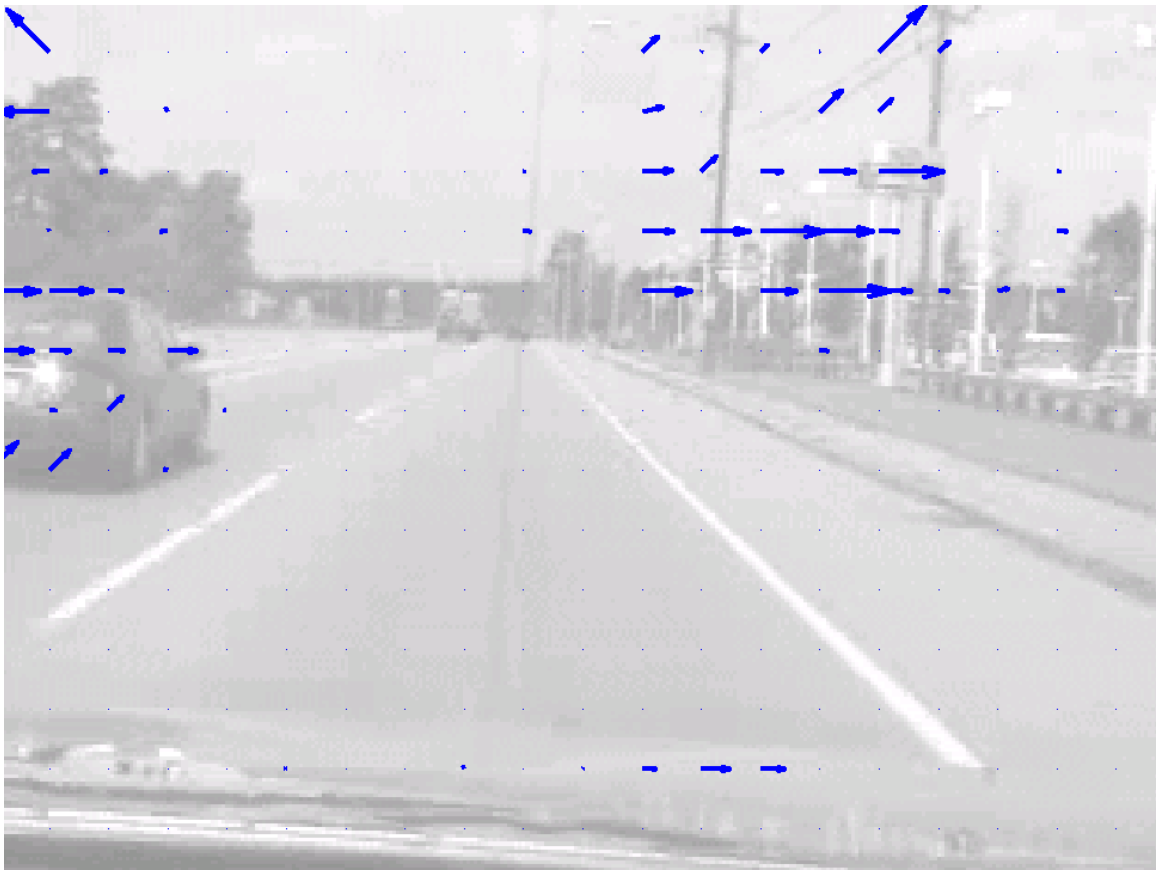

Figure 5: Model MSTv outputs when processing video taken in a car while driving. The camera is moving straight forward. Motion signals are present where there is a difference between an object and its surrounding. The overtaking car on the left, the tree on the left ,and the telephone poles on the right are represented as moving objects with their relative directions of motion accurately assessed. The overtaking car is moving faster than the camera. The static trees and telephone poles have a significant depth discontinuity between themselves and the sky, thus creating a differential motion signal as the car moves towards them.

In all video sequences tested with the model, MT$^-$ and MSTv provide reasonable speed and direction estimates of objects in the sequence. Directional competition in MSTv ensures that motion estimates are in the direction of the maximally active cell at that spatial position, and removes activation from areas of high motion ambiguity, resulting in motion-based object segmentation. Direction estimates in MT$^-$ are less strictly resolved. At any spatial position, multiple directions can be active although, in general, a single direction will dominate. This allows model MT$^-$ to represent direction at a finer resolution than the eight that are presently implemented

The motion estimates in both MT$^-$ and MSTv track moving contours in the object texture and therefore the density of the motion estimate depends on the nature of the

object and its texture. Motion estimates in model MSTv track objects in the video sequences with around 150-200ms delay.

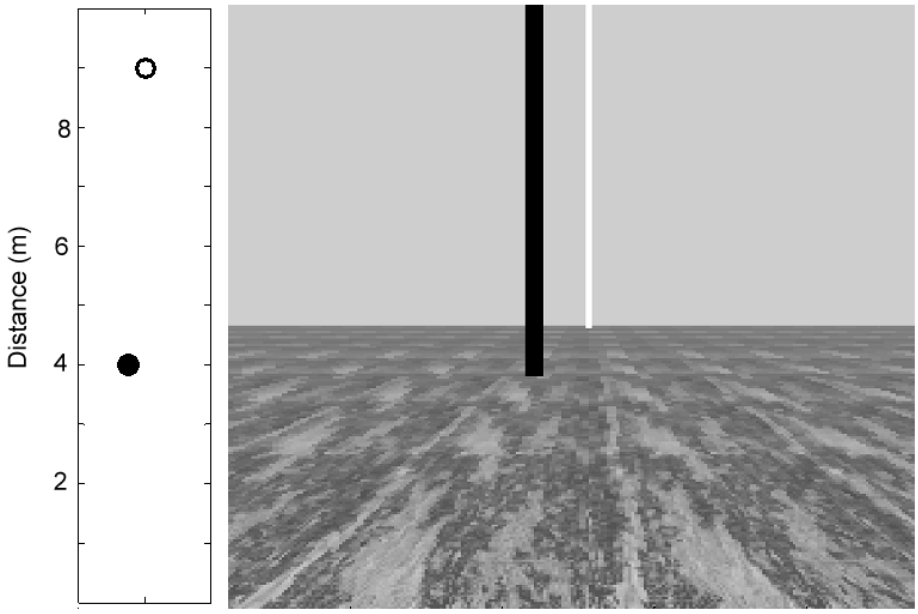

Figure 6: Virtual environment layout. Left panel, overhead view: the solid black circle represents the position of the obstacle, and the empty circle represents the position of the goal. Right panel, model view: the black column is the obstacle, and the white column is the goal.

