Pergamon

0042-6989(93)E0065-F

Vision Res. Vol. 34, No. 17, pp. 2259–2275, 1994

0042-6989/94 $7.00 + 0.00

Integration of Stereopsis and Motion Shape Cues

ELIZABETH B. JOHNSTON,* BRUCE G. CUMMING,† MICHAEL S. LANDY‡

Received 16 March 1992; in revised form 2 December 1993

A global shape judgement task was used to investigate the combination of stereopsis and kinetic depth. With both cues present, there were no distortions of shape perception, even under conditions where either cue alone did show such distortions. We suggest that the addition of motion information overcomes the stereo distance scaling problem. However, when incongruent combinations of disparity and motion were used, the results did not match predictions of a number of combination theories. These data could be described by a model which used weighted linear combination *after* correctly scaling disparities for viewing distance. When the motion cue was weakened by presenting only two frames of each motion sequence, stereo was weighted more heavily.

Stereopsis Structure-from-motion Three-dimensional shape perception Integration of depth cues

INTRODUCTION

When exploring the integration of depth cues experimentally there are a wide variety of cues to be considered. Three dimensional shape can be specified by stereopsis, relative motion, and a number of pictorial cues including perspective, shading and textural variation. The focus of this paper is the means by which two strong depth cues—stereopsis and relative motion—interact. The main question addressed is whether it is appropriate to consider stereo and motion as independent depth modules which are linearly combined to yield veridical depth estimates, or whether some more complex nonlinear interaction takes place.

Rationale for studying stereo-motion combination

Stereopsis and structure-from-motion are considered powerful cues in isolation, meaning that for most observers both sources of information independently provide compelling sensations of depth. von Helmholtz (1910) likened the depth percept from motion parallax to "a good stereoscopic view". In addition, stereopsis affords exquisitely precise depth judgements, as evidenced by stereoacuity thresholds of a few seconds of arc under the best conditions (Berry, 1948). However, stereoacuity does decline exponentially with distance from the fixation plane (Blakemore, 1970). Thus, obtaining high quality depth information from stereopsis may be restricted to fixated targets of small three-dimensional size (McKee, Levi & Bowne, 1990). Motion can also provide fine depth information. Speed discrimination thresholds, which provide a basic limitation on the processing of structure-from-motion, are less than 5% at speeds greater than 3 deg/sec (McKee, 1981). There are few data on the precision of structure-from-motion judgements (but see Rogers & Graham, 1982; Todd & Norman, 1991).

Both stereopsis and motion parallax can yield an *absolute* measure of depth at each point in the scene where there is a surface marking (e.g. in meters from the observer), given some additional information about the observer's position. In the case of motion the additional information required is knowledge of egomotion and eye rotation, which could be obtained from nonvisual sources. For stereopsis to provide absolute depth values the distance from the observer to the fixation point must be known as well as the observer's interocular separation. In principle, combining the information from stereo and motion allows three-dimensional shape to be extracted without the need for additional information about viewing distance or egomotion (some possible schemes are discussed below).

A number of psychophysical observations point to close links between the motion and stereopsis processing systems. Rogers and Graham (1982) documented extensive similarities between depth from motion parallax and stereopsis. The shape of the sensitivity functions for depth modulation as a function of spatial frequency are highly similar for stereopsis and motion parallax (Rogers & Graham, 1982). However, for some of their observers absolute sensitivity to binocular disparity was

*Department of Psychology, Sarah Lawrence College, 1 Mead Way, Bronxville, NY 10708, U.S.A.
†University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, England.
‡Department of Psychology and Center for Neural Science, New York University, 6 Washington Place, 8th Floor, New York, NY 10003, U.S.A.

greater than that for relative motion. Successive contrast effects were found to be similar in both domains (Graham & Rogers, 1982). Rogers and Graham proceeded to demonstrate that not only were the two sources of information very similar but there were also interactions in their processing. An ambiguous percept of a surface specified by either stereopsis or motion parallax could be biased by prior exposure to an unambiguously perceived surface specified by the other source of information. Nawrot and Blake (1991) recently demonstrated the same biasing effect of stereograms upon ambiguous kinetic depth effect (KDE) stimuli.

Physiological studies also reveal considerable overlap in the processing of stereo and motion. A number of cells, both in striate and extrastriate cortex, are sensitive both to disparity and to motion. The processing of motion-in-depth by these cells has been extensively investigated (Cynader & Regan, 1978; Poggio & Talbot, 1981), but these studies do not rule out other possible interactions between motion and stereo. In areas MT and MST, where motion selective cells predominate, recent studies show that cells vary in their sensitivity to different components of the flow field such as rotation and dilation (Saito, Yukie, Taneka, Hikosaka, Fukada & Iwai, 1986; Taneka, Hikosaka, Saito, Yukie, Fukada & Iwai, 1986; Andersen, Graziano, Snowden & Treue, 1990; Orban, Lagae, Verri, Raiguel, Xiao, Maes & Torre, 1992). This suggests that they may play an important part in extracting structure-from-motion. Roy, Komatsu and Wurtz (1992) recorded from neurons in area MST which altered their directional selectivity depending upon the sign of binocular disparity present. They suggested that these neurons contribute to a signal about the direction of self-motion, but such cells could participate in other interactions between stereo and motion (Judge, 1990).

In short, there are a number of reasons to focus on the combination of stereo and motion. Each provides a strong depth cue in isolation, yet combining them solves problems arising from the incompleteness of the individual cues. In addition, a number of psychophysical and physiological interactions have already been demonstrated.

Previous work on stereo-motion combination

A major issue in the literature on depth cue combination is whether the interaction between cues can be described as weighted linear combination (Dosher, Sperling & Wurst, 1986; Bruno & Cutting, 1988; Maloney & Landy, 1989; Landy, Maloney & Young, 1991a; Bülthoff, 1991; Johnston, Cumming & Parker, 1993). In a *weak fusion* model (Clark & Yuille, 1990; Bülthoff, 1991) the individual depth cues are regarded as modules that provide independent measures of depth. These independent estimates are then combined by weighted averaging of the depths. In contrast, a combination process is described as *strong fusion* if the cues interact prior to yielding depth estimates. The difficulty with this weak–strong dichotomy is that classifying a particular cue interaction as strong simply means that the depth cues cannot be considered entirely independent, but it does not limit the form of the interaction between cues. To further differentiate types of cue interaction a third descriptive category *modified weak fusion* was proposed by Landy, Maloney, Johnston and Young (1991b) which is a specific combination of weak and strong processes. Modified weak fusion incorporates *promotion* (Maloney & Landy, 1989)—the use of one cue to provide missing information required by another cue to yield absolute depth measures. This initial stage of promotion to complete each depth cue is followed by a linear combination of the depths specified by the promoted cues.* Modified weak fusion is a hybrid two stage model which incorporates useful features of both weak and strong fusion. In the first promotion stage the allowable interactions are specified. The second stage involves weighted averaging of the depth cues taking account of relative cue reliability. In experimental studies of stereo and motion combination little attention has been paid to the question of promotion, but the linearity of integration has been examined.

