Perception, 2008, volume 37, pages 79 – 95

doi:10.1068/p5501

The intrinsic constraint model for stereo-motion integration

Hadley Tassinari, Fulvio Domini¶
Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, RI 02912, USA; e-mail: Fulvio_Domini@brown.edu
Corrado Caudek
Department of Psychology, University of Florence, Piazzo San Marco 4, I 50121 Florence, Italy
Received 17 August 2005, in revised form 8 January 2007

Abstract. How the visual system integrates the information provided by several depth cues is central for vision research. Here, we present a model for how the human visual system combines disparity and velocity information. The model provides a depth interpretation to a subspace defined by the covariation of the two signals. We show that human performance is consistent with the predictions of the model, and compare them with those of another theoretical approach, the modified weak-fusion model. We discuss the validity of each approach as a model for human perception of 3-D shape from multiple cues to depth.

1 Introduction

Human vision makes use of different kinds of image measurements (binocular and motion parallax, pictorial cues, etc) to estimate the properties of the 3-D layout of objects and scenes (eg depth slant, curvature). The problem of cue integration is to determine how observers combine information from multiple image measurements (cues to depth) into a unique 3-D percept. In principle, cue integration is simplest when each cue provides commensurable estimates of a quantity of interest in a scene. The problem is complicated, however, when different cues provide qualitatively different kinds of information (eg ordinal depth and absolute depth) about the properties of a visual scene (Clark and Yuille 1990).

In the present investigation, we examined visual cue integration by considering two cues to depth: binocular disparity and image motion. The empirical results are discussed in light of two different normative frameworks for depth cue combination: the modified weak-fusion model and the intrinsic constraint model.

1.1 *Modified weak-fusion model*

The problem of any model of depth-cue integration is to determine how an observer can combine information from all available sources so as to compute the statistically best (though not necessarily veridical) estimates of the relevant scene variables. If we select depth as the target scene variable, then the modified weak-fusion (MWF) model is currently the most widely accepted framework in the literature (Landy et al 1995). This model accounts for cue combination in terms of a modular architecture of the visual system: each scene variable (eg depth) is processed in a separate module whose output is an individual estimate from the relevant image measurements (eg stereo or motion). It is then necessary to find an optimal rule for combining such separate estimates.

For the MWF model, each module (indexed by i) produces a depth estimate of the true depth D that is equal to $\hat{D}_i = f_i(D)$. The goal of the MWF model is to find the D estimate with minimal variance. If the cues are conditionally independent given the scene variable (depth), then this optimal estimate is a weighted average

$$\hat{D} = \sum_i w_i \hat{D}_i \,, \tag{1}$$

¶ Author to whom all correspondence should be addressed.

where the weights

$$w_i = \frac{1}{\sigma_i^2} \Big/ \sum_i \frac{1}{\sigma_i}$$

are inversely proportional to the noise variance of each cue (ie more reliable cues are weighted more). The important characteristic of this cue combination rule is that the variance of $\hat{D}$ is smaller than the variance σ_i^2 of the individual estimates $\hat{D}_i$:

$$\frac{1}{\sigma^2} = \sum_i \frac{1}{\sigma_i^2} \,. \tag{2}$$

The MWF model is a viable alternative to *strong fusion*, that is, the attempt to compute the best estimate (in some statistical sense) of the scene variables of interest by jointly considering all the available image measurements (Clark and Yuille 1990). So far, strong fusion models have not been pursued because they become mathematically intractable as the number q of cues grows. In fact, the solution space for such models rapidly becomes prohibitive, as it would require the evaluation of functions of q image measurements, to search for maxima over q image measurements, etc.

1.1.1 *Assumptions of the modified weak-fusion model.* It is easy to understand why the MWF model considers the optimal combination to be the $\hat{D}$ having minimum variance. The MWF model attempts to explain human performance in terms of an inverse-optics approach that migrated from computer vision (eg Duda et al 2000; Faugeras 1993; Forsyth and Ponce 2003; Horn 1986). Like those models, the MWF approach also assumes that "the expected value of each cue in isolation is the true value of the property, s, ie the estimate available from each cue is unbiased" (Oruç et al 2003, page 2451). As a consequence, the goal of the MWF model becomes that of maximizing performance in the sense of increasing the precision of the estimates of the scene properties. This can be achieved by reducing the variance of the estimates (see also Atkins et al 2001; Van Ee et al 1999). Both the theoretical foundation of the MWF model and the methodology used to test it (perturbation analysis—see Young et al 1993) are based on the assumption of unbiased estimates. In the presence of largely discrepant estimates, the statistical principle of robustness is invoked (Landy et al 1995).

1.1.2 *Missing parameters.* A large number of theoretical investigations have demonstrated that, in principle, the retinal projections contain sufficient information to recover an unbiased description of the geometrical properties of the 3-D layout of objects and scenes. We shall focus here on the case of disparity information (eg Garding et al 1995; Mayhew and Longuet-Higgins 1982) and motion information (eg Hildreth 1984; Koenderink and van Doorn 1976; Longuet-Higgins and Prazdny 1980; Ullman 1979). For either cue, empirical support has been found showing that, in some stimulus conditions, perceptual performance can be veridical (Backus et al 1999; Banks et al 2002; Dijkstra et al 1995; Garding et al 1995; Lappin and Craft 2000; Mayhew and Longuet-Higgins 1982; Rogers and Bradshaw 1995; Wexler et al 2001). Within other stimulus conditions, however, reports have revealed that distortions often occur (eg Bingham et al 2004; Domini and Caudek 2003; Johnston 1991).

One reason distortions occur is that a correct Euclidean interpretation of each cue in isolation requires the knowledge of some parameters not specified in the optical information. The issue of missing parameters is critical for both approaches to cue integration discussed here.

For stereo, the relationship between horizontal disparity and depth depends on the viewing distance; a Euclidean metric solution requires a correct estimate of this parameter. The viewing distance could be estimated, for example, from ocular vergence and vertical disparities (eg Mon-Williams et al 2000). Also, with a sufficiently large

field of view, vertical disparities may be used to scale the perceived depth (eg Rogers and Bradshaw 1993).[1] If the viewing distance is unknown, and vertical disparities are not effective, 3-D reconstruction from stereo information is defined only up to an affine transformation.

A similar problem arises in the perceptual analysis of optic flow. Even though, in principle, a veridical Euclidean interpretation can be assigned to the instantaneous optic flow (Bennett et al 1989; Dijkstra et al 1994; Hoffman 1982; Koenderink and van Doorn 1976, 1991; Longuet-Higgins and Prazdny 1980), such a solution is based on the analysis of second-order optic flow (acceleration), and the human visual system is scarcely sensitive to such information (Domini et al 1997; Hogervorst and Eagle 1998). If the second-order information is not available, 3-D reconstruction from the optic flow is also defined only up to an affine transformation. In this case, the missing parameter (for a local analysis) is the component of angular rotation ω about an axis in the image plane (or, analogously, the slant of the distal surface—eg Domini and Caudek 1999).

In conclusion, neither stereo nor motion information in isolation allows a veridical 3-D metric Euclidean interpretation of the proximal stimulus information, in the absence of an unbiased estimate of viewing distance and ω. The MWF theory has dealt with this problem by postulating a process of *promotion*.

1.1.3 *Promotion.* Landy et al (1995) hypothesized that the mutual constraints conjointly provided by different cues are sufficient for disambiguating the missing parameters. The case of disparity and motion information, for example, has been discussed by Richards (1985), who showed that two binocular views of three feature points are sufficient to specify the Euclidean structure of a projected 3-D shape.

