Published in final edited form as:

Neurocomputing. 2008 March ; 71(7-9): 1629–1641. doi:10.1016/j.neucom.2007.04.006.

A neural model for the integration of stereopsis and motion parallax in structure-from-motion

Julian Martin Fernandez and Bart Farell

Institute for Sensory Research, Syracuse University, Syracuse, New York 13244-5290, USA

Abstract

We introduce a model for the computation of structure-from-motion based on the physiology of visual cortical areas MT and MST. The model assumes 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's MT neurons are connected to each other laterally to form modulatory receptive-field surrounds that are gated by feedback connections from area MST. This allows the building up of a depth map from motion in area MT, even in absence of disparity in the input. Depth maps from motion and from stereo are combined by a weighted average at a final stage. The model's predictions for the interaction between motion and stereo cues agree with previous psychophysical data, both when the cues are consistent with each other or when they are contradictory. In particular, the model shows nonlinearities as a result of early interactions between motion and stereo before their depth maps are averaged. The two cues interact in a way that represents an alternative to the "modified weak fusion" model of depth-cue combination.

Introduction

Structure from motion (SFM) is the brain's recovery of depth information from the 2D-velocity pattern falling on the retina. A growing body of research provides evidence that humans do not perceive SFM as veridically as we used to think. Humans are able to detect most types of non-rigid deformations from motion cues alone, and to accurately discriminate structural properties that are invariant over affine transformations, but have considerably more difficulty in tasks that require an accurate perception of Euclidean metric properties (Todd, 1998; Todd & Perotti, 1999). Thus, observers' appreciation of Euclidean metric structure may be more cognitive than perceptual (Todd & Perotti, 1999). Velocity relations in a moving pattern constrain its possible rigid interpretation to a one-parameter family of structures (under orthographic projection), and thus humans can only recover *relief structure from motion*, i.e. an object's shape can be determined up to an affine stretching transformation along the line of sight. Thus, the problem is to understand how the brain recovers relief structure, and to do so using a model based, as far as possible, on physiologically plausible components.

In this article we introduce a model for relief-SFM computation. It is an expanded version of an earlier model (Fernandez, Watson & Qian, 2002). The earlier version explained several experimental results, including recovery of relief SFM, the Ramachandran illusions from coaxial rotating cylinders (Ramachandran, Cobb & Rogers-Ramachandran, 1988), and the correlation between the activity of disparity-tuned MT cells and the perceived depth polarity

of ambiguous stimuli (Bradley, Chang & Andersen, 1998). The model had one major limitation: it had no mechanism for combining the depth maps constructed from different cues. The modification present here overcomes this constraint problem. This will enable us to study the problem of motion-stereo integration and compare the model's predictions against psychophysical data. To this end we will use the data of Rogers & Collett (1989). Rogers & Collett assessed perceived depth when subjects were shown surface shapes in which motion and stereo cues specified the same depth or different depths. The combined depth in the final percept was compared to that obtained when either cue, motion or stereo, was presented in isolation. We chose this work because it is a well-implemented representative study of motion- and stereo-cue combination that covers a large range of parameters—both when the cues agree and when they are in conflict. Also important in this choice was that Rogers & Collett directly measured perceived depth, rather than discrimination thresholds, and therefore provide data in a format that can be directly compared with the output of our model.

We introduce the model below, with separate sections devoted to the physiological properties used by the model, and an anatomical and functional overview of the model. We then describe the computations that give the model's predictions for the Rogers & Collett experiment. A review of the psychophysics of SFM, and a mathematical analysis of the SFM problem can be found in Fernandez *et al.* (2002) and will not be repeated here.

Physiological motivation of model assumptions

We will review here properties of neurons in visual areas V1, MT and MST. We will review here those properties that bear on our model.

Visual area V1 represents the earliest stage in which binocular disparity is coded. Neurons in this area are broadly tuned to different values of binocular disparity, forming what is essentially a continuous distribution (LeVay & Voigt, 1988; Ohzawa *et al.*, 1996). Disparity is widely represented outside V1, both in the dorsal stream and the ventral stream. Of particular concern to the model is the disparity sensitivity of neurons in area MT, which is highly responsive to motion.

Area MT (V5) seems to play a major role in SFM. Neurons in this area respond to direction and speed of motion (Maunsell & Van Essen, 1983a) and also to binocular disparity (Maunsell & Van Essen, 1983b) and are organized in direction (Albright, Desimone & Gross, 1984) and disparity (DeAngelis & Newsome, 1999) columns. The responses of many MT neurons to velocity to stimuli within the central or “classical” receptive field (CRF) can be modulated by appropriate simultaneous stimulation of a larger surrounding region of the visual field (up to 10 times larger than the CRF)—the “non-classical” surround receptive field (SRF) (Allman, Miezin & McGinnes, 1985; Tanaka *et al.*, 1986; Xiao *et al.*, 1998). Stimulation of the surround alone elicits no neuronal response. The modulation produced by surround stimulation can be antagonistic or synergistic (Born & Tootell, 1992), and the geometry of the surround can be symmetric, axially symmetric, or asymmetric (Xiao *et al.*, 1998). Similar to this center-surround RF organization for velocity, primate area MT also shows a center-surround antagonism in the RF organization for disparity (Bradley & Andersen, 1998).

Area MT projects to area MST (Ungerleider & Desimone 1986; Maunsell & Van Essen, 1983c). MST cells have larger receptive fields (about ten times greater area) than MT cells, and are tuned to different patterns of motion (including rotations, expansions/contractions, spirals, translations) (Graziano, Andersen & Snowden, 1994; Duffy & Wurtz, 1991, 1997). In a large proportion of MST cells, the response to motion in the center of the receptive field is increased when the surround stimulus is stationary (Eifuku & Wurtz, 1998). This contrasts with the minimal effect of stationary surrounds in MT neurons. These and other data are consistent with a role for MST neurons in the segmentation of objects using motion cues. Thus,

we hypothesize that one task of MST is the segmentation and identification of motion pattern from its local-speed MT input. The traffic goes both ways, however; area MST projects back extensively to area MT (Maunsell & Van Essen, 1983c). These feedback connections could in principle modify neural responses in order to code higher-order properties of complex scenes beyond what is possible with the typical center-surround interaction, analogous to what happens in V1 due to its feedback connections from extra-striate visual areas (Zipser, Lammer & Schiller, 1996).

Using ambiguous bistable stimuli, Bradley, Chang and Andersen (1998) observed that responses of MT cells modulated in phase with reversals of the perceived surface order. In this experiment, monkeys were trained to fixate their gaze while viewing a 2-dimensional projection of a transparent revolving cylinder. To humans, such a stimulus projection appears 3-dimensional, with spontaneously reversing perceived surface order and direction of rotation. These alternative perceptual outcomes are equally consistent with the physical evidence of the stimulus, which is ambiguous. Monkeys in the Bradley, Chang and Andersen study reported the direction of rotation (clockwise or counterclockwise) they perceived, and this could be consistent with one perceived surface order or the other. In many of the MT neurons tested, there was a reproducible change in activity that coincided with reversals of the perceived surface order. The change in activity was consistent with the tuning of the cell to disparity—for instance, enhanced responsiveness of a cell tuned to left motion and near disparity when the surface with left motion was perceived in front.