Stereo scaling problem

One of the main aims of this paper is to determine the veridicality of the depth percepts that arise from the combination of stereopsis and kinetic depth. The initial motivation for this study came from the finding that shape from stereopsis alone is not veridically perceived (Gogel, 1960; Foley, 1980; Johnston, 1991; Parker, Johnston, Mansfield & Yang, 1991). At far viewing distances depth is underestimated and at close viewing distances it is overestimated. There is some intermediate distance where depth perception is veridical which varies between subjects, averaging 80 cm. These shape distortions† are most simply explained by the subjects using a misestimate of the viewing distance which tends towards an intermediate default value. Gogel (1972) described the specific distance tendency—in the absence of good distance information visual objects will tend to be located at a default viewing distance. This leads to distortions of perceived shape: while angles subtended by fronto-parallel extents (the height and width of the shape) scale inversely with viewing distance, disparities scale approximately inversely with the *square* of the viewing distance. Consequently, when the viewing distance to a fixed disparity field is doubled, while the portrayed depth quadruples, the angular subtense of the width or height of the shape specifies only double the original extent (see Fig. 1). When distance is misestimated, the differential scaling leads to calculation of a

*Modified weak fusion also specifies other factors determining cue interactions, such as weighting according to the reliability of individual cues and robust depth estimation (see Landy *et al.*, 1991b).

†The naming conventions used here are depth = an absolute measure of distance in cm from the front to back of a stimulus surface, shape = a relative measure of the depth/height ratio of a stimulus surface (all cylindrical in the following experiments), size = an absolute measure of three-dimensional size in cm in both the height and depth dimensions (e.g. for a fixed shape).

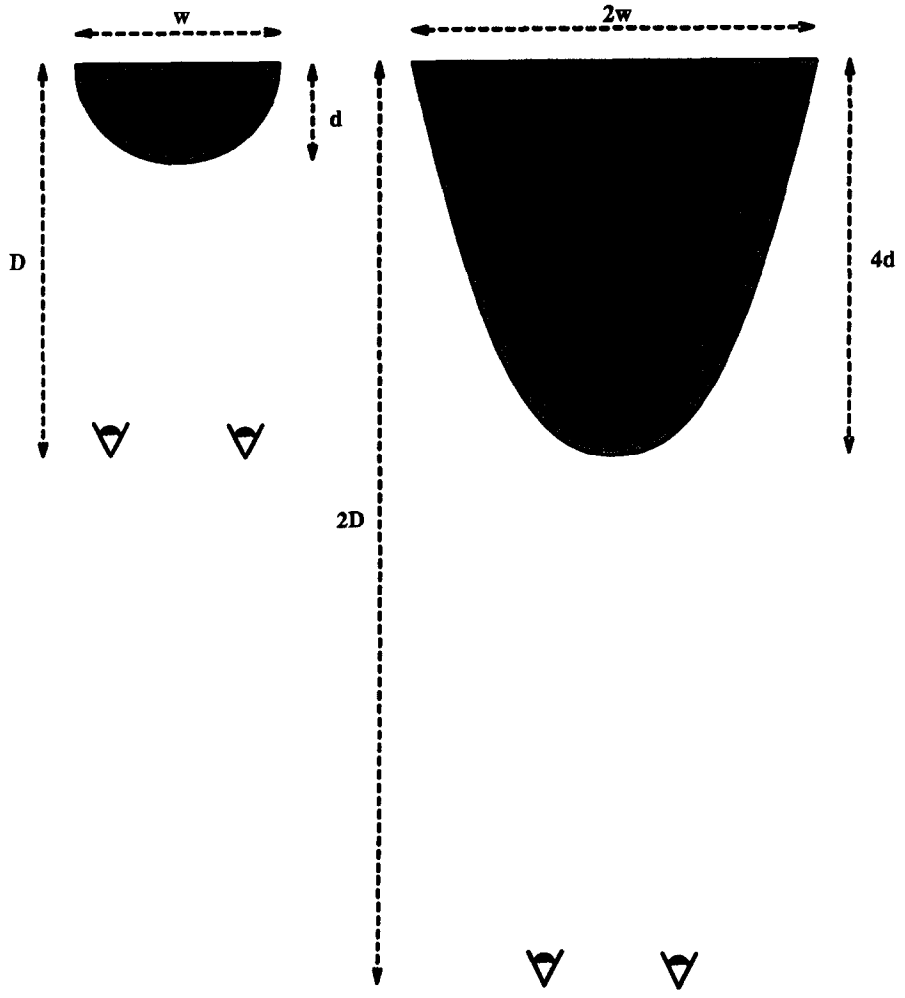

FIGURE 1. Stereo scaling problem. The filled arcs illustrate cross-sections through the cylinders depicted by the same disparity field viewed from two distances. When the viewing distance (D) is doubled for the same disparity field the depths depicted by those disparities are quadrupled. The width calculated from the visual angle subtended by the fronto-parallel extent of the shape is only doubled. Thus, there is a change in the depth/width ratio of the cylinder when the viewing distance used to scale the disparity field is changed.

different depth/height ratio than if the information had been correctly scaled. This differential scaling allows us to use a global shape judgement task to assess veridicality of three-dimensional shape reconstruction. Using this task, we explored the effects on perceived shape of various combinations of depths specified by binocular disparity and relative motion.

If a misestimate of viewing distance causes nonveridical scaling from disparities to depth, it is important to determine what information is used by human observers to specify viewing distance. Viewing distance is specified by both nonvisual and visual sources of information, in the form of the convergence angle of the eyes and the gradient of the vertical disparity field (Mayhew & Longuet-Higgins, 1982) respectively. Cumming, Johnston and Parker (1991) and Sobel and Collett (1991) discovered that vertical disparity gradients which specify different viewing distances had no effect on the perceived shape of cylindrical surfaces.* In contrast, varying the vergence angle affected depth perception. Since shape from stereopsis alone is not veridically perceived, this implies that the scaling as a result of varying vergence angle was not complete. This finding provides support for the position that the vergence cue to viewing distance is not sufficiently well calibrated to produce correct depth estimates from disparities. The veridicality of stereopsis is poor in circumstances where observers are unable to obtain accurate viewing distance estimates. Thus, stereopsis is a good instance of a cue in need of *promotion.*

Structure-from-motion provides a good candidate for this interaction because for motion the consequences of misestimating viewing distance are different from those described above for stereopsis. For polar projection, the retinal velocities available to observers to determine structure-from-motion require a measure of the viewing distance to be translated into three-dimensional distances.† The consequence of using an incorrect distance estimate to scale relative velocities is that the shape is uniformly scaled. For a shape judgement task (such as

*In later work Rogers and Bradshaw (1993) did find an effect of the distance specified by vertical disparities, but only with a large field of view (80 × 70 deg displays).

†Ullman (1979) proved that three views of four corresponding points are sufficient to determine the structure of a rigid object. This proof is for orthographic projection, so it does not consider the effect of viewing distance which is of importance here.

ours, discussed below), such uniform changes in scale would have no effect on the judgement made as the calculated depth/height ratio would not change. To be more concrete, an incorrectly, but uniformly, scaled sphere still appears spherical, regardless of changes in the overall apparent size.

