PERGAMON

Vision Research 42 (2002) 883–898

Vision Research

www.elsevier.com/locate/visres

Computing relief structure from motion with a distributed velocity and disparity representation

Julián Martín Fernández [1], Brendon Watson, Ning Qian *

Center for Neurobiology and Behavior, Columbia University, Annex Room 730, 722 W 168th Street, New York, NY 10032, USA

Received 20 April 2001; received in revised form 7 January 2002

Abstract

Recent psychophysical experiments suggest that humans can recover only relief structure from motion (SFM); i.e., an object's 3D shape can only be determined up to a stretching transformation along the line of sight. Here we propose a physiologically plausible model for the computation of relief SFM, which is also applicable to the related problem of motion parallax. We assume that the perception of depth from motion is related to the firing of a subset of MT neurons tuned to both velocity and disparity. The model MT neurons are connected to each other laterally to form modulatory interactions. The overall connectivity is such that when a zero-disparity velocity pattern is fed into the system, the most responsive neurons are not those tuned to zero disparity, but instead are those having preferred disparities consistent with the relief structure of the velocity pattern. The model computes the correct relief structure under a wide range of parameters and can also reproduce the SFM illusions involving coaxial cylinders. It is consistent with the psychophysical observation that subjects with stereo impairment are also deficient in perceiving motion parallax, and with the physiological data that the responses of direction- and disparity-tuned MT cells covary with the perceived surface order of bistable SFM stimuli.

Keywords: Kinetic depth effect; Motion parallax; Area MT; Binocular disparity; Motion-stereo integration; Computational modeling

1. Introduction

We can perceive vivid depth from appropriate retinal motion patterns. This phenomenon is known as the kinetic depth effect (Wallach & O'Connell, 1953) or structure from motion (SFM) (Ullman, 1979). Many models for computing SFM have been proposed. The best known are the class of models based on Ullman's theorem (and its variations) that under orthographic projection, three views of four non-coplanar points from a rigid object are sufficient for uniquely determining the 3D structure of the points (up to a mirror reflection) (Ullman, 1979). Ullman later proposed an "incremental rigidity algorithm" for recovering structure from input data over time (Ullman, 1984). While these studies are extremely interesting in their own right, they do not tell us how neurons in the brain could solve the problem. In fact, there is no evidence that visual neurons could explicitly track positions of fine image features over time, as implied by the input representation of these models.

More recent psychophysical experiments indicate that humans seem to use velocity information instead of positional locations of image features for surface interpolation in SFM tasks (Treue, Andersen, Ando, & Hildreth, 1995). This result has led to a major modification of the incremental rigidity algorithm (Hildreth, Ando, Andersen, & Treue, 1995). While the new method is consistent with a wide range of psychophysical data, it relies on an explicit velocity representation as the input. In reality, the velocity information is only coded in a distributed fashion by a population of cells with broad tuning curves.

In addition to the input representation, another problem with most of the existing models is that they compute the Euclidean metric structure of the object as the output (Ullman, 1979, 1984; Hildreth et al., 1995). Psychophysical evidence indicates that while observers are able to accurately judge an object's topological or

* Corresponding author. Tel.: +1-212-543-5213; fax: +1-212-543-5161.

E-mail addresses: fernande@venus.fisica.unlp.edu.ar (J.M. Fernández), bow4@columbia.edu (B. Watson), nq6@columbia.edu (N. Qian).

[1] Present address: Departamento de Física, Universidad Nacional de La Plata, CC 727 (1900) La Plata, Argentina.

0042-6989/02/$ - see front matter
PII: S0042-6989(02)00023-8

ordinal properties, they have considerably more difficulty on tasks that require an accurate perception of Euclidean metric structure (Todd, 1998; Todd & Perotti, 1999). For instance, observers required to make judgments about lengths or angles of visible objects in 3D space resort to using ad hoc heuristics, which typically produce low levels of accuracy and reliability, and which can vary unpredictably among different individuals or for different stimulus configurations (Todd & Perotti, 1999). These experiments suggest that humans can only recover relief SFM, i.e. an object's shape can only be determined up to a stretching transformation along the line of sight (Todd & Perotti, 1999).

There have also been some suggestions in the literature on how SFM may be computed with physiologically plausible mechanisms, but these suggestions remain largely descriptive. For example, Nawrot and Blake (1991) argued that SFM is processed by a network of disparity and motion selective units, but they only simulated the transition between bistable percepts and did not consider how to compute the perceived structure. Buracas and Albright (1996) modeled MT receptive field surrounds and found that their properties resemble those of differential motion operators that could characterize the 3D shape of a smooth moving surface (Droulez & Cornilleau-Perez, 1990; Koenderink & van Doorn, 1992); however, they did not show how to use those MT receptive fields to compute SFM.

In this paper, we propose a specific model for computing relief SFM using a subset of area MT/V5 neurons. MT seems to play an important role in SFM: Neurons in this area are tuned to velocity gradients typically found in SFM stimuli (Treue & Andersen, 1996; Xiao, Marcar, Raiguel, & Orban, 1997a), and a selective lesion of this area can impair SFM perception in monkeys (Andersen & Siegel, 1990). In addition, the responses of MT cells covary with the reversal of the perceived surface-order in bistable SFM stimuli, in such way that is consistent with the disparity and direction tuning of the cells (Bradley, Chang, & Andersen, 1998). This last finding supports the notion that stereo processing and motion processing overlap in the brain, and that SFM perception is mediated by cells tuned to both disparity and motion (Rogers & Graham, 1982; Nawrot & Blake, 1989, 1991; Qian, 1994; Qian & Andersen, 1997). Further support comes from the psychophysical report that amblyopic subjects perform much worse than normal subjects not only on depth discriminations based on disparity, as expected, but also on depth discriminations based on motion parallax (Thompson & Nawrot, 1999). We have therefore constructed our SFM model based on a subset of disparity- and motion-tuned MT cells and the interactions among them. Some preliminary results have been reported previously in abstract form (Fernández & Qian, 2000).

2. Methods

For simplicity, we assume that the motion patterns are generated by an object rotating about an axis perpendicular to the line of sight (but not necessarily vertical), or by a relative translational motion between the observer and the object (i.e., motion parallax). We also assume that the viewing distance is large compared with the object size so that the retinal image can be approximated as a scaled orthographic projection. This last assumption is justified since human observers do not appear to take advantage of the additional information in the perspective projection during perceptual analysis of 3D structure (Todd, 1984, 1998). Under these assumptions, the image velocity vectors are all parallel to each other, and we only need to consider the speed of motion (with the sign of the speed indicating the direction of motion). We will denote the common motion axis as the x-axis.

For each retinal location, we consider a population of model MT neurons tuned to various combinations of speed and disparity. For the convenience of mathematical analysis, we assume Gaussian tuning curves for the classical receptive fields (but we have also verified our model with log-normal speed tuning curves through simulations; see Section 3). At location i, the firing rate A of a neuron preferring speed v_i^0 and disparity d_i^0 in response to image speed v_i and disparity d_i is thus given by:

$$A(i, v_i^0, d_i^0) = A_0 \exp\left(-\frac{(v_i^0 - v_i)^2}{2\sigma_v^2}\right) \exp\left(-\frac{(d_i^0 - d_i)^2}{2\sigma_d^2}\right) \tag{1}$$

where A_0 is the maximum firing rate, and σ_v and σ_d determine the speed and disparity tuning widths, respectively. The preferred speed (v_i^0) and disparity (d_i^0) are constants for a fixed cell, but varies among different cells at location i. We next consider modulatory influences to a cell from L other cells tuned to different retinal locations. Specifically, we assume the overall firing rate of a cell is computed as:

$$f(\alpha) = GA(\alpha)\left[1 + \sum_{\beta \neq \alpha}^{L} \epsilon_{\alpha\beta} A(\beta)\right] \tag{2}$$

where we have defined $\alpha \equiv (i, v_i^0, d_i^0)$ to denote the triplet of parameters that identify each cell in Eq. (1), and β denotes the triplet of parameters indexing another cell. $\epsilon_{\alpha\beta}$ represents the strength of the modulatory connection between cells α and β. G is a gain-control factor for normalizing excessive activity (Heeger, 1992), and is defined as:

$$G = \frac{g}{\sum_{\alpha}^{m} f(\alpha)} \tag{3}$$

where g is a constant, and the summation is local, over m neurons tuned to the same retinal position i but different v_i^0 and d_i^0. When there is no stimulation in the classical receptive field of a cell (i.e., $A(\alpha) = 0$), the modulatory connections from other cells in Eq. (2) do not contribute.

It should be noted that as a simplification, here we start with response tuning curves (Eq. (1)) instead of with detailed receptive field profiles. Therefore, we will only specify the center locations (index by i) of the receptive fields but not the other aspects. For this reason, v_i and d_i should really be viewed as the mean stimulus speed and disparity sensed by the cells' classical receptive fields. It should be possible to construct a more detailed model in the future by using appropriate spatiotemporal receptive field profiles that can indeed generate responses tuned to both disparity and speed (Qian & Andersen, 1997; Chen, Wang, & Qian, 2001).

We would like to choose the pattern of connectivity $\epsilon_{\alpha\beta}$ among the speed- and disparity-tuned cells in such a way that when a zero-disparity motion pattern is fed into the system, cells with preferred disparities consistent with the relief depth structure of the velocity pattern are maximally activated. To determine the connectivity, we first examine how the image speed and disparity are related to each other. For an object rotating with angular speed Ω about an axis perpendicular to the line of sight, it can be shown that the projected speed of a point on the object is given by (see Appendix A):

$$v \simeq -\Omega \frac{(Z - Z_0)}{Z_0} \tag{4}$$

where Z is the distance of the point along the line of sight, and Z_0 is the distance of the axis of rotation. Thus, for two arbitrary points on the object with a relative depth ΔZ, their relative image speed is:

$$\Delta v \simeq -\Omega \frac{\Delta Z}{Z_0} \tag{5}$$

If the object is viewed stereoscopically, with a fixation distance approximately equal to Z_0, the relative disparity between any two points is related to their relative depth ΔZ according to:

$$\Delta d \simeq a \frac{\Delta Z}{Z_0^2} \tag{6}$$

where a is the inter-ocular distance (see, e.g., Howard & Rogers, 1995). Combining Eqs. (6) and (5) to eliminate ΔZ we obtain:

$$\Delta v \simeq -\frac{\Omega Z_0}{a} \Delta d \tag{7}$$

That is, the relative image-speed and relative image-disparity between any two points on an rotating object are proportional to each other.