An anatomical link between the perception of SFM and the perception of depth from binocular disparity comes from the study of amblyopic subjects. Amblyopia refers to reduced vision in one eye caused by visual deprivation in childhood. Amblyopic subjects perform much worse than normal subjects not only when discriminating depth using disparity as the cue, as expected, but also when discriminating depth using motion parallax as the cue (Thompson & Nawrot, 1999). This supplements a variety of psychophysical interactions and similarities between binocular stereopsis and motion parallax in the perception of depth (see Thompson & Nawrot (1999) and references therein). Evidence from fMRI responses in humans suggests that hMT +/V5 and the lateral occipital complex are involved in the combination of depth cues from binocular disparity and perspective (Welchman *et al.*, 2005).

Together these studies present suggestive evidence that the analysis of binocular disparity and SFM share common or interacting neural processes that lead to or support the perception of depth from motion seen in the absence of disparity cues, and that areas MT and/or MST have a major role in SFM computation and depth-cue combination. The model described below attempts to integrate this evidence. The model is not intended to be used as a basis for implementing a machine vision system. Thus, the model as described here lacks a number of fundamental components that would be needed for it to serve as a schematic guide a practical application (e.g., an input preprocessing module that performs such low-level tasks as noise reduction and image segmentation, and dedicated computational modules for computing disparity and speed at each image position). Instead, the model is meant to provide a physiologically-informed understanding of how the human visual turns stereo and motion depth cues into a represent the depth structure of objects.

The model

Anatomical Overview

The model's architecture is sketched in Fig. 1. It consists of three functional units: 1) Two input areas, which we identify with cortical areas V1 and MT, 2) An association area, identified with cortical area MST and reciprocally connected to MT, and 3) A depth-cue combination

area—provisionally identified as the lateral occipital complex (LOC)—that receives input from both V1 and MT.

In our model, area MT receives direct visual input, but physiologically most of the forward input to MT comes from area V1 (Sincich *et al.*, 2004). Our strategy here was to directly make use of the known tuning properties of MT cells while avoiding the complication of explicitly modelling these properties from the V1 input (Simoncelli & Heeger, 1998), which is not the concern of our model.

Neurons in area MT are tuned to both velocity and disparity. As seen in Fig. 1, MT is modelled as consisting of two different layers, both of which receive direct visual input. In both layers cells are arranged retinotopically and cells representing a given position on the retina are arranged according to their tuning for disparity and speed. The two layers are similarly tuned, but differ in their connectivity. Whereas cells in MT Layer 1 receive only direct visual input, Layer 2 cells receive both direct visual input and modulatory input from Layer 1 cells. Layer 1 cells also send excitatory connections to all MST cells. Layer 2 cells, by contrast, send excitatory connections only to area LOC. Area LOC also receives excitatory connections from V1. Like MT cells, V1 and LOC cells are arranged retinotopically and according to their disparity tuning. As noted earlier, MST cells, because of their input from MT Layer 1, are best tuned to global configurations in the input. Thus, MST works as a global pattern recognition, or template matching, area.

Modulatory connections from MT Layer 1 to MT Layer 2 affect the firing rate of active cells, but do not by themselves activate an inactive cell. Modulatory connections are arranged in clusters corresponding to different branches of the axon of the modulatory cell, and cells within a modulated cluster are related in their tuning parameters. Let v^0 and d^0 be the preferred speed and disparity, respectively, of a cell in Layer 1 or 2. The modulatory connections from Layer 1 to Layer 2 are non-zero only for cells whose differences in preferred speed (Δv^0) and in preferred disparity (Δd^0) are related by:

$$\Delta v^0 = -K \Delta d^0 \quad (1)$$

where K is a constant that differs across clusters. Eq. 1 mimics a similar relationship between differences in disparity and speed that is satisfied between any two points in a rotating or translating object (Fernandez *et al.*, 2002).

MST cells in turn send gating feedback, shown by dashed lines in Fig. 1, to selected axon branches of MT Layer 1 cells. These axon branches carry modulatory signals to different clusters of MT Layer 2. Thus, the modulatory connections from MT Layer 1 to MT Layer 2 require gating feedback signals from MST to be active. Only a subset of these modulatory connections, if any, will be active at a given moment. The active subset corresponds to a particular value of K and is selected by the pattern of activity in the MST layer, thus providing a dynamical remapping of the modulatory architecture. Examples of the computation of K are given later in this paper.

Functional Overview

The model functions as follows. Visual input activates cells in MT Layers 1 and 2, in accordance with their disparity and speed tuning properties. Activity in MT Layer 1 simply reflects the disparity and speed information in the stimulus. But activity in Layer 2 is influenced by modulation from Layer 1 cells, as described by Eq. 1. Because of these connections—which mimic a similar relationship satisfied between any two points in a rotating or translating object—the activity of some cells is enhanced at the expense of that of others. The modulation shifts the Layer-2 population activity, making it more consistent to the depths given by the motion cue. For instance, feeding the network an input simulating a rotating or translating object, the

resulting pattern of activity in MT Layer 2 will reveal that cells tuned to disparities different from zero are active, even if the original pattern had zero disparity (a flat, fixation-plane stimulus) or no disparity at all (monocular viewing). Taking population averages at each retinal position, we get a depth map in terms of these motion-derived illusory disparities. This is what we call the depth map from motion.

When the input disparity is different from zero, stereo information interacts with motion in the build up of the depth map from motion. The depth map from the binocular disparity of the stimulus, however, is assumed to be represented in area V1, and to be represented independently from the depth map from motion in MT Layer 2. These two depth maps—the depth map from binocular disparity and the depth map from motion—are averaged in area LOC. Averaging of other depth cues (like shading and texture) is also assumed but is not modelled here.

There are two features of the model that make it possible to compute a depth map in MT Layer 2. These are the modulatory connections from MT Layer 1, and the gating feedback from MST. These work as follows. First, as mentioned earlier, for a given rotating or translating object, the difference in depth between any two points projecting on the retina can be calculated from by combining information from two sources. One source is the difference in retinal speeds between the two points. The other source is parameters, like speed of rotation and object distance, that depend on the whole stimulus and not on the particular retinal positions; these parameters are common to all pairs of object points (Fernandez *et al.*, 2002). Thus, MT Layer 1 neurons excite MT Layer 2 neurons at other retinal positions that satisfy the same depth-retinal speed relationship. Any particular depth-speed relationship is represented by a straight line in the disparity-speed space, as shown by the diagonal row of shaded dots in Fig. 1b (see also Eq. 1). The slope of this line depends on global stimulus parameters, including the object's rotational or translational speed and distance. Thus, a given MT Layer 1 cell should excite a set of MT Layer 2 cells that satisfy this linear disparity-speed relationship. Of those Layer 2 cells satisfying the relationship, however, only those tuned to a particular speed—the retinal speed of the stimulus at that retinal position—should be excited. This is accomplished by having modulatory instead of standard excitatory connections; the modulation affects active cells only. In other words, modulatory connections are equivalent to an AND operation, selecting only active cells that satisfy the linear disparity-speed relationship of Eq. 1.

From a single set of modulatory connections, representing a fixed slope in the disparity-speed space, the correct relief of the object would be recovered in MT Layer 2. However, the extent of depth recovered would, in general, be non-veridical—in particular, depth would be recoverable only up to an affine transformation. This problem can be managed by having several sets of modulatory connections from each cell in MT Layer 1, each set corresponding to a different slope, and allowing feedback to select among them. For a particular stimulus, only one slope value should be selected, with selection accomplished by the MST feedback gating connections. A gating connection has the effect of enabling or disabling other connections between two neurons, and we assume it to be axo-axonic, this is, to act on a connection between two cells, not on a cell body. Thus, each MST cell sends feedback to a particular cluster of modulatory MT Layer 1 to Layer 2 axonal connections, allowing only one cluster, representing a particular slope, to remain active. A winner-take-all process within MST gives the most active MST cell control over selecting the slope of the disparity-speed relationship.