Since the effects of misestimating viewing distance are quite different for stereo and motion, the combination of both cues will allow the scaling problem to be solved. Information from stereo alone is compatible with a set of different shapes, whose sizes are all known. Motion alone allows for a set of objects of different sizes, but all the same shape. These sets will only intersect at one point—an object of the correct shape and size. If only two motion frames are available, it may not be possible to calculate object shape correctly (Todd & Norman, 1991). Nonetheless, given only two pairs of frames from a *binocular* motion sequence it is still possible to calculate veridical depths (e.g. Richards, 1985). Richards' scheme measures the disparity of individual points in each frame, and uses the ratio of these disparities to extract shape information. An interesting consequence of this scheme is that if all the depth values corresponding to the image disparities were doubled as a result of overestimating the viewing distance, there would be no change in perceived shape, since the disparity ratios have not changed. In other words, in this hypothetical case the object would be perceived as larger overall and viewed from a farther distance, but the perceived shape would remain the same. Thus, even when structure-from-motion alone does not provide sufficient information, the combination of stereo and motion may lead to a percept where the shape (but not three-dimensional size) is fixed by the motion flow field, and is unchanged by rescaling the disparity field. A more formal presentation of one of these "distance scaling" stereo-motion integration schemes is given in the Appendix.

The experimental questions addressed in this paper are: does the addition of relative motion increase the veridicality of shape from stereopsis, and what form does this interaction take? Four experiments were performed to investigate these questions. In the first experiment shape perception from kinetic depth alone, stereo alone, and a congruent combination of the two cues was examined. In the second experiment incongruent combinations of relative motion and binocular disparity were used to examine the linearity of the interaction, and the relative weights assigned to the cues. In the third experiment the interaction of stereo and motion was investigated when the motion cue was weakened by only presenting two frames of the sequence to each eye. In the last experiment the interaction of stereo and two frame motion specifying inconsistent depths was examined.

METHODS

Stimulus generation

Volumetric ray-tracing. The stimuli were stereo-motion sequences of horizontally oriented cylinders which rotated about a vertical axis. They were generated using the ray-tracing and volumetric representation software described in Johnston *et al.* (1993). Cylinders are defined in a volumetric representation of texture. Perspective projections of the surfaces are produced by finding the intersection of a line which goes from the image plane to the viewpoint with the textured surface. Figure 2 illustrates the geometry used. The back plane of the volume was the image plane, thus hemicylinders protruding from the monitor were depicted. The texture was composed of randomly positioned nonoverlapping spheres. The background grey level of the block was 127, and the spheres varied in grey level between 1–117 and 137–254. The spheres varied randomly in radius from 0.17 to 0.42 cm. Figure 3 shows a stereo pair which formed one frame of a stereo movie. Three images are displayed to allow both crossed and uncrossed fusion.

Separating stereo and motion. In Expt 2, stimuli were required in which the depth specified by motion and the depth specified by stereo differed. The depth specified by motion and the depth specified by stereo were varied

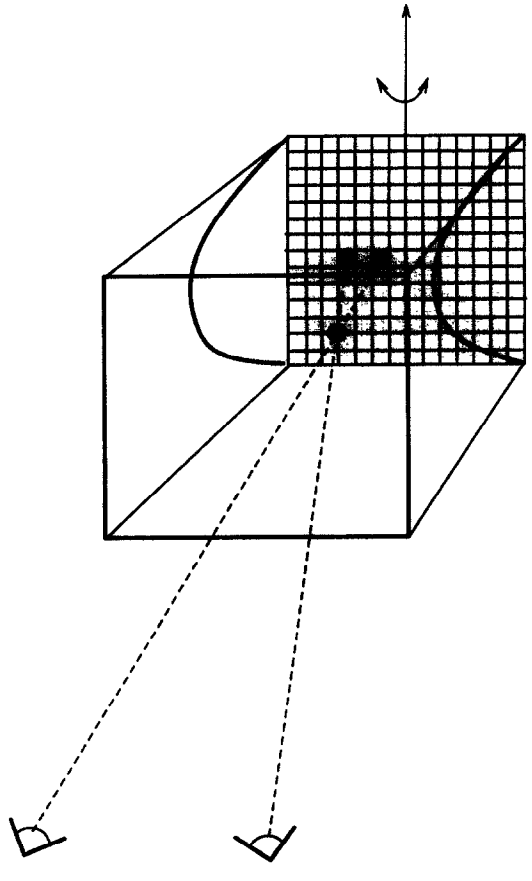

FIGURE 2. The geometry used in stimulus generation. The emboldened arcs indicate the occluding contours of a cylindrical surface. The surface is embedded in a volume filled with texture elements. The back plane of the volume is the image plane, where the pixel array is indicated by hatching. To generate each eye's image a line is traced from that eye's view through the surface to an image pixel (shown as a dashed line). That pixel is set to the colour of the surface at the point of intersection (marked with a solid circle). This process is repeated for each image pixel. The disparity between the two eyes' views can be seen from the separation of the two marked pixels on the image plane. The vertical axis of surface rotation is marked.

independently within the same stimulus by varying the interocular separation parameter supplied to the ray-tracing routine.

The logic of this manipulation is illustrated in Fig. 4(A).

Making the small angle approximation, the equation for horizontal binocular disparity (η) can be expressed:

$$\eta = \frac{Id_s}{I^2 + D^2 - Dd_s}, \quad (1)$$

where I is the observer's interocular separation, D is the viewing distance and d_s is the depth specified by this stereoscopic disparity cue. We require that the ray-tracing program produces this same value of η while interpreting a volumetric representation of a surface of depth d_m, where d_m is the depth specified by the motion cue. We do this by supplying the program with a fallacious value of the interocular separation i satisfying:

$$\frac{Id_s}{I^2 + D^2 - Dd_s} = \frac{id_m}{i^2 + D^2 - Dd_m}. \quad (2)$$

This equation is solved for the unknown i, and this value of i is supplied to the ray-tracing routine to produce a surface where the depth specified by motion is d_m but the depth specified by disparity is d_s (as viewed with the actual interocular separation I).* From equation (2) it can be seen that for a given i the ratio d_s/d_m varies slightly depending on the value of d_m used. However, over the range of values for d_m used in these experiments, this produced negligible differences in disparities. The ratio of the true disparity to the manipulated disparity is also slightly different for points away from the midline. Again this effect was negligible for the object sizes used here.

Although this method for manipulating disparities does not scale the whole disparity field by exactly the same factor, it does generate a disparity field which is exactly appropriate for an object of the same shape (but different size) at another viewing distance [see Fig. 4(B)]. This is important, because if stereo and motion were used together to calculate three-dimensional shape, then this stimulus is exactly appropriate to produce no change in perceived object shape, but a change in apparent size and distance. An important detail is that the stimulus display configuration must exactly correspond to the geometry used for ray-tracing. Consequently, when generating a stereogram using a large interocular separation, careful account is taken of where the stimuli will actually be displayed, relative to the subjects' eyes. These manipulations affect only horizontal disparities—the vertical disparity field was always appropriate for the real viewing distance.