In the case of a relative translation between the observer and the object, an expression similar to Eq. (7) can be derived (see Appendix A):

$$\Delta v \simeq -\frac{T}{a} \Delta d \tag{8}$$

where T is the component of translational velocity in the frontoparallel plane. In the following, we will mainly discuss the rotation case, but keep in mind that equivalent results also hold for motion parallax; one only needs to replace the product of the angular speed Ω and viewing distance Z_0 by the relative motion speed T.

To better visualize connectivity, we arrange the MT cells tuned to a given retinal position on a plane, with the horizontal axis representing preferred disparity and the vertical axis representing preferred speed (Fig. 1). For example, a neuron positioned in the upper-right corner represents a cell tuned to a large positive speed and a large positive disparity. Eq. (7) suggests that we should introduce connections (i.e., $\epsilon_{\alpha\beta} \neq 0$) between cells for different retinal locations only if the difference between their preferred disparities and the difference between their preferred speeds are proportional to each other:

$$\Delta v^0 = -K \Delta d^0 \tag{9}$$

To see the implication of this connectivity intuitively, consider a zero-disparity input stimulus with different speeds at two different locations. The stimulus will best activate cells with different preferred speeds at the two locations, i.e., $\Delta v^0 \neq 0$. The connection pattern specified by Eq. (9) then ensures $\Delta d^0 \neq 0$, i.e., the most activated cells will also have different preferred disparities. Therefore, although the stimulus has the same, zero disparity everywhere, the most activated cells will not all prefer the same disparity. In other words, the connectivity pattern can convert a speed gradient into a disparity gradient. A more quantitative analysis will be presented in Section 3 below.

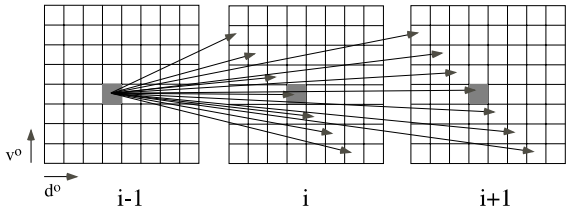

Fig. 1. A schematic illustration of the modulatory connections among MT cells tuned to three adjacent retinal locations $i-1$, i, and $i+1$. At each location, cells preferring different disparity d^0 and speed v^0 are arranged into a 2D array for better visualization. Each cell is represented as a small square. Only the connections from a cell at $i-1$ to cells at i and $i+1$ are shown in the figure. The cells whose d^0 and v^0 parameters are related through Eq. (9) are connected. Therefore, the connectivity pattern follows a straight line with slope $-K$ and passing through the corresponding cells (marked gray) at different locations with the same speed and disparity preference.

In Fig. 1, Eq. (9) means that a cell at a given location is only connected to cells at another location that fall along a straight line with slope $-K$ and passing through the corresponding cell (marked gray) with the same speed and disparity preference. We will also consider in our simulations the cases where the connection strength falls off gradually around this straight line, and where the connection strength decays with the retinal distance between cells' receptive field locations. To account for the bistability of SFM perception, we further assume that there are actually two sets of connections corresponding to $K = \pm K_0$, for the two opposite senses of object rotation, and that they compete with each other through mutual inhibition so that only one set of connections (i.e., one K value) is functional at a given time (see Bistable Ambiguous Percepts).

The connectivity defined above gives rise to synergistic surrounds because $\epsilon_{\alpha\beta}$ is always positive (i.e., excitatory). Both synergistic and antagonistic surrounds exist in MT (Born & Tootell, 1992; Born, 2000). Our model only makes use of the synergistic surrounds. We will also show later, in connection with Eq. (15), that antagonistic surrounds may be an emergent property of the model.

We can fix the value of K in Eq. (9) instead of letting it depend on Ω and Z_0 since we are only interested in encoding the relief structure by our model MT cells; if $K \neq \Omega Z_0/a$ for a given stimulus, the computed structure will be related to the true structure by a stretching transformation along the line of sight. Other MT cells or other brain areas such as MST might be involved in estimating the scale factor along the line of sight, based on retinal or extra-retinal cues. It is also possible for the modulatory connections specified by Eq. (9) to be dynamically remapped (Anderson & Van Essen, 1987) for each viewing distance such that K is proportional to the viewing distance; such a remapping could be performed by the feedback connections from MST to MT. (In the case of motion parallax, the remapping should be based on the translational motion of the observer.) Since human subjects are very poor at judging 3D distances in SFM tasks (Todd, 1998; Todd & Perotti, 1999), the estimation of the scale factor and the remapping of the connections must be crude. We will not deal with the scaling or remapping problem in this paper. For the purpose of showing the simulation results, we simply chose a scaling factor such that rotating circular cylinders are computed near veridically. This is mathematically equivalent to setting K to the value (see Appendix A):

$$K = \frac{3\pi\langle|v|\rangle Z_0}{4a\theta} \tag{10}$$

where $\langle|v|\rangle = \sum_i^N |v_i|/N$ is the averaged absolute value of speed over the stimulus, and θ is the angular size of the cylinder. We will use angled brackets to denote an average over a stimulus on the retina throughout the paper.

Therefore, for a given input motion field, we can apply Eqs. (1)–(3) and the connectivity pattern specified by Eq. (9) to determine, for each retinal location i, the population activity $f(i, v_i^0, d_i^0)$ of all cells tuned to different speed v_i^0 and disparity d_i^0 at that location. We can then take the population average of the preferred disparity (weighted by the firing rate) at location i as the "perceived" disparity reported by our model:

$$\overline{d}_i = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d_i^0 f(i, v_i^0, d_i^0)\mathrm{d}(d_i^0)\mathrm{d}(v_i^0)}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(i, v_i^0, d_i^0)\mathrm{d}(d_i^0)\mathrm{d}(v_i^0)} \tag{11}$$

To simplify the calculations, we will reduce the above expression by cutting a slice through $f(i, v_i^0, d_i^0)$ at v_i^0 equal to stimulus speed v_i, and then taking a one-dimensional average on the slice according to:

$$\overline{d}_i = \frac{\int_{-\infty}^{\infty} d_i^0 f(i, v_i, d_i^0)\mathrm{d}(d_i^0)}{\int_{-\infty}^{\infty} f(i, v_i, d_i^0)\mathrm{d}(d_i^0)} \tag{12}$$

Here $f(i, v_i, d_i^0)$ denotes $f(i, v_i^0 = v_i, d_i^0)$, i.e., $f(i, v_i^0, d_i^0)$ evaluated at $v_i^0 = v_i$. Our simulations show that there is practically no difference between the results produced by Eqs. (11) and (12). We will use an over-line to denote an average over a population of cells for a given retinal location throughout the paper.

In the simulations involving transparent cylinders with a front surface and a back surface, we assume that at a given location, the responses of classical receptive fields are determined by velocity vectors from only one surface, and that receptive fields for different locations represent velocity vectors belonging to different surfaces. This assumption is consistent with our previous finding that transparent motion appears to be detected through locally unbalanced motion signals (Qian, Andersen, & Adelson, 1994a,b; Qian & Andersen, 1994), and allows us to avoid the problem of interference between opposite velocity vectors (from the front and back surfaces) at each location.

3. Results

We now present our analytical and simulation results. We first consider the simple case where $\sigma_v \to 0$ (i.e, model MT cells have infinitely sharp speed tuning), and the modulatory connection between two cells is nonzero only when their preferred parameters satisfy Eq. (9) exactly. (These assumptions will be relaxed below.) Thus, at each retinal position i only the cells with preferred speed v_i^0 equal to stimulus speed v_i will be active, and the total number (L) of cells modulating a given cell at i is equal to the number (N) of sampled retinal positions that send modulatory connections to position i. The overall firing rate of an MT cell defined in Eq. (2) to

a zero-disparity motion pattern is then given by (see Appendix A):

$$f\left(i, v_i, d_i^0\right) = A_0 G \exp\left(-\frac{\{d_i^0\}^2}{2\sigma_d^2}\right) \times \left[1 + \sum_{j\neq i}^{N} A_0 \epsilon_{ij} \exp\left(-\frac{(d_i^0 + v_{ij}/K)^2}{2\sigma_d^2}\right)\right] \quad (13)$$

where $v_{ij} \equiv v_i - v_j$ is the difference of image speeds at locations i and j, and $f(i, v_i, d_i^0) \equiv f(i, v_i^0 = v_i, d_i^0)$. The strength of the modulatory connections between neurons is denoted ϵ_{ij} because its value in this case only depends on the preferred speeds $v_i^0 = v_i$ and $v_j^0 = v_j$ of the two cells under consideration.