The earlier version of the model (Fernandez, Watson & Qian, 2002) lacked modulation from MST. Area MT had only one set of modulatory connections and therefore was modelled as a single interconnected layer. The earlier version accounted well for several experimental results, but it had no mechanism for combining depth maps constructed from different cues. In that

version, the scaling factor for transforming depth maps to perceived depth differed across maps. Thus, the depth defined by the disparity-coding cells in the motion map would be different from the depth defined by the disparity cells in the stereo map. This made a final weighted average meaningless and a quantitative study of interactions between motion and stereo cues impossible. Introducing area MST into the network solves this problem. Through dynamical remapping, accomplished by the gating connections feedback to area MT, MST provides a common metric for the disparities in different depth maps.

Methods¹

For simplicity, we assume, as is standard in SFM studied, 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). The relationship between relative disparity and relative speed is linear under orthographic projection in both of these cases (Fernandez *et al.*, 2002). 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 a neuron in area V1 with receptive field (RF) centered at retinal location i , the firing rate A_{V1} in response to image disparity d_i describes the neuron's tuning properties:

$$A_{V1}(i, d_i^0) = A_0 \exp\left(-\frac{(d_i^0 - d_i)^2}{2\sigma_d^2}\right) \quad (2)$$

where A_0 is the maximum firing rate, d_i^0 is the preferred disparity (which is constant for a particular cell, but varies among different cells with RF's at location i), and σ_d describes the disparity tuning width.

At retinal location i , the tuning properties of neurons in Layers 1 and 2 in MT are given by the firing rate A in response to image speed v_i and disparity d_i :

$$A_{MT1}(i, v_i^0, d_i^0) = A_0 \exp\left(-\frac{(v_i^0 - v_i)^2}{2\sigma_v^2}\right) \exp\left(-\frac{(d_i^0 - d_i)^2}{2\sigma_d^2}\right) \quad (3)$$

where A_0 is the maximum firing rate, v_i^0 and d_i^0 are respectively the preferred speed and disparity (again constant for a particular cell, and varying across cells with RF location i), and σ_v and σ_d describe the speed and disparity tuning widths, respectively. The response of cells in Layer 2 to the external input, A_{MT2} , is described by the same tuning function as those of Layer 1 cells. But, whereas the firing rate of Layer 1 cells is given by A_{MT1} , that of neurons in Layer 2 is modulated by MT Layer 1, and is given by:

$$f_{MT2}(\alpha) = A_{MT2}(\alpha) \left[1 + \sum_{\beta \neq \alpha} \varepsilon_{\alpha\beta} A_{MT1}(\beta) \right] \quad (4)$$

where we have defined $\alpha \equiv (i, v_i^0, d_i^0)$ to denote the triplet of parameters that identify each cell (and similarly for β). The modulatory connections $\varepsilon_{\alpha\beta}$ are different from 0 only between cells

¹A Matlab program is available upon request to those who want to run simulations of the model.

whose differences between preferred speed (Δv^0) and between preferred disparity (Δd^0) are related by Eq. 1, which we repeat here:

$$\Delta v^0 = -K \Delta d^0 \quad (1)$$

and is set to 0 otherwise. Between cells satisfying this relationship, $\varepsilon_{\alpha\beta}$ is given by:

$$\varepsilon_{\alpha\beta} = \varepsilon_0 \exp\left[-\frac{\Delta d^0{}^2}{2\sigma_d}\right] \quad (5)$$

where ε_0 is a constant. This value for $\varepsilon_{\alpha\beta}$ was introduced to help find analytical solutions, but the main results presented here change only negligibly if we simply set $\varepsilon_{\alpha\beta}=1$.

The value of K is given by the most active cell in the MST layer (always the same value for a given MST cell); thus, this winning MST cell gates the particular set of modulatory connections that will be active for a given stimulus. As discussed earlier, MST cells receive their input from Layer 1 MT cells and are tuned to global patterns in the input.

To study how depth from motion and from stereo combine, we do not need to know explicitly the shape of the tuning function of MST cells. The reason is that we are not interested in how close to veridical the perceived depth of a stimulus is, but only in how the perceived depths from the two individual cues combine. Thus, we do not need to assess what specific value of K is obtained for a particular stimulus, but instead can assume K is given and then assess how the individual depth maps combine. It will only be necessary to make two general assumptions. First, we assume that the cells' tuning properties are such that binocular disparity determines depth order, this is, that the sign of K gated by the winning MST cell is consistent with the depth order present in the stimulus. For a flat stimulus (i.e., one in which disparity is zero everywhere), it is assumed that there is no single stable winning MST cell, but rather that two cells representing opposite signs of K alternate in a bistable manner. In this case the sign of K is initially chosen at random, and allowed to alternate in sign thereafter. Second, we assume that only the sign of the disparity in the input influences the winning MST cell. Thus, regardless of the disparity *magnitude* in the input (i.e., the amount of depth from the stereo cue), the winning cell will be determined only by the motion cue.

The activity of neurons in MT Layer 2 defines the depth map from motion. For each retinal location i this population of neurons reports a disparity that can be computed numerically as the population average:

$$d_i^{motion} = \frac{\sum_{d_i^0} \sum_{v_i^0} d_i^0 f_{MT2}(i, v_i^0, d_i^0)}{\sum_{d_i^0} \sum_{v_i^0} f_{MT2}(i, v_i^0, d_i^0)} \quad (6)$$

where the sum goes over all cells at retinal position i , each cell being defined by its tuning parameters d_i^0 and v_i^0 . This average is not explicitly computed in the model, but we can use it as a gauge of the depth map from motion represented in layer 2 of MT. This distribution of activity in MT Layer 2 (Eq. 4) combines with that in V1 (Eq. 2) by means of a weighted average in area LOC as follows:

$$A_{LOC}(i, d_i^0) = p \sum_{v_i^0} f_{MT2}(i, v_i^0, d_i^0) + (1 - p) A_{V1}(i, d_i^0) \quad (7)$$

where $A_{LOC}(i, d_i^0)$ and $A_{V1}(i, d_i^0)$ are the activities of cells at corresponding retinal location i in areas LOC and V1, respectively (we assume that under monocular viewing $p=1$). Note that

LOC and V1 neurons are assumed to be tuned to disparity, with a preferred value of d_i^0 , but not to speed. Finally, the disparity at position i can be “read” as:

$$d_i^{final} = \frac{\sum_{d_i^0} d_i^0 A_{LOC}(i, d_i^0)}{\sum_{d_i^0} A_{LOC}(i, d_i^0)} \quad (8)$$

This disparity map can be interpreted as a depth map only after a simple scaling operation is applied to it. The details of where and how this scaling takes place is not an issue particular to SFM; it is an unsolved problem in binocular vision generally, so we will not deal with it here. In order to match the psychophysical data, the only assumption that we make here is that under binocular viewing, this scaling factor, λ , set to 1.0 in our simulations, is half as large as its value under monocular viewing.