*As a reviewer pointed out, it would also be possible to provide the program with a different value of viewing distance than the one used in the experimental set-up and with all depths scaled accordingly. This would be equivalent, if appropriate adjustments were made to account for the fact that the display monitor was not placed at the distance for which images were traced. To avoid these complications, we chose to manipulate the interocular separation.

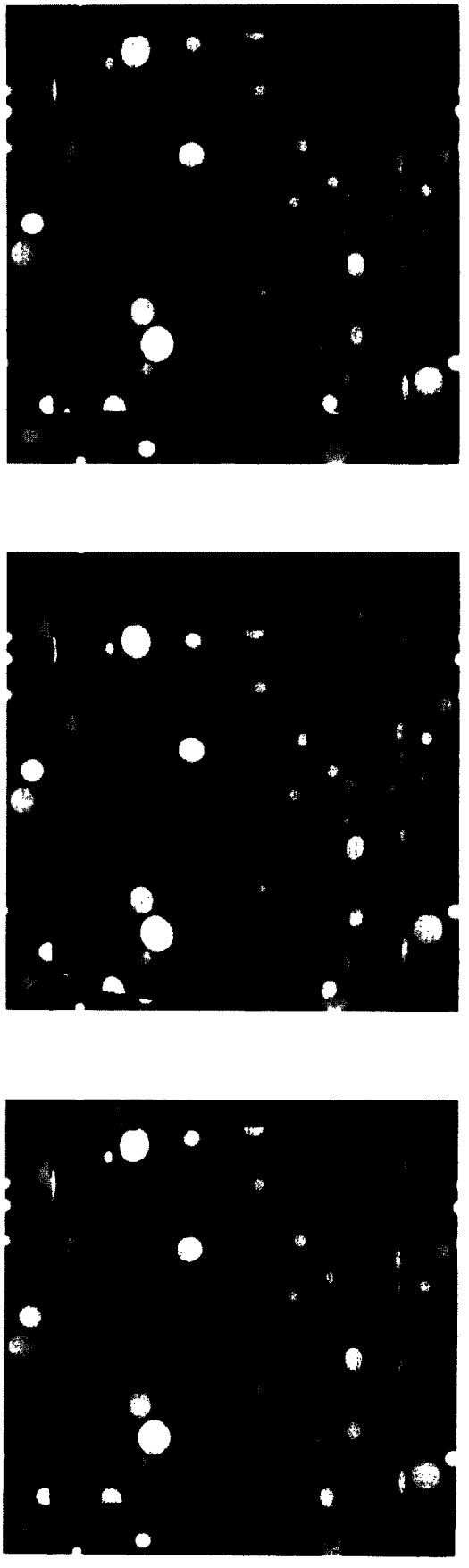

FIGURE 3. Example stimulus. A stereo pair which formed one frame of a motion stimulus. The images were traced for a circular cylinder to be displayed at 50 cm, rotated 3 deg anti-clockwise. The images are not exactly to scale, thus the disparity field is not as it would be in the experiments.

As the depth specified to the ray-tracing program was always the depth specified by motion, the shape-from-texture cue was always consistent with the motion cue. This choice was made because the texture cue always carries some shape information even if it specifies a flat surface (as in a conventional random dot stereogram). Thus, when the weight assigned to motion is discussed below it is really a combination of motion and texture. Since the effectiveness of the texture cue is small relative to stereo (Johnston *et al.*, 1993), this texture cue should not lead to a substantial misestimate of the importance of the motion cue.

Stimulus dimensions. The stimulus depicted was always 6 cm in both height and width. At the two viewing distances of 50 and 200 cm, the stimulus subtended 6.82 and 1.72 deg respectively. The individual sphere texels in the volume used to create stimuli varied in size from 0.17 to 0.42 cm, subtending visual angles of 11.68–28.88 min arc at 50 cm and 2.92–7.22 min arc at 200 cm. The stimuli were always presented on a background made from slices through the appropriate textured block, so a comparison of the textured three-dimensional surface with a flat surface specified by the same texture was always available.

Controls for speed and rotation angle. In this task, where the stimuli vary along the depth dimension, it is important that there is not some source of information other than depth which allows the subjects to order the stimuli, allowing them to perform the task on that basis. Two possible cues which might covary with the depth depicted are the speed of rotation and the maximum displacement of the occluding contour. These factors were randomized by varying the sequence of frames which were presented on each trial. Stereo-motion sequences consisted of frames covering ± 15 deg in steps of 1 deg. All 31 frames of the relevant stereo motion sequence were loaded into the display memory at the beginning of each trial. The individual 256×256 pixel frames could then be cycled through at frame rate (66 Hz). The speed of rotation was randomly varied from trial to trial by varying the number of times that each frame was presented. Frames were displayed either one, two or three times consecutively.

To prevent use of the total movement of the occluding contour, the extent of rotation of the stimulus was randomly varied from trial to trial, between 16 and 30 deg. This was achieved by varying the subset of frames presented to the subject. The frames displayed were always symmetrically arranged about 0 deg of rotation. The total number of frames was also randomized from trial to trial so that the duration of the sequence did not indicate to the subject which control condition was used. The surface could rotate back and forth between three and eight times.

Procedure

Experimental task. A global shape judgement task was used. The experimental task was identical to the task used previously to assess the veridicality of three-dimensional shape perception from cylindrical surfaces displayed as random-dot stereograms (Johnston, 1991). Subjects were presented with a series of horizontally oriented elliptical cylinders which varied in their elongation in depth from trial to trial. On each trial, observers decided if the cylinder was more or less extended in depth than a cylinder of circular cross section. Expressed differently, they were asked to determine if the ratio of depth to half-height was greater or less than 1. The measure obtained was the point of subjective equality, which corresponds to the apparently circular cylinder (ACC). In the figures, the depth/half-height for the ACC is plotted (labelled depth/height). A value of 1 means that equal portrayed height and depth yield a surface which appears circular to subjects, thus perception is veridical. ACC depth is inversely related to perceived depth so a value of less than 1 means that the perceived depth is overestimated.

Staircase method. A modified one-up one-down staircase method was employed. Up and down staircases were randomly interleaved (Cornsweet, 1962). The 15 stereo-motion sequences, which varied in the depth depicted, were rank-ordered. The stepsize between consecutively presented stimuli was initially set to three. When a reversal occurred the stepsize was decreased by one. Reversals were recorded once the stepsize was reduced to one. In order to collect only independent reversals, consecutive reversals in the same staircase were rejected. On each experimental run three reversals of up staircases and three of down were collected. Each run was repeated four times, making a total of 24 reversals for each data point presented here. In all of the following data plots the error bars shown are standard deviations calculated from the 24 values for the ACC collected for each data point.

Apparatus

The stimuli were presented on a Trinitron GDM-1955A15 monitor. The display screen measured 35×27 cm, and 1192×900 display pixels were available. The stimuli were presented on a TAAC-1 graphics accelerator on a Sun Microsystems SPARC 370. Stereoscopic presentation was achieved by means of a modified Wheatstone mirror stereoscope as described in Johnston *et al.* (1993). The vergence angle was correctly set by monocularly aligning stereo fixation crosses with a physical fixation cross set at the viewing distance required.