The available empirical evidence, however, does not support the hypothesis that motion information can constrain stereo interpretation: by adding motion, the perceptual interpretation of horizontal disparities does not necessarily improve (eg Johnston et al 1994; Tittle et al 1995). On this issue, Landy and Brenner (2001) concluded: "it appears unlikely that the interaction between the motion and disparity cues leads to an improvement in the estimate of the fixation distance used to scale disparities and other aspects of the 3-D percept" (page 132).

In general, perceptual judgments are not veridical, neither when stereo or motion are presented in isolation, nor when they are presented together. Thus, we cannot conclude that promotion guarantees unbiased estimates of scene properties, as hypothesized by the MWF model. In the following section, we will propose a more plausible model of depth-cue integration.

1.2 *Rationale for an alternative approach*

The motivation for the new approach that we propose stems from the hypothesis that neither promotion nor extra-visual cues may be sufficient to guarantee an unbiased estimate of the missing parameters (viewing distance, ω) necessary for a Euclidean reconstruction of 3-D shape from stereo and motion information. The goal of the present investigation is to gauge human performance against both the MWF model and the alternative approach that we propose. Both models are based on the idea that the visual system uses an 'optimal' combination rule, in some statistical sense. According to the MWF model, optimal performance means minimum variance of the combined estimate $\hat{D}$, by assuming independent and unbiased estimates from each cue. Instead, according to our alternative model, optimal performance means the most plausible 3-D estimate that can be computed by combining evidence from different sources, without assuming independent and unbiased estimates from each cue.

[1] Since we used displays which subtended a small visual angle, vertical disparities and the issue of their effectiveness for human vision are not relevant for the present discussion.

2 Intrinsic constraint model

We propose that different signals to depth are not analyzed separately, but rather conjointly define an underlying dimension from which 3-D shape can be directly recovered. In our proposal, the problem of 3-D reconstruction follows two steps: (i) a multidimensional input space is transformed into a unidimensional manifold which specifies the affine structure of the distal object; (ii) within the recovered affine space, a maximum-likelihood Euclidean solution is found.

Our proposal capitalizes on the strong covariation among the depth signals due to the fact that they are generated by the same 3-D object properties. To clarify this point, let us consider the case of an observer moving while fixating an object. In these circumstances, the object's features project a pattern of velocities on the retina, and the relative displacement of those same features with respect to the fixation point define a pattern of disparities.

In general, one disparity field can be associated with many different velocity fields, since the latter depend on the properties of the observer's motion. For any rigid transformation, however, the disparity and velocity signals are always in a linear relation to each other. This has an important implication: even if we need a 2-D space to represent the velocity – disparity input signals, these signals are maximally informative about the distal z-depth values if they are projected onto an appropriate 1-D subspace (ie the line that they identify within the velocity – disparity space). In this proposal, this underlying dimension will be termed the intrinsic constraint (IC) line. With more available sources of depth information (eg under more natural viewing conditions), the input signals define a more complex q-dimensional space, with q indicating the number of depth cues. We propose that the visual system reduces the dimensionality of the input signal-space through a process akin to principal-components analysis (PCA). But, for the present purposes, let us consider the simplified situation with only two signals.

2.1 *Dimensionality reduction*

Let us consider an observer moving with planar motion. If the viewing angle is small, the velocity and disparity fields are characterized by parallel vectors, since the vertical disparity and velocity components are negligible. If we indicate with Δz_i the depth of a point P_i with respect to the fixation point, the disparity d_i expressed in terms of visual angle can be approximated by (see figures 1a and 1b):

$$d_i \approx \frac{E}{z_\mathrm{f}^2}\Delta z_i + \varepsilon_{d_i}\,, \tag{3}$$

where z_f is the fixation distance, E is the interocular distance (IOD), and ε_{d_i} is the noise of disparity measurements. We hypothesize that ε_{d_i} is Gaussian and independent of the measurement noise of other signals. If we define the scaled depth as $z_i = \Delta z_i / z_\mathrm{f}$, and the vergence angle as $\mu \approx E/z_\mathrm{f}$,[(2)] then the previous equation becomes:

$$d_i \approx \mu z_i + \varepsilon_{d_i}\,. \tag{4}$$

If the observer moves while keeping fixation at the same point, the object will produce a retinal motion that is equivalent to the optic flow produced by the projection of an object undergoing a pure rotation. Since we assume that the visual angle subtended by the object is small (less than 8°), we can simplify the equation describing the optic flow by considering a parallel projection. In these circumstances, the velocity v_i (defined in terms of angular displacement) of a generic point P_i is:

$$v_i \approx \omega z_i + \varepsilon_{v_i}\,, \tag{5}$$

[(2)] This is a good approximation for objects at a distance of at least 50 cm from the observer.

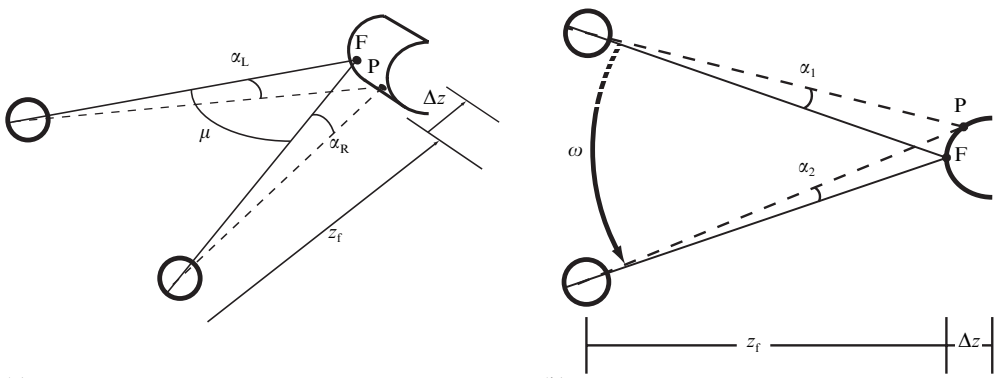

Figure 1. (a) Schematic illustration of an observer fixating a horizontally oriented hemicylinder at a viewing distance (z_f). Point F represents the fixation point, vergence angle is given by μ. α_L and α_R represent the angles formed by point P with respect to the fixation point. The front-to-back depth of the cylinder (along the line of sight) is given by Δz. (b) Schematic illustration of a monocular observer rotating by an amount ω while fixating point F. α_1 and α_2 represent the different angles subtended by F and P.

where ε_{v_i} is the noise of velocity measurements, and ω is the angular rotation undergone in the small time interval during which the optic flow was registered. Lappin and Craft (2000) hypothesized that this time window is approximately 150 ms.

Let us now consider the measurements of the disparity and velocity values produced by the projection of n feature points P_i that belong to a local region of a 3-D surface. These measurements will be correlated, since each of the two signals is linearly related to the same scaled depth map z_i. We propose that the visual system exploits this covariation in order to reduce the noise in the disparity and velocity measurements and to provide a lower-dimensional description of the input signals.

Although several computational techniques can be used for the purpose of dimensionality reduction, here we will use a standard PCA. Let us start by scaling the input signals. This is achieved by dividing equations (4) and (5) by the standard deviations of the measurement noise of the disparity d_i and velocity v_i signals, respectively:[(3)]

$$\bar{d}_i \approx \bar{\mu} z_i + \varepsilon_d \tag{6}$$

$$\bar{v}_i \approx \bar{\omega} z_i + \varepsilon_v \tag{7}$$

where $\bar{\mu}$ and $\bar{\omega}$ are the scaled vergence angle and angular velocity, and ε_d and ε_v are the measurement errors scaled by σ_d and σ_v.