To assess the model’s computed depth, either for each individual cue (stereo in V1, or motion in MT Layer 2) or for the final percept (in LOC), it is better to define a global measure of depth that does not depend on the particular shape of the stimulus. In order to do this, we will first take advantage of the fact that the relationship between computed disparity in the depth map from motion in MT Layer 2 (Eq. 6) and input speed is linear—as assessed numerically, and also analytically demonstrable for a model in which the number of cells tends to infinite. This is a fundamental desired property of the model, as it implies that the correct relief structure is recovered (see Eq. 1 and Fernandez *et al.*, 2002). Thus, we can define the characteristic depth in the depth map from motion (Z_{motion}) to be the fitted slope in the curve (d_i^{motion} ’s are obtained from Eq. 6):

$$\Delta d_i^{motion} = -Z_{motion} \Delta v_i \quad (9)$$

Similarly, as mentioned earlier, the relationship between disparity and speed in the input is linear under orthographic projection (Fernandez *et al.*, 2002), and thus for the characteristic depth defined by stereo we can use Z_{stereo} defined by (d_i ’s are the input disparities)

$$\Delta d_i = -Z_{stereo} \Delta v_i \quad (10)$$

Similarly, we define the perceived depth Z_{final} (computed in LOC), by the fitting of (d_i^{final} ’s are obtained from Eq. 8):

$$\Delta d_i^{final} = -Z_{final} \Delta v_i \quad (11)$$

To compute $Z_{parallax}$, the depth from motion under monocular conditions, we set $Z_{stereo}=0$ and $p=1$ (Eq. 7) and then compute it as $Z_{parallax}=2Z_{final}$. This follows from our assumption, mentioned above, that for this case $p=1$ and $\lambda=2$

The justification of using $Z_{stereo}=0$, which represents disparity signalling a flat stimulus, for the monocular condition—which lacks disparity—is as follows.

We assume that under monocular stimulation MT cells will still fire differentially according to their tuning to speed, and that all cells tuned to a particular speed will respond equally, regardless of their tuning to disparity. Thus, the population response for disparity elicited by a monocular stimulus (i.e., a flat profile) would be the same as the response elicited by a binocular stimulus with a well-defined disparity in a model in which $\sigma_d \rightarrow \infty$. But Z_{motion} , the depth from the motion depth map, is independent of σ_d when $Z_{stereo}=0$. This means that for stimuli having $Z_{stereo}=0$, the model’s Z_{motion} output would be the same as that of a model in which $\sigma_d \rightarrow \infty$, and thus similar to the Z_{motion} output in the monocular condition. Therefore,

we can use $Z_{\text{stereo}}=0$ when simulating the monocular condition, because it will result in the same depth as when the input is truly monocular. Using $Z_{\text{stereo}}=0$ for the monocular condition implies that:

$$Z_{\text{parallax}} = 1/K \quad (12)$$

This results was obtained from the above assumption that $Z_{\text{parallax}}=2Z_{\text{final}}$ and of Eqs. 15 and 17 of Fernandez *et al.* (2002).

Parameter values used were $A_0=1$ (maximum firing rate), $\varepsilon_0=1$ (strength of the modulatory connections), $2\sigma_d^2=1$ (disparity tuning width), $\sigma_v=0$ (speed tuning width), $i=20$ (number of retinal positions) and $p=1/(1+9Z_{\text{stereo}}/Z_{\text{motion}})$ (weight of the depth map from motion in LOC, which will be discussed in more detail later). The input speeds v_i 's (20 values) were randomly chosen from $[0,1]$ (but using $[-1,1]$ gives the same results)—a random input is used because, in our model, retinal location has no influence on the results. At each i , 80 cells tuned to a preferred disparity d_i^0 from -4 to 4 were used. We used $\sigma_v=0$ in our simulations to reduce computational complexity. $\sigma_v=0$ ensures that, at each retinal position, only cells tuned to the input speed are able to fire. Thus, at each retinal position, we only used cells tuned to a preferred speed v_i^0 similar to the input speed at that retinal position, rather than a population of cells tuned to a full spectrum of speeds. In previous work (Fernandez *et al.*, 2002) we showed that using $\sigma_v=0$ gives similar results to the general case $\sigma_v \neq 0$.

In order to study depth cue combination, we feed the model with inputs in which the depths from motion (Z_{parallax}) and from stereo (Z_{stereo}) vary independently. To this end we generate stimuli with different values for Z_{stereo} —from which we compute the input disparities using the stimulus speeds (Eq. 10)—and for K (from which we compute Z_{parallax} , see Eq. 12).

Because we did not model the effect of noise, only perceived depths are computed, this is, the current model makes no threshold predictions.

Results

The basic problem in motion-stereo integration is to understand how the visual system computes the final percept when motion and stereo specify the same or different depths of the same surface shape. To this end our main goal is to reproduce the psychophysical data from Rogers & Collett (1989), which is perhaps the most complete study on this issue and at the same time provides data in a format directly comparable to the output of our model, this is, in the form of perceived depths.

The model computes the depth map from motion in MT Layer 2. Here is solved the most difficult part of the problem. The population activity of MT layer 2 model cells for a single retinal position is shown in Fig. 1c. This population activity is used to compute the perceived disparity from motion at that retinal position (Eq. 6). Perceived depth is later computed in LOC, but that stage is simply a linear combination of the depth maps computed in Layer 2 and in V1. It is in Layer 2 where the non-linear interactions between stereo and motion are generated. These early non-linearities are a fundamental aspect of the modified weak fusion model of depth cue combination (Landy, Maloney & Johnston, 1995). Notice that under monocular viewing the depth map computed in Layer 2 matches that representing perceived depth in LOC.

We ran the model for different combinations of the depth given by stereo and motion parallax. Before showing the results, let us call Z to any of the object's depth parameters (e.g., its mean depth). We can then define ω , the weight given to motion parallax in the depth map from motion, by the equation:

$$Z_{\text{motion}} = \omega Z_{\text{parallax}} + (1 - \omega) Z_{\text{stereo}} \quad (13)$$

where Z_{motion} is the depth of the object in the depth map from motion, Z_{parallax} is the perceived depth from motion parallax under monocular viewing, and Z_{stereo} is the perceived depth from binocular disparity alone. We can always define ω by Eq. 13, but for it to be properly called a weight, it must take values between zero and one, which is by no means guaranteed a priori. We can compute ω from the model's output using Eq. 13, after computing Z_{motion} from Z_{parallax} and Z_{stereo} (see Eqs. 9, 10 and 12 in *Methods*). Computed values of ω outside the $[0,1]$ interval would imply that Z_{motion} , as computed by the model, cannot be regarded as a "standard" average of Z_{parallax} and Z_{stereo} . In addition, Eq. 13 will be linear in Z_{parallax} and Z_{stereo} only if ω is a constant, but not if it varies with Z_{parallax} and Z_{stereo} . Simulation results are shown in Fig. 2. We can see that the variation of ω with Z_{parallax} and Z_{stereo} has a rich structure. This figure appears to justify our expectation of ω as a weighting factor that takes on values between 0 and 1 that vary with Z_{parallax} and Z_{stereo} . There is a small region where ω is slightly larger than one, but this departure from unity is very small in magnitude. Regions where ω is close to 1 or 0 are regions where the depth map from motion reflects monocular or binocular cues almost exclusively. From Fig. 2 we can see that ω is not a constant but a non-linear function of Z_{parallax} and Z_{stereo} , which in turn implies that Z_{motion} (Eq. 13) is not a linear function of Z_{parallax} and Z_{stereo} .