Subjects

The subjects were the first author and two others who had no knowledge of the experimental manipulations and aims of the experiments. Two of the subjects (EBJ and RBC) are slightly myopic and wore their optical correction to perform the experiments.

EXPERIMENT 1

This experiment examines the perception of three-dimensional shape from stimuli defined by stereo alone, motion alone, and a congruent combination of stereo

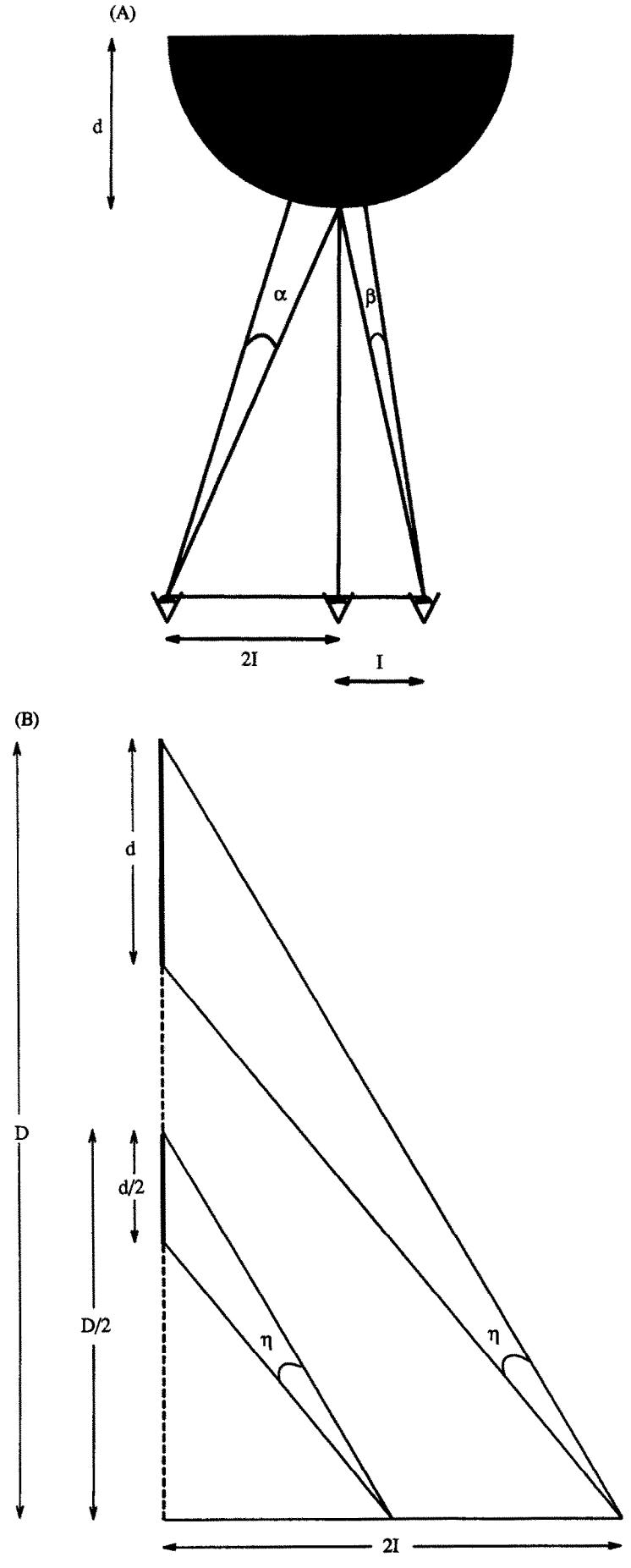

FIGURE 4. How stereo and motion depths are independently manipulated. The interocular separation parameter used in ray-tracing is varied. (A) The consequence of doubling the interocular separation (I) is that disparity generated by the depth (d) is doubled—α is twice the size of β. Manipulating the interocular separation to change the depth specified by disparity does not alter the depth depicted by each eye's flow field. (B) The disparity (η) generated by doubling the interocular distance is exactly equivalent to that generated by an object of half the size at half the distance (proven here by similar triangles). Changing the interocular separation is equivalent to changing the viewing distance in terms of the size of disparity generated.

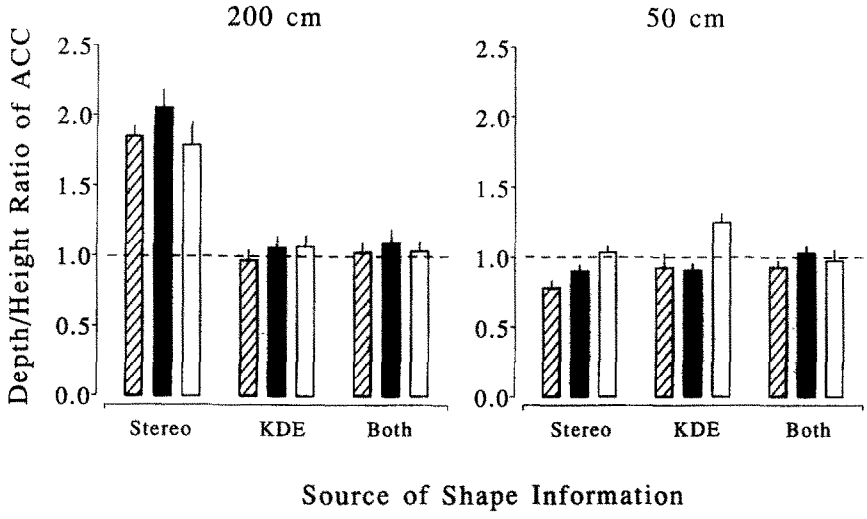

FIGURE 5. Portrayed depth/height for cylinders which appeared circular to the three subjects. Height actually refers to half-height of the cylinder which was fixed at 3 cm. The data are plotted in terms of depth/height yielding an inverse scale—*larger* values represent *decreased* perceived depth. The dashed line indicates veridical perception where portrayed height equals depth. Hatched bars, subject EBJ; solid bars, subject RBC; open bars, subject MJY. Data for viewing distances of 200 and 50 cm are shown.

and motion. The viewing conditions were kept as similar as possible for the three conditions. All stimuli, including those with motion alone, were viewed through the mirror stereoscope. For the motion alone stimuli, viewing was monocular to rule out the conflict with stereopsis which would arise if there were no disparities. The stereo alone stimuli were viewed face-on (unrotated). The purpose of this experiment is to compare the veridicality of the combined stereo and motion percept with the percepts derived from either cue alone.

The data are shown in Fig. 5. At a viewing distance of 200 cm (left-hand panel) depth from stereopsis alone is underestimated by the three subjects, as was found previously (Johnston, 1991). Depth from kinetic depth is veridically perceived, as is depth from the congruent combination of stereo and kinetic depth. At the close viewing distance of 50 cm (right-hand panel) depth from stereo is slightly overestimated as previously reported. Again depth from monocularly viewed KDE stimuli is close to veridical, showing the lack of distance dependence of structure from motion. (However, subject MJY shows a slight underestimation of depth from KDE.) The consistent combination of stereo and kinetic depth also yields percepts close to veridicality.