What is interesting to note is that the ρ_i scores on the first principal component (PC_1) will have a higher correlation with the scaled depth z_i than either of the signals considered in isolation (see figure 2). In other words, *the scores on* PC_1 *provide the best estimate of the affine structure of the projected object, given the available disparity and velocity signals.*

This method is not the only method that can be used for estimating the affine structure. A well-known method referred to as total least squares (TLS) can perform an equivalent analysis, and only requires knowledge about the ratio between the variances of the measurement noise of the disparity and velocity signals (see Appendix B). In fact, the PCA described above can be performed on signals scaled by any factor proportional to their standard deviation as long as this factor is the same for both signals.

[(3)] In the following, we limit our analysis to the case of a local patch, by assuming that velocity and disparity noise is approximately constant. The present results, however, can be extended to the general case in which measurement noise depends on the intensity of the velocity and disparity signals.

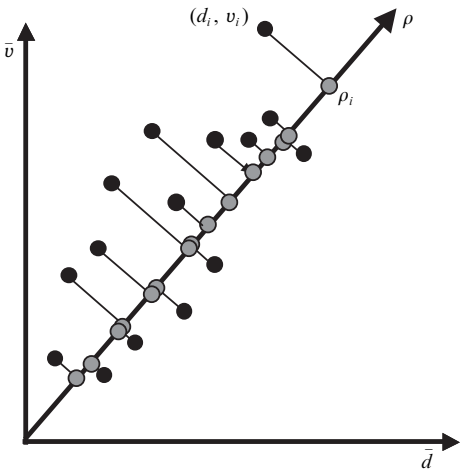

Figure 2. The intrinsic constraint subspace: black points represent noisy disparity and velocity signals scaled by σ_d and σ_v, respectively. The first eigenvector of the disparity – velocity covariance matrix determines ρ, which is the intrinsic constraint subspace. Grey points thus represent the projections of the scaled signals ρ_i on the first principal component.

2.2 *Depth interpretation*

The goal of the second stage of the proposed model is to estimate (up to a scale factor) the Euclidean depth map z_i from the scores ρ_i on PC_1 and from the first eigenvector $\boldsymbol{e}_1$. Since the disparity and velocity measurements are noisy, ρ_i and $\boldsymbol{e}_1$ will be noisy as well. Repeated viewing of the same 3-D structure at the same distance and undergoing the same 3-D rotation may, in fact, produce slightly different patterns of velocities and disparities. Consequently, the score ρ_i associated with the same distal point P_i will take on slightly different values. It can be shown that:

$$\rho_i = z_i(\bar{\mu}^2 + \bar{\omega}^2)^{1/2} + \varepsilon_\rho \,, \tag{8}$$

$$\boldsymbol{e} = \begin{bmatrix} \bar{\mu}/(\bar{\mu}^2 + \bar{\omega}^2)^{1/2} \\ \bar{\omega}/(\bar{\mu}^2 + \bar{\omega}^2)^{1/2} \end{bmatrix} + \varepsilon_{\boldsymbol{e}} \,, \tag{9}$$

where ε_ρ is Gaussian noise with zero mean and standard deviation σ_ρ, and $\varepsilon_{\boldsymbol{e}}$ is Gaussian noise with zero mean and standard deviation $\sigma_{\boldsymbol{e}}$. Equation (8) therefore shows that the scores ρ_i are linearly related to the scaled depth map z_i and, hence, specify the affine structure of the distal object.

The goal of the second stage of processing is to choose one Euclidean interpretation in the affine subspace, given the information provided by the scores ρ_i and the first eigenvector $\boldsymbol{e}_1$. This goal can be achieved in several manners. Here, we will describe a maximum-likelihood procedure. Through Bayes's rule

$$p(z_i|\rho_i, \boldsymbol{e}_1) = \frac{p(\rho_i, \boldsymbol{e}_1|z_i)p(z_i)}{p(\rho_i, \boldsymbol{e}_1)} \,; \tag{10}$$

and, therefore, the estimated scaled depth map $\hat{z}_i$ is:

$$\hat{z}_i = \text{argmax}_{z_i}\, p(\rho_i, \boldsymbol{e}_1|z_i) \,. \tag{11}$$

The likelihood function $p(\rho_i, \boldsymbol{e}_1|z_i)$ can be calculated by integrating over the unknown parameters $\bar{\mu}$ and $\bar{\omega}$:

$$p(\rho_i, \boldsymbol{e}_1|z_i) \propto \int_{\bar{\mu}\bar{\omega}} p(\rho_i|z_i, \bar{\mu}, \bar{\omega})p(\boldsymbol{e}_1|\bar{\mu}, \bar{\omega})p(\bar{\mu})p(\bar{\omega})\mathrm{d}\bar{\mu}\mathrm{d}\bar{\omega} \,. \tag{12}$$

The two probability distributions $p(\rho_i|z_i, \bar{\mu}, \bar{\omega})$ and $p(\boldsymbol{e}_1|\bar{\mu}, \bar{\omega})$ will be assumed to be Gaussian. We will also assume uniform distributions for $p(\bar{\mu})$ and $p(\bar{\omega})$, limited by $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$, respectively.

The Euclidean solution found in this way is not, in general, veridical. But it is not in any way arbitrary: it represents the best possible guess about the depth z_i, given the input signals and the assumptions that we have introduced in the interpretation process. It remains a task of psychophysical research to establish whether the constraints that we have used here are adequate to describe the actual functioning of the human perceptual system.

2.3 *Implementation of the IC model*

Before describing how to implement the above procedure, we need to briefly present the method used to collect the data. Observers were asked to perform an apparently circular cylinder (ACC) task (Johnston 1991) in three conditions: cylinders viewed monocularly (motion-only), static cylinders (stereo-only), and cylinders specified by both cues congruently. In the combined condition, disparity and motion information always specified the same amount of simulated depth.

For each observer, in the experiment described below we found the simulated depths giving rise to an ACC in each condition. These simulated depths, in turn, define the values ρ_m, ρ_s, and ρ_c for the motion-only, stereo-only, and combined conditions, respectively.

The predictions of the IC model in the combined condition are based on the observers' settings in the single-cue conditions. These predictions were formulated by running a simulation where the integral of equation (12) was computed (see Appendix A). To compute this integral, the parameters specifying the distributions $p(\bar{\mu})$ and $p(\bar{\omega})$ are required. Since we assume that these distributions are uniform and bounded, the needed parameters correspond to the maximum values $\bar{\mu}_{max}$ and $\bar{\omega}_{max}$.

The unknown values $\bar{\mu}_{max}$ and $\bar{\omega}_{max}$ were estimated by using the observers' settings in the single-cue conditions. For the stereo-only condition, equation (12) includes only the prior $p(\bar{\mu})$ and, therefore, the only unknown parameter is $\bar{\mu}_{max}$. This parameter was empirically estimated so that $p(\rho_s, \boldsymbol{e}_1 = 0|z_0)$ peaked at $z_0 = 25$ mm. Similarly, we estimated $\bar{\omega}_{max}$ from the settings of each observer in the motion-only condition.