To use model MST outputs for steering, video sequences were created in a virtual environment that was designed to mimic the Fajen and Warren (2003) experimental environments as closely as possible (Figure 6). Video sequences were generated for each of the seven trajectories shown in Figure 2, panels A and B. These video sequences were processed by the model, and object positional representations in MSTv were assessed. Figure 7 demonstrates how the MSTv activation distributions cluster around object boundaries.

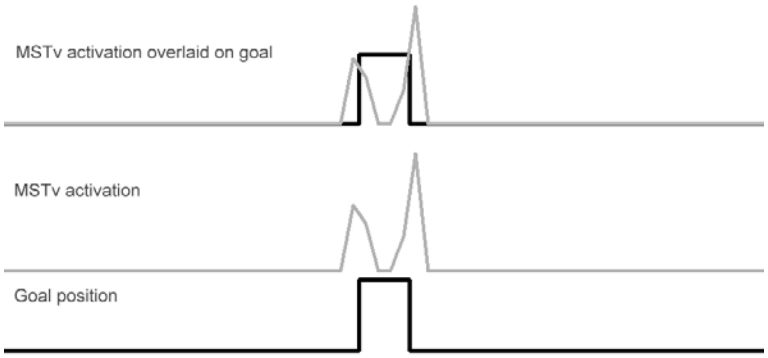

Figure 7: Model MSTv activates around the boundaries of the object. Representative results taken from frame time step 700 when processing trajectory 3 (4m depth, 2^{o} visual angle) from Fajen and Warren (2003).

Figure 8 illustrates the goal object position overlaid with MSTv activation across all seven Fajen and Warren (2003) trajectories. MSTv activations track the boundaries of the objects and, when the object is close to the observer, omission errors become more

prevalent due to the larger texture-free central region of the object. Figure 8 also shows that MSTv activity can occur at positions where no object is present. Such positional errors are generally small. Over 98% of active model cells are within 4 pixels of the object, where 4 pixels represent less than 5% of the width of the input sequence. Many of these positional errors are an artifact of the distributed spatial representation in MSTv, shown in Figure 7.

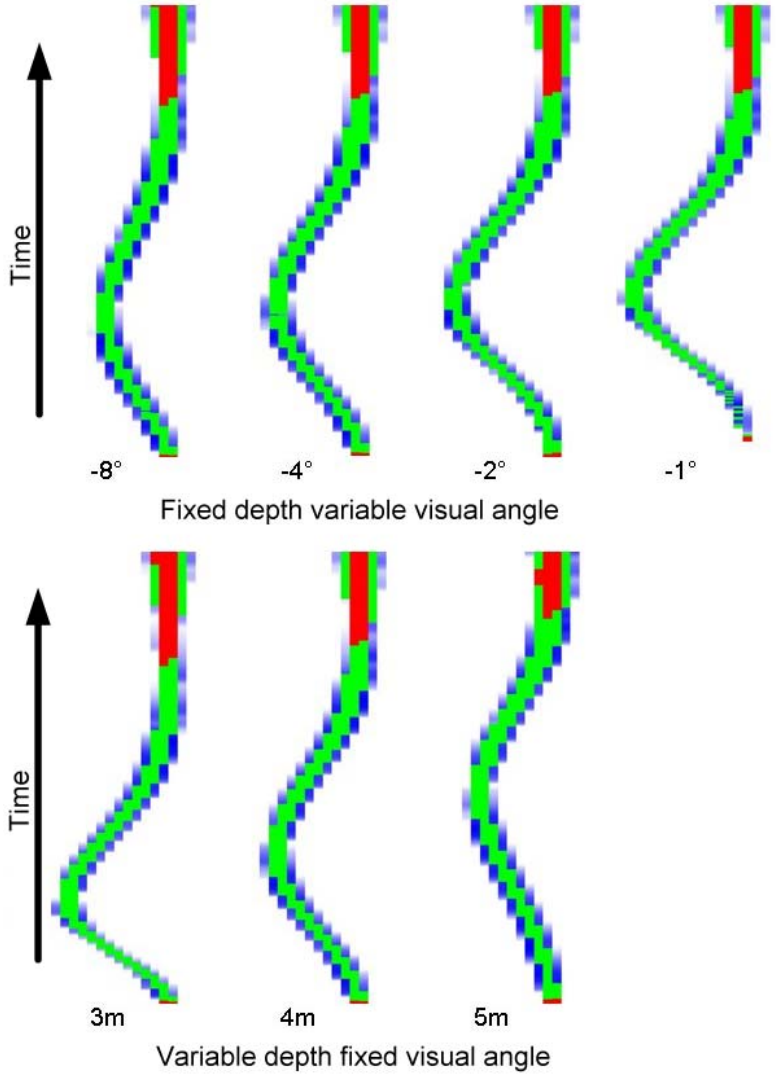

Figure 8: A comparison of model MSTv activations with the position of the goal in the visual input as the observer travels along the trajectories shown in Figure 2 panels A and B. Each horizontal slice displays the outputs from model MSTv to the steering field, color coded to represent whether or not MSTv activation is congruent with the goal position. Green indicates where MSTv activation occurs congruent with the pixel positions of the goal. Blue indicates positional errors, where MSTv activation occurs in positions not occupied by the goal. Red indicates omission errors, where MSTv is not active but the goal occupies those pixel positions. As the observer turns rightward to avoid the obstacle (not shown), the goal moves to the left on the visual image. Once the obstacle is circumnavigated, the observer turns and moves back towards the goal. Positional errors cluster around the boundaries of the goal and, as such, are often an artifact of distributed spatial representations in MSTv. Omission errors occur predominantly when the observer is close to the goal, in these cases the central region of the goal does not move on the visual image and therefore does not generate motion signals.

In the tested stimuli, the goal and obstacle objects have no texture, so only their boundaries generate motion signals. 82% of the object pixels are represented in the MSTv activations, with 18% missed due to positional or omission errors. As the observer approaches the objects, they appear larger in the input image. Therefore central regions of the objects no longer produce motion signals.

To test if these ViSTARS estimates are sufficient for human-like steering performance, the model was tested in a virtual environment using STARS steering dynamics (Elder et al., 2007), as defined in Appendix equations 9.0-9.7. Steering parameters were hand-tuned to emulate qualitatively similar steering behaviors to humans, as described by Fajen and Warren (2003). Parameters for the model from its Retina to MST remained the same as in prior simulations. Results are shown in Figure 2, panels E and F. Figure 2 panel E demonstrates that, for obstacles at a fixed depth, smaller visual angles between observer heading and object position result in larger deviations from a straight line trajectory. Figure 2 panel F demonstrates that, for fixed visual angle, closer objects result in faster, and slightly wider, steering trajectories. These steering properties are also exhibited by humans.

