Communicated by David Heeger

Functional Consequences of an Integration of Motion and Stereopsis in Area MT of Monkey Extrastriate Visual Cortex

Markus Lappe
Department of General Zoology and Neurobiology,
Ruhr-University-Bochum, D-44780 Bochum, Germany

Experimental evidence from neurophysiological recordings in the middle temporal (MT) area of the macaque monkey suggests that motion-selective cells can use disparity information to separate motion signals that originate from different depths. This finding of a cross-talk between different visual channels has implications for the understanding of the processing of motion in the primate visual system and especially for behavioral tasks requiring the determination of global motion. In this paper, the consequences for the analysis of optic flow fields are explored. A network model is presented that effectively uses the disparity sensitivity of MT-like neurons for the reduction of noise in optic flow fields. Simulations reproduce the recent psychophysical finding that the robustness of the human optic flow processing system is improved by stereoscopic depth information, but that the use of this information depends on the structure of the visual environment.

1 Introduction

Visual tasks are often defined with reference to only one specific visual modality, such as, for instance, motion. But in natural situations, biological vision systems usually have access to a number of additional visual or extraretinal signals that could support each other's functionality and add to solving the task. This has traditionally been largely ignored by many computational vision schemes, which focused on the understanding of specific vision mechanisms. But now a merging of different modalities or visual maps is receiving more attention in the computer vision and robotics communities. It is referred to as "sensor fusion" (Clark and Yuille 1990). In this paper, a biologically plausible neural model of the functional consequences of an integration of motion and stereopsis is presented, which uses depth information to increase its robustness against noise in a visual navigation task.

For visual navigation in an unknown environment, the optic flow field has long been considered a major source of information (Gibson 1950). In the case of a general self-motion in a rigid environment, i.e., observer translation and rotation with respect to a static scene, the task of heading detection from optic flow mathematically involves solving a large number of equations in a large number of unknowns (Koenderink and van Doorn 1987): Besides the observer's rotation and translation direction, the distances of the visual objects in the scene are also unknown. An optimization scheme implemented in a neural network can solve this task within the psychophysically measured limits of human observers. This network shares a number of features with the known properties of neurons in primate extrastriate areas MT and MST (Lappe and Rauschecker 1993, 1995).

However, it has also been demonstrated that human heading detection is heavily influenced by other visual modalities, most notably extraretinal eye movement signals (Warren and Hannon 1990; van den Berg 1993; Lappe *et al.* 1994). But the problem of heading detection from optic flow could also be simplified when the depths of the visible objects were explicitly known (Ballard and Kimball 1983), signaled for instance by the stereoscopic system. Mathematically, the optic flow field is a function of the observer's translation, his (eye) rotation, and the distances of the visible objects. Prior knowledge of any of these parameters would simplify the determination of the rest. Indeed, a recent psychophysical study was the first to show an effect of disparity on heading judgments in that the robustness against noise is strongly increased when optic flow fields are presented stereoscopically (van den Berg and Brenner 1994b). This poses the question of the neuronal mechanisms that support this implicit use of stereoscopic depth. The aim of this work is to investigate whether the recently observed disparity dependence of MT neurons could serve this function.

2 Neurophysiological Findings of Disparity Sensitivity in Area MT

It has been known for some time now that motion-selective neurons in area MT of the macaque monkey also exhibit broad disparity selectivity, but are insensitive to motion in depth (Maunsell and Van Essen 1983a). But in a recent study, Bradley *et al.* (1995) found an interesting specific disparity dependence of MT responses to transparent motion. Previously it was demonstrated (Snowdon *et al.* 1991) that the response of an MT cell to the motion of random dots in the cell's preferred direction is strongly reduced when a second, transparent dot pattern moves in the opposite direction. Bradley *et al.* (1995) now showed that in most neurons, this response reduction occurs only when the disparity difference between the two countermoving dot patterns is within a certain limited range. When both patterns are clearly separated in depth, no response reduction is observed.