Without the modulatory connections (i.e., $\epsilon_{ij} = 0$), the above population response as a function of d_i^0 will simply be a Gaussian peaked at zero, and the "perceived" disparity according to Eq. (12) will be zero, as expected. With the modulatory connections, however, the peak location of the population response will be shifted. As we show in Appendix A, if we make the specific choice of

$$\epsilon_{ij} = \epsilon \exp\left[\left(\frac{v_i^0 - v_j^0}{2K\sigma_d}\right)^2\right] = \epsilon \exp\left[\left(\frac{v_{ij}}{2K\sigma_d}\right)^2\right] \quad (14)$$

then a simple analytical result can be obtained for the "perceived" disparity (Eq. (12)):

$$\overline{d}_i = \frac{1}{\hat{K}}(\langle v\rangle - v_i) \quad (15)$$

where

$$\hat{K} = \frac{2K\left(\sqrt{2} + (N-1)A_0\epsilon\right)}{NA_0\epsilon} \quad (16)$$

and $\langle v\rangle$ is the stimulus speed averaged over the N sampled retinal positions that have connections to position i and include position i. Eq. (15) means that the computed equivalent disparity at a location is proportional to the difference between the stimulus speed v_i at that location (which, as we mentioned in Section 2, should really be viewed as the mean speed in the cells' classical receptive field at that location) and the average speed $\langle v\rangle$ of the stimulus over a larger area including both the classical receptive field and the modulatory surround. The model can therefore compute the relief structure of the input motion pattern (cf. Eq. (7)). For the special case where the input motion field is constant everywhere the mean speed v_i sensed by the cells' classical receptive fields will be equal to the mean speed $\langle v\rangle$ pooled over the larger area, and the "perceived" disparity will be zero according to Eq. (15). For input patterns with a spatial speed gradient, on the other hand, the "perceived" disparity will not be zero in general. In particular, for a transparent cylinder with opposite front and back motion fields, v_i sensed by the classical receptive fields will be dominated by only one of the two motion fields at a given position (see Section 2) while $\langle v\rangle$ will be zero because it pools over many positions, with half of them dominated by the front motion fields, and the other half dominated by the back motion fields. Therefore, the "perceived" disparity will simply follow v_i. For an opaque cylinder with only the front motion field visible, $\langle v\rangle$ will not be zero, but it will be smaller than v_i at positions near the center of the cylinder and larger than v_i at positions near the edges.

Note that although we have computed the "perceived" disparity $\overline{d}_i$ from the population activity in Eq. (15) to make sure that the model works, the brain may not have to do so explicitly because the distributed representation in the population activity might directly correspond to the perception of $\overline{d}_i$. On the other hand, if there *are* cells in area MT that represents $\overline{d}_i$ explicitly, then according to Eq. (15) these cells should show center-surround antagonism because the stimulus speed v_i in the cells' classical receptive fields at a given location i is subtracted from the average stimulus speed $\langle v\rangle$ over a larger region containing i. The implication is that cells with antagonistic surrounds should be at a later stage of processing than cells with synergistic surrounds. Both symmetric and asymmetric surrounds have been hypothesized (Buracas & Albright, 1994, 1996) and found (Xiao, Raiguel, Marcar, Koenderink, & Orban, 1995; Xiao, Raiguel, Marcar, & Orban, 1997b) in area MT. In our model, the surround area is determined by all the other locations that send modulatory connections to location i; its exact shape is not important because the spatial sampling used for computing $\langle v\rangle$ is not critical.

Also note that the proportionality constant $\hat{K}$ in Eq. (15) is not equal to K in Eq. (9) for specifying connectivity. This does not represent a problem since we are only interested in relief structure. If $NA_0\epsilon \gg 1$, which is the case for the parameters used in our simulations, then we have:

$$\hat{K}/K = 2 \quad (17)$$

This means that the computed disparity is half of the value specified by the connectivity pattern. This makes sense because the input stimulus has zero disparity and thus excite cells tuned to zero-disparity best, while the connectivity pattern favors cells tuned to a non-zero disparity determined by the K parameter. Therefore, the peak location of the population activity is halfway between that specified by the input and that specified by the connectivity.

We have made extensive numerical simulations to confirm the above analytical result. An example of our simulation is shown in Fig. 2. Here, the input pattern is the projected motion of a transparent rotating cylinder, and the input disparity is zero everywhere. Fig. 2a shows the population activity profile of the model cells tuned

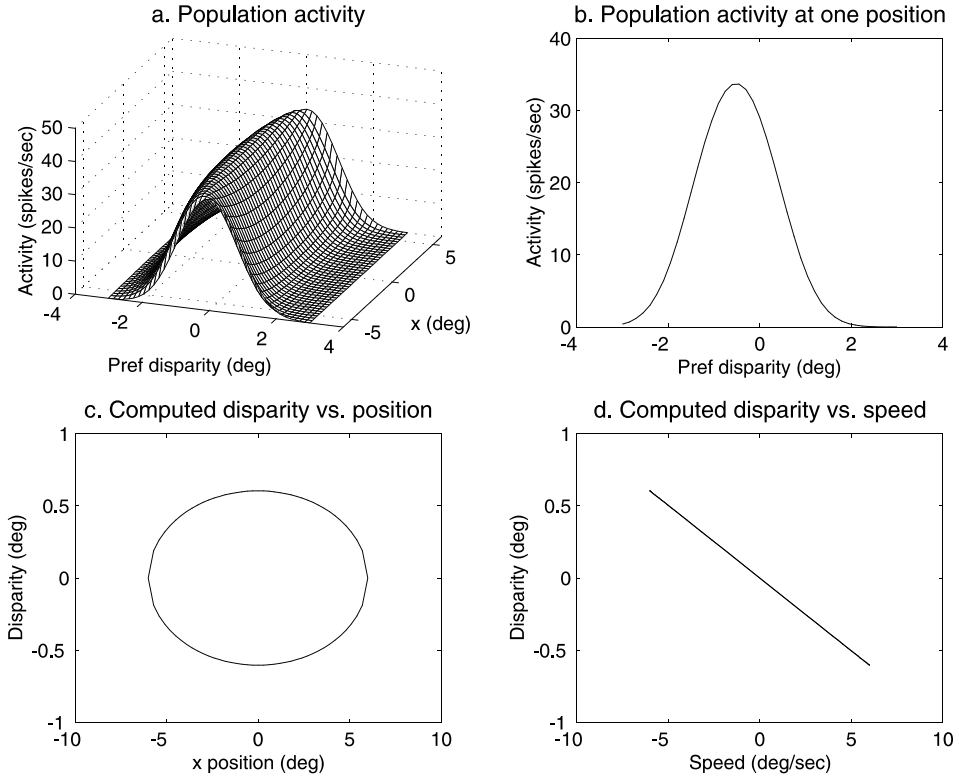

Fig. 2. Simulation results for a transparent rotating cylinder. (a) Population activity profile of the model cells tuned to different disparities and retinal locations (x) perpendicular to the axis of rotation. For clarity, only the results for the near half of the cylinder are shown. The contour of the most active cells resembles the relief structure of a cylinder. (b) The section through the population activity profile at a fixed position 2.7° from the center of the cylinder. The figure shows that although the input stimulus has zero disparity, the most active cell is tuned to a non-zero disparity. (c) Computed disparities (Eq. (12)) at different retinal positions resemble the cylinder's shape. (d) Computed disparity (same as in (c)) as a function of the input speed. The straight line confirms that the correct relief structure is computed by our model (see Eq. (7)). For this simulation, the cylinder had a radius of 6°, and the maximal retinal speed at the center of the projection was $v_{\max} = 6°/s$. The model parameters used were $\sigma_d = 1°$, $\sigma_v = 0$, $A_0 = 50$ spikes/s, $\epsilon = 1$, $K = 5/s$, and $g = 500$. We considered $m = 41$ neurons at each retinal position tuned to disparities between $-3°$ and 3°, and sampled the cylinder at $N = 41$ retinal positions perpendicular to the axis of rotation.

to different disparities and retinal locations. For clarity, only the results representing the near half of the cylinder are shown. (The activity representing the far half of the cylinder is a mirror reflection of Fig. 2a with respect to the zero-disparity plane.) The activity profiles from the two halves of the cylinder do not interfere with each other in our model because we assume (see Section 2) that only the velocity vector from a single surface is represented by a given classical receptive field (Qian et al., 1994a,b; Qian & Andersen, 1994). The section of the activity in Fig. 2a at a fixed position is plotted in Fig. 2b. It is clear that although the input stimulus has zero disparity, the most active cell is tuned to a non-zero disparity. Fig. 2c shows that the computed disparities (Eq. (12)) at different retinal positions resemble the cylinder's cross-section. We also plotted the computed disparity as a function of the input speed (Fig. 2d). The straight line confirms that the relief structure is indeed computed according to Eq. (7).

Very similar results can be obtained if we relax the condition $\sigma_v = 0$. As we show analytically in the Appendix A, Eq. (15) is still valid although the expression for $\hat{K}$ has to be modified. With a non-zero σ_v, the active neurons for a given retinal position i are not limited to those tuned exactly to input speed v_i. Instead, cells tuned to nearby speeds will also respond, albeit to a lesser extent. Therefore, unlike Fig. 2b for the zero-σ_v case, now the population activity of all cells for a given location is not only a graded function of the preferred disparity but also a graded function of the preferred speed. This population activity computed at one position is shown in Fig. 3. To show such population activity for all positions would require a four dimensional figure. To simplify the matter and to facilitate comparison with the zero-σ_v case in Fig. 2 we eliminated the preferred-speed (v_i^0) dimension by cutting a slice through $v_i^0 = v_i$, the most responsive cell along that dimension, and the reduced population response as a function of the preferred disparity for all positions is shown in Fig. 4a. This figure, and the rest of the simulation results in Fig. 4, are very similar to those in Fig. 2, demonstrating the proper computation of the relief structure.

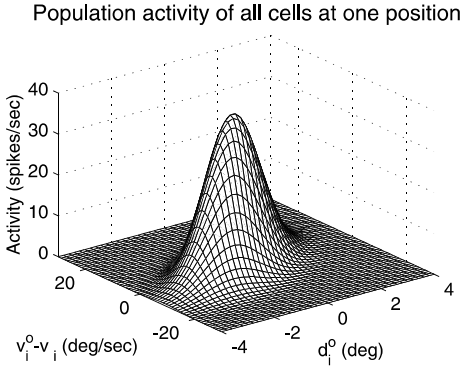

Fig. 3. The computed population activity profile of all cells tuned to different disparities and speeds for a given retinal location 2.7° from the center of the cylinder. The simulation parameters were the same as in Fig 2, except that $\sigma_v = 0$ was replaced by $\sigma_v = v_{\max}$, the maximal retinal speed in the stimulus.