4. Discussion

The ViSTARS model unifies and extends two streams of modeling work: the 3D FORMOTION motion processing model (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg et al., 2001) and the STARS navigational model (Elder et al., 2007). Until now neither model processed natural image sequences to control reactive steering. ViSTARS successfully demonstrates these capabilities.

In particular, ViSTARS demonstrates that the steering mechanisms of the STARS model are capable of human-like performance in response to image-based inputs. STARS used a log-polar transformation in V1, as occurs *in vivo* (Schwartz, 1977; Wagner, Polimeni, & Schwartz, 2005). ViSTARS produced qualitatively human-like steering performance in the absence of the log-polar transformation. Its future inclusion may allow for a closer match to human data.

The 3D FORMOTION model is herein extended to include the parallel cortical streams MT^+/MSTd and MT^-/MSTv, whose complementary properties enable visually based navigation and object tracking, respectively. Additional work is required to integrate this enhanced motion model with the form-motion feedback interactions between V1, V2 and MT that are needed to deal with many situations (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Francis & Grossberg, 1996; Grossberg et al., 2001).

Although ViSTARS offers a good qualitative match to the human data, the differences between steering commands in environments with obstacles at a fixed visual angle (Figure 2, panel F) are relatively small compared with those at a variable visual angle (Figure 2, panel E). These differences are due to how depth is represented in the model. Depth in the ViSTARS model is encoded in two ways: speed (faster objects in a static world are closer), and size (for objects of fixed size, farther objects are smaller). When the trajectory consists of a curved path, speed can be an unreliable measure of depth, since rotations produce the same motion vectors irrespective of depth (Longuet-Higgins & Prazdny, 1980). The influence of rotational information is shown in data wherein humans tend to report curvilinear heading unless specifically instructed to report

tangential heading (Banks, Ehrlich, Backus, & Crowell, 1996; Kelly, Beall, Loomis, Smith, & Macuga; Li & Warren, 2004; Royden, Crowell, & Banks, 1994; Stone & Perrone, 1997b). The ViSTARS representation of speed by the magnitude of MSTv activation is a reliable measure of depth only on straight line trajectories.

Due to the resolution available in our immersed environment (1024 X 64 pixels) and the resulting aliasing, the differences between the widths of the near and far objects are in the range of 1-2 pixels on each side. Thus the depth-from-size information available to ViSTARS is also limited.

Both the STARS model and the Fajen and Warren (2003) model had high precision measures of depth available to them. STARS utilized highly accurate optic flow to provide precise speed measurements, reliably predicting depth in static environments. Fajen and Warren (2003) explicitly incorporated precise distance measurements into their model. ViSTARS demonstrates that only coarse depth estimates are required to compute heading and segment moving objects in order to steer in a human-like fashion, including obstacle avoidance and goal approach. ViSTARS processes low resolution noisy natural image sequences to produce accurate motion estimates using only 3 speeds and 8 directions.

A number of models exhibit properties comparable to portions of ViSTARS. Heading based models and navigation models are discussed in Browning, Grossberg and Mingolla (2008b) and Elder et al. (2007). Here we concentrate on object motion estimation models.

Bayerl and Neumann (2004) modeled how the primate magnocellular pathway processes optic flow. Inputs to their model, in the form of artificial stimuli and natural image sequences, are processed by Elaborated Reichardt detectors (van Santen & Sperling, 1985). Directional normalization in model V1 reduces activity in areas of high ambiguity, and model MT performs long-range directional filtering of motion signals, as in Chey et al. (1997) and Grossberg et al. (2001). Feedback from MT to V1 resolves the aperture problem in V1, which in turn results in a resolved signal in MT. Good object motion and observer motion estimates are hereby generated when the camera is stationary but the system cannot segment object motion from observer motion when the camera (or eye or body) is moving. We believe that dealing with a moving observer is a basic evolutionary pressure that caused the branching of processing in MT, with MT^- computing object motion estimates, and MT^+ developed to allow for estimates of observer motion. The Bayerl and Neumann model (2004) is implemented by solving differential equations at equilibrium and analytically computing the solution. Network dynamics are simulated via an iteration scheme. Their model predicts that V1 should have an aperture-resolved motion estimate on a similar time course to that demonstrated in MT (Pack & Born, 2001), as predicted in Chey et al. (1997). Some cells in V1, such as end-stopped cells, demonstrate an aperture-resolved motion signal over time (Pack, Livingstone, Duffy & Born, 2003). However, at present there is no evidence that these unambiguous signals in V1 propagate to overcome the aperture problem in other regions of V1, as suggested by the Bayerl and Neumann (2004) model. Thus, although Bayerl and Neumann (2004) present interesting results, the model would need to be modified to account for the human behavioral and primate neurophysiological data that ViSTARS explains.

Zemel and Sejnowski (1998) presented a model in which MSTd codes not only observer motion with respect to the environment (heading) but also the relative motion between an observer and a particular object. Inputs were ray-tracings of image sequences depicting observer motion, eye movements, and object motion. MT was defined to represent accurate optic flow describing the scene. A neural network was trained through an unsupervised optimization procedure to encode a compressed representation of motion represented in MT. The resulting compressed representation had receptive fields similar to various MSTd cells. The unsupervised optimization algorithm they used computed connection weights between input, hidden, and output layers such that the output was the same as the input. The hidden layer modeled MSTd and compressed motion in the input into object and environmental motion structure. Their model pooled across all MSTd cells to determine heading. Radial cells in primate MSTd generally have receptive fields covering greater than 50 degrees of the visual field, with many having receptive fields greater than 90 degrees (Duffy & Wurtz, 1997), that are suitable for processing observer motion. However, some MSTd cells have receptive fields of less than 10 degrees; ViSTARS does not presently explain their function. While neither the method used to define flow fields nor the optimization scheme used to tune MT-to-MST connections is biologically plausible, the Zemel and Sejnowski (1998) model does provide evidence that small receptive field MSTd cells could process optic flow to calculate object time-to-contact, or relative depth. The use of optic flow to estimate object motion in depth, or time-to-contact, has been theorized in a number of models, some of which have also been related to MST (Gibson, 1950; Grossberg et al., 1999; Lidén & Pack, 1999; Nakayama & Loomis, 1974).