Having estimated $\bar{\mu}_{max}$ and $\bar{\omega}_{max}$ from the observers' settings in the single-cue conditions, it was then possible to compute the integral of equation (12) in the combined condition for any ρ value. By keeping $\bar{\mu}_{max}$ and $\bar{\omega}_{max}$ fixed, the distribution $p(\rho_0, \boldsymbol{e}|z_0)$ peaks at different z-depth values if we vary the value of ρ_0. We looked for the ρ_0 value for which $p(\rho_0, \boldsymbol{e}|z_0)$ peaked at $z_0 = 25$ mm. This ρ_0 value, in turn, was the ACC depth predicted by the IC model in the combined condition (see Appendix A).

3 Experiment

Having described the implementation of the IC model, we next describe an experiment designed to investigate how well IC predictions compare with human observers' performance. Subjects participated in an ACC task (see section 3.1.4 below). Performance in the single-cue conditions provided the basis for the IC model's predictions of performance when both cues are presented congruently.

3.1 *Method*

3.1.1 *Observers.* Six observers (aged 22 to 37 years) with normal or corrected-to-normal vision participated. Four observers were naive and two were authors.

3.1.2 *Apparatus.* Stereoscopic stimuli were displayed on a haploscope consisting of two CRT monitors (0.22 mm dot pitch) located on swing arms pivoting directly beneath the observer's eyes (figure 3). Antialiasing and spatial calibrating procedures allow spatial precision of dot location greater than hyperacuity levels. Each monitor is seen in a mirror through one eye. Head position was fixed with a chin-and-forehead locating apparatus. The actual distance from each eye to the corresponding monitor was 95 cm.

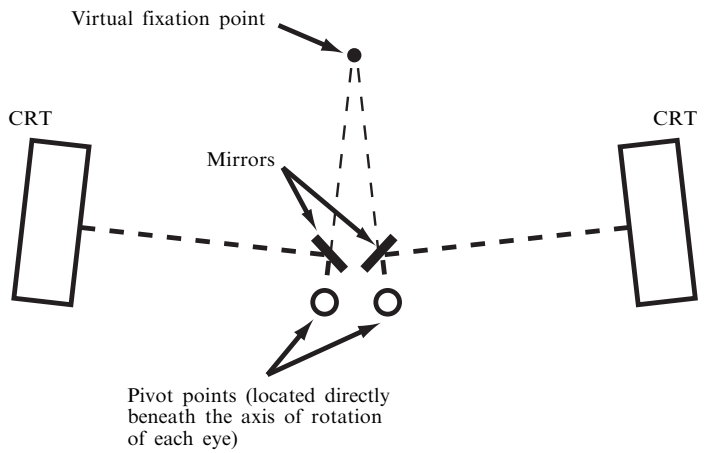

Figure 3. Schematic illustration of the stereo viewing apparatus (haploscope). CRT monitors project through mirrors to each eye individually, and are moved about pivots which coincide with the axis of rotation of each eye. The virtual fixation point is therefore specified by vergence angle, which was adjusted properly for each observer's IOD and viewing distance.

The only cue to simulated distance was the vergence of the eyes, which was directly manipulated by physically moving the monitors on their swing arms. Since the monitors and mirrors pivot rigidly about the axis of rotation of the eyes, the retinal images always remain the same for all positions of the two armatures. Thus, changes in eye position can be dissociated from changes in retinal images.

3.1.3 *Stimuli.* Elliptical hemicylinders were simulated with random-dot stereograms on each CRT screen. Cylinders consisted of 200 dots; the 2-D arrangement of dots was randomly defined with each stimulus presentation. The dots were projected on the two eyes such that horizontal disparities specified a protruding elliptical cross-section of varying depth, parallel to the line of sight. Cylinders were oriented horizontally and rotated 20° (up 10°, down 10°) about the horizontal axis perpendicular to the line of sight.

Cylinder height and width were constant, while disparity and motion information specified a cross-section (Δz) that was less deep, more deep, or equal to the height of the cylinder. Vertical subtense was sufficiently small, such that vertical disparities were not a reliable cue to depth (Cumming et al 1991; Rogers and Bradshaw 1993).

3.1.4 *Procedure.* Observers' IODs were precisely measured prior to starting the experiment. The haploscope was adjusted for each observer's IOD and the proper vergence angle for simulating 100 cm viewing distance. We ran a two-alternative forced-choice ACC task (as in Johnston 1991): on each trial, observers decided if a cylinder was more or less elongated in depth than an apparently circular cylinder.

Five different values of cylinder elongation (Δz) were presented in random order to the observer 40 times at a simulated 100 cm distance. Observers viewed the stimuli for as much time as they required to make their decisions. A response automatically triggered the next stimulus presentation. Feedback was not provided.

In the calibration stage, observers were presented with a large range of cylinder depths in order to estimate the proper range of elongation values. Over the first five values of elongation, stimuli ranged from 0.5 to 1.5 times the height-to-depth ratio of a circular cylinder. Each elongation value was presented 10 times in random order, for a total of 50 trials. The height-to-depth ratio at the 10% and 90% points of the resulting psychometric function were then used to define a smaller range of stimulus values in the experimental stage. This was done separately for stereo-only and motion-only conditions.

With a smaller range determined from the calibration stage (eg 0.8 to 1.2 times the depth-to-height of a circular cylinder), in the experimental stage the point of subjective equality could be better determined from the data. Observers first viewed stereo-only stimuli, for a total of 40 presentations of each cylinder elongation (200 trials). This was broken into four blocks of 50 trials, as in the calibration stage procedure. Next, the dominant eye was determined for each observer, and then used monocularly for the second condition: motion-only. Observers did not wear an eye patch; the CRT corresponding to the non-dominant eye was blank. The procedure was identical to the stereo-only experimental condition.

In the third and final stage, stereo-only and motion-only stimuli were randomly interleaved with stereo-motion stimuli. Stereo information always specified a depth congruent with motion information in the stereo-motion stimuli. Observers viewed 40 presentations of each cylinder elongation for all three conditions (600 trials). This was broken into four blocks of 150 trials (10 presentations of each cylinder elongation over the three conditions).

3.2 *Results and discussion*

The psychometric functions for all observers in each condition are reported in figure 4. The predictions of the IC model can be formulated by implementing the procedure discussed in section 2. The predictions of the MWF model, conversely, require the estimation of the weights associated to each separate cue, discussed below.

3.2.1 *Points of subjective equality and the IC model.* In section 2, we described how the IC model provides a 3-D Euclidean depth interpretation to the combined-cue stimulus on the basis of ρ_i (the score on PC_1) and $\boldsymbol{e}_1$ (the first eigenvector of the disparity–velocity covariance matrix).

The amount of simulated depth z_0 necessary for the IC model to predict a regular cylinder ($\hat{z}_0 = 25$ mm) can be compared with the PSEs obtained in the combined-cues condition (ie the amount of simulated depth that, in the psychophysical experiment, gave rise to an ACC). For observer ABB, for example, by using the values of $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$ estimated from the single-cue conditions, the IC model needs a simulated depth $z_0 = 18.21$ mm for producing $\hat{z}_0 = 25$ mm. In other words, according to the model, an ACC is the most likely interpretation of the disparity–velocity field produced by a cylinder with $z_0 = 18.21$ mm. This prediction is in very good agreement with the observed PSE: for observer ABB, in fact, the ACC was perceived in the combined-cues condition for a simulated depth $z_0 = 18.02$ mm. In figure 5, the results from subject ABB are reproduced along with each model's prediction for the combined-cues ACC.