The main difference between the the case of $\sigma_v = 0$ and the case of $\sigma_v \neq 0$ is the width of the population response as a function of the cells' preferred disparity (cf. Figs. 2b and 4b). Analytical (see Appendix A) and numerical results both show that in the limit of $NA_0\epsilon \gg 1$, the standard deviation of the population activity along the disparity dimension averaged over all retinal positions is

$$\sigma^2 \simeq \frac{\sigma_d^2}{2} + \frac{\sigma_{\text{vel}}^2 + \sigma_v^2}{4K^2} \tag{18}$$

where σ_{vel} is the standard deviation of the input speed distribution profile. Thus, the effect of adding a finite tuning in speed ($\sigma_v \neq 0$) is to broaden the population activity. However, the peak locations which determine the computed disparity do not change significantly (cf. Figs. 2c and 4c).

We used Gaussian disparity and speed tuning curves above for their analytical convenience. However, the qualitative features of our model are insensitive to the details of the tuning curve shape. This is because for any population response, the connectivity pattern in the model will tend to convert an appropriate motion field into an equivalent disparity representation by shifting the response along the disparity dimension. Since the speed tuning of visual cortical cells is often skewed towards lower speed (Maunsell & Van Essen, 1983), we have performed additional simulations with a log-normal speed tuning function, and obtained very similar results (not shown). The only difference is that the equivalent disparity computed with the log-normal function is somewhat smaller than that computed with the corresponding normal function. The reason is that

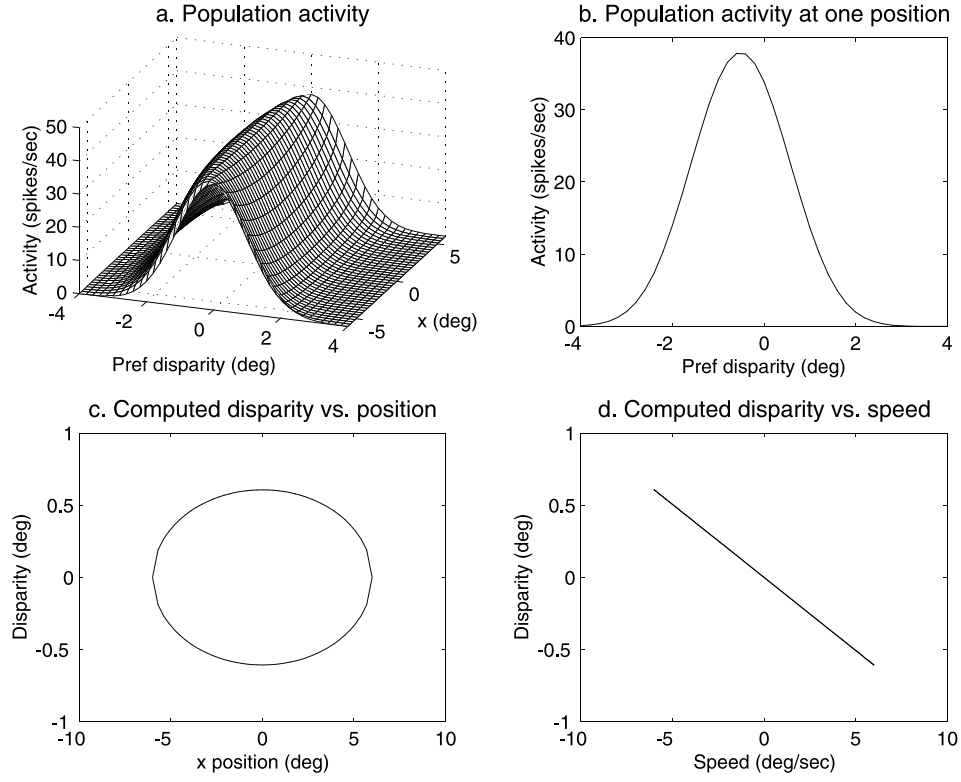

Fig. 4. Same as Fig 2, except that $\sigma_v = 0$ was replaced by $\sigma_v = v_{\max}$. The main effect is a broadening of the population responses (cf. Fig. 2b). The computed structure is nearly identical to that of Fig. 2. For each retinal position i we used a population of 29 cells tuned around the input speed v_i between $v_i \pm \sigma_v$. We also made a more physiologically plausible simulation in which σ_v is proportional to v_i^0 instead of a constant for all cells, and nearly identical results (not shown) were obtained.

with the log-normal tuning function, a cell's response is stronger on the right side of the peak (the preferred speed) than on the left side. Therefore, for a given stimulus speed, cells preferring lower speed contribute more than cells preferring higher speed. Since cells preferring lower speeds connect to cells preferring lower disparities as well, according to the connectivity pattern, the computed equivalent disparity is smaller. Of course, the relief structure is not affected by the smaller disparity values.

We assumed above a sharp pattern of modulatory connections according to Eq. (9) exactly. We have also made more realistic simulations by using a graded connectivity pattern around the straight-line pattern determined by Eq. (9) (see Fig. 1). Specifically, we used a Gaussian connection strength distribution centered on the straight-line and with a $\sigma_c = 1°$ along the preferred-disparity dimension; we sampled 41 different preferred-disparity values within $\pm 2\sigma_c$ of each central value. The results (not shown) are nearly identical to those in Fig. 2.

In the above simulations, we assumed that the connection strength is independent of the difference between receptive field locations of the cells. It may be more plausible to let the connection strength decay as a function of the distance between the cells' receptive field locations. We therefore considered the following function for the connection strength:

$$\epsilon_{ij} = \epsilon \exp\left[\left(\frac{v_{ij}}{2K\sigma_d}\right)^2\right] \exp\left(\frac{-\Delta x^2}{2\sigma_x^2}\right) \exp\left(\frac{-\Delta y^2}{2\sigma_y^2}\right) \quad (19)$$

where ϵ is a constant and x, y are retinal position coordinates. Again, our simulations produced nearly identical results (not shown) to those in Fig. 2. In addition, the results are not sensitive to the values of σ_x and σ_y.

It should be clear from the above that our model is quite robust against parameter variations. In fact, if Eq. (14) is used for ϵ_{ij}, practically any parameter set can be used. The main difference among different sets of parameters is in the widths and heights of the population activity, but the correct relief structure is always obtained. Our simulations also demonstrate that Eq. (14), which is for simplifying the analysis, is not critical. Indeed, we can obtain very similar results (not shown) by simply letting ϵ_{ij} be a constant so long as the product $K\sigma_d$ is large enough such that the expression on the right hand side of Eq. (14) does not vary much over the stimuli. It is also worth noting that although we simulated a cylinder above, our analytical result (Eq. (15)) does not depend on a particular object shape.

Finally, we would like to note that our model is highly robust against noise. This is because unlike models using differential operators that may amplify noise, our model mainly rely on summation of activities from many cells, thus effectively smoothing out the noise in individual responses. For this reason, noise added to the responses inside the summation in Eq. (2) will have negligible effects to the final responses. Obviously, noise should have the greatest impact when it is added to the final responses (left hand side of Eq. (2)) directly. We considered this worst case scenario by assuming that the final responses used in the calculation of equivalent disparity are drawn from a Gaussian distribution with a mean given by Eq. (2) and a variance equal to two times the mean. The simulation results corresponding to Fig. 2b and c is shown in Fig. 5a and b. It is clear from the Fig. 5b that the relief structure is only degraded somewhat by the noise, but not destroyed. Note that here the equivalent disparity was computed with only a total of 41 sampled retinal positions. If we increase the number of sampled positions to 201, the effect of noise is much reduced (Fig. 5c), as expected.

3.1. Some illusions

Ramachandran, Cobb, and Rogers-Ramachandran (1988) reported some interesting SFM illusions using projections of two rotating coaxial cylinders. In the first demonstration, the two cylinders have the same radius so that their surfaces occupy the same locations in the 3D space, but one cylinder is rotated at twice the speed of the other. Perceptually, however, human observers

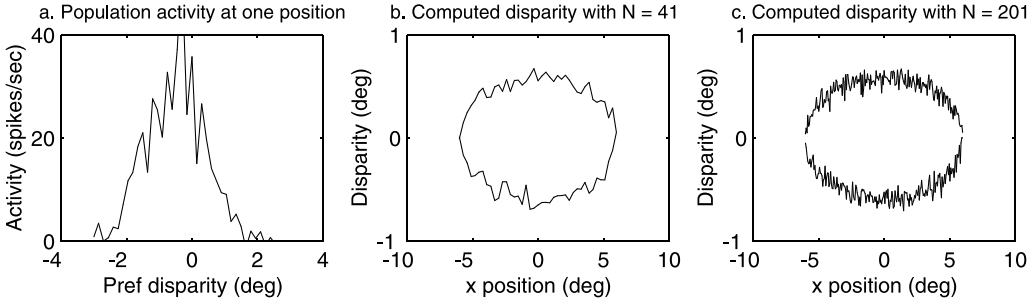

Fig. 5. Same as Fig. 2, except that Gaussian noise with a variance proportional to the mean has been added to the final neuronal responses. (a) Noise-added population activity at a fixed position 2.7° from the center of the cylinder (cf. Fig. 2b). (b) Computed disparities at different retinal positions with $N = 41$ sampled retinal positions (cf. Fig. 2c). (c) Computed disparities at different retinal positions with $N = 201$ sampled retinal positions.