This property of MT neurons might serve as the basis for the increased robustness against noise when optic flow stimuli are presented stereoscopically to human subjects. For optic flow fields simulating self-motion in a static, structured environment, visible objects close to each other in space usually give rise to similar optical velocities while objects separated in depth move at different optical velocities. Thus, a spatial averaging of the visual motion signals within a restricted disparity range might improve the representation of the optic flow field in area MT, and provide an enhanced, noise-reduced input to optic flow processing neurons in the medial superior temporal (MST) area.

The structure of the optic flow representation in MT is an important parameter for the modeling of system capabilities of heading detection (Lappe and Rauschecker 1995). In the following, the consequences of the disparity dependence of the motion signal averaging in MT for the flow field analysis are explored. To this end, a simple functional model of the integration of motion and stereopsis for the task of determining heading will be used. The main concern of this model is the representation of the optic flow field in area MT, taking into account the observed disparity dependence. To evaluate its implications for the determination of self-motion, presumably taking part in area MST, a heading detection scheme developed earlier (Lappe and Rauschecker 1993) is adopted.

3 Visual Computation of Heading

The algorithm for heading detection determines the most likely direction **T** by minimizing a certain residual function $R(\mathbf{T})$ (Heeger and Jepson 1992). The neural implementation of this scheme solves the minimization by determining $R(\mathbf{T})$ for various candidates $\mathbf{T}_j$, and then choosing the optimum $\mathbf{T}_j$ by a winner-take-all mechanism. This computation involves only two layers of neurons. The first layer forms a representation of the optic flow input. This representation shall be based on the properties of MT neurons, and will be presented in more detail later. Its basic structure consists of sets of motion-selective neurons with different preferred velocities $\mathbf{e}_m$ and velocity tuning functions s_m, which are assumed to form a population encoding of an optic flow vector θ:

$$\theta = \sum_m s_m \mathbf{e}_m \tag{3.1}$$

These direction-selective neurons connect to a second layer, which contains cell populations that implement the computations necessary to determine $R(\mathbf{T}_j)$ and become maximally excited when $R(\mathbf{T}_j) = 0$. The peak of neuronal population activity in this layer signals the best matching direction of heading (Lappe and Rauschecker 1993).

4 Functional Model of the Representation of Motion and Stereopsis in Area MT

Our investigation here is concerned with the functional consequences of the specific combination of motion and disparity signals in area MT that has been found experimentally. For this reason, the question of how this combination is generated from the inputs of visual processing stages preceding area MT (Qian 1994; Wilson and Kim 1994; Nowlan and Sejnowski 1995) is not explicitly considered. Rather, a simple functional model of the representation of the flow field in area MT, serving as the input to the heading detection stage, is introduced.

4.1 Distributed Representation of Velocity. Most MT neurons are tuned for speed and direction (Maunsell and Van Essen 1983b). In single neurons, the speed and direction tuning are independent of one another (Rodman and Albright 1987). Here, the direction tuning is assumed to follow a rectified cosine function. A neuron's direction-specific response to a movement into direction ϕ is

$$s_{\mathrm{dir}}(\phi) = \begin{cases} \cos(\phi - \phi_{\mathrm{p}}) & \text{if } \cos(\phi - \phi_p) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

The speed tuning is modeled as a gaussian of the logarithm of the ratio between actual speed v and preferred speed v_{p}:

$$s_{\mathrm{speed}}(v) = \exp\left\{-\left[\log_2(v/v_{\mathrm{p}})\right]^2\right\} \tag{4.2}$$

the response s of the neuron is

$$s(v. \phi) = s_{\mathrm{speed}}(v) s_{\mathrm{dir}}(\phi) \tag{4.3}$$

The responses of groups of neurons are used to form a distributed representation of visual motion. In the simulations, four equally spaced direction preferences, $\phi_{\mathrm{p}} = \pi k/2.\ k = \{1. \ldots .4\}$, and eight speed preferences, $v_{\mathrm{p}} = 2^l$ deg/sec. $l = \{-1. \ldots .6\}$ are used. Preferred speeds between 0.5 and 64 deg/sec are within the range of preferred speeds in MT (Maunsell and Van Essen 1983b). A distributed representation of velocity is obtained by summating the neuronal activities weighted by the speed and direction preferences of the neurons:

$$\theta = \frac{1}{2} \sum_{k=1}^{4} \sum_{l=-1}^{6} s_{kl}(v. \phi) 2^l \begin{bmatrix} \cos(\pi k/2) \\ \sin(\pi k/2) \end{bmatrix} \tag{4.4}$$