Nolan and Sejnowski (1993, 1995) presented a model of motion segmentation and velocity integration by primate MT. Like our MT^-/MSTv stream, their model does not attempt to generate a dense veridical motion estimate. Rather, a sparse coarse motion representation is computed. Features are selected based on form information and are then used to determine regions for motion tracking. Their feature tracking implementation allows the model to segment objects where there is conflicting motion information in random dot displays or transparent plaid gratings. Input comes from spatiotemporal energy filters (Adelson & Bergen, 1985) and regions are selected based on global assessments of the input structure. The model produces some interesting results when segmenting objects and resolving conflicting motion estimates. However, the authors do not describe how the neural structures of primate MT^+ and MT^- relate to their model or discuss how the model could be adapted to process natural image sequences.

Wang (1997) presented a model which demonstrated how the differing receptive field properties in MT^+ and MT^- could be learned via competitive learning (Grossberg, 1976a, 1978; Rummelhart & Zipser, 1986). The response properties were found to closely resemble many known features of MT neurons, illustrating how the MT stages of ViSTARS may self-organize.

The 3D FORMOTION model (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg et al., 2001) exploits depth planes for object segmentation via interactions between the form (what) and motion (where) streams. We have not implemented any form-motion interactions for object segmentation in ViSTARS. Instead moving objects are represented by activations in MT^-/MSTv, which are directly used for steering. The form stream is needed to select, and coherently bind, motion signals that

are compatible with form boundaries, notably boundaries wherein multiple forms are separated from each other via figure-ground separation (Baloch & Grossberg, 1997; Berzhanskaya et al., 2007; Grossberg, 1994; Grossberg et al., 2001; Kelly & Grossberg, 2000). Cao, Grossberg and Zaydens (2008) have recently demonstrated how the 3D LAMINART model can generate 3D boundary and surface representations, in cortical areas V1, V2, and V4 of the form stream, in response to natural images. Future work will integrate it with ViSTARS to incorporate the computational benefit of form-motion interactions.

The STARS model (Elder et al., 2007) did not explain the interception of moving targets. Fajen and Warren (2004) have demonstrated that humans tend to take an interception strategy that is somewhere between a pursuit strategy, where the target path is followed, and a constant bearing strategy, where a minimal path to the intersection point of the observer path with the target path is taken. Fajen and Warren (2004) required extensive updates to their 2003 model in order to explain these data. Dessing, Peper, Bullock & Beek (2005) demonstrated how visually derived information can be used to catch a moving ball. In order to match human data, their model utilized object speed and direction of motion estimates to modulate interception movements. We hypothesize that by incorporating speed and direction of object motion into the STARS steering component, rather than relying solely on positional information, ViSTARS may have enough information to explain interception data. This hypothesis will be tested in future work.

Finally, model computation time will have to improve before it is feasible to implement our model as a robotic control system. We have reduced computation time by using image resizing rather than multiple filter sizes, and through coarse integration time steps. In our MATLAB implementation, 8 seconds of input still require roughly 2.5 hours of simulation time. In contrast, STARS ran at frame-rate on a GPU when implemented carefully and in a slightly pared-down form (Elder, Grossberg, & Mingolla, 2005). We have replaced STARS visual processing layers with more complicated recurrent neural networks, but at the same time GPU power has increased dramatically since 2005. Subsequent work will investigate the feasibility of operating the ViSTARS model at frame-rate on a GPU. Reduced versions of the model are currently capable of running on a modern GPU at 25 frames-per-second for input resolutions of 256x256, indicating good potential for the production of a real time system.

5. Conclusion

The ViSTARS neural model processes natural image streams for the purposes of object segmentation, reactive steering, and navigation. The model elucidates motion representations in cortical areas V1, MT, and MST, and describes how these representations may be generated from noisy inputs. The model has been tested using video streams from a variety of sources. In each case, model MSTv produces accurate positional, directional and speed estimates of objects in the scene at sufficient accuracies to emulate human-like reactive navigation. The ViSTARS model, along with the STARS and 3D FORMOTION models, fits into an emerging framework for heading estimation, navigation, object tracking, eye movements, and form-motion interactions that together contribute to designing a large-scale neural controller for a visually-responsive autonomous mobile robotic system.

6. 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).

7. Appendix A. Model equations, parameters and implementation

All stages of the model are defined by differential equations and were numerically integrated using Euler's method. The resolution and frame rate were defined by the input source. Table 1 describes each input source.

Table 1: Input sequence resolution, frame rate and sequence length. Note that the frame rate for Hamburg Taxi sequence is defined based on model parameterization.

	Resolution (pixels)	*Frame rate (fps)*	*Length (frames | s)*
Hamburg taxi	191x256	15	41 | 2.7
Overtaking driving video	320x240	15	26 | 1.7
Fajen & Warren (2003) trajectories	256x256	100	750 | 7.5
Immersed environment	1024x64	47	376 | 8

Calibration was performed by fitting the time-course of the transient cell layer (level 1 of the model) such that the peak of the burst response occurred after 70-75ms, in accordance with primate non-directional transient cells (Benardete & Kaplan, 1999). Integration was not performed to equilibrium. Instead, the model activations ebb and flow as the input stream evolves. Figure 1 describes the functional stages of the model with respect to their equation numbers and variable labels.

Model stages are designed to elucidate how biological neurons work. Each differential equation specifies the activation state of individual neurons or neuron populations. Model cells are typically controlled by shunting, or membrane, equations (Grossberg, 1973) that perform 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 spiking neuron.

In the equations that follow, lowercase letters correspond to variables, and uppercase letters correspond to output signals. For example, r corresponds to the activity of MSTd cells, whereas R corresponds to an output signal from MSTd cells. Subscript indices correspond to spatial positions. Superscript indices correspond to non-spatial dimensions, such as speed or direction. Uppercase indices denote dummy indices. Parameters are shown 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.

The multi-scale input transformation and model levels 1 to 6 were used to compute heading in Browning, Grossberg & Mingolla (2008b). Our treatment closely follows their description and we use identical parameters.

Input (g). Video input is converted to grayscale and scaled between 0 and 1. Videos tested were: Fajen and Warren (2003) trajectories, a driving video stimulus set, the Hamburg taxi sequence, and an immersive environment. Details on the resolution, frame-rate and length of each source are shown in Table 1. Function $g_{ij}^{\delta}(t)$ represents the video input; it is indexed by spatial position (*i,j*) and disparity (δ). In the virtual environment, input was generated at two disparities, near depth, and fixation depth. In all other tests the input stream was defined at a single disparity.

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).

All resizing algorithms introduce some form of aliasing, although some are more benign than others. 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.