As discussed in the Introduction, schemes which combine stereo and motion to solve the distance scaling problem (Richards, 1985; the Appendix to this paper) predict that perceived shape is determined by the shape specified by motion. This is consistent with the results of Expt 1. At both distances the portrayed depth/height of the ACC is similar for KDE and combined stereo and KDE—veridical for both—but the depth/height of the ACC for stereopsis alone depends upon the viewing distance. In terms of the shape judgement the contribution from stereo is ignored in the combined percept.* This type of interaction is predicted by any scheme which uses the motion information to overcome the stereo viewing distance scaling problem.

As stated in the Introduction, scaling the disparity data with an incorrect estimate of the viewing distance results in a misperception of the depth/height ratio, because height and depth scale differently with viewing distance. However, the effect of misestimating the viewing distance on KDE information is *not* shape distortion, but a *uniform* scaling of the shape which maintains the depth/height ratio. Thus, the type of stereo/KDE interaction scheme considered here would hold that the depth/height ratio is fixed by KDE and the viewing distance parameter is altered until there is agreement between the depth/height ratios specified by stereo and motion.

To clarify the predictions of such an interaction scheme consider a stimulus with incongruent shapes specified by stereo and KDE. If the stimulus is such that the binocular disparities specify twice the depth specified by the relative velocities then the depth/height ratio portrayed by stereo will be twice that specified by KDE. A weighted average would then produce some overall perceived depth somewhere between the two values. However, if the type of stereo-motion combination scheme discussed above is used to determine the combined percept the depth/height ratio is anchored by KDE. Thus, the viewing distance parameter estimated from the KDE/stereo combination will be half the actual viewing distance. If the observer is using such a scheme a cylinder with KDE depth/height = 1 and stereo depth/height = 2 will be perceived as circular, because the disparities will be scaled by a halved estimate of the viewing distance. The prediction of this scheme is that subjects should perceive cylinders as circular whenever the KDE information specifies a depth/height ratio of 1, whatever the depth/height ratio specified by stereo. This

*A reviewer pointed out that for the combined stereo-KDE stimulus stereo could affect the percept by determining the perceived distance of the object. This was not measured in our experiments.

produces veto-like behaviour in that the perceived depth/height ratio is determined solely by the KDE information. Note that this is not a true veto of stereo by motion—if for some reason structure from motion gave rise to non-veridical percepts (e.g. if there are too few frames to constrain the perception of depth), the combination of stereo and motion should nonetheless lead to veridical perception of the shape *specified* by the motion field. We will use the term veto here to describe the way in which perceived depth/height should be independent of disparity (when both motion and stereo are available), but stress that this does not mean simply that the stereo information has no effect. Another perceptual consequence of varying the depth specified by stereo (in spite of this "veto") is that it may affect the perceived size of the cylinder. The hypothesis that perceived shape is independent of disparity is tested directly in the next experiment which explores incongruent combinations of motion and stereo.

EXPERIMENT 2

In this experiment the ACC was measured for stimuli in which different depth/height ratios were specified by stereo and motion. Differences between the depths specified by motion and stereo were generated by varying the interocular separation parameter supplied to the ray-tracing algorithm as described in the Methods section. The vetoing hypothesis predicts that the ACC will be determined solely by the depth/height portrayed by motion. A weighted linear combination hypothesis predicts that as the depth/height specified by stereo is increased, the depth/height specified by motion will decrease linearly for the cylinder which appears circular.

Figure 6 shows the data for the two viewing distances. Each data point corresponds to an apparently circular cylinder with depth/height specified by stereo plotted on the abscissa and depth/height specified by motion on the ordinate. It is clear that the vetoing hypothesis is not borne out. As the depth/height specified by stereo is increased the depth/height specified by motion must be decreased in order for the cylinder to appear circular to the subjects. This is consistent with the findings of Tittle and Braunstein (1991) on interactions of incongruent stereo and kinetic depth.

The relative weights assigned to stereo and motion can be calculated from the equation:

$$d = \alpha_s f_s(d_s) + \alpha_m f_m(d_m) \quad (3)$$

where d is perceived depth (assumed equal to the half-height given the task), α_s and α_m are the weights assigned to stereo and motion, d_s and d_m are the depths portrayed by stereo and motion and f_s and f_m are the functions relating perceived depth from stereo and motion to the portrayed depths. In order to calculate weights from the data shown in Fig. 6 some assumptions must be made. We deal with depth rather than depth/height below for ease of exposition, this means that we assume that the height is veridically perceived for KDE. The weights α_s and α_m are taken to sum to 1 (Maloney & Landy, 1989), this means that the full depth percept is accounted for by the contributions of stereo and motion. Another assumption made is that the functions f_s and f_m are linear. Further, based on the results of Expt 1 for the KDE alone condition, it is assumed that f_m is veridical. Given these assumptions it is possible to solve for f_s, α_s and α_m. In all cases the estimate of f_s was approximately the identity function, that is veridical estimation of depth from stereo disparity, thus the functions f_s and f_m can be dropped from equation (3).*

Note that this value of f_s is different from that obtained when stereo was the only depth cue (when depth was misestimated). This supports an interaction in the form of promotion, in which the addition of motion information enables veridical estimates of depth from disparity ($f_s = 1$). Table 1 shows the values of α_s and α_m calculated. The calculation of weights in this manner assumes a linear interaction between cues, thus it is only valid to apply the analysis to the linear portions of the data. In order to calculate the weights for the 50 cm viewing distance the nonlinearity was discounted by omitting the points for the smallest stereo depth when calculating regression lines. The weights were calculated from the fitted regression lines shown in Fig. 6.

At the far viewing distance the weight assigned to motion is much larger than that assigned to stereo, accounting for an average of 77% of the depth percept. At the close distance stereo and motion are given approximately equal weight. This may be due to assigning weights according to the reliability of the individual cues (Maloney & Landy, 1989). Stereo can be considered less reliable at far viewing distances as the disparities generated are small—e.g. a maximum of 1.6 min arc for a 3 cm depth cylinder viewed from 200 cm. Since small disparities correspond to large shifts in depth at far distances, small misestimates of disparity can lead to large errors in calculated depth.

Although these calculations suggest that depth from stereo, $f_s(d_s)$, is close to veridical in the presence of motion, the observations with stereo alone (Fig. 5) showed that $f_s(d_s)$ was far from veridical. Therefore it is necessary to propose that in addition to the linear combination, the presence of motion signals is modifying the scaling of disparity data. In other words, there is clear evidence of promotion in the integration of these cues. Reconsidering the data from Expt 1 (Fig. 5), the only weighting of stereo and motion which would explain those data is a weight of 0 assigned to stereo, that is a vetoing of stereo by motion. This type of vetoing was predicted from the schemes considered for cooperation between stereo and motion, using the combination either to calculate viewing distance or to remove the stereo dependence on viewing distance. However, we find here

*A reviewer pointed out if f_m is the identity function then the necessary and sufficient condition to have f_s also the identity function is for the regression lines to pass through the point (1, 1). Inspection of the figures shows that this approximately holds for all of the fitted lines.