4.2 Spatial Integration at Different Scales by Extended Receptive Fields. Instead of using a single flow vector as input for an individual neuron, the two-dimensional spatial integration provided by the extended receptive field of the cell is incorporated first. It is assumed that the total response $s_{kl.i}$ for a single neuron i is obtained by averaging its responses $s_{kl}(v_j. \phi_j)$ to all flow vectors j that fall inside its receptive field R:

$$s_{kl.i} = \sum_{j \in R} s_{kl}(v_j. \phi_j) \Big/ \sum_{j \in R} 1 \tag{4.5}$$

Such an averaging over the response distributions has several properties that approximate the spatial integration performed by MT cells. Similar to MT cells (Britten *et al.* 1993), the response of an individual neuron is a monotonic function of the amount of correlated motion inside its receptive field. Responses are maximal for 100% correlated motion into the preferred direction and reduce to a medium level for 0% correlated motion. Further response reduction is obtained for correlated motion into the null direction. Transparent motion in preferred and null direction elicits a response of 50% of the maximum response, also similar to MT cells (Snowdon *et al.* 1991).

The area of spatial integration for a specific neuron is given by the size of its receptive field. In the visual motion pathway of primates, the receptive field sizes increase from V1 to MT to MST. Within each area, receptive field size is a function of retinal eccentricity ϵ of the receptive field center, and usually increases toward the periphery of the visual field. In MT, the average size of the receptive field and its dependence on eccentricity are empirically described by Albright and Desimone (1987):

$$\text{RFSize} = 1.04\text{deg} + 0.61\epsilon \tag{4.6}$$

The increase of the receptive field size with retinal eccentricity is also a useful property for the reduction of noise in optic flow fields that arise from self-motion. Typically, during self-motion the singular point of the optic flow field is near the center of the visual field (Lappe and Rauschecker 1995). Therefore, the center of the visual field contains many different local motion directions that are important for the analysis of the flow field. In contrast, in the periphery the flow becomes more lamellar, allowing spatial averaging over a larger scale without losing too much information about the local motion directions.

4.3 Disparity Dependence of the Spatial Integration within the Receptive Field. During self-motion, the optical velocity of a visible object depends on the distance of the object from the observer. Objects close to each other in space generate similar visual motion. Objects separated in depth result in different optical velocities. This "motion parallax" affects only the component of the optic flow field that is due to the translation of the observer, not the component due to rotation of the observer's eye or head. It provides a major cue for the visual system to differentiate both components and to correctly perceive the direction of self-motion (Warren and Hannon 1990). Averaging motion signals from different depths removes this very important cue.

The next step therefore involves simulating the specific disparity dependence of the MT responses. To this end, the spatial averaging within the receptive field is weighted by disparity. In its simplest form, this weighting can be implemented as a cutoff at an upper disparity limit D. For each flow vector inside the receptive field, the disparity is compared to the disparity of the flow vector in the receptive field center, which serves as a reference value or a preferred disparity of the cell. Then, if the disparity difference δ is less than D, the motion signal of this flow vector contributes to the spatial averaging, otherwise it is excluded from the calculation of the neuron's response. Thus, when adding disparity information in the representation of the flow field in MT, the response of a single neuron is given by

$$s_{kl,i} = \sum_{j\in R,\delta<D} s_{kl,j}(v_j, \phi_j) \Bigg/ \sum_{j\in R,\delta<D} 1 \tag{4.7}$$

instead of equation 4.5. The choice of a reference value for each neuron can be effectively interpreted as a winner-take-all selection within an ensemble of neurons with identical receptive fields but different preferred disparities. In this view, sets of disparity and velocity tuned neurons determine estimates of average velocity within defined disparity bands. Then a selection mechanism identifies the disparity band that corresponds to the disparity in the receptive field center. The averaged velocity signal of this disparity band is transmitted to the subsequent heading detection stage.