The grayscale intensity values of the resized input stream are defined in the ON-*channel* (equation 0.2)*,* and the complement of its activity is defined in the OFF-*channel* (equation 0.3). Complement coding to define the OFF channel was first described in the DISCOV model (Chelian & Carpenter, 2005). Equations (0.2) and (0.3) define $I_{ij}^{\delta ps}$, the multi-scale input, indexed by spatial position (*i, j*), disparity (δ), ON/OFF channel (p=1,2), and scale (s = 1,2,3):

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

and

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

In equation 0.2, g_{XY}^{δ} is the input intensity at location (*X,Y)* and disparity (δ), i = 1, 2, …, $\frac{i_{\max}}{n}$, j = 1, 2, …, $\frac{j_{\max}}{n}$, i_{max} = horizontal resolution, 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:

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

In equation (1.1), $a_{ij}^{\delta ps}$ is the cell activity at position (*i,j*), disparity (δ), 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 $D_1 = 0.25$. Function $I_{ij}^{\delta ps}$ is the input from equations (0.2) and (0.3). F_{ijXY} is a Gaussian inhibitory surround kernel, truncated to a 7x7 filter:

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

where F_1 scales the inhibitory kernel gain, and σ_1 is the inhibitory kernel variance. In our simulations, $F_1 = 10.225$, and $\sigma_1 = 1$. This value of F_1 was chosen by Browning, Grossberg & Mingolla (2008b) through parameter search to provide the best results across a range of input stimuli. The output signal $\gamma_{ij}^{\delta ps}$ is a sigmoid function of activity $a_{ij}^{\delta ps}$:

. (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 $\varphi_1 = 0.1$.

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

$$b_{ij}^{\delta ps} = x_{ij}^{\delta ps} z_{ij}^{\delta ps}, \quad (2.1)$$

where cell activities, $x_{ij}^{\delta ps}$, perform leaky integration on their inputs $\gamma_{ij}^{\delta ps}$ (equation 1.3):

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

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

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

(Grossberg, 1968, 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, $x_{ij}^{\delta ps}$ approaches C_2 at a rate proportional to $(C_2 - x_{ij}^{\delta ps})$. 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 $x_{ij}^{\delta ps}$. When $x_{ij}^{\delta ps}$ is zero, activity at $z_{ij}^{\delta ps}$ recovers to 1 at rate D_2. In our simulations, $D_2 = 0.01$, and $K_2 = 20$.

Activity $x_{ij}^{\delta ps}$ is gated by the habituative transmitter $z_{ij}^{\delta ps}$ to generate transient non-directional cell activities $b_{ij}^{\delta ps}$. For visual inputs with a short dwell time, such as moving boundaries, activities $b_{ij}^{\delta ps}$ respond throughout their duration. A static input on the other hand, produces only a weak response after an initial transient burst of activation.

Level 3: Directionally selective transient cells (E). This model level defines directionally selective cells that can retain their sensitivity in response to variable speed inputs (Chey et al., 1997). Eight directions were implemented at 45 degree increments. A key design that enables variable speed selectivity is the use of directional inhibitory interneuron activities, $c_{ij}^{\delta psd}$:

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

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

Activity $c_{ij}^{\delta psd}$ increases proportionally to input $b_{ij}^{\delta ps}$ with coefficient A_3C_3 and decays to zero with rate $A_3B_3\, c_{ij}^{\delta psd}$. The strength of opponent inhibition is $K_3\left[c_{XY}^{\delta psD}\right]^+$. 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, $e_{ij}^{\delta psd}$, combine transient input $b_{ij}^{\delta ps}$, with inhibitory interneuron activity, $c_{ij}^{\delta psd}$. Their dynamics are similar to those of $c_{ij}^{\delta psd}$:

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

Activity $e_{ij}^{\delta psd}$ increases proportionally to transient input $b_{ij}^{\delta ps}$, passively decays with a 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 $e_{ij}^{\delta psd}$:

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

Level 4: Directional competition (f). Due to the 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). Cross-directional normalizing competition enhances the least ambiguous regions and suppresses the most ambiguous regions, thus strengthening feature tracking signals which help reduce the effects of the aperture problem (Bayerl & Neumann, 2004; Berzhanskaya et al., 2007; Chey et al., 1997; Lucas & Kanade, 1981; Mingolla, Todd & Norman, 1992). ON and OFF channel directional transient cell inputs are added together at this stage, and competitively normalized across direction:

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

In equation (4.1), activity, $f_{ij}^{\delta sd}$, integrates excitatory input from the directional transient cells across channels (p) at the same disparity (δ), position (i, j), scale (s), and directional preference (d), and is suppressed by directional transient cells at the same scale and position, from both channels, with directional preferences $D \neq d$. 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 scale $s = 3$. Variable $m_{ij}^{\delta sd}$ computes the mean activity across groups of cells with the same scale and directional selectivity:

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

where $N_s = 2^{3-s}$, for 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}^{δ}. Note that this definition of N_s in (4.2) is different from that of n_s equation (0.2). In equation (0.2), scale 1 is at the original input resolution, scale 2 is reduced by a factor of 2, and scale 3 is reduced by a factor of 4. In equation (4.2), scale 1 is reduced by a factor of 4, scale 2 is reduced by a factor of 2, and scale 3 does not change.

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

$$\frac{dq_{ij}^d}{dt} = -A_6 q_{ij}^d + \left(B_6 - q_{ij}^d\right)\left(\left(\sum_{XY} L_{ijXY}^d \left(\sum_s N_s \sum_{\delta} m_{XY}^{\delta sd}\right)\right)\left(1 + \frac{C_6}{M_6} \sum_z R_z w_{ijz}^d\right) + D_6 Q_{ij}^d\right) - q_{ij}^d \sum_D v_{dD} Q_{ij}^D \quad (5.1)$$

In equation (5.1), excitatory input signals $m_{ij}^{\delta sd}$ from equation (4.2) are added across disparity (δ), and scale (s) with weights $N_s = 2^{3-s}$ to account for the low density of signals at lower scales, 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 output from the cell population with heading z (equation (6.2)), and w_{ijz}^d defines the flow filter associated with heading z, at spatial position (i, j) and directional selectivity (d); see Level 6. Recurrent connections Q_{ij}^D within equation (5.1) implement a choice network via self-excitation and lateral inhibition across direction from cells in the same position. The inhibitory strength is governed by the kernel v_{dD} (equation 5.4). 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, and D_6 is 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^d_{ijXY}, is an anisotropic Gaussian elongated along the filter's direction of selectivity:

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

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 $\sigma_y = 2$. Output from level 5 is half-wave rectified and squared:

$$Q^d_{ij} = \left(\left[q^d_{ij} - \theta_6\right]^+\right)^2, \tag{5.3}$$

where θ_6 = 0.2 is the signal threshold. The lateral inhibition weighting function is defined as follows:

$$v_{dD} = \begin{cases} 0 & D = d \\ 0.5 & D = d \pm 45° \\ 1 & D = d \pm 90° \\ 1 & D = d \pm 135° \\ 10 & D = d \pm 180° \end{cases}. \tag{5.4}$$

Function v_{dD} represents a distributed opponent inhibition function. Browning, Grossberg & Mingolla (2008b) demonstrated that many types of opponent inhibition can produce accurate and robust results. The function shown in equation (5.4) produced the best reported results across a range of stimuli.