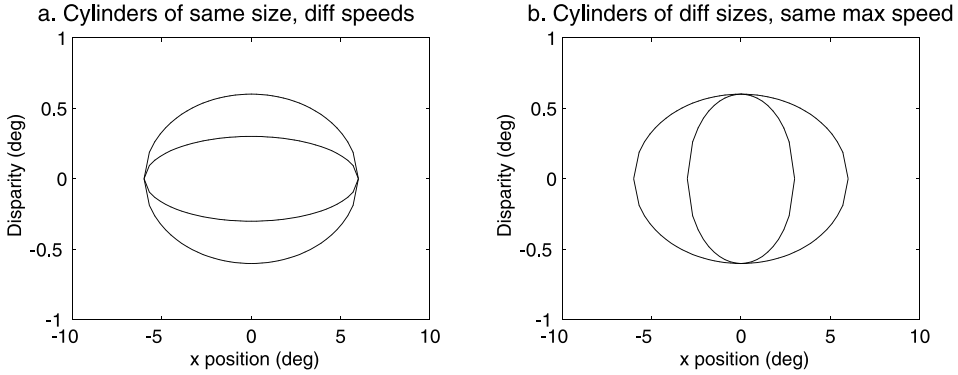

Fig. 6. Simulations of illusions of coaxial cylinders. (a) The rotation speed of one cylinder is twice as fast as that of the other. When this stimulus is shown to human observers, the slower cylinder appears to have a lesser depth. As shown in the figure, our model can reproduce this illusion because the connectivity pattern from a single K value is used for the whole stimulus. (b) The radius of one cylinder is twice as large as that of the other. The smaller cylinder, however, rotates twice as fast as the larger one such that the maximum projected image speeds at the centers of the two cylinders are the same. Human observers perceive two surfaces that merge together in the center. As shown in the figure, our model can reproduce this illusion again because the connectivity pattern from a single K value is used for the whole stimulus.

see two surfaces that are separated in depth. Our model reproduces this illusion, as shown in Fig. 6a. This is a simple consequence of imposing the same K in Eq. (9) for the two objects. The parameters for the faster cylinder are the same as those in Fig 2, and $v_{\max}$ for the slower cylinder is half the value of the faster one.

In the second demonstration, one of the cylinders has half the radius of the other, but its rotation is twice as fast. Therefore, the projected image speed is the same for the points in the center of both surfaces. In this case, human observers perceive two surfaces as merging together in the center. Again, our model is able to reproduce this illusion by imposing the same value of K in Eq. (9) to both surfaces. This is shown in Fig. 6b. The parameter values are the same as in Fig. 2 for each surface except that the smaller cylinder has half the radius. Although the model of Hildreth et al. (1995) can also explain these perceptual illusions, our model is significantly simpler and physiologically more plausible.

3.2. Integration of stereo and motion shape cues

We have assumed above that the input stimuli have zero disparity. We now show that our model also works for stimuli in which both motion and disparity cues are present. When the input disparity field is different from zero, the overall firing rate of a cell at location i becomes (see Appendix A)

$$f\left(i, v_i^0 = v_i, d_i^0\right) = A_0 \exp\left(-\frac{\left(d_i^0 - d_i\right)^2}{2\sigma_d^2}\right) \times \left[1 + \sum_{j \neq i}^{N} A_0 \epsilon_{ij} \exp\left(-\frac{\left(d_i^0 - d_j + v_{ij}/K\right)^2}{2\sigma_d^2}\right)\right] \tag{20}$$

where d_i and d_j are the stimulus disparities at retinal positions i and j, respectively. Define K_s according to:

$$v_i - v_j = -K_s(d_i - d_j) \tag{21}$$

For the stimuli we consider in this paper, K_s is equal to either $\Omega Z_0/a$ or T/a (see Eqs. (7) and (8)), and does not depend on the positional indices i and j. We demonstrate in the Appendix A that for the special case of $K_s = K$ the following analytical solution for the computed disparity is obtained:

$$\overline{d}_i = \frac{v_0 - v_i}{K} \tag{22}$$

where v_0 is the retinal speed at a reference point with zero disparity. The condition $K_s = K$ simply means that the structure specified by the stimulus disparity happens to be consistent with that specified by the connectivity pattern. It is therefore not surprising that under this condition, the proportionality constant in Eq. (22) is K instead of $\hat{K}$ for the zero stimulus disparity case in Eqs. (16) and (17). We have confirmed Eq. (22) through simulations (results not shown).

Since K is a fixed constant that determines the connectivity pattern, and does not change with the stimulus, the condition $K_s = K$ obviously cannot hold in general. We have performed extensive numerical simulations for the general case of $K_s \neq K$ using Eq. (20), and found that the computed disparity still satisfy the relief-structure relationship:

$$\overline{d}_i = \frac{v_0 - v_i}{\hat{K}} \tag{23}$$

but now the proportionality constant is $\hat{K}$ that can be approximated by

$$\frac{1}{\hat{K}} \simeq \frac{1}{2}\left(\frac{1}{K} + \frac{1}{K_s}\right) \tag{24}$$

That is, the computed disparity is the average of the structure specified by the input stimulus disparity and that specified by the connectivity pattern. Eq. (17)

derived previously is a special case of Eq. (24) when the input disparity is 0 (i.e., $K_s \to \infty$).

3.3. Bistable ambiguous percepts

The depth order of the front and back surfaces of a rotating transparent cylinder is ambiguous in the absence of real disparity cues, and our percept alternates between the two possibilities. Although we did not model this bistability explicitly, the phenomena could be explained in the framework of our model by assuming that there are two sets of connections corresponding to $K = \pm K_0$ for the two opposite senses of object rotation, and that they compete with each other through mutual inhibition so that only one set of connections (i.e., one K value) is functional at a given time. The strength of the dominant connections decreases with time due to adaptation, and eventually the other set of connections wins the competition (Nawrot & Blake, 1991). When the switching between the two sets of connections happens, there should also be a corresponding change in the population activity of the model MT cells. Specifically, a highly active cell should now become less active because the new connection pattern no longer supports its activity. Likewise, some weakly active cells should now become more active because their preferred disparities and velocities now match the new connection pattern. This prediction is consistent with the physiological observations of Bradley et al. (1998). In their experiment, monkeys were trained to fixate while viewing 2D projections of transparent revolving cylinders and reporting spontaneous reversals of the perceived surface order. For many of the MT neurons tested, there was a change in the activity that coincided with the reversals of the perceived surface order, even though the stimulus remained identical. The enhancement or diminution of activity was consistent with the disparity and direction tuning of the cells and the perceived surface order (enhancement of the response if the cell was tuned to near disparity and the surface was perceived at front, and vice versa).

Our model is also consistent with the fact that when there is a real disparity cue in the stimulus, the SFM percept is biased toward the one that agrees with the cue (Braunstein, Andersen, Rouse, & Tittle, 1986; Dosher, Sperling, & Wurst, 1986). This is simply because the stimulus disparity reinforces the activity generated by the consistent set of connections and makes it less likely for the other, inconsistent set of connections to win the competition. After prolonged viewing of an SFM stimuli with disparity, the subsequent perception of a SFM stimuli without disparity is biased in the opposite direction (Nawrot & Blake, 1989) because the consistent set of connections has been strongly adapted.

4. Discussion

We have shown in this paper that model MT cells with broad velocity- and disparity-tuning can interact with each other through modulatory connections to compute the relief depth structure from retinal motion patterns. The connectivity among the cells in our model is based on the simple observation that for an object rotating about an axis perpendicular to the line of sight, or translating relative to the observer with a velocity component in a frontoparallel plane (i.e., motion parallax), the relative speed and the relative disparity between the projections of any two points on the object are proportional to each other. In this sense, the relief structure can be viewed as already contained in the input motion field, and we have simply proposed a mechanism for re-coding it as equivalent disparity responses. We have demonstrated through both analyses and simulations that our model can indeed compute the correct relief SFM, and is highly robust. In addition, the model can naturally explain the SFM illusions involving coaxial cylinders. Our work indicates that if we reduce the goal of SFM computation from the Euclidean metric structure to the relief structure, as suggested by the psychophysical evidence (Todd, 1998; Todd & Perotti, 1999), then the SFM problem can be solved with simple and physiologically plausible mechanisms.

Most SFM models make some form of rigidity assumption. The original motivation was mathematical, namely that the 3D structure of an object cannot be determined based on retinal images alone, and that additional assumptions have to be introduced to constrain the problem. To deal with motion patterns produced by non-rigid objects, it has been suggested that image features first be divided into rigid subsets through a testing procedure, and then the structure for each subset be computed (Ullman, 1979). Alternatively, one may simply require that the rigidity be maximized but not strictly enforced (Ullman, 1984; Hildreth et al., 1995). However, to either enforce or maximize rigidity among a set of features, the brain has to measure the 3D distances between the features accurately (Ullman, 1984; Hildreth et al., 1995). As we mentioned in the Introduction, psychophysical evidence suggests that humans are very poor at judging 3D distances along the line of sight (Todd, 1998; Todd & Perotti, 1999). It is thus doubtful that a rigidity assumption is actually employed by the human visual system during SFM computations. In our model, although the derivation of Eqs. (7) and (8) also depends on the rigidity of objects, once the fixed parameter K is chosen in Eq. (9) to determine the connectivity pattern, there is no rigidity assumption during the computation of relief structure. Indeed, no step in our model requires the measurement of 3D distances between the features of the objects.