A more elaborate population encoding could use the activities in several disparity bands to determine an estimate of the disparity itself, similar to the encoding of speed and direction of motion. However, the actual value of the disparity is not explicitly used in the heading detection scheme. Thus, for the purpose of the present work the simple disparity weighting seems sufficient. Figure 1 summarizes the structure of the assumed representation of velocity in MT.

The above procedure results in an improved representation of the flow field in the presence of noise (Fig. 2). The following section explores the consequences of this representation for the heading detection system.

5 Results

The network was tested with simulations of the psychophysical experiments by van den Berg and Brenner (1994b). Self-motion with respect to a three-dimensional cloud of random dots or a ground plane was simulated. The flow fields contained additional eye rotation appropriate to track a point in the environment (Lappe and Rauschecker 1995). Noise was added to the flow field in the following way: Each flow vector was disturbed by a noise vector, the direction of which was taken at random from a uniform distribution over the interval $[0, 2\pi]$. The magnitude of the noise vector was proportional to the magnitude of the flow vector. The proportionality constant defined the signal-to-noise ratio SNR. When such flow fields were presented to human subjects, either stereoscopically, preserving the three-dimensional layout of the scene, or synoptically, without stereoscopic depth information, van den Berg and Brenner found that the results in the heading detection task depended on

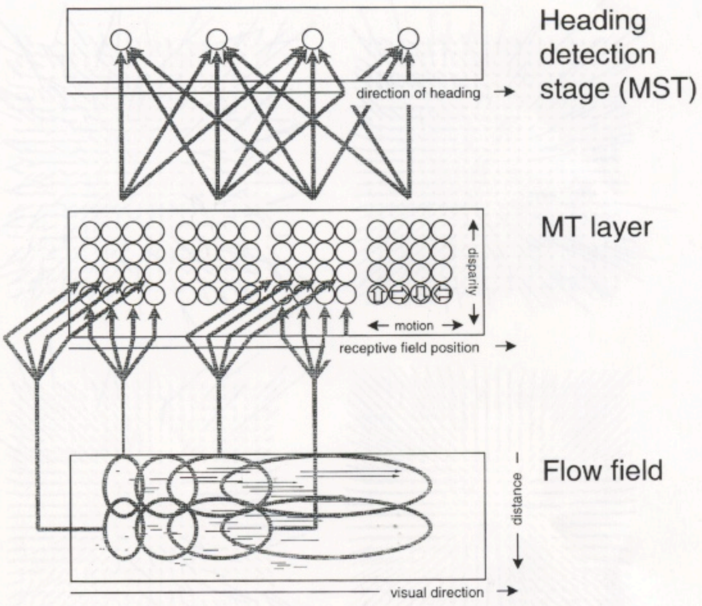

Figure 1: Schematic representation of the flow field computations assumed to take place in areas MT and MST of monkey extrastriate cortex. The optic flow field consists of the motion vectors of visible points located in various distances from the observer. The MT layer is a retinotopic map of visual motion. Each map position contains ensembles of neurons with different preferred velocities and preferred disparities. Each neuron averages motion from within a restricted spatial receptive field and a restricted disparity band. Receptive field sizes grow with eccentricity of the receptive field center. The averaged motion signal from the disparity band that corresponds to the disparity of the flow vector in the receptive field center is then fed into a biologically plausible heading detection scheme presumably located within area MST. Details of the heading detection stage can be found in Lappe and Rauschecker (1993).

scene geometry. For the cloud, heading errors increased with decreasing SNR, but were always much lower in the stereoscopic as compared to the synoptic condition. For the ground plane, little difference between the two presentation conditions was observed, and only a modest variation with SNR occurred.