Level 6: Heading filter (R). Flow filters, or templates, w^d_{ijz}, were generated to match 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 Browning et al. (2008b), this is consistent with filters learned using a self-organizing map (Cameron et al., 1998; Elder et al., 2007). A row of flow filters was created corresponding to headings at 1/2 the vertical resolution of the input. Within the row, heading cells were spaced at every third pixel, starting at pixel 2. Heading cell activity, r_z, results from matching the flow filters, w^d_{ijz}, against inputs Q^d_{ij} from level 5 (equation 5.3):

$$\frac{dr_z}{dt} = -A_7 r_z + \left(B_7 - r_z\right)\left(\frac{C_7}{N_7}\sum_d\sum_{ij} w^d_{ijz}Q^d_{ij} + D_7 R_z\right) - r_z\left(E_7\sum_{\varepsilon\neq z} R_\varepsilon\right), \tag{6.1}$$

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

flow filter, defined as the sum of all values in the filter. Self-excitation and mutual inhibition via a sigmoid feedback signal

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

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 $E_7 = 0.25$. In equation (6.2), parameter $G_7 = 0.1$ defines the value at which the sigmoid signal attains one-half of its maximum value, and $\theta_7 = 0.2$ is a signal threshold. For simplicity, the feedback and output signals R_z are the same.

Level 7: Differential motion filter (Ω). To determine motion discontinuities in the input, differential motion filters process input from level 4 via a directionally-tuned ON-center OFF-surround network:

$$\begin{aligned}\frac{d\omega_{ij}^{\delta sd}}{dt} = -A_8\omega_{ij}^{\delta sd} + \left(B_8 - q_{ij}^{\delta sd}\right)\left(\left(\sum_{XY} L_{ijXY}K_s m_{XY}^{\delta sd}\right)\left(1 + C_8\Psi_{ij}^{\delta d}\right) + D_8\Omega_{ij}^{\delta sd}\right) \\ - \omega_{ij}^{\delta sd}\left(E_8\sum_D \frac{u_{dD}}{u_d}\sum_{XY} G_{ijXY}m_{XY}^{\Delta sD} + F_8\sum_D v_{dD}\Omega_{ij}^{\Delta sD}\right)\end{aligned}. \tag{7.1}$$

In equation (7.1), cell activity $\omega_{ij}^{\delta sd}$ computes motion boundaries. The ON-center, $\sum_{XY} L_{ijXY}K_s m_{XY}^{\delta sd}$, receives excitatory input $m_{ij}^{\delta sd}$ from level 4 (see equation (4.2)), across positions (X,Y), at scale (s), direction selectivity (d), and disparity (δ). These inputs are integrated by the Gaussian filter

$$L_{ijXY} = \frac{L_8}{2\pi\sigma_x^2}\exp\left(-0.25\left(\frac{(X-i)^2 + (Y-j)^2}{\sigma_x^2}\right)\right), \tag{7.2}$$

where $L_8 = 0.25$ is the filter gain, and $\sigma_x = 0.5$ is its variance. L_{ijXY} is weighted by scale parameter K_s, where $K_1 = 2$, $K_2 = 5$, and $K_3 = 9$. Top-down modulatory feedback from the object motion level 8, $\Psi_{ij}^{\delta d}$, enhances congruent object motion estimates. The OFF-surround, $E_8\sum_D \frac{u_{dD}}{u_d}\sum_{XY} G_{ijXY}m_{XY}^{\Delta sD}$, competes across depths $\Delta \neq \delta$, and positions (X,Y), within each scale (s) via Gaussian kernel

$$G_{ijXY}^d = \frac{G_8}{2\pi\sigma_y^2}\exp\left(-0.25\left(\frac{(X-i)^2 + (Y-j)^2}{\sigma_y^2}\right)\right) \tag{7.3}$$

where $G_8 = 0.57$ is the filter gain, $\sigma_y = 1.5$ is its variance, and directional kernel

$$u_{dD} = \begin{cases} 10 & D = d \\ 1 & D = d \pm 45^{\circ} \\ 0.25 & D = d \pm 90^{\circ} \\ 0 & otherwise \end{cases} \tag{7.4}$$

where u_{dD} is normalized via division by the sum u_d over all u_{dD}, which equals 12.50. Recurrent shunting ON-center OFF-surround feedback signals

$$\Omega_{ij}^{\delta sd} = \left(\left[\omega_{ij}^{\delta sd} - \theta_8 \right]^+ \right)^2, \tag{7.5}$$

where $\theta_8 = 0.1$ is the signal threshold, implement a choice network via directional competition from cells at the same position. The directional inhibitory coefficients v_{dD} are defined by equation (5.4). In our simulations, $A_8 = 0.5$, $B_8 = 1$, $C_8 = 0.05$, $D_8 = 0.05$, $E_8 = 0.25$, and $F_8 = 0.05$.