Note that ω is not the final weight given to the motion cue (Z_{parallax}), as there is another step in which the depth map from motion (Z_{motion}) and the depth map from stereo (Z_{stereo}) are averaged (see Eq. 7). According to the model, the final perceived depth (Z_{final}) is computed in area LOC as a weighted average of the depth map from motion and that built from stereo alone (Eq. 7). Thus, we have, for the final combined map:

$$Z_{\text{final}} = pZ_{\text{motion}} + (1 - p)Z_{\text{stereo}} \quad (14)$$

with $0 < p < 1$. Written in terms of Z_{parallax} and Z_{stereo} this becomes:

$$Z_{\text{final}} = p\omega Z_{\text{parallax}} + (1 - p\omega)Z_{\text{stereo}} \quad (15)$$

Thus, $p\omega$ is the effective weight given to the motion cues. As the product $p\omega$ is often less than 1/2 (see below), this result is consistent with the observed direction of between-cue cancellation aftereffects (Graham & Rogers, 1982), in which more motion parallax is needed to null the effects of adaptation to disparity-defined surfaces than vice versa. It is also consistent with reports indicating that disparity is weighted more heavily than motion parallax (Rogers & Collet, 1989; Tittle & Braunstein, 1993).

As will be discussed later (see *Discussion*), there are reliability issues involved in determining the value of p . For present purposes though, we will use the following equation:

$$p = \frac{1}{1 + 9 \frac{Z_{\text{stereo}}}{Z_{\text{motion}}}} \quad (16)$$

This is a reasonable heuristics that will be shown to provide good quantitative agreement with psychophysical data. This heuristics simply states that the weight given to motion grows with the depth-from-motion to depth-from-stereo ratio, and vice versa.

In the simulation, model observers were shown a surface in which motion and stereo cues specified the same depth or different depths. The combined depth in the final percept was compared to that obtained when either cue, motion or stereo, is presented in isolation, through monocular viewing or static stereo images. We calculated the perceived depth of the stimulus (Z_{final} in Eq. 15) as Z_{parallax} and Z_{stereo} were varied. Fig. 3a shows results. Perceived depth, Z_{final} , is shown as a function of Z_{parallax} , with Z_{stereo} as the parameter. We can see that perceived depth is given by a combination of the depth specified by both cues in isolation, in a highly non-linear fashion. It is important to stress here that even when the depths from the two cues

were in conflict (this is, they both indicate the same relief, but with different depth intensities Z_{parallax} and Z_{stereo}), the correct relief structure is recovered in the depth map from motion. This was tested by confirming that in all cases (not shown) the obtained relationship between $\Delta d_i^{\text{final}}$ and Δv_i in Eq. 11 is linear, as is that in the input stimulus defined by Eqs. 9 and 10.

In the study that we modelled here (Rogers and Collet, 1989), observers adjusted the depth of a 0.2 cycles/deg corrugated surface specified by different amounts of motion parallax and disparity. The goal was to make the depth of this adjusted surface match the depth of a corrugated test surface specified by equal values of disparity and parallax-defined depth. To this end, observers were required to track the translating surfaces as they oscillated to and from across the line of sight. Using a potentiometer, observers adjusted the peak-to-trough depth of the corrugation until the amplitudes of both surfaces appeared equal. The test surface was shown in alternation (with a period of 6 s) to the adjusted surface for as long as the subjects needed to be satisfied with the match. They tested a range of peak-to-trough disparities between 0' and 8' and parallax amplitudes between 0' and 11' (given in units of equivalent depth-defined disparity). Rogers and Collet (1989) also included a condition in which the adjusted surface was presented monocularly.

The data of Rogers and Collet (1989) in their Fig. 5 are shown in Fig. 3b. This data are directly comparable with ours shown in Fig. 3a. The quantitative similarity between the two plots is exceptional.

Potential extensions of the model

We will discuss here ways in which the model could be developed to explain additional psychophysical data.

The model's early interactions between motion and stereo could in principle support between-cue threshold elevation data and subthreshold summation (Bradshaw & Rogers, 1996). Between-cue threshold elevation—the elevation of the depth discrimination thresholds for the motion depth cue after adaptation to the stereo depth cue, and vice versa—demonstrates that the mechanisms processing depth from stereo and motion are not completely independent, though they do not distinguish between late, linear interactions and early, non-linear interactions between the two cues. On the other hand, early interactions between stereo and motion do seem to be necessary in order to explain why the subthreshold summation effect found experimentally (Bradshaw & Rogers, 1996) is greater than that predicted by most models of probability summation (Graham, 1989). Even if our model is not intended to compute thresholds—additional assumptions are needed to compare noisy population responses across trials and to model the effects of adaptation—the model provides a mechanism that, with some further elaboration, could naturally explain the data.

In extreme cases one cue can veto the other. This can happen when the two cues define different surfaces, such as orthogonal gratings. The result is rivalry (Ichikawa & Saida, 1996; Norman & Todd, 1995). Rivalry is represented in the model as the combined weight product $p\omega$ having a value of either 1 or 0. We did not model rivalry explicitly because it is not a problem particular to SFM but to vision in general.

In the model, stereo information is combined with motion information in the formation of the depth map from motion in MT layer 2. In this sense the model is in agreement with some aspects of the modified weak fusion model of cue combination (Landy, Maloney & Johnston, 1995), in particular with the idea that depth cues can interact, in a limited way, before the final weighted averaging, which occurs in model's area LOC.

In addition, the modified weak fusion model proposes the possibility of “promotion” between different depth cues. That is, one modality has the ability to promote or specify missing parameters needed by another modality. However, there is still no consensus about the existence of promotion in depth cue combination (Bradshaw, Parton & Glennester, 2000; Norman *et al.*, 1996; Tittle & Perotti, 1997; Tittle *et al.*, 1998). Because of this, we did not include results related to promotion in this article, although our model can easily incorporate it, as we will show below.

In addition to the nonlinearities shown in Fig. 2, the model also allows for other forms of non-linear interaction between the depths defined by motion and stereo. This happens when the sign of the depth cued by motion parallax is ambiguous. If no stereo cues are present, the depth sign either is decided by some heuristic or becomes bistable. But stereo cues, when present, determine the sign of the depth defined by motion parallax in most cases (Rogers & Collet, 1989; Ichikawa & Saida, 1996). As shown next, the model can reproduce this behaviour because of the tuning properties of the MST neurons.

First, let us introduce some examples of MST tuning functions, in order to show explicitly how the model can explain additional features such as bistability, sign of depth from stereo, and promotion. A simple example of a MST cell’s tuning function is given by:

$$A_{\text{MST}}(\langle v \rangle, \theta, Z_0) = A_0 \exp\left(-\frac{(\langle v \rangle^0 - \langle v \rangle \text{sign}(\langle v \rangle))^2}{2\sigma_{\langle v \rangle}^2}\right) \exp\left(-\frac{(\theta^0 - \theta)^2}{2\sigma_{\theta}^2}\right) \exp\left(-\frac{(Z_0^0 - Z_0)^2}{2\sigma_{Z_0}^2}\right) \quad (17)$$

where $\langle v \rangle$ is the average speed in the stimulus, θ is the retinal size of the stimulus in the preferred direction of motion (i.e., perpendicular to the axis of rotation), and Z_0 is the stimulus’ absolute distance (e.g., as given by vergence angle and/or lens accommodation). *Sign* refers to the sign function and reflects tuning to a specific direction of rotation. The superscript 0 indicates the preferred parameter values for this cell, and the σ ’s indicate the respective tuning widths. The estimated distance information is probably not available from the input from MT, so it is assumed to come from another (unspecified) area. (The dependence of MST cell responses on binocular disparity is not shown in Eq. 17, as this equation gives only a first approximation.) In addition, the model establishes that the particular value of K gated by a given MST cell should be directly related to its tuning properties. An instance for the tuning function given by Eq. 17 would be:

$$K = \frac{3\pi \langle v \rangle^0 Z_0^0}{4a\theta^0} \quad (18)$$

where a is the inter-ocular distance. This particular choice would result in a rotating cylinder’s depth being veridically perceived (see Fernandez *et al.* (2002) for a derivation). Note that for transparent stimuli, MST cells indicating equal magnitudes but opposite signs for K will be equally active, and competition between them would produce a bistable percept, in agreement with psychophysical results (Bradley *et al.*, 1998).