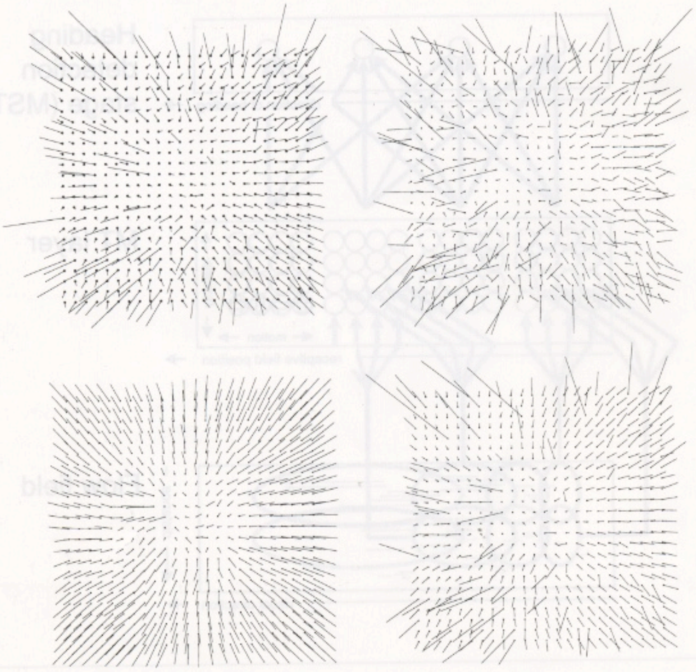

Figure 2: Robust representation of the flow field in the presence of noise by MT neurons. Top left: Optic flow field for movement through a cloud of random dots (see Section 5 for parameters). Because of the random depth distribution, there is no apparent global structure, but the flow field is noise-free. Top right: The same flow field with added noise (SNR = 2). Bottom left: Effect of spatial averaging of the flow field disregarding disparity. This heavily smoothed flow field lacks much of the original motion parallax information. Bottom right: Representation in MT when the disparity dependence is accounted for. Spatial averaging within a restricted depth range results in significant noise reduction.

The stereoscopic and synoptic conditions were recreated in model simulations. For the stereoscopic condition, the disparity range over which an individual neuron spatially averaged the motion signals within its receptive field was set to an intermediate value of $D = 0.4$ deg, which is the range suggested by the physiological data. For the synoptic condition, the spatial averaging was performed for all motion vectors regardless of disparity. In accordance with the psychophysical experiments, visual field size was set to 54 by 54 deg. Simulated translational speed was 1.5 m/sec. For the cloud stimulus, eye rotation appropriate to track a point 8 m away from the observer was simulated. Depth of the cloud ranged from 2 to 20 m. For the ground plane, eye rotation was appropriate to track a point on the plane, the depth of which depended on the simulated heading and on the simulated eye level (0.65 m). As in the experimental conditions of van den Berg and Brenner only the horizontal component of the computed self-motion was evaluated. The input layer of the network consisted of 18,432 neurons arranged on a 24 by 24 grid. Each grid position contained 32 neurons, 8 × 4 speed and direction preferences. The output layer consisted of 14,440 neurons arranged on a 19 by 19 grid of heading directions covering the central 40 by 40 deg of the visual field. Each grid position contained 40 neurons forming a population encoding of the direction of heading (see Lappe and Rauschecker 1993 for details).

The results of the simulation are shown in Figure 3. Each data point is the average of 100 simulation runs. Similar to the results of van den Berg and Brenner, the robustness of the model depends on both, the simulated viewing condition and the geometry of the scene. For the ground plane, mean errors for both viewing conditions are roughly equivalent, down to an SNR of 2. Errors vary moderately, as SNR decreases. The results are different for the cloud. There, the errors depend strongly on SNR, but in all cases the errors in the stereoscopic condition are much lower than in the synoptic condition. Also similar to the results of van den Berg and Brenner, the errors in the simulated stereoscopic condition are similar for both environments down to an SNR of 2. Taken together, the results show that the model draws on stereoscopic information in the case where it is most needed, namely for movement in a cluttered, noisy environment.