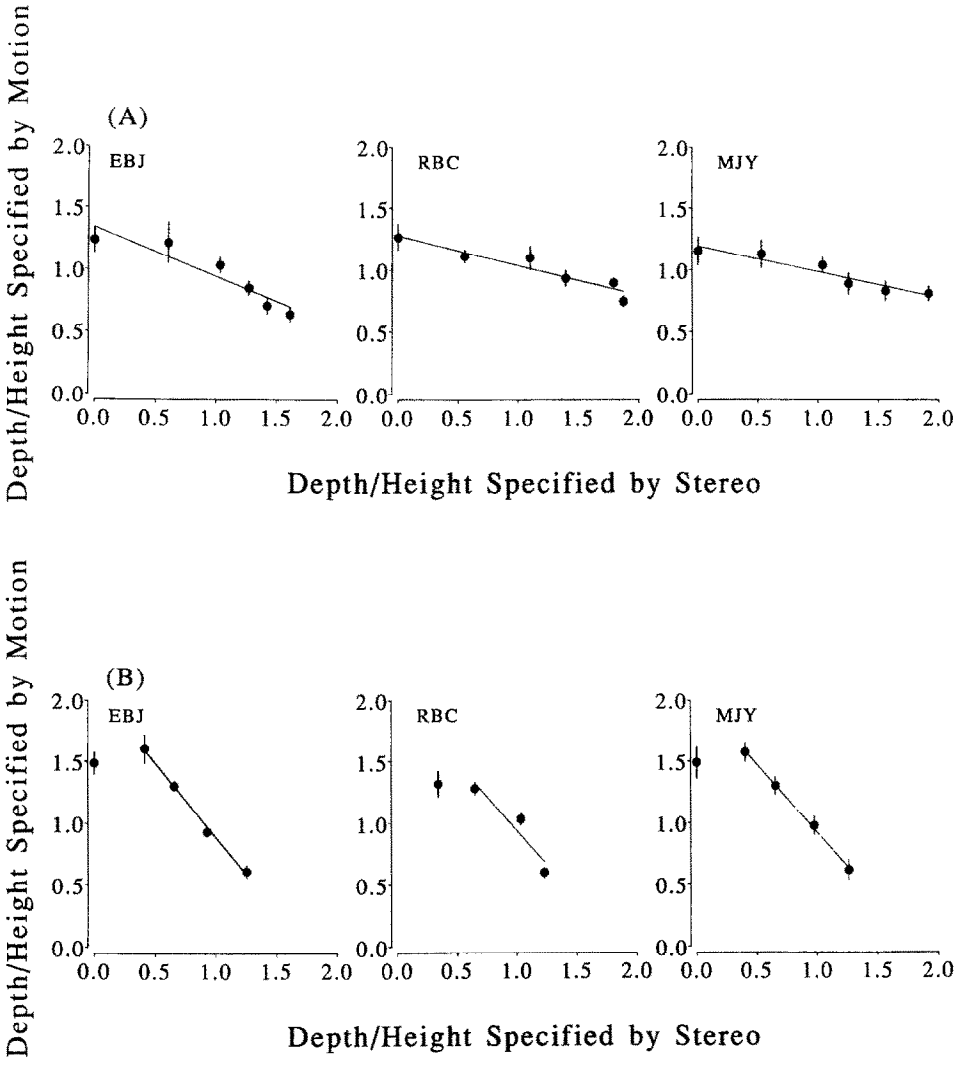

FIGURE 6. Apparently circular cylinder data for surfaces with incongruent depths specified by kinetic depth and stereo. The data are plotted in terms of the depth/height portrayed by stereo and the depth/height portrayed by motion for the cylinders which were judged to have equal apparent depth and height. In terms of the abbreviations used in the text the ordinate is d_m/h, and the abscissa is d_s/h. Each data point corresponds to a different portrayed d_s/d_m ratio. The error bars show standard deviations. (A) Viewing distance 200 cm. The lines shown are regression lines. The goodness of fit of the regression lines can be estimated from the closeness of the squared correlation coefficient to 1. The r^2 values are EBJ, 0.872; RBC, 0.892; MJY, 0.891. (B) Viewing distance 50 cm. The lines shown are regression lines fit to all points but the leftmost, thus taking account of only the linear portion of the data set (see text). The r^2 values are EBJ, 0.996; RBC, 0.892; MJY, 0.997.

that the weight applied to the stereo cue is nonzero. In effect, the motion information was used to scale stereo disparities to be nearly veridical, and depth from disparity was then given nonzero weight in the final percept. This initial scaling interaction appears to be a form of promotion or interaction between stereo and motion. Thus, the data follow the pattern described as modified weak fusion.

TABLE 1. Experiment 2—weights assigned to motion and stereo

Subject	Distance	KDE	Stereo
EBJ	200	0.70	0.30
RBC	200	0.81	0.19
MJY	200	0.82	0.18
EBJ	50	0.42	0.58
RBC	50	0.46	0.54
MJY	50	0.46	0.54

Calculated assuming that the weights sum to 1, and that the perceived depths from motion and stereo are veridical.

EXPERIMENT 3

In Expts 1 and 2 the number of different frames of each motion sequence presented ranged from 16 to 31. Thus, it was always possible to perform the task correctly from the motion information alone. In this experiment the number of frames in each motion sequence was reduced to only two in order to examine stereo/motion combination under conditions where the motion cue alone is not sufficient. In this case it would be necessary to combine information from stereo and motion to derive a veridical depth percept. When only two frames of motion are presented, there is not enough information to extract structure using the scheme presented in

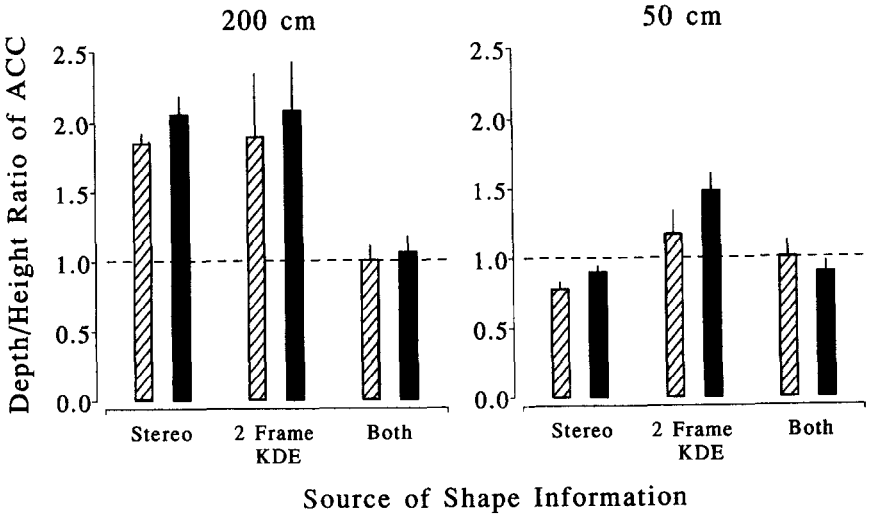

FIGURE 7. The depth/height ratio for the apparently circular cylinder data plotted for three conditions. *Increased* depth/height ratio indicates *decreased* perceived depth. Condition 1: stereo alone replotted from Fig. 5. Condition 2: two frame KDE, monocularly viewed. Condition 3: two frame KDE plus stereo, four frames in total, two to each eye. Hatched bars, subject EBJ; solid bars, subject RBC. Data for viewing distances of 200 and 50 cm are shown.