A better way to define the tuning properties of MST cells would be to define the cells’ receptive field profiles in terms of different patterns in speed and disparity (i.e., templates), and then to compute the activity of a given cell by filtering (convolving) the input by the corresponding receptive field profile for that cell. For instance, cells could be tuned to respond best to cylinders of different sizes rotating at different speeds (the templates), and then K could be computed for an arbitrary object by finding the most active MST cell. Thus, for an arbitrary object the best matching MST cell will not necessarily represent a good match, which implies that depth would not in general be veridically perceived for an arbitrary object. Current psychophysical data are insufficient to test any hypothesis concerning the model’s MST tuning properties, but physiological data could be of some help. The tuning properties of some MST cells suggest a

sensitivity to rotating patterns in depth. Disparity-dependent direction-selective (called ‘DDD’) neurons found in area MST have preferred direction of stimulus motion that reverses as the disparity of the stimulus reverses (Roy, Komatsu & Wurtz, 1992). For instance, a DDD cell that responded best to rightward motion for near stimuli responded best to leftward motion for far stimuli. These cells also responded, to a lesser extent, to stimuli having zero disparity regardless of (left or right) direction. Thus DDD cells’ features make them well suited to account for the disambiguation of depth sign from motion when there is disparity in the stimulus. In addition, the dependence of MST receptive field profiles on binocular disparity provides a mechanism for promotion in depth-cue combination: When multiple cues are present, the value of K determined by the best-tuned MST cell could be different from the value determined by the best cell when only motion is present. The matching in the case of multiple cues should be better. Thus the perceived depth obtained from two cues will be more reliable than that obtained from a single cue in isolation. This is a promotion effect. Notice, however, that in the simulations of this article we assumed that the best responding MST cell was the same regardless of the disparity in the input: K did not change with Z_{stereo} and promotion was not implemented.

Discussion

We presented a model for depth cue combination between stereo and motion in structure from motion. The model’s structure is based on simple assumptions. In overview, the model assumes that a depth map from motion is computed in area MT, with the help of feedback information from area MST. Then this depth map from motion and that from stereo—assumed to be computed in V1—are linearly combined as the basis for a final percept in area LOC. In addition to this linear combination of maps from motion and stereo, there are early non-linear interactions between motion and stereo in MT itself, whose neurons are tuned to both speed and binocular disparity.

We found, using analytical and numerical techniques, that when both stereo and motion define the same depth surface, the model recovers this structure in the depth map from motion. We also found, using numerical analysis, that when depth from both cues are in conflict (the same relief, but different depth intensities), the correct relief structure is recovered in the depth map from motion and the total depth is given by a combination of the depth specified by both cues in isolation. Quantitatively, these results were in good agreement with the findings of Rogers & Collet (1989). Also, the model seems to agree qualitatively with additional psychophysical effects of a variety of kinds.

As already mentioned, there are reasons for believing the visual system might be able to recover the depth of a particular stimulus with only limited veracity. The model MST cells are best tuned to global motion configurations in the input, especially rotating or translating configurations. However, because they are finite in number, MST cells cannot provide an optimal tuning to all possible rotating objects, only a limited number of them. Even so, the scaling in depth can in the general case be close to veridical. That closeness depends on the capacity of MST cells to find a good matching template for the presented stimulus. In the simplified version of the model presented here, the winning (most active) MST cell feeds back to gate a *particular* set (i.e., cluster) of MT Layer 1-to-Layer 2 modulatory connections. In the full version of the model, there is no single winning MST cell, but a unimodal population of cell responses. In this version, the gating of clusters in MT is not all-or-none but proportional to the activity of the corresponding MST cells. It can be demonstrated analytically that the full version of the model requires only a small number of clusters to mimic the results of the winner-take-all version implemented in this article, thus providing a more physiologically plausible implementation.

Psychophysical evidence shows that the reliability of each cue influences the final weight given to it. This has been clearly shown for the cues of motion and texture (Jacobs & Fine, 1999), and evidence suggests that it is valid for motion and stereo, too (Johnston, Cumming & Landy, 1994). In the model, this weight is $p\omega$ (Eq. 15). ω is already computed on the basis of the internal parameters of the model, and thus is not easily modulated by cue reliability. However, there are at least two ways for reliably to influence the value of the factor p . One is by changing the levels of activity of cells in the motion and/or the stereo maps themselves². The other is by changing the relative gain of the respective inputs to area LOC, where both stereo and motion are combined. Notice that in both cases the depth contribution from each cue does not change, as the shape of the population activity before combination does not change. What changes is the relative intensity of each population. Both populations are combined (i.e., superimposed) in area LOC and a final population average is computed there. This average will depend on the relative intensities of the respective input populations coming from the stereo and motion maps (V1 and MT layer 2 in our model). Thus, regardless of which of the two methods is implemented, the final outcome and the value of p remain the same.

The next natural step in improving the model would be to introduce self-organized learning from input images for both the MST tuning templates (perhaps as a Kohonen map (Kohonen, 1995)) and the feedback gating connectivity. This addition will increase the model's plausibility from a biological perspective.

One attractive feature of the model is that it provides a simple architecture (dynamical remapping of modulatory connections through higher level feedback) for solving the problem of SFM computation. In addition, it should not be difficult to extend its use to include other cues to depth. Moreover, because this kind of architecture is very general, it is quite plausible to use it in a more generic way to solve a variety of problems that would be difficult to solve using current architectures.

Finally, let us discuss the relationship between our model and other models of depth cue combination. Landy *et al.* (1995), classify all potential cue combination models into three categories: weak fusion, modified weak fusion (MWF), and strong fusion. In a weak fusion model, the depth (or any other variable that is being considered) from the different cues are processed independent from each other before combination. The final percept is obtained as a weighted average of the depths obtained from the different cues (for instance, using a maximum-likelihood estimation, or MLE, model). In a MWF model, cues are allowed to interact in restricted ways before the final averaging. According to the MWF model, each depth cue is first processed in an independent module that computes a relative depth map. These depth maps are generally specified in different units, and thus cannot be meaningfully averaged. Then, in a process termed *promotion*, an interaction between units is allowed to bring the different maps to the same common units (i.e., to absolute depth units). Finally, the promoted depth maps are averaged in an optimal fashion, such as in MLE. Finally, in a strong fusion model, the depth percept is obtained by an arbitrary combination rule. Here the computation is not necessarily modular, and it is not clear that there is any meaningful definition of "depth cue in isolation" for such a case.

According to the classification scheme outlined above, our model should be considered a strong-fusion model, because it does not match the definition of a weak fusion or a MWF model. In our model, motion and stereo interact in a nonlinear way before they are averaged, in a way different than that allowed by MWF. Recent work (Domini, Caudek & Tassinari, 2006) shows that depth estimates from a combination of stereo and motion deviates from the

²For instance, an easy way to regulate the intensity of the motion depth map is to introduce a second V1 layer. This additional V1 Layer 2 would explicitly compute Z_{stereo} from its V1 layer 1 input, and globally inhibit MT Layer 2 as a function of Z_{stereo} .

predictions of the MWF model, although they do not make quantitative predictions that could be compared with those of our model.