Level 8: Object motion (Ψ). Motion boundary outputs, $\Omega_{ij}^{\delta sd}$ (equation (7.4)), are grouped across scale to produce activations, $\psi_{ij}^{\delta d}$, that are proportional to object speed:

$$\frac{d\psi_{ij}^{\delta d}}{dt} = -A_9 \psi_{ij}^{\delta d} + \left(B_9 - \psi_{ij}^{\delta d} \right) \left(C_9 \sum_s w_s \Omega_{ij}^{\delta sd} + D_9 \Psi_{ij}^{\delta d} \right) - E_9 \psi_{ij}^{\delta d} \sum_{D \neq d} \Psi_{ij}^{\delta D} . \tag{8.1}$$

In equation (8.1), excitatory input from $\Omega_{ij}^{\delta sd}$ at the same position (*i,j*), directional selectivity (*d*), and depth (δ) is summed across scale with weights w_s, such that $w_1 = \frac{1}{6}$, $w_2 = \frac{1}{3}$ and $w_3 = \frac{1}{2}$. This weighting function ensures that objects in higher scales are always seen to be moving faster than those in lower scales. A recurrent ON-center OFF-surround network with directional competition and sigmoid signals

$$\Psi_{ij}^{\delta d} = \frac{\left(\left[\psi_{ij}^{\delta d} - \theta_8 \right]^+ \right)^2}{G_8^2 + \left(\left[\psi_{ij}^{\delta d} - \theta_8 \right]^+ \right)^2}, \tag{8.2}$$

implements a contrast enhancing network (Grossberg, 1973), with $G_8 = 0.1$, and $\theta_8 = 0.2$, within equation (8.1). Parameters, $A_9 = 0.5$, $B_9 = 1$, $C_9 = 2$, $D_9 = 1$, $E_9 = 2$.

STARS steering module

The STARS (Elder et al., 2007) steering module was adapted for embedding within ViSTARS. Steering commands to the left or right depend on the horizontal positions of objects and heading. Object motion outputs are therefore computed along horizontal lines and summed across direction:

$$p_j^{\delta} = \sum_d \Psi_{\lambda j}^{\delta d} . \tag{9.0}$$

In equation (9.0), activity in p_j^{δ} at horizontal position *j,* and disparity δ, groups object activations across direction at vertical position $\lambda = 12$.

STARS defines objects at fixation depth, $\delta = F$, as goals and at nearer depths, $\delta = N$, as obstacles. Therefore:

$$g_j = p_j^F \tag{9.1}$$

and

$$o_j = p_j^N, \tag{9.2}$$

where *g* denotes goal, *o* denotes obstacle, and *p* is defined in equation (9.0). Heading cell output R_z (equation (6.2)), is relabeled: $h_j = R_z$ where horizontal position *j* maps directly on to the horizontal position of heading *z*. The steering field, s_j, can then be defined as:

$$s_j = A\sum_Y G_{jY}^g g_Y + Bh_j - C\sum_Y G_{jY}^o o_Y \text{ .} \tag{9.3}$$

Goals, g_Y, excite the steering field via the Gaussian kernel

$$G_{jY}^g = \frac{1}{\sigma^g\sqrt{2\pi}}\exp\left(-\left(\frac{j-Y}{\sigma^g}\right)^2\right) \tag{9.4}$$

where $\sigma^g = 1.25$. Heading, h_j, excites the steering field. Heading is represented as a Gaussian-like activation distribution across position *j*, and as such no additional kernel is required. Obstacles, o_Y, inhibit the steering field via a Gaussian kernel

$$G_{jY}^o = \frac{1}{\sigma^o\sqrt{2\pi}}\exp\left(-\left(\frac{j-Y}{\sigma^o}\right)^2\right) \tag{9.5}$$

where $\sigma^o = 0.25$. Parameter *A* scales the goal excitation, *B* scales the heading excitation, and *C* scales the obstacle inhibition. In our simulations, *A* = 4, B = 1, and C = 4.

The peak of the steering field:

$$\tilde{S} = \arg\max_j(s_j) \tag{9.6}$$

is used to define the steering command from the rate of change of heading angle,

$$\frac{d\phi}{dt} = \tilde{S} - N, \tag{9.7}$$

where *N* = 128.5 is the position of the mid-line of the stimulus input. The virtual environment is simulated with a horizontal resolution of 256 pixels at area MST (1024 / 4); the horizontal midline therefore lies midway between pixel 128 and pixel 129. Heading changes of less than 1 pixel on any given time step are suppressed:

$$\frac{d\phi'}{dt} = \begin{cases} 0 & \left|\frac{d\phi}{dt}\right| < 1 \\ \mathrm{H}\frac{d\phi}{dt} & \text{otherwise,} \end{cases} \tag{9.8}$$

where H scales the steering command in relation to observer translation speed. In our simulations, the observer moves forward at a speed of $1ms^{-1}$, there are 47 image frames per second, and rotations are scaled by a factor of 5, so that $H = \frac{5}{47}$.

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.

Aggarwal, J., & Nandhakumar, N. (1988). On the computation of motion from sequences of images-A review. *Proceedings of the IEEE*, *76*(8), 917-935. doi: 10.1109/5.5965.

Albright, T.D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, 52(6), 1106-30.

Allman, J., Miezin, F., & McGuinness, E. L. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience*, *8*, 407-430.

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

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.

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

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.

Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey. *J Neurophysiol*, *84*(5), 2658-69.

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. doi: 10.1038/357497a0.

Bradley, D. C., & Andersen, R. A. (1998). Center-Surround Antagonism Based on Disparity in Primate Area MT. *Journal of Neuroscience*, *18*(18), 7552-7565.

Bradski, G. R. & Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library.* O'Reilly Press, Cambridge, MA.

Browning, N. A., Mingolla, E., & Grossberg, S. (2007a). Heading from optic flow of natural scenes during motion processing by cortical areas MT and MST. *Program*

No. 337.12. 2007 Neuroscience Meeting Planner. San Diego, CA: Society for Neuroscience. Online.

Browning, N. A., Mingolla, E., & Grossberg, S. (2007b). Heading from optic flow in a neural model of primate motion processing and navigation. In *Eleventh International Conference on Cognitive and Neural Systems Proceedings. May 2008.* Boston, MA.

Browning, N. A., Mingolla, E., & Grossberg, S. (2007c). Computing heading and optic flow in MT-MST: A neural model of primate motion processing. *Computational Cognitive Neuroscience Conference. San Diego, CA: November 2007*. Retrieved from http://www.ccnconference.org/page10.html.

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

Browning, N. A., Mingolla, E., & Grossberg, S. (2008b). A neural model of how the brain computes heading from optic flow in realistic scenes. *(submitted for publication)*

Cao, Y., Grossberg, S., & Zaydens, E. (2008). A laminar cortical model of stereopsis and 3D surface perception of complex natural scenes [Abstract]. *Journal of Vision*, *8*(6):850, 850a, http://journalofvision.org/8/6/850/, doi:10.1167/8.6.850.

Callaway, E. M. (2005). Structure and function of parallel pathways in the primate early visual system. *J Physiol*, *566*(1), 13-19. doi: 10.1113/jphysiol.2005.088047.

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, with applications to image processing and classification. *In Proceedings of AIC05: 10th congress of the international colour association,* Granada, Spain, May.