Although motion is a monocular depth cue and SFM perception does not require binocular viewing, our model relies on the link between motion and stereopsis (Rogers & Graham, 1982; Nawrot & Blake, 1989). The model cells are connected according to their relative disparity and velocity tuning such that when a zero-disparity stimulus with an appropriate velocity pattern is fed into the system, the most responsive neurons are not those tuned to zero disparity, but instead are those having preferred disparities consistent with the relief structure of the velocity pattern. In the framework of our model, any input motion pattern, regardless of whether it contains zero or non-zero disparity, is processed by a population of binocular cells tuned to both disparity and motion. Therefore, our model is consistent with the psychophysical observation that subjects with stereo impairments are also deficient in perceiving motion parallax (Thompson & Nawrot, 1999), and with the physiological data that the responses of direction- and disparity-tuned MT cells covary with the perceived surface order of bistable SFM stimuli (Bradley et al., 1998). To our knowledge, our model provides the first SFM algorithm that relies on an interaction among motion- and disparity-tuned units. Previous models on motion-stereo integration (Nawrot & Blake, 1991; Qian, 1994; Qian et al., 1994b; Qian & Andersen, 1997) do not involve SFM computation. The re-coding of motion as equivalent binocular disparity through modulatory interactions proposed here might be a generic mechanism applicable to some of the other monocular depth cues.

Our model is symmetric with respect to disparity and speed, and as such, it not only predicts that motion should influence disparity (and thus depth perception) but also predicts that disparity should influence motion perception as well. However, although a zero-disparity motion pattern can generate depth perception, as demonstrated by typical SFM stimuli, a static disparity pattern does not seem to generate any motion. This problem can be solved by assuming that among the cells connected according to Eq. (9), there are cells tuned to zero disparity but no cells tuned to zero speed. Alternatively, we can break the symmetry by assuming that there are two separate populations of cells, the first population tuned to motion, and the second population tuned to disparity. We can then assume there are connections (again according to Eq. (9)) from the first population to the second, but not from the second to the first. This way, motion patterns could generate disparity response but not vise versa.

When we considered the interactions among the cells, the contributions from other cells to a given cell were multiplied by the classical receptive field response of the cell (Eq. (2)). We used this multiplicative interaction to simulate the modulatory effects of the non-classical surround of MT receptive fields. For completeness, we also explored using additive interaction in the model. Our computer simulations (results not shown) confirmed that the model works with the additive interaction as well.

As we explained in the Methods section, our model can compute relief structure from projections generated by objects undergoing either rotational or translational motions. A major limitation of our model, however, is that in the rotational case, it assumes the axis of rotation is in the image plane, i.e., perpendicular to the line of sight. Extension of our model to the situation of an arbitrary axis of rotation is not straightforward because the simple relationship between relative speed and relative disparity does not hold in general, and a very complex pattern of connectivity among the cells would be required to convert a motion pattern into an appropriate disparity response. However, this problem may be avoided since any object rotation can be decomposed into a rotation about an axis perpendicular to the line of sight, followed by a rotation about the line of sight. Furthermore, Ullman (1983) has shown that for any non-planar object under orthographic projection, the rotation about the line of sight can be uniquely determined from the image velocity field itself, and therefore can be removed to produce a pattern of parallel motion vectors used by our current model. Since the rotation about the line of sight only generates a concentric pattern of image motion that probably does not contribute to the perception of 3D structure, our visual system might have learned to discard this component and only use the parallel motion pattern for SFM computation as suggested by our model. Much further experimental and theoretical work is needed for resolving these issues in the future.

Acknowledgements

We would like to thank Dr. Yuzhi Chen, Dr. Nestor Matthews, and anonymous reviewers for their very helpful comments. This work was supported by a research grant from the McDonnell-Pew Program in Cognitive Neuroscience and NIH grant # MH54125 to N.Q., and by the Research Fellowship Program of the Universdad Nacional de La Plata (Argentina) to J.M.F.

Appendix A.

A.1. Derivation of Eqs. (4) and (8)

Fig. 7a shows the viewing geometry of a rotating object, with the axis of rotation perpendicular to the page. U is the component of the 3D velocity vector in a frontoparallel plane and is given by:

$$U = \Omega R \cos\alpha = \Omega(Z_0 - Z) \tag{A.1}$$

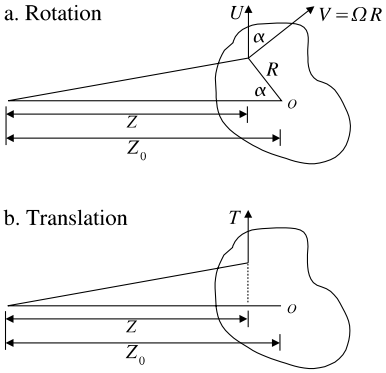

Fig. 7. (a) Viewing geometry of a rotating object. The axis of rotation is perpendicular to the page through o. (b) Viewing geometry of an object with a relative translational motion with respect to the observer.

Assuming that the object is far enough so that $Z \simeq Z_0$, and only keeping the first order term of $Z_0 - Z$, we obtain the projected velocity v as:

$$v \simeq \frac{U}{Z} = \Omega \frac{Z_0 - Z}{Z} \simeq \Omega \frac{Z_0 - Z}{Z_0} \tag{A.2}$$

which is Eq. (4).

A similar situation for an object with a relative translation to the observer is shown in Fig. 7b where T is the component of the translation velocity in a fronto-parallel plane. The projected retinal velocity is:

$$v \simeq \frac{T}{Z} \tag{A.3}$$

Differentiating this equation and again keeping only the first order term of $Z - Z_0$, we have:

$$\Delta v = -T \frac{\Delta Z}{Z_0^2} \tag{A.4}$$

Combining this equation with Eq. (6), we obtain Eq. (8).

A.2. Derivation of Eq. (10)

For a circular cylinder, the ratio between its depth (R_z) and width (R_x) is 1. We have,

$$R_x = \theta Z_0 \tag{A.5}$$

and, according to Eq. (6):

$$R_z = \frac{Z_0^2 (\Delta d)_{\max}}{a} \tag{A.6}$$

where $(\Delta d)_{\max}$ is the maximum relative disparity between the nearest and the furthest points on the cylinder.

According to Eq. (7),

$$(\Delta d)_{\max} = \frac{2 v_{\max} a}{\Omega Z_0} \tag{A.7}$$

where $v_{\max}$ is the retinal speed of the nearest point on the cylinder, and $2v_{\max}$ is the maximum relative speed between the nearest and the furthest points on the cylinder. Then, in order to obtain $R_x/R_z = 1$, we have

$$\Omega = \frac{2 v_{\max}}{\theta} = \frac{3\pi \langle v \rangle}{2\theta} \tag{A.8}$$

where $\langle v \rangle = 4 v_{\max}/3\pi$ is the mean absolute speed of motion pattern. Thus, the corresponding $\hat{K}$ is:

$$\hat{K} = \frac{3\pi Z_0 \langle v \rangle}{2\theta a} \tag{A.9}$$

Because of Eq. (17), we have

$$K = \frac{3\pi Z_0 \langle v \rangle}{4\theta a} \tag{A.10}$$

as the condition for a circular cylinder to be computed as circular by the model.

A.3. Derivation of Eqs. (13) and (15)

The firing rate of a neuron tuned to velocity v_i^0 and disparity d_i^0 is give by Eq. (1) in the absence of modulation from other neurons. When there is no input disparity, $d_i = 0$, and Eq. (1) becomes:

$$A\left(i, v_i^0, d_i^0\right) = A_0 \exp\left(-\frac{\left(v_i^0 - v_i\right)^2}{2\sigma_v^2}\right) \exp\left(-\frac{d_i^{02}}{2\sigma_d^2}\right) \tag{A.11}$$

Now, consider the modulatory interactions between neurons related by Eq. (9). Assume $\sigma_v \to 0$ (we will relax this condition later), then at every retinal position i only the cells with tuning parameter $v_i^0 = v_i$ will be firing, and the total number (L) of cells modulating a given cell at i is equal to the number (N) of sampled retinal positions that send modulatory connections to position i. Eq. (9) then becomes $v_i^0 - v_j^0 = v_i - v_j = -K(d_i^0 - d_j^0)$, or $d_j^0 = d_i^0 + v_{ij}/K$, where $v_{ij} \equiv v_i - v_j$ is the difference of image speeds at locations i and j. The overall firing rate of a neuron after considering the modulatory connections is thus given by:

$$f\left(i, v_i, d_i^0\right) = A\left(i, v_i, d_i^0\right)\left[1 + \sum_{j \neq i}^{N} \epsilon_{ij} A\left(j, v_j, d_i^0 + \frac{v_{ij}}{K}\right)\right] \tag{A.12}$$

where $f(i, v_i, d_i^0) \equiv f(i, v_i^0 = v_i, d_i^0)$ and $A(i, v_i, d_i^0) \equiv A(i, v_i^0 = v_i, d_i^0)$. We have omitted here the gain factor G introduced in the text (see Eq. (3)) because it only scales the population activity at each location and does not affect any of the results presented here.

Combining Eqs. (A.11) and (A.12), we have

$$f\left(i, v_i, d_i^0\right) = A_0 \exp\left(-\frac{\left\{d_i^0\right\}^2}{2\sigma_d^2}\right) \times \left[1 + \sum_{j \neq i}^{N} A_0 \epsilon_{ij} \exp\left(-\frac{\left(d_i^0 + v_{ij}/K\right)^2}{2\sigma_d^2}\right)\right] \tag{A.13}$$

which is Eq. (13) in the text. Note that the speed terms disappear from this expression because we only need to consider cells with $v_i^0 = v_i$. By choosing ϵ_{ij} according to Eq. (14) we can rearrange terms to get:

$$f(i, v_i, d_i^0) = A_0 \exp\left(-\frac{\{d_i^0\}^2}{2\sigma_d^2}\right) + A_0^2 \epsilon \sum_{j\neq i}^{N} \exp\left(-\frac{\left(d_i^0 + v_{ij}/(2K)\right)^2}{\sigma_d^2}\right) \tag{A.14}$$

It is worth mentioning that for the parameter used in our simulations $\exp[(v_{ij}/2K\sigma_d)^2]$ is always very close to one, and thus ϵ_{ij} can be replaced by a constant without affecting the results.