Acknowledgements

This research was supported by NEI Grant EY12286 (B.F.) and EY015637 (J.M.F.). Correspondence and requests for materials should be addressed to J.M.F. (e-mail: julian_fernandez@isr.syr.edu).

References

- Albright TD, Desimone R, Gross CG. Columnar organization of directionally selective cells in visual area MT of the macaque. *J Neurophysiol* 1984;51:16–31. [PubMed: 6693933]
- Allman J, Miezin F, McGuinness E. Stimulus-specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Ann Rev Neurosci* 1985;8:407–430. [PubMed: 3885829]
- Born RT, Tootell RBH. Segregation of global and local motion processing in primate middle temporal visual area. *Nature* 1992;357:497–499. [PubMed: 1608448]
- Bradley DC, Andersen RA. Center-surround antagonism based on disparity in primate area MT. *J Neurosci* 1998;18:7552–7565. [PubMed: 9736673]
- Bradley DC, Chang GC, Andersen RA. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* 1998;392:714–717. [PubMed: 9565031]
- Bradshaw MF, Parton AD, Glennester A. The task dependent use of binocular disparity and motion parallax information. *Vis Res* 2000;40:3725–3734. [PubMed: 11090665]
- Bradshaw MF, Rogers BJ. The interaction of binocular disparity and motion parallax in the computation of depth. *Vis Res* 1996;36:3457–3468. [PubMed: 8977012]
- DeAngelis GC, Newsome WT. Organization of disparity-selective neurons in macaque area MT. *J Neurosci* 1999;19:1398–1415. [PubMed: 9952417]
- Domini F, Caudek C, Tassinari H. Stereo and motion information are not independently processed by the visual system. *Vision Research* 2006;46:1707–1723. [PubMed: 16412492]
- Duffy CJ, Wurtz RH. Sensitivity of MST neurons to optic flow stimuli. I A continuum of response selectivity to large-field stimuli. *J Neurophysiol* 1991;65:1329–1345. [PubMed: 1875243]
- Duffy CJ, Wurtz RH. Medial superior temporal area neurons respond to speed patterns in optic flow. *J Neurosci* 1997;17:2839–2851. [PubMed: 9092605]
- Eifuku S, Wurtz RH. Response to motion in extrastriate area MSTl: center-surround interactions. *J Neurophysiol* 1998;80:282–296. [PubMed: 9658050]
- Fernandez JM, Watson B, Qian N. Computing relief structure from motion with a distributed velocity and disparity representation. *Vis Res* 2002;42:883–898. [PubMed: 11927353]
- Graham ME, Rogers BJ. Interactions between monocular and binocular depth aftereffects. *Invest Ophthalmol Vis Sci* 1982;22(Abs):272.
- Graham, NVS. *Visual pattern analyzers*. Oxford Univ. Press; New York: 1989.
- Graziano M, Andersen RA, Snowden RJ. Tuning of MST neurons to spiral motion. *J Neurosci* 1994;14:54–67. [PubMed: 8283251]
- Howard, IP.; Rogers, BJ. *Seeing in depth*. 2. I. Porteus; Toronto: 2002.
- Ichikawa M, Saida S. How is motion disparity integrated with binocular disparity in depth perception? *Percept Psychophys* 1996;58:271–282. [PubMed: 8838169]
- Jacobs RA, Fine I. Experience-dependent integration of texture and motion cues to depth. *Vis Res* 1999;39:4062–4075. [PubMed: 10748939]
- Johnston EB, Cumming BG, Landy MS. Integration of motion and stereopsis cues. *Vis Res* 1994;34:2259–2275. [PubMed: 7941420]
- Kohonen, T. *Springer Series in Information Sciences*. 30. Springer; New York: 1995. *Self-Organizing Maps*.
- Landy MS, Maloney LT, Johnston EB, Young MJ. Measurement and modelling of depth cue combination: in defence of weak fusion. *Vis Res* 1995;35:389–412. [PubMed: 7892735]

- LeVay S, Voigt T. Ocular dominance and disparity coding in cat visual cortex. *Vis Neurosci* 1988;1:395–414. [PubMed: 3154808]
- Maunsell JHR, Van Essen DC. Functional properties of neurons in the middle temporal visual area of the macaque monkey I. Selectivity for stimulus direction speed and orientation. *J Neurophysiol* 1983a;49:1127–1147. [PubMed: 6864242]
- Maunsell JHR, Van Essen DC. Functional properties of neurons in the middle temporal visual area of the macaque monkey II. Binocular interactions and sensitivity to binocular disparity. *J Neurophysiol* 1983b;49:1148–1167. [PubMed: 6864243]
- Maunsell JHR, Van Essen DC. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J Neurosci* 1983c;3:2563–2586. [PubMed: 6655500]
- Norman JF, Todd JT. The perception of 3-d structure from contradictory optical patterns. *Percept Psychophys* 1995;57:826–834. [PubMed: 7651807]
- Norman JF, Todd JT, Perotti VJ, Tittle JS. The visual perception of 3-d length. *J Exp Psychol: HPP* 1996;22:173–186.
- Ohzawa I, DeAngelis GC, Freeman RD. Encoding of binocular disparity by simple cells in the cat's visual cortex. *J Neurophysiol* 1996;75:1779–1805. [PubMed: 8734580]
- Ramachandran VS, Cobb S, Rogers-Ramachandran D. Perception of 3-d structure from motion: the role of velocity gradients and segmentation boundaries. *Percept Psychophys* 1988;44:390–393. [PubMed: 3226888]
- Rogers BJ, Collet TS. The appearance of surfaces specified by motion parallax and binocular disparity. *Quart J Exp Psychol* 1989;41A:697–717.
- Roy JP, Komatsu H, Wurtz RH. Disparity sensitivity of neurons in monkey extrastriate area MST. *J Neurosci* 1992;12:2478–2492. [PubMed: 1613542]
- Simoncelli EP, Heeger DJ. A model of neuronal responses in visual area MT. *Vision Res* 1998;38:743–761. [PubMed: 9604103]
- Sincich LC, Park KF, Wohlgenuth MJ, Horton JC. Bypassing V1: a direct geniculate input to area MT. *Nat Neurosci* 2004;7(10):1123–1128. [PubMed: 15378066]
- Tanaka K, Hikosaka K, Saito H, Yukie M, Fukada Y, Iwai E. Analysis of local and wide-field movements in the superior visual areas of the macaque monkey. *J Neurosci* 1986;6:134–144. [PubMed: 3944614]
- Thompson AM, Nawrot M. Abnormal depth perception from motion parallax in amblyopic observers. *Vis Res* 1999;39:1407–1413. [PubMed: 10343851]
- Tittle JS, Braunstein ML. Recovery of 3-D shape from binocular disparity and structure from motion. *Percept Psychophys* 1993;54:157–169. [PubMed: 8361830]
- Tittle JS, Perotti VJ. The perception of shape and curvedness from binocular stereopsis and structure from motion. *Percept Psychophys* 1997;59:1167–1179. [PubMed: 9401452]
- Tittle JS, Norman JF, Perotti VJ, Phillips F. The perception of scale-independent surface structure from binocular disparity, texture and shading. *Perception* 1998;27:147–166. [PubMed: 9709448]
- Todd, JT. Theoretical and biological limitations on the visual perception of 3d structure from motion. In: Watanabe, T., editor. *High-level motion processing, computational, neurophysiological and psychophysical perspectives*. MIT Press; Cambridge, MA: 1998. p. 359–380.
- Todd JT, Perotti VJ. The visual perception of surface orientation from optical motion. *Percept Psychophys* 1999;61:1577–1589. [PubMed: 10598471]
- Ungerleider LG, Desimone R. Cortical connections of visual area MT in the macaque. *J Comp Neurol* 1986;248:190–222. [PubMed: 3722458]
- Welchman AE, Deubelius A, Conrad V, Bulthoff HH, Kourtzi Z. 3D shape perception from combined depth cues in human visual cortex. *Nat Neurosci* 2005;8(6):820–827. [PubMed: 15864303]
- Xiao DK, Raiguel S, Marcar V, Orban GA. Influence of stimulus speed upon the antagonistic surrounds of area MT/V5 neurons. *Neuroreport* 1998;9:1321–1326. [PubMed: 9631422]
- Zipser K, Lammer VAF, Schiller PH. Contextual modulation in primary visual cortex. *J Neurosci* 1996;16:7376–7389. [PubMed: 8929444]