The procedure described above was repeated for each observer. The results of all six observers are shown in figure 6. Note the very good agreement between the predictions and the data. It is important to remember that the simulations were carried out with no free parameters, once $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$ were chosen for the single-cue conditions.

It should be pointed out that this is not the first investigation which found judged depth was greater with combined stereo and motion than with either cue alone. For example, Tittle and Braunstein (1993) found that adding rotation to a binocularly specified random-dot cylinder increased judged depth. Importantly, they invoked the process of cue promotion to explain their result; such increases are directly predicted from the IC model.

3.2.2 *Points of subjective equality and the MWF model.* The PSEs predicted by the MWF model in the combined-cues condition were computed by a weighted average of

the PSE estimates in the single-cue conditions. If the variances estimated from the psychometric functions of the stereo-only and motion-only conditions are σ_{zd}^2 and σ_{zv}^2, respectively, then the predicted depth for the combined-cues condition is

$$z_c = \frac{\sigma_{zv}^2}{\sigma_{zd}^2 + \sigma_{zv}^2} z_d + \frac{\sigma_{zd}^2}{\sigma_{zd}^2 + \sigma_{zv}^2} z_v \,, \tag{13}$$

where z_d and z_v are the PSEs for the stereo-only and motion-only conditions, respectively.

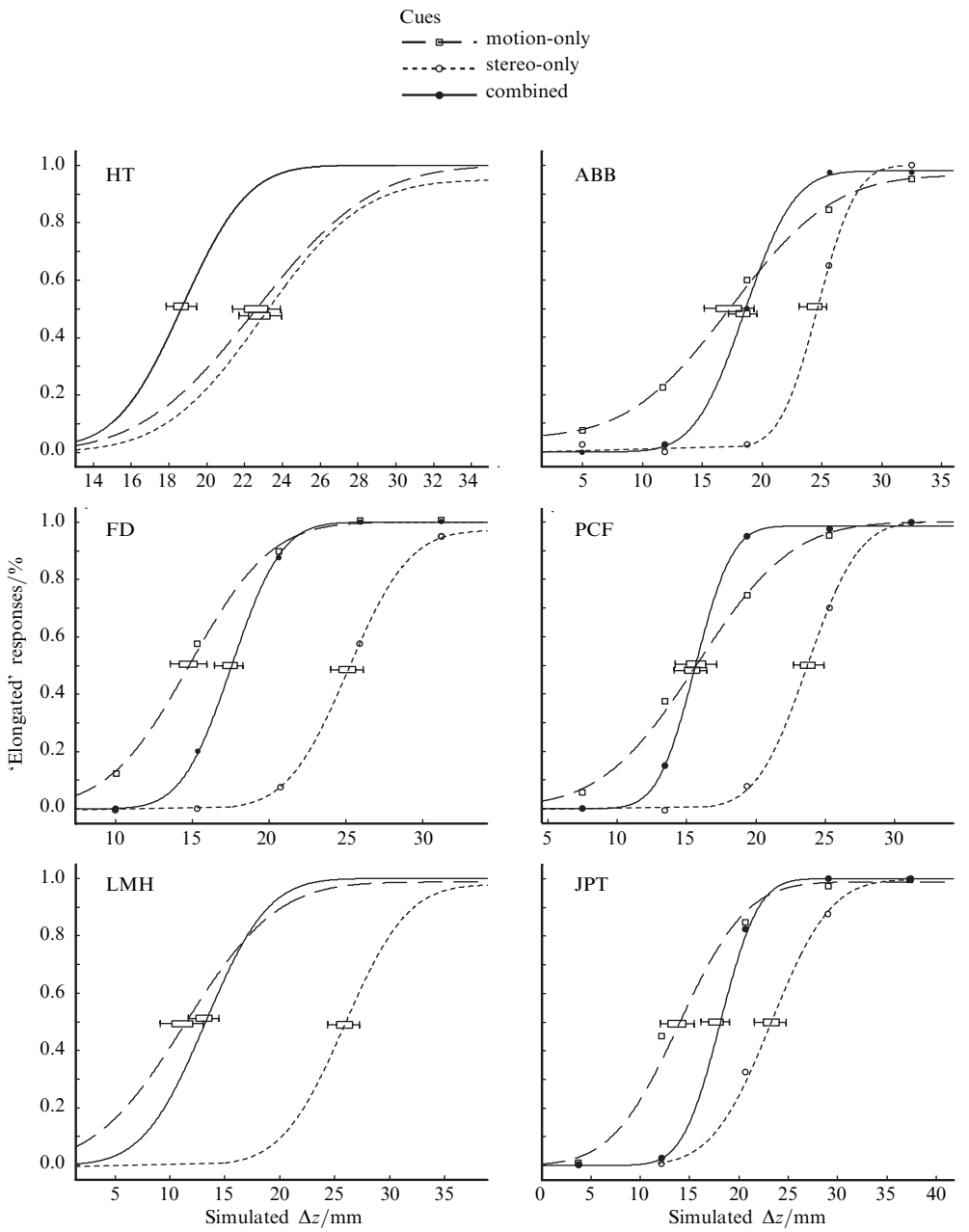

Figure 4. Psychometric functions for six observers in each condition. The 50% level represents the Δz corresponding to a perceived ACC.

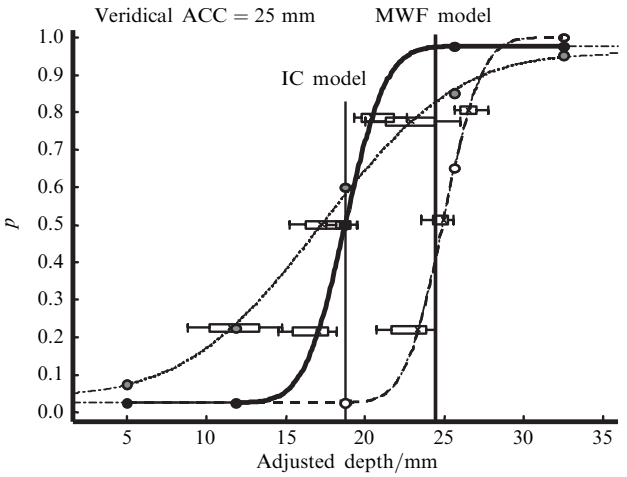

Figure 5. Subject ABB's results reproduced from figure 4. Vertical lines represent the MWF and IC model predictions for the combined-cues condition. The IC model prediction falls directly on the PSE for the combined condition (solid line), while the MWF prediction is closer to the PSE for the stereo-only condition (dashed line).

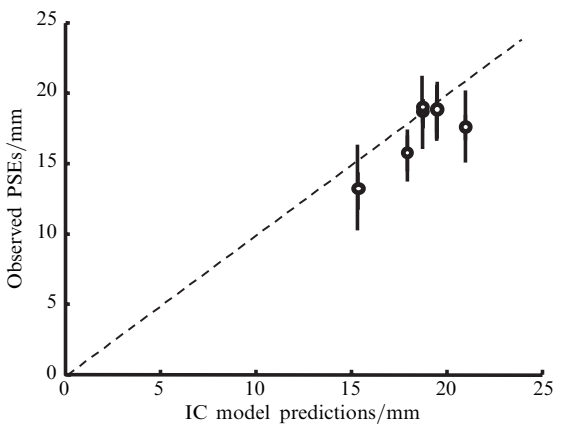

Figure 6. IC predictions plotted against the data. Points falling on the dashed line indicate a perfect fit. Error bars represent the 95% CI.