The "perceived" disparity from the model is computed as the population average according to Eq. (12). Instead of a brute force calculation, we can make use of a shortcut: Since Eq. (A.14) is a sum of Gaussians, the mean d_i^0 is simply the mean of the Gaussian centers, each weighted by the total area under the corresponding Gaussian:

$$\overline{d}_i = \frac{-\sum_j^N A_0^2 \epsilon \sqrt{\pi}\sigma_d \frac{v_{ij}}{2K}}{A_0\sigma_d\sqrt{2\pi} + (N-1)A_0^2\epsilon\sqrt{\pi}\sigma_d} = \frac{NA_0\epsilon}{2K\left(\sqrt{2} + (N-1)A_0\epsilon\right)}(\langle v\rangle - v_i) \tag{A.15}$$

Here $\langle v\rangle = \sum_j^N v_j/N$ is the averaged stimulus velocity over all retinal positions that send modulatory connections to i (and include position i itself). We see that $\overline{d}_i = 0$ for $v_i = \langle v\rangle$ and non-zero otherwise. This completes the derivation of Eq. (15).

A.4. Derivation of Eq. (15) for $\sigma_v \neq 0$

We now relax the condition of $\sigma_v \to 0$. When σ_v is finite, we should add to the right hand side of Eq. (A.12) those terms corresponding to the modulation from neurons at positions j that are not exactly tuned to velocity v_j but that also fire:

$$f(i, v_i, d_i^0) = A(i, v_i, d_i^0) \times \left[1 + \sum_{j\neq i}^{N}\sum_{k=-M}^{M} \epsilon_{ik}^j A\left(j, v_j^k, d_i^0 + \frac{v_{ij}^k}{K}\right)\right] \tag{A.16}$$

Here we only considered cells at position i whose preferred speed v_i^0 is equal to the stimuli speed v_i. We have included the influence of the $2M+1$ most responsive neurons from each position j, which are tuned around stimulus speed v_j. We have defined $v_{ij}^k = v_i - v_j^k$, with $v_j^k = v_j + k\delta$ (note that $v_j^0 = v_j$), and δ is a fixed velocity sampling step. Combining Eqs. (A.11) and (A.16), we have:

$$f(i, v_i, d_i^0) = A_0 \exp\left(-\frac{d_i^{02}}{2\sigma_d^2}\right)\left[1 + \sum_{j\neq i}^{N}\sum_{k=-M}^{M} A_0\epsilon_{ij}^k \times \exp\left(-\frac{(v_j^k - v_j)^2}{2\sigma_v^2}\right)\exp\left(-\frac{(d_i^0 + v_{ij}^k/K)^2}{2\sigma_d^2}\right)\right] \tag{A.17}$$

By choosing:

$$\epsilon_{ij}^k = \epsilon \exp\left[\left(\frac{v_{ij}^k}{2K\sigma_d}\right)^2\right] \tag{A.18}$$

we can again rearrange and complete the squares for the Gaussians, and then calculate $\overline{d}_i$ as the mean of the weighted Gaussian centers to obtain:

$$\overline{d}_i = -\frac{1}{K'}\sum_{j\neq i}^{N}\left[v_{ij} + \sum_{k=1}^{M}(v_{ij}^k + v_{ij}^{-k})\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)\right] \tag{A.19}$$

where

$$K' = \frac{2K\left[\sqrt{2} + \sum_{j\neq i}^{N}\sum_{k=-M}^{M} A_0\epsilon\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)\right]}{\epsilon A_0} \tag{A.20}$$

Since $v_{ij}^k + v_{ij}^{-k} = 2v_{ij}$, we get:

$$\overline{d}_i = -\frac{1}{K'}\left[1 + 2\sum_{k=1}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)\right]\sum_{j\neq i}^{N} v_{ij} \tag{A.21}$$

or

$$\overline{d}_i = \frac{\langle v\rangle - v_i}{\hat{K}} \tag{A.22}$$

where

$$\hat{K} = \frac{K'}{N\left[1 + 2\sum_{k=1}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)\right]} \tag{A.23}$$

Therefore, we find again a linear relationship between $\overline{d}_i$ and v_i as in Eq. (15) but with a different proportionality constant $\hat{K}$.

A.5. Derivation of Eq. (18)

We will use the following two conventions introduced in the Section 2: (1) An over-line such as $\overline{d}_i$ means average over the population of cells with different preferred disparities d_i^0 for a given location i, and (2) Brackets such as $\langle v\rangle$ denote an average over the different retinal positions i.

The standard deviation of disparity at a given position i is:

$$\sigma_i^2 = \overline{d_i^{02}} - \{\overline{d_i^0}\}^2 \tag{A.24}$$

and the averaged standard deviation over the different positions is:

$$\sigma^2 \equiv \langle \sigma_i^2 \rangle = \langle \overline{d_i^{02}} \rangle - \langle \overline{d_i^0}^2 \rangle \tag{A.25}$$

From Eq. (A.22), we have:

$$\overline{d_i^0}^2 = (\langle v \rangle^2 + v_i^2 - 2v_i \langle v \rangle)/\hat{K}^2 \tag{A.26}$$

Therefore

$$\langle \overline{d_i^0}^2 \rangle = (\langle v^2 \rangle - \langle v \rangle^2)/\hat{K}^2 = \sigma_{\text{vel}}^2/\hat{K}^2 \tag{A.27}$$

We now calculate $\langle \{\overline{d_i^0}\}^2 \rangle$. First note that:

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right) \mathrm{d}x = \sqrt{2\pi}\sigma\left(\sigma^2 + \overline{x}^2\right) \tag{A.28}$$

By definition, we have:

$$\{\overline{d_i^0}\}^2 = \frac{\int_{-\infty}^{\infty} \{d_i^0\}^2 f(i, v_i, d_i^0) \mathrm{d}(d_i^0)}{\int_{-\infty}^{\infty} f(i, v_i, d_i^0) \mathrm{d}(d_i^0)} \tag{A.29}$$

Since $f(i, v_i, d_i^0)$ is a sum of Gaussians, we can apply Eq. (A.28) to the numerator of Eq. (A.29). The denominator is calculated as before.

If $\sigma_v = 0$, we have (see Eq. (A.14)):

$$\{\overline{d_i^0}\}^2 = \frac{A_0\sigma_d^3\sqrt{2\pi} + \sum_j^N A_0^2\epsilon\sqrt{\pi}\sigma_d\left(\frac{\sigma_d^2}{2} + \frac{v_{ij}^2}{4K^2}\right)}{A_0\sigma_d\sqrt{2\pi} + (N-1)A_0^2\epsilon\sqrt{\pi}\sigma_d} \tag{A.30}$$

Using the fact that:

$$\sum_j^N v_{ij}^2 = \sum_j^N (v_i - v_j)^2 = N(\langle v^2 \rangle + v_i^2 - 2v_i\langle v \rangle) \tag{A.31}$$

we can rewrite Eq. (A.30) as:

$$\overline{d_i^{02}} = A + B(\langle v^2 \rangle + v_i^2 - 2v_i\langle v \rangle) \tag{A.32}$$

where

$$A = \frac{\sigma_d^2(1 + (N-1)A_0\epsilon/(2\sqrt{2}))}{1 + (N-1)A_0\epsilon/\sqrt{2}} \tag{A.33}$$

and

$$B = \frac{NA_0\epsilon}{4\sqrt{2}K^2(1 + (N-1)A_0\epsilon/\sqrt{2})} \tag{A.34}$$

Averaging Eq. (A.32) over retinal positions, we obtain:

$$\langle \{\overline{d_i^0}\}^2 \rangle = A + 2B\sigma_{\text{vel}}^2 \tag{A.35}$$

Finally, by replacing Eqs. (A.27) and (A.35) in Eq. (A.25) we obtain:

$$\sigma^2 = A + \sigma_{\text{vel}}^2(2B - 1/\hat{K}^2) \tag{A.36}$$

When $A_0\epsilon N \gg 1$, we have $A \simeq \sigma_d^2/2$, $B \simeq 1/(4K^2)$ and $\hat{K} \simeq 2K$, then:

$$\sigma^2 \simeq \frac{\sigma_d^2}{2} + \frac{\sigma_{\text{vel}}^2}{4K^2} \tag{A.37}$$

If $\sigma_v \neq 0$, we have (see Eq. (A.17)):

$$\{\overline{d_i^0}\}^2 = \frac{A_0\sigma_d^3\sqrt{2\pi} + \sum_j^N \sum_{k=-M}^{M} A_0^2\epsilon\sqrt{\pi}\sigma_d \exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)\left(\frac{\sigma_d^2}{2} + \frac{(v_{ij}-k\delta)^2}{4K^2}\right)}{A_0\sigma_d\sqrt{2\pi} + (N-1)A_0^2\epsilon\sqrt{\pi}\sigma_d \sum_{k=-M}^{M} \exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)} \tag{A.38}$$

Simplifying and separating terms we get:

$$\begin{aligned}\{\overline{d_i^0}\}^2 = {} & \frac{\sigma_d^2 + A_0\epsilon/\sqrt{2}\sum_j^N\left(\frac{\sigma_d^2}{2} + \frac{v_{ij}^2}{4K^2}\right)\sum_{k=-M}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)}{1 + (N-1)A_0\epsilon/\sqrt{2}\sum_{k=-M}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)} \\ & + \frac{A_0\epsilon/\sqrt{2}\sum_j^N\sum_{k=-M}^{M}\frac{(k\delta)^2}{4K^2}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)}{1 + (N-1)A_0\epsilon/\sqrt{2}\sum_{k=-M}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)} \\ & - \frac{A_0\epsilon/\sqrt{2}\sum_j^N\sum_{k=-M}^{M}\frac{2v_{ij}k\delta}{4K^2}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)}{1 + (N-1)A_0\epsilon/\sqrt{2}\sum_{k=-M}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)}\end{aligned} \tag{A.39}$$

It is easy to see that the last term vanishes because every term in the sum is canceled by the corresponding one for k of the opposite sign. The first term is almost identical to Eq. (A.30), but instead of ϵ we have $\hat{\epsilon} = \epsilon\gamma$ with $\gamma = \sum_{k=-M}^{M}\exp(-(k\delta)^2/2\sigma_v^2)$. Thus, in the limit of $A_0\epsilon N \gg 1$ the contribution of this term to σ^2 is identical to the contribution of Eq. (A.30) just calculated.