These results suggest that the lack of differences in the responses of the human subjects in the he ground plane condition stems from the smooth depth variations in this stimulus. In the cloud condition, dots within the receptive field of any given cell could show very large disparity differences. In the ground plane condition many dots within a receptive field have similar disparities, because the distances change smoothly from one point on the ground plane to the next. In this case, limiting the spatial integration to a certain disparity range has little effect on the choice of motion signals that contribute.

6 Discussion

A simple model of the functional integration of motion and stereopsis for the task of visual heading detection was presented. This model incorporates many important features of neurons in the visual motion pathway of primates. In simulations, it reproduces the psychophysically observed dependencies of the human heading detection system. Stereo-

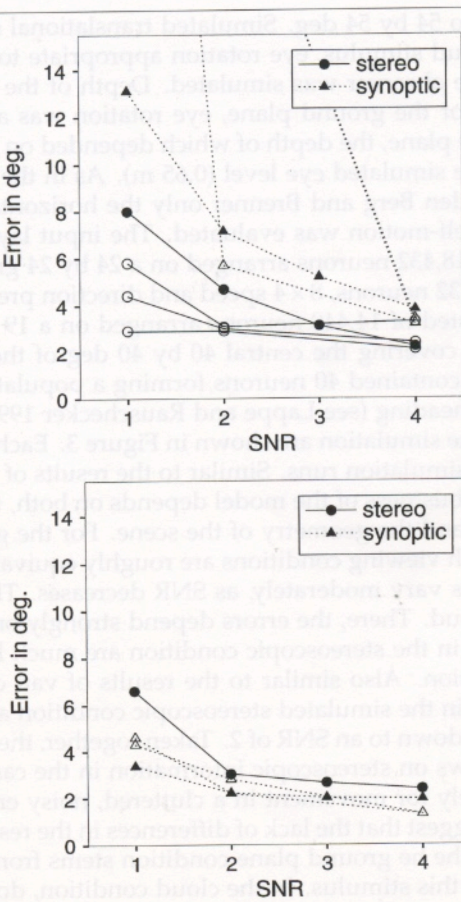

Figure 3: Mean heading errors as a function of the signal-to-noise ratio (SNR) for simulated stereoscopic and synoptic flow fields. For movements through a three-dimensional cloud of dots (upper panel), the model is much more robust against noise in the simulated stereoscopic condition. For movements over a ground plane (lower panel) there is little difference between the two conditions. Filled circles are simulation results for the stereoscopic condition. Filled triangles are simulation results for the synoptic condition. Human data for the corresponding situations are shown by open symbols. Open circles and open triangles in the cloud condition show the results of three individual subjects of van den Berg and Brenner (1994b) in the stereoscopic and synoptic conditions, respectively. For the plane condition, only data for two of these subjects in the synoptic case were available from an earlier paper (van den Berg and Brenner 1994a). These data are shown by open triangles.

scopic depth information is used to support the two-dimensional motion information and to reduce flow field noise. This approach results in an increased robustness for movements in a cluttered environment. Consistent with the human data, there is no advantage to using stereo in smooth environments such as the ground plane.

The use of disparity information in the model is an implicit one. Depth does not directly contribute to the computations involved in determining the direction of heading. Rather it is used only to enhance the representation of the flow field. This is in line with the psychophysical results, since the experiments by van den Berg and Brenner (1994b) provide evidence that it is not the motion in depth but the relative depth of the scene that is used by human subjects.