Figure 7 shows the results for observer ABB, together with the predictions of the IC model and MWF model. Note that, for this particular observer (likewise for most of the other observers in this experiment), the depth estimate was much more reliable in the stereo-only condition than in the motion-only condition. In the stereo-only condition, moreover, the PSE was closer to the veridical value than in the motion-only condition. In spite of these facts, in the combined-cues condition the PSE of observer ABB is closer to the PSE of the motion-only condition. This result is inconsistent with the MWF model, since we should expect a larger weight for the (more reliable) stereo signal. In fact, the predictions of the MWF model for the combined-cues condition almost coincide with the PSE in the stereo-only condition (see figure 7).

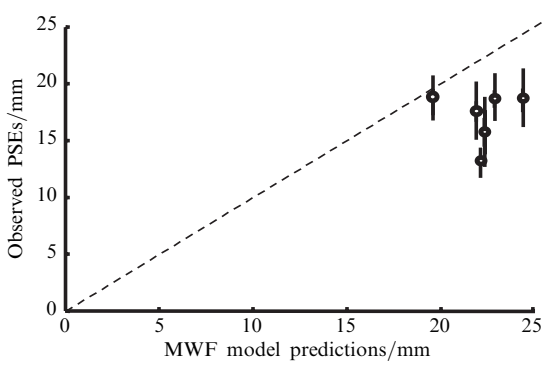

Figure 7. MWF predictions plotted against the data. Error bars represent the 95% CI.

Figure 7 shows the predicted ACC depths for the MWF model, together with the observed ACC depths for all six observers. Clearly the predictions of the IC model shown in figure 7 are in much closer agreement with the observers' settings than the predictions of the MWF model.

3.2.3 *Validity of single-cue measurements for two-cue condition.* One possible criticism that could be raised by proponents of the MWF model is that the apparently poor fit of figure 7 is due to an imprecise estimate of the weights of the motion and stereo signals in the single-cue conditions. According to this criticism, the standard deviation observed in the motion-only condition may be high (low reliability) because of unmodelled depth cues affecting responses (eg cue to flatness). Moreover, the motion and stereo signals may promote each other so that the reliability of the motion cue may increase when both cues are present, relative to when motion is shown alone.

If observers failed to ignore conflicting cues such as the blur gradient, accommodation, and the phosphor grid of the CRTs (see Hillis et al 2004), then the variances that we measured would be higher than the true variances associated with disparity and motion. To determine whether this was the case, we computed the variance of the combined-cues condition according to equation (2), and compared it with the empirical variance of the observers' judgments. From figure 8 we can see that the observers' accuracy in the combined-cues condition was very close to optimal [according to equation (23)]. We can therefore rule out the alternative explanation according to which the poor fit of the MWF model to the PSEs is due to the presence of unmodelled depth cues which may have biased our estimates of the stereo and motion weights.

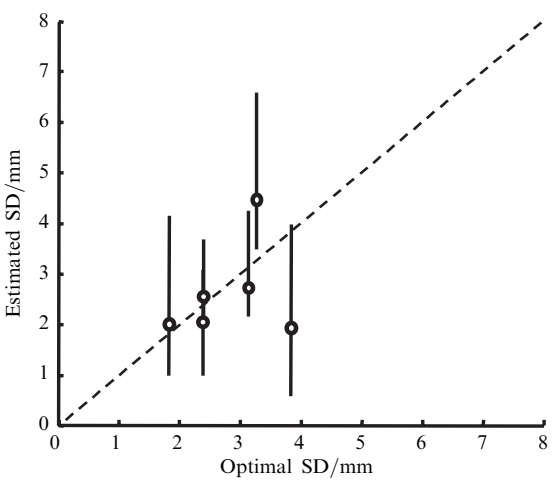

Figure 8. Standard deviations predicted by the MWF optimal integration rule [equation (13)] plotted against the standard deviation of each observer. The close relationship indicates that unmodelled depth cues did not bias estimates of cue weights.

4 General discussion

The MWF model deals with the problem of 3-D reconstruction by first separating each source of depth information (such as disparity, motion, texture, shading) and then putting back together the Euclidean estimates computed from each source separately. This approach overlooks two fundamental problems: (i) it is not obvious how cue promotion can guarantee a veridical Euclidean solution for each separate depth cue—in fact, perceptual performance is not necessarily veridical even for real objects in fully illuminated natural environments (eg Bradshaw et al 2000; Cuijpers et al 2000; Hecht et al 1999; Koenderink et al 2002; Loomis et al 1992; Loomis and Philbeck 1999; Norman et al 2000); and (ii) the process of dynamic cue re-weighting (postulating that different weights are assigned to the same cue, depending on the other cues available in the scene) is unspecified. Thus the problem of how different 3-D estimates may be combined is left unsolved.

Here we propose a different approach, where the covariation among the different goals is used to define an underlying 1-D manifold which directly specifies the affine structure of the distal 3-D shapes. The advantage of such a description is that the correlation between the points belonging to this underlying dimension (the IC line) and the distal 3-D property will be higher than the correlation between each single signal and the distal 3-D property.

Once the affine structure of the 3-D shapes has been recovered, a 3-D Euclidean interpretation can be provided through a maximum-likelihood procedure [see equation (11)]. Such a Euclidean interpretation is not generally unbiased, but rather aims at reproducing both veridical interpretations and biases in human performance.

4.1 *IC model and optimal combination*

Despite the theoretical limitations of the MWF model, there are many empirical investigations that support this approach. The vast majority of these results confirm the prediction of optimal cue combination expressed by equation (2). In this section we show that the IC model described above produces the same predictions as the MWF model, insofar as the accuracy of the observers' judgments is concerned.

To determine whether the visual system combines two cues in a statistically optimal fashion (according to the MWF definition of optimality), the discrimination accuracy of an observer can be experimentally measured for each cue shown in isolation, and also when they are shown simultaneously. In the case of stereo and motion cues, for example, the experimenter may choose a certain baseline value of simulated depth, z_0, and measure the discrimination accuracy for this depth value in three conditions: stereo-only, motion-only, and stereo-plus-motion. The goal of the experiment is to estimate the standard deviations σ_{z_s}, σ_{z_v}, and σ_{z_c}. If the MWF model holds, then σ_{z_c} must be related to σ_{z_s}, σ_{z_v} according to equation (2).

The discrimination accuracy in each single-cue condition depends on the measurement noise of the disparity and velocity values, and on the viewing parameters μ and ω. According to equation (14), if the noise of the disparity measurements has a standard deviation σ_d, then $\sigma_{z_s} = \sigma_d/\mu$. Clearly, the further away an object is (small μ), the noisier are the depth estimates. Similarly, if the standard deviation of the noise of the velocity measurements is σ_v, then $\sigma_{z_v} = \sigma_v/\omega$. For the MWF model [see equation (2)]:

$$\sigma_{z_c}^2 = \frac{\sigma_{z_s}^2 \sigma_{z_v}^2}{\sigma_{z_s}^2 + \sigma_{z_v}^2}, \tag{14}$$

and, therefore:

$$\sigma_{z_c}^2 = \frac{(\sigma_d/\mu)^2 (\sigma_v/\omega)^2}{(\sigma_d/\mu)^2 + (\sigma_v/\omega)^2}. \tag{15}$$

To show that the predictions of the IC model for the discrimination accuracy are identical to those of the MWF model, let us consider again the equations for standardized velocities and disparities [equations (6) and (7)]. From trial to trial, the same simulated depth z_0 will produce a distribution of scaled disparities and velocities (with $\bar{d}_0$ and $\bar{v}_0$ means, and unit standard deviations). In the combined-cue condition, therefore, the noise can be described as a bivariate Gaussian random variable. Since we assume independent errors scaled by their standard deviations, the constant-probability density contours of such a distribution will be circular.