Cao, Y., Grossberg, S., & Zaydens, E. (2008). A laminar cortical model of stereopsis and 3D surface perception of complex natural scenes [Abstract]. *Journal of Vision*, *8*(6):850, 850a, http://journalofvision.org/8/6/850/, doi:10.1167/8.6.850.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*(1), 54-115.

Carpenter, G., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, *21*(3), 77-88. doi: 10.1109/2.33.

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.

Cleland, B. G., Dubin, M. W., & Levick, W. R. (1971). Sustained and transient neurones in the cat's retina and lateral geniculate nucleus. The Journal of Physiology, 217(2), 473-96.

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

Del Viva, M. M., & Morrone, M. C. (1998). Motion analysis by feature tracking. *Vision Research*, *38*(22), 3633-3653.

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.

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. (2007). A neural model of visually guided steering, obstacle avoidance, and route selection. *Boston University Technical Report - CAS/CNS-TR-07-009*.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. J Physiol, 187(3), 517-552.

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.

Francis, G., & Grossberg, S. (1996). Cortical dynamics of form and motion integration: Persistence, apparent motion, and illusory contours. Vision Research, 36(1), 149-173.

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.

Grossberg, S. (1968). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Science**s,*** 59 , 368-372.

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

Grossberg, S. (1976a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*(3), 121-134. doi: 10.1007/BF00344744.

Grossberg, S. (1976b). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*(4), 187-202.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), Progress in theoretical biology, Volume 5. New York: Academic Press, pp. 233-374.

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., (1994). 3-D vision and figure-ground separation by visual cortex,. *Perception & Psychophysics*, 55, 48-121.

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

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., 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.

Hildreth, E. C. (1992). Recovering heading for visually-guided navigation. *Vision Research*, *32*(6), 1177-1192.

Hildreth, E. C., & Royden, C. S. (1998). Computing observer motion from optical flow. *High-Level Motion Processing*, 269–293.

Hodgkin, A. L. (1964). *The Conduction of the Nerve Impulse*, C C. Thomas, Springfield, Illinois.

Hubel, D. H. and Wiesel, T. N. (1959). *Receptive fields of single neurones in the cat's striate cortex.* J Physiol, 148, 574-591.

Hubel, D. H., Wiesel, T. N. (1962). *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.* J. Physiol. Lond. 160: 106-154.

Hubel, D. H., Wiesel, T. N. (1968). *Receptive fields and functional architecture of monkey striate cortex.* J. Physiol. Lond. 195: 215-243.

Institut fuer Algorithmen und Kognitive Systeme. . Retrieved May 23, 2008, from http://i21www.ira.uka.de/image_sequences/#taxi.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science*. Appleton & Lange.

Kaplan, E., & Benardete, E. (2001). The dynamics of primate retinal ganglion cells. *Prog Brain Res*, *134*, 17-34.

Kelly, J. W., Beall, A. C., Loomis, J. M., Smith, R. S., & Macuga, K. L. Simultaneous measurement of steering performance and perceived heading on a curving path. *ACM Trans. Appl. Percept.*, *3*(2), 83-94. doi: 10.1145/1141897.1141898.

Kelly, F. & Grossberg, S., (2000). Neural Dynamics of 3-D Surface Perception: Figure-Ground Separation and Lightness Perception. *Perception and Psychophysics*, 62(8), 1596-1618.

Langer, M. S., & Mann, R. (2003). Optical snow. *International Journal of Computer Vision*, *55*(1), 55-71. doi: 10.1023/A:1024440524579.

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.

Lidén, L., & Pack, C. C. (1999). The role of terminators and occlusion in motion integration and segmentation: A neural solution. *Vision Research*, *39*, 3301–3320.

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

Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J. Neurosci.*, *7*(11), 3416-3468.

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.

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*.

Maunsell, J. H., & Van Essen, D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *Journal of Neurophysiology*, *49*(5), 1127-1147.

McOwan, P. W., Benton, C., Jason, D., & Johnston, A. (1999). A multi-differential neuromorphic approach to motion detection. *International Journal of Neural Systems*, *9*(5), 429-434.

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. doi: 10.1167/8.6.1154.

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.

Nakayama, K., & Loomis, J. M. (1974). Optical velocity patterns, velocity-sensitive neurons, and space perception: a hypothesis. *Perception*, *3*(1), 63-80.

Nowlan, S. J., & Sejnowski, T. (1993). Filter selection model for generating visual motion signals. *Advances in Neural Information Processing Systems*, *5*, 369-376.

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, *15*(2), 1195-1214.

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

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., Livingstone, M. S., Duffy, K. R., & Born, R. T. (2003). End-stopping and the aperture problem: Two-dimensional motion signals in Macaque V1. *Neuron*, *Vol 39*, 671-680.

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.

Ponce, C. R., Lomber, S. G., & Born, R. T. (2008). Integrating motion and depth via parallel pathways. *Nature Neuroscience*, *11*(2), 216-223.

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

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

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

Rumelhart, D. E., & Zipser, D., (1986). *Feature discovery by competitive learning.* In Rumelhart & McClelland (Eds.), Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. MIT Press Cambridge, MA, USA. pp. 151-193.

Rushton, S. K., Bradshaw, M. F., & Warren, P. A. (2007). The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition*, *105*(1), 237-245.

Rushton, S. K., & Warren, P. A. (2005a). Perception of object movement during self-movement. *Proceedings of SPIE*, *5666*, 473.

Rushton, S. K., & Warren, P. A. (2005b). Moving observers, relative retinal motion and the detection of object movement. *Current Biology*, *15*(14), 542-543.

Sachtler, W. L., & Zaidi, Q. (1995). Visual processing of motion boundaries. *Vision Research*, *35*(6), 807-826.

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

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. I. Spatiotemporal organization of receptive fields. *J Neurophysiol*, *39*(6), 1288-1319.

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.

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.

Tanaka, K., Sugita, Y., Moriya, M., & Saito, H. (1993). Analysis of object motion in the ventral part of the medial superior temporal area of the macaque visual cortex. *J Neurophysiol*, *69*(1), 128-142.

Valois, R., Albrecht, R., & Thorell, L. G. (1982). Spatial Frequency Selectivity of Cells in Macaque Visual Cortex. Vision Research, 22, 545-559.

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.

Wang, R. (1997). A Network Model of Motion Processing in Area MT of Primates. *Journal of Computational Neuroscience*, *4*(4), 287-308.

Warren, P. A., & Rushton, S. K. (2007). Perception of object trajectory: Parsing retinal motion into self and object movement components. *Journal of Vision*, *7*(11), 2.

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

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.