Let us now study the contribution of the second term, to be denoted as C. If $A_0\epsilon N \gg 1$ we have:

$$C \simeq \frac{1}{4K^2}\frac{\sum_{k=-M}^{M}(k\delta)^2\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)}{\sum_{k=-M}^{M}\exp\left(-\frac{(k\delta)^2}{2\sigma_v^2}\right)} \tag{A.40}$$

We can approximate the sums by the corresponding integrals, and obtain $C \simeq \sigma_v^2/(4K^2)$ (see Eq. (A.28)).

Putting all contributions together we get:

$$\sigma^2 \simeq \frac{\sigma_d^2}{2} + \frac{\sigma_{\text{vel}}^2 + \sigma_v^2}{4K^2} \tag{A.41}$$

For our simulation parameters, the second term is smaller than the first one. These results are confirmed by the simulations.

A.6. Derivation of Eq. (22)

Now the input disparity field d_i is different from zero. In such a case, Eq. (A.13) becomes:

$$\begin{aligned}f(i, v_i, d_i^0) = {} & A_0\exp\left(-\frac{(d_i^0 - d_i)^2}{2\sigma_d^2}\right) \\ & \times\left[1 + \sum_{j\neq i}^{N} A_0\epsilon_{ij}\exp\left(-\frac{(d_i^0 - d_j + v_{ij}/K)^2}{2\sigma_d^2}\right)\right]\end{aligned} \tag{A.42}$$

Recall the definition of K_s in Eq. (21). If we take as a reference some position i for which $d_i = 0$ and $v_i = v_0$, then we have $d_j = (v_0 - v_j)/K_s$, and Eq. (A.42) becomes:

$$f\left(i, v_i, d_i^0\right) = A_0 \exp\left(-\frac{\left(d_i^0 - \frac{v_0 - v_i}{K_s}\right)^2}{2\sigma_d^2}\right) \times \left[1 + \sum_{j \neq i}^{N} A_0 \epsilon_{ij} \exp\left(-\frac{\left(d_i^0 + \frac{v_i}{K} + v_j\left(\frac{1}{K_s} - \frac{1}{K}\right) - \frac{v_0}{K_s}\right)^2}{2\sigma_d^2}\right)\right] \tag{A.43}$$

In the special case when $K_s = K$, we have:

$$f\left(i, v_i, d_i^0\right) = A_0 \exp\left(-\frac{\left(d_i^0 - \frac{v_0 - v_i}{K}\right)^2}{2\sigma_d^2}\right) \times \left[1 + \sum_{j \neq i}^{N} A_0 \epsilon_{ij} \exp\left(-\frac{\left(d_i^0 + \frac{v_i - v_0}{K}\right)^2}{2\sigma_d^2}\right)\right] \tag{A.44}$$

If we compute $\overline{d}_i$ as before, we get:

$$\overline{d}_i = \frac{-\left[A_0 \sigma_d \sqrt{2\pi} \frac{v_i - v_0}{K} + \sum_j^N A_0^2 \epsilon \sqrt{\pi} \sigma_d \frac{v_i - v_0}{K} \exp\left(\frac{v_i - v_j}{2K\sigma_d}\right)^2\right]}{A_0 \sigma_d \sqrt{2\pi} + \sum_j^N A_0^2 \epsilon \sqrt{\pi} \sigma_d \exp\left(\frac{v_i - v_j}{2K\sigma_d}\right)^2} = \frac{v_0 - v_i}{K} \tag{A.45}$$

and thus the correct depth structure is obtained.

References

Anderson, C. H., & Essen, D. C. (1987). Shifter circuits—a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Science USA*, *84*, 6297–6301.

Andersen, R. A., & Siegel, R. M. (1990). Motion processing in the primate cortex. In G. M. Edelman, W. L. Gall, & W. M. Cowan (Eds.), *Signal and Sense: Local and Global Order in Perceptual Maps*. New York: Wiley, pp. 163–184.

Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey. *Journal of Neurophysiology*, *84*, 2658–2669.

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

Bradley, D. C., Chang, G. C., & Andersen, R. A. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, *392*, 714–717.

Braunstein, M. L., Andersen, G. J., Rouse, M. W., & Tittle, J. S. (1986). Recovering viewer-centered depth from disparity, occlusion, and velocity gradients. *Percept. Psychophys.*, *40*, 216–224.

Buracas, G. T., & Albright, T. D. (1994). The role of MT neuron receptive field surrounds in computing object shape from velocity fields. *Adv. Neural Info. Proc. Sys.*, *6*, 969–976.

Buracas, G. T., & Albright, T. D. (1996). Contribution of area MT to perception of three-dimensional shape: a computational study. *Vision Research*, *36*, 869–887.

Chen, Y., Wang, Y., Qian, N., (2001). Modeling V1 disparity tuning to time-dependent stimuli. *Journal of Neurophysiology*.

Dosher, B. A., Sperling, G., & Wurst, S. A. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. *Vision Research*, *26*, 973–990.

Droulez, J., & Cornilleau-Perez, V. (1990). Visual perception of surface curvature: the spin variation and its physiological implementation. *Biol. Cybern.*, *62*, 211–224.

Fernández, J. M., & Qian, N. (2000). A physiologically-based model for computing relief structure from motion. *Soc. Neurosci. Abs.*, *26*, 250.2.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197.

Hildreth, E. C., Ando, H., Andersen, R. A., & Treue, S. (1995). Recovering three-dimensional structure from motion with surface reconstruction. *Vision Research*, *35*, 117–137.

Howard, I. P., & Rogers, B. J. (1995). *Binocular Vision and Stereopsis*. New York, NY: Oxford University Press.

Koenderink, J. J., & Doorn, A. J. (1992). Second order optic flow. *J. Opt. Soc. Am. A*, *9*,, 530–538.

Maunsell, J. H. R., & 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*, 1127–1147.

Nawrot, M., & Blake, R. (1989). Neural integration of information specifying structure from stereopsis and motion. *Science*, *244*, 716–718.

Nawrot, M., & Blake, R. (1991). A neural network model of kinetic depth. *Visual Neuroscience*, *6*, 219–227.

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

Qian, N., & Andersen, R. A. (1994). Transparent motion perception as detection of unbalanced motion signals II: Physiology. *Journal of Neuroscience*, *14*, 7367–7380.

Qian, N., & Andersen, R. A. (1997). A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena. *Vision Research*, *37*, 1683–1698.

Qian, N., Andersen, R. A., & Adelson, E. H. (1994a). Transparent motion perception as detection of unbalanced motion signals I: Psychophysics. *Journal of Neuroscience*, *14*, 7357–7366.

Qian, N., Andersen, R. A., & Adelson, E. H. (1994b). Transparent motion perception as detection of unbalanced motion signals III: Modeling. *Journal of Neuroscience*, *14*, 7381–7392.

Ramachandran, V. S., Cobb, S., & Rogers-Ramachandran, D. (1988). Perception of 3-d structure from motion: The role of velocity gradients and segmentation boundaries. *Percept. Psychophys.*, *44*, 390–393.

Rogers, B. J., & Graham, M. E. (1982). Similarities between motion parallax and stereopsis in human depth perception. *Vision Research*, *22*, 216–270.

Thompson, A. M., & Nawrot, M. (1999). Abnormal depth perception from motion parallax in amblyopic observers. *Vision Research*, *39*, 1407–1413.

Todd, J. T. (1984). The perception of three-dimensional structure from rigid and nonrigid motion. *Percept. Psychophys.*, *36*, 97–103.

Todd, J. T. (1998). Theoretical and biological limitations on the visual perception of 3d structure from motion. In T. Watanabe (Ed.), *High-level motion processing—computational, neurophysiological and psychophysical perspectives* (pp. 359–380). Cambridge, MA: MIT Press.

Todd, J. T., & Perotti, V. J. (1999). The visual perception of surface orientation from optical motion. *Percept. Psychophys.*, *61*, 1577–1589.

Treue, S., & Andersen, R. A. (1996). Neural responses to velocity gradients in macaque cortical area MT. *Visual Neuroscience*, *13*, 797–804.

Treue, S., Andersen, R. A., Ando, H., & Hildreth, E. C. (1995). Structure-from-motion: Perceptual evidence for surface interpolation. *Vision Research*, *35*, 139–148.

Ullman, S. (1979). *The Interpretation of Visual Motion*. Cambridge, MA: MIT Press.

Ullman, S. (1983). Recent computational studies in the interpretation of structure from motion. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and Machine Vision* (pp. 459–480). New York: Academic Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-d structure from rigid and nonrigid motion. *Perception*, *13*, 255–274.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psychol.*, *45*, 205–217.

Xiao, D. K., Marcar, V. L., Raiguel, S. E., & Orban, G. A. (1997a). Selectivity of macaque MT/V5 neurons for surface orientation in depth specified by motion. *European Journal of Neuroscience*, *9*, 956–964.

Xiao, D. K., Raiguel, S. E., Marcar, V., & Orban, G. A. (1997b). Spatial distribution of the antagonistic surround of MT/V5 neurons. *Cereb. Cortex*, *7*, 662–677.

Xiao, D. K., Raiguel, S. E., Marcar, V., Koenderink, J., & Orban, G. A. (1995). Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proceedings of National Academy of Science USA*, *92*, 11303–11306.