In each trial, we can compute a $\hat{\rho}_0$ score on PC_1 for the depth value z_0; repeated measurements yield a distribution of such scores. The $\hat{\rho}_0$ score is therefore a random variable with $\rho_0 = z_0(\bar{\mu}^2 + \bar{\omega}^2)^{1/2}$ mean and unit standard deviation. Of interest here is how the observer's accuracy relates to the z-values. Therefore, we must consider the distribution of the depth estimates z_0' for each distal depth z_0. Since the $\hat{\rho}_0$ scores have

unit variance, and the relationship between z_0' and ρ_0 is given by

$$z_0' = \frac{\rho_0}{(\bar{\mu}^2 + \bar{\omega}^2)^{1/2}} ,$$

it follows that the variance of inferred depth values (z_0') is

$$\sigma^2_{z_0'} = \frac{1}{\bar{\mu}^2 + \bar{\omega}^2} , \tag{16}$$

since $\bar{\omega} = \omega/\sigma_v$ and $\bar{\mu} = \mu/\sigma_d$. Note that the previous equation is identical to equation (14).

5 Conclusions

Why would the visual system adopt a strategy that may lead to biased estimations rather than following an inverse-optics approach for each signal separately that guarantees a veridical solution? There may be many reasons for that. First, the parameters of vergence angle and angular velocity (required for an unbiased Euclidean estimate) may not be easily estimated [see equations (6) and (7)]. Second, even if all the necessary parameters are available, the signals necessary for an unbiased Euclidean interpretation may not be measurable by the visual system at the required level of precision (see the discussion on second-order temporal information in the perceptual interpretation of the optic flow—eg Domini and Caudek 2003).

In the terminology of Clark and Yuille (1990), our approach can be characterized as a strong-fusion model of 3-D shape recovery from multiple cues. So far, no model of this kind has been proposed in the literature, since such a model aiming at an unbiased 3-D estimate is certainly mathematically intractable. Our contribution has been to provide a first step in this direction, by recognizing that a psychophysically plausible model does not necessarily require unbiased estimates.

References

Atkins J E, Fiser J, Jacobs R A, 2001 "Experience-dependent visual cue integration based on consistencies between visual and haptic percepts" *Vision Research* **41** 449 – 461

Backus B T, Banks M S, Ee R van, Crowell J A, 1999 "Horizontal and vertical disparity, eye position, and stereoscopic slant perception" *Vision Research* **39** 1143 – 1170

Banks M S, Backus B T, Banks R S, 2002 "Is vertical disparity used to determine azimuth?" *Vision Research* **42** 801 – 807

Bennett B, Hoffman D, Nicola J, Prakash C, 1989 "Structure from two orthographic views of rigid motion" *Journal of the Optical Society of America A* **6** 1052 – 1069

Bingham G P, Crowell J A, Todd J T, 2004 "Distortions of distance and shape are not produced by a single continuous transformation of reach space" *Perception & Psychophysics* **66** 152 – 169

Bradshaw M F, Parton A D, Glennerster A, 2000 "The task-dependent use of binocular disparity and motion parallax information" *Vision Research* **40** 3725 – 3734

Clark J, Yuille A, 1990 *Data Fusion for Sensory Information Processing Systems* (Boston, MA: Kluwer)

Cuijpers R H, Kappers A M L, Koenderink J J, 2000 "Investigation of visual space using an exocentric pointing task" *Perception & Psychophysics* **62** 1556 – 1571

Cumming B G, Johnston E B, Parker A J, 1991 "Vertical disparities and perception of 3-dimensional shape" *Nature* **349** 411 – 413

Dijkstra T M H, Cornilleau-Peres V, Gielen C C A M, Droulez J, 1995 "Perception of 3D shape from ego- and object-motion: comparison between small and large field stimuli" *Vision Research* **35** 453 – 462

Dijkstra T M H, Snoeren P R, Gielen C C A M, 1994 "Extraction of three-dimensional shape from optic flow: a geometric approach" *Journal of the Optical Society of America A* **11** 2184 – 2196

Domini F, Caudek C, 1999 "Perceiving surface slant from deformation of optic flow" *Journal of Experimental Psychology: Human Perception and Performance* **25** 426 – 444

Domini F, Caudek C, 2003 "3-D structure perceived from dynamic information: a new theory" *Trends in Cognitive Sciences* **7** 444 – 449

Domini F, Caudek C, Proffitt D R, 1997 "Misperceptions of angular velocities influence the perception of rigidity in the Kinetic Depth Effect" *Journal of Experimental Psychology: Human Perception and Performance* **23** 1111 – 1129
Duda R O, Hart P E, Stork D G, 2000 *Pattern Classification* 2nd edition (New York: John Wiley and Sons)
Ee R van, Banks M S, Backus B T, 1999 "Perceived visual direction near an occluder" *Vision Research* **24** 4085 – 4097
Faugeras O, 1993 *Three-Dimensional Computer Vision* (Cambridge, MA: MIT Press)
Forsyth D, Ponce J, 2003 *Computer Vision: A Modern Approach* (Upper Saddle River, NJ: Prentice Hall)
Garding J, Porrill J, Mayhew J E, Frisby J P, 1995 "Stereopsis, vertical disparity and relief transformations" *Vision Research* **35** 703 – 722
Hecht H, Doorn A van, Koenderink J J, 1999 "Compression of visual space in natural scenes and in their photographic counterparts" *Perception & Psychophysics* **61** 1269 – 1286
Hildreth E C, 1984 "Computations underlying the measurement of visual motion" *Artificial Intelligence* **23** 309 – 354
Hillis J M, Watt S J, Landy M S, Banks M S, 2004 "Slant from texture and disparity cues: Optimal cue combination" *Journal of Vision* **4** 967 – 992
Hoffman D D, 1982 "Inferring local surface orientation from motion fields" *Journal of the Optical Society of America* **72** 888 – 892
Hogervorst M A, Eagle R A, 1998 "Biases in three-dimensional structure-from-motion arise from noise in the early visual system" *Proceedings of the Royal Society, Section B* **265** 1587 – 1593
Horn B K P, 1986 *Robot Vision* (Cambridge, MA: MIT Press)
Johnston E B, 1991 "Systematic distortions of shape from stereopsis" *Vision Research* **31** 1351 – 1360
Johnston E B, Cumming B G, Landy M S, 1994 "Integration of stereopsis and motion shape cues" *Vision Research* **34** 2259 – 2275
Koenderink J J, Doorn A J van, 1976 "Local structure of movement parallax of the plane" *Journal of the Optical Society of America A* **66** 717 – 723
Koenderink J J, Doorn A J van, 1991 "Affine structure from motion" *Journal of the Optical Society of America A* **8** 377 – 385
Koenderink J J, Doorn A J van, Kappers A M L, Todd J T, 2002 "Pappus in optical space" *Perception & Psychophysics* **64** 380 – 391
Landy M S, Brenner E, 2001 "Motion-disparity interaction and the scaling of stereoscopic disparity", in *Vision and Attention* Eds L R Harris, M R M Jenkin (New York: Springer) pp 129 – 151
Landy M S, Maloney L T, Johnston E B, Young M J, 1995 "Measurement and modeling of depth cue combination: in defense of weak fusion" *Vision Research* **35** 389 – 412
Lappin J S, Craft W D, 2000 "Foundations of spatial vision: From retinal images to perceived shapes" *Psychological Review* **107** 6 – 38
Longuet-Higgins H C, Prazdny K, 1980 "The interpretation of a moving retinal image" *Proceedings of the Royal Society of London, Series B* **208** 385 – 397
Loomis J M, Dasilva J A, Fujita N, Fukusima S S, 1992 "Visual space-perception and visually directed action" *Journal of Experimental Psychology: Human Perception and Performance* **18** 906 – 921
Loomis J M, Philbeck J W, 1999 "Is the anisotropy of perceived 3-D shape invariant across scale?" *Perception & Psychophysics* **61** 397 – 402
Mayhew J E W, Longuet-Higgins H C, 1982 "A computational model of binocular depth perception" *Nature* **297** 376 – 378
Mon-Williams M, Tresilian J R, Roberts A, 2000 "Vergence provides veridical depth perception from horizontal retinal image disparities" *Experimental Brain Research* **133** 407 – 413
Norman J F, Lappin J S, Norman H F, 2000 "The perception of length on curved and flat surfaces" *Perception & Psychophysics* **62** 1133 – 1145
Oruç I, Maloney L T, Landy M S, 2003 "Weighted linear cue combination with possibly correlated error" *Vision Research* **43** 2451 – 2468
Richards W, 1985 "Correlation between stereo ability and the recovery of structure-from-motion" *American Journal of Optometry A: Optics, Image Science and Vision* **2** 343 – 349
Rogers B J, Bradshaw M F, 1993 "Vertical disparities, differential perspective and binocular stereopsis" *Nature* **361** 253 – 255
Rogers B J, Bradshaw M F, 1995 "Disparity scaling and the perception of frontoparallel surfaces" *Perception* **24** 155 – 179
Tittle J S, Braunstein M L, 1993 "Recovery of 3D shape from binocular disparity and structure from motion" *Perception & Psychophysics* **54** 157 – 169