Appendix

Biographies

Julian Martin Fernandez is a Research Assistant Professor at the Institute for Sensory Research, Syracuse University, USA. He received a B.S. degree in Astronomy in 1993, a B.S. degree in Physics in 1994, and a PhD in Physics in 1998, all from National University of La Plata, Argentina. He was a Research Associate at Columbia University from September 1999 to September 2000. His research interests are physiologically plausible modeling of the visual system and psychophysical experiments on visual perception.

Bart Farell is a Research Associate Professor at the Institute for Sensory Research, Syracuse University, USA. He received a B.S. degree in Psychology in 1968, from the University of California at Berkeley, USA, a M.Ed. in Elementary Education in 1971, from Ohio University, USA, and a PhD in Psychology in 1977 from McGill University, Canada. He was a Research Associate at Bell Labs, USA, between 1985 and 1986, and at New York University, USA, between 1986 and 1990. His research interests are psychophysical experiments on visual perception, especially stereopsis.

Figure 1.

a) Model's architecture. Visual input excites area V1 and MT layers 1 and 2. The two MT layers are identical, consisting of a retinotopic arrangement of cells simultaneously tuned to disparity and speed. Each cell in layer 1 sends modulatory connections to all cells in layer 2 and excitatory connections to all MST cells. Modulatory connections have only the effect of changing the activity of an already active cell, but do not have by themselves the capacity to activate a cell. Modulatory connections are arranged in clusters corresponding to different axon's branches of the modulatory cell, and cells in a given modulated cluster share some relationships between their tuning parameters. In turn, MST cells send feedback to MT, consisting of a gating connection onto the layer 1-to-layer 2 modulatory connection clusters. Feedback from each MST cell connects to a fixed set of clusters (different for each MST cell, which depends on the tuning parameters of the MST cell), and only the gating connections of the most active MST cell are active, thus providing a dynamical remapping of the modulatory architecture. Cells with the same tuning to disparity in V1 and MT layer 2 send excitatory connections to similarly tuned cells in area LOC, in a retinotopically topographic fashion. b) Detail of the modulatory connections among MT cells at three different retinal locations ($i-1$, i , $i+1$). At each location in each layer, cells preferring different disparity d^0 and speed v^0 are arranged into a 2D array for easy visualization. Each cell is represented as a small square. Only the connections from a particular Layer 1 cell at $i-1$ are shown in the figure, but all Layer 1 cells send connections according to the pattern shown here: Connections are sent to Layer 2 cells the cells whose d^0 and v^0 parameters are related through Eq. 1. Therefore the connectivity pattern follows a straight line (marked grey) with slope $-K$ and passing through the corresponding cells (marked dark grey) with the same speed and disparity preference in different retinal positions. Only the active connections gated by the most active MST cell (i.e., corresponding to a particular value of K) are shown in the figure: inactive connections (corresponding to other values of K) are not shown. c) Population activity of MT layer 2 model cells computed from Eq. 4, corresponding to a single retinal position. This population activity is used to compute the perceived disparity from motion at that retinal position. This is done by computing the population average as prescribed in Eq. 6.

Figure 2.

Variation of ω The weight ω given to the motion cue in the depth map from motion is shown as a function of Z_{parallax} and Z_{stereo} (arbitrary units). Note that there are regions in which its value is close to 1, and others in which it is close to zero. The intersection between the surface and the plane $Z_s=0$ is a straight line with an average value of 0.494. It approaches asymptotically a horizontal straight line valued 0.5 as the number of cells used in the simulation grows.

b

Figure 3.

a) Model's predictions for perceived depth (Z_{final}) obtained for different specified amounts of depth from motion parallax (Z_{parallax}) (abscissa) and from disparity (Z_{stereo}) (parameter of different curves). Dashed line: monocular viewing, dotted line: $Z_{\text{stereo}}=0$ (binocular), full lines from bottom to top: Z_{stereo} equal to 2, 4 and 8. All arbitrary units. b) Psychophysical results, as they appear in Howard & Rogers (2002) p. 476 (adapted from Rogers & Collet, 1989). Observers matched the depth in corrugations specified by different amounts of motion parallax

(abscissa) and disparity (different curves) to depth in corrugations specified by equal values of disparity and parallax defined depth (dashed line).

Table 1

Numerical data used in drawing Fig. 3a (only data for $Z_{parallax} \geq 0$ are shown).

$Z_{parallax}$	$Z_{final} (Z_{ste\ reo}=0)$	$Z_{final} (Z_{ste\ reo}=2)$	$Z_{final} (Z_{ste\ reo}=4)$	$Z_{final} (Z_{ste\ reo}=8)$
0	0	1.9483	3.9043	7.8736
0.4000	0.1910	1.9518	3.9067	7.8720
0.8000	0.3820	1.9583	3.9098	7.8702
1.2000	0.5730	1.9681	3.9138	7.8682
1.6000	0.7640	1.9819	3.9189	7.8660
2.0000	0.9550	2.0000	3.9255	7.8636
2.4000	1.1460	2.0231	3.9341	7.8610
2.8000	1.3370	2.0519	3.9451	7.8582
3.2000	1.5280	2.0869	3.9592	7.8555
3.6000	1.7190	2.1289	3.9772	7.8529
4.0000	1.9100	2.1785	4.0000	7.8508
4.4000	2.1010	2.2365	4.0285	7.8497
4.8000	2.2920	2.3035	4.0636	7.8500
5.2000	2.4830	2.3802	4.1065	7.8526
5.6000	2.6739	2.4672	4.1581	7.8583
6.0000	2.8649	2.5653	4.2191	7.8679
6.4000	3.0559	2.6750	4.2904	7.8823
6.8000	3.2469	2.7967	4.3725	7.9022
7.2000	3.4379	2.9309	4.4657	7.9282
7.6000	3.6289	3.0781	4.5703	7.9606
8.0000	3.8199	3.2383	4.6862	7.9995
8.4000	4.0109	3.4120	4.8133	8.0452
8.8000	4.2019	3.5990	4.9515	8.0973
9.2000	4.3929	3.7995	5.1003	8.1557
9.6000	4.5838	4.0132	5.2593	8.2201
10.0000	4.7747	4.2400	5.4280	8.2901
10.4000	4.9656	4.4795	5.6058	8.3652
10.8000	5.1563	4.7313	5.7919	8.4449
11.2000	5.3468	4.9946	5.9856	8.5288
11.6000	5.5370	5.2685	6.1859	8.6163
12.0000	5.7266	5.5522	6.3918	8.7069