The model is concerned with the features of the flow field representation in area MT used as input to a later heading detection stage. The simulation results show that these features can have a profound influence on the performance of the whole system. In modeling the input representation more closely, one can thus expect to gain more insights into the functioning of the system. It needs to be emphasized that the presented network is not intended as a model of how the observed features of MT neurons are generated. Within the scope of this work, only the functional properties of MT neurons have been used. The question of how the transparent motion detection and disparity tuning of MT neurons can be achieved given the sensory inputs from the retina is a different problem that has already received much consideration on its own (Qian 1994; Wilson and Kim 1994; Nowlan and Sejnowski 1995). However, the presented work shows that the functional consequences of these properties are consistent with the requirements of the human heading detection system. This complements a number of other features of area MT which are also beneficial for the representation of self-motion-induced optic flow fields. These features include the increase of preferred speed and receptive field size with retinal eccentricity and the predominance of centrifugal direction preferences in the peripheral visual field (Lappe and Rauschecker 1995). For future research, this poses the interesting question of whether even more visual modalities, such as color, or higher level non-Fourier motion signals, could also provide additional supportive information for this task (Braddick 1995).

References

Albright, T. D., and Desimone, R. 1987. Local precision of visuotopic organization in the middle temporal area (MT) of the macaque. *Exp. Brain Res.* **65**, 582–592.

Ballard, D. H., and Kimball, O. A. 1983. Rigid body motion from depth and optical flow. *Comp. Vis. Graph. Image Pro.* **22**, 95–115.

Braddick, O. 1995. Visual perception: Seeing motion signals in noise. *Curr. Biol.* **5**, 7–9.

Bradley, D., Qian, N., and Andersen, R. 1995. Integration of motion and stereopsis in middle temporal cortical area of macaques. *Nature (London)* **373**, 609–611.

Britten, K. H., Shadlen, M. S., Newsome, W. T., and Movshon, J. A. 1993. Responses of neurons in macaque MT to stochastic motion signals. *Vis. Neurosci.* **10**, 1157–1169.

Clark, J. J., and Yuille, A. L. 1990. *Data Fusion for Sensory Information Processing Systems.* Kluwer, Boston, MA.

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

Heeger, D. J., and Jepson, A. 1992. Subspace methods for recovering rigid motion I: Algorithm and implementation. *I. J. Comput. Vision* **7**(2), 95–117.

Koenderink, J. J., and van Doorn, A. J. 1987. Facts on optic flow. *Biol. Cybern.* **56**, 247–254.

Lappe, M., and Rauschecker, J. P. 1993. A neural network for the processing of optic flow from ego–motion in higher mammals. *Neural Comp.* **5**, 374–391.

Lappe, M., and Rauschecker, J. P. 1995. Motion anisotropies and heading detection. *Biol. Cybern.* **72**, 261–277.

Lappe, M., Bremmer, F., and Hoffmann, K.-P. 1994. How to use non-visual information for optic flow processing in monkey visual cortical area MSTd. In *ICANN 94 — Proceedings of the International Conference on Artificial Neural Networks*, M. Marinaro and P. G. Morasso, eds., pp. 46–49. Springer, Berlin.

Maunsell, J. H. R., and Van Essen, D. C. 1983a. Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity. *J. Neurophysiol.* **49**(5), 1148–1167.

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

Nowlan, S., and Sejnowski, T. 1995. A selection model for motion processing in area MT of primates. *J. Neurosci.* **15**, 1195–1214.

Qian, N. 1994. Computing stereo disparity and motion with known binocular cell properties. *Neural Comp.* **6**, 390–404.

Rodman, H. R., and Albright, T. D. 1987. Coding of visual stimulus velocity in area MT of the macaque. *Vis. Res.* **27**(12), 2035–2048.

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

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

van den Berg, A. V., and Brenner, E. 1994a. Humans combine the optic flow with static depth cues for robust perception of heading. *Vis. Res.* **34**, 2153–2167.

van den Berg, A. V., and Brenner, E. 1994b. Why two eyes are better than one for judgements of heading. *Nature (London)* **371**, 700–702.

Warren, W. H., Jr., and Hannon, D. J. 1990. Eye movements and optical flow. *J. Opt. Soc. Am. A* **7**(1), 160–169.

Wilson, H., and Kim, J. 1994. A model for motion coherence and transparency. *Vis. Neurosci.* **11**, 1205–1220.

Received September 27, 1995; accepted March 19, 1996.