Tittle J S, Todd J T, Perotti V J, Norman J F, 1995 "The systematic distortion of perceived 3D structure from motion and binocular stereopsis" *Journal of Experimental Psychology: Human Perception and Performance* **21** 663 – 678

Ullman S, 1979 "The interpretation of structure from motion" *Proceedings of the Royal Society of London, Series B* **203** 405 – 426

Wexler M, Panerai F, Lamouret I, Droulez J, 2001 "Self-motion and the perception of stationary objects" *Nature* **409** 85 – 88

Young M J, Landy M S, Maloney L T, 1993 "A perturbation analysis of depth perception from combinations of texture and motion cues" *Vision Research* **33** 2685 – 2696

Appendix A

In equation (11), z_i is the simulated depth of the ith point on the cylinder, and $\hat{z}_i$ is the depth interpretation provided by the model to ρ_i. In the following, we will call z_0 the simulated depth of the point on the cylinder's surface corresponding to the maximum front-to-back depth extent ($z_0 = 25$ mm for a regular cylinder). In the simulation, we seek the value of z_0 for which the interpretation of the model, $\hat{z}_0$, corresponds to a cylinder with a regular cross-section. The value of z_0, in turn, defines the model's prediction for the PSE in the combined-cues condition.

In order to find z_0, we must estimate the likelihood $p(\rho_0, \boldsymbol{e}_1 | z_0)$. Since this distribution is calculated by integrating over the unknown parameters $\bar{\mu}$ and $\bar{\omega}$ [see equation (16)], its shape depends on the priors $p(\bar{\mu})$ and $p(\bar{\omega})$. Here, we assume that these distributions are uniform and limited by $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$. To compute the integral of equation (12), therefore, these two parameters are needed. In the simulation, the values of $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$ were empirically chosen so that, in each single-cue condition, the likelihood function $p(\rho_0, \boldsymbol{e}_1 | z_0)$ peaked at $z_0 = 25$ mm.

Once $\bar{\mu}_{\max}$ and $\bar{\omega}_{\max}$ are found from the single-cue conditions, we can then predict performance in the combined-cues condition through the following steps:

(1) the velocity (v) and disparity (d) values are computed for 200 points randomly placed on a hemicylinder;

(2) the v and d values are scaled by the standard deviation of the psychometric functions computed in the single-cue condition of one observer;

(3) random noise $\sim N$ (0, 1) is added to the scaled v and d values, so as to simulate the internal noise of the system (see figure 3);

(4) a PC analysis is performed to estimate the first eigenvector $\boldsymbol{e}_1$, and the ρ_0 score associated to z_0;

(5) the simulation is repeated 500 times in order to estimate the distributions of $\boldsymbol{e}_1$ and ρ_0, that is, $p(\boldsymbol{e}_1 | \bar{\mu}, \bar{\omega})$ and $p(\rho_0 | z, \bar{\mu}, \bar{\omega})$;

(6) finally, to compute $p(\rho_0, \boldsymbol{e}_1 | z)$, we calculate the integral equation (16) by Monte Carlo integration. An example of the resulting likelihood function is shown in figure 6.

Appendix B

The analysis of signals that we described in the introduction can also be performed by the well-known method known as total least squares (TLS). According to this method, the data are assumed to be the pairs (v_i, d_i) and each is the value of a normally distributed random variable $v_i \approx N(\xi_i, \sigma_v)$, respectively $d_i \approx N(\beta\xi_i, \sigma_d)$. The goal of TLS is to estimate the parameters β, σ_d, σ_v, ξ_i. The MLEs ξ_i are a multiple of the depth values z_i. The idea in TLS is that projections of the data to the desired line are perpendicular to the line, not vertical as in ordinary least squares. This procedure is equivalent to a PCA, at least when the error variances are assumed to be the same.

The main problem is that the likelihood grows without bound as $\sigma_v \to 0$ and $\xi_i = v_i$, and, therefore, there are no unrestricted MLEs. This problem can be overcome

if it is assumed that $\sigma_v^2 = \lambda\sigma_d^2$ where the ratio λ is known. In this case an estimate of the parameter β is given by:

$$\hat{\beta} = \frac{1 - (S_{vv} - \lambda S_{dd}) + [(S_{vv} - \lambda S_{dd})^2 + 4\lambda S_{vd}^2]^{1/2}}{2\lambda S_{vd}}, \tag{B1}$$

where $S_{vd}^2 = \sum_{i=1}^{n} (v_i - \bar{v})(d_i - \bar{d})$ is the sample covariance and S_{dd}^2 and S_{vv}^2 are the sample sums of squares.

In this case the estimated depth vector is:

$$\hat{\xi}_i = \frac{\lambda\hat{\beta}d_i + v_i}{1 + \lambda\hat{\beta}^2}. \tag{B2}$$

This method is more general than the method described in section 1 since it only requires knowledge of the ratio λ between the variances of the measurement noise of the disparity and velocity signals. However, the depth estimate is known only up to an unknown scaling factor.

© 2008 a Pion publication

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 36 2007

www.perceptionweb.com

Conditions of use. This article may be downloaded from the Perception website for personal research by members of subscribing organisations. Authors are entitled to distribute their own article (in printed form or by e-mail) to up to 50 people. This PDF may not be placed on any website (or other online distribution system) without permission of the publisher.