Ullman (1979). Ullman did devise a scheme for perspective projection which requires only two views, but it is highly susceptible to noise, and thus is unlikely to be used in human vision. In the two frame motion case there is more reason to use a combination of relative velocities and disparities to produce one depth estimate, as neither cue is complete and reliable alone.

Another reason that it is of interest to examine two frame motion is the recent suggestion of Todd and others (Bennett, Hoffman & Prakash, 1989; Todd & Bressan, 1990; Todd & Norman, 1991) concerning the representation produced by structure from motion. Their hypothesis is that metric structure is not available from kinetic depth. Rather, structure is only extracted up to a family of affinely equivalent surfaces. These are related by a stretching transformation along the line of sight (similar to the set of possible surfaces compatible with a given disparity field). They argue that all the information extracted from the kinetic depth cue is available in two frames, as higher order derivatives are not computed. The interpretation of any perceived differences between two frame motion and many frame motion is made complicated by the fact that the effectiveness of a two frame sequence may depend upon which two views of an object are used (Todd & Norman, 1991). Using two frame motion combined with stereo, we can investigate whether the human visual system can recover metric structure, even when the motion stimulus alone only permits recovery of affine structure. This avoids some of the difficulties inherent in many frame motion sequences. This experiment examines two frame motion alone and consistent two frame motion and stereo.

Method

The method is essentially identical to that used in the previous experiments. However, as there are only two frames, the same controls for speed and angle of rotation cannot be used. To break the linkage between depth, speed, and extent of rotation, the angle of rotation between the two frames presented was varied randomly from trial to trial. When there are only two frames it is necessary to have a larger step in angular increment between stimuli than the 1 deg used in the many frame experiments, in order to obtain a reasonable depth impression. The possible values of angular increment varied between 5 and 7 deg. Each frame was presented either 4, 5 or 6 times consecutively (i.e. SOAs of 61–91 msec) and the surface rotated back and forth 8 times. The number of times a single frame was presented had to be increased in comparison with the many frame case, otherwise the large angular increment would result in surfaces which were moving too quickly to form a good depth impression. The random variations in extent of angular motion and speed prevented observers from ordering the depths portrayed by the stimuli based on speed or maximum movement of the occluding contour.

Results

Figure 7 shows the data. As in Expt 1, three conditions are shown—stereo alone, two frame motion alone (monocularly viewed) and a congruent combination of stereo and two frame motion. The stereo data are simply replotted from Fig. 5, and show that perceived depth from stereo alone is underestimated by both subjects at 200 cm, and slightly overestimated at 50 cm. The data for two frame KDE alone show a substantial underestimation of perceived depth at 200 cm, similar to that obtained with stereo. At 50 cm both subjects still underestimate depth from two frame KDE alone, although less severely.* When two frame motion is combined with consistent stereo disparities, perception is

*We are not certain why depth from motion is affected by change of viewing distance in the two frame case. It may be related to the smaller image velocities produced by more distant targets. Perhaps the visual system interprets slower retinal velocities as resulting from slower absolute velocities (produced by an object with less depth).

close to veridical at both viewing distances. At the 200 cm viewing distance, this provides a dramatic illustration of the part of the cue interaction that cannot be described by linear weighted averaging. Although depth is substantially underestimated by both cues alone, perception is close to veridical when they are combined, at 200 cm.

Young, Landy and Maloney (1993) studied combinations of texture and motion cues to depth and introduced the notion of cues to "flatness". In a display which the experimenter gives depth by manipulating texture, stereo and motion, there are other, extraneous cues (vergence, accommodation, motion parallax if the head is free to move, prior knowledge) all of which may signal that the display is flat (which, in fact, it is). They suggest that when the manipulated cues are of low quality, a weighted averaging of depth cues is still used, but more weight is given to these extraneous cues, resulting in a display which appears flattened. In Expt 3 one might want to interpret the results as a lowering of the weight given to a default "flatness" cue when both stereo and motion are present (i.e. when display quality is high). However, the data follow the opposite pattern at 50 cm. At 50 cm the combined percept has an equal or smaller depth than that of stereo alone. Lowering of the weight given to a flatness cue would predict the opposite, that more depth should be perceived in the combined stereo-motion stimulus than in the stereo alone stimuli. This lends further support to the claim that motion does not veto stereo—when motion alone does not generate a veridical depth percept, the difference between the motion-defined percept and that produced by both cues becomes clear.

In this experiment obtaining veridical perception relied critically upon the combination of stereo and motion. This contrasts with Expts 1 and 2, when veridical perception was possible in the absence of stereo. Exploring the effects of different combinations of motion and stereo using only two motion frames may therefore provide a better insight into how the two cues are combined.

EXPERIMENT 4

In this experiment the comparison between two and many frame motion is extended to stimuli with incongruent cues. The data of Expt 3 show that reducing the motion sequence to only two frames has a profound effect on perceived depth from KDE alone. Here we examined whether this weakening of the motion cue would have an effect on the relative weighting of stereo and motion. The method for randomizing speed and extent of rotation described above for Expt 3 was used again. The inconsistency between stereo and motion was produced by varying the interocular separation parameter supplied to the ray-tracing routine, as described in the general Methods section.

The data are shown in Fig. 8—as the depth specified by stereo is increased, the depth specified by two frame motion has to be decreased considerably for the cylinder to appear circular to the subject. This effect of stereo is stronger than that seen in Expt 2. The stereo and motion weights were calculated by the same method as described above, and are listed in Table 2. In these calculations we assumed that both f_m and f_s were veridical. Although neither stimulus alone gave veridical perception, the results of Expt 3 showed that when the cues were combined promotion occurred, so that veridical depth estimates were available. This is similar to the procedure in Expt 2, where it was shown that f_s became veridical when combined with motion. The figures given in Table 3 demonstrate that stereo makes a larger contribution with two frame motion than with many frame motion.

DISCUSSION

One of the most significant findings reported here is that the interaction of stereo and structure-from-motion cannot be described by a linear combination rule alone, so the interaction cannot simply be weak fusion. The human visual system utilizes a stereo-motion combination scheme to solve the stereo scaling problem as outlined in the Introduction. Some stereo-motion combination schemes predict that perceived shape from incongruent stereo-motion stimuli will be independent of disparity. For example, Richards' (1985) scheme predicts that when stereo and kinetic depth are put in conflict, the perceived depth/height ratio will be entirely determined by relative motion whatever the depth/height calculated from binocular disparities. The same result follows from the scheme presented in the Appendix. This prediction is not borne out by the data presented here—the depth/height ratio specified by stereo disparities affects the perceived depth/height ratio of the combined stimulus even when stereo and motion signal inconsistent depths (Fig. 6). Perhaps the lack of vetoing shown in the experimental data is not entirely surprising given that the schemes discussed carry the assumption that the viewing distance is calculated *only* from the stereo-motion interaction. Vetoing schemes are equivalent to a rescaling of all image data by the particular viewing distance which will produce the same depth/height ratio as specified by KDE. In the case of a stereo/motion depth ratio of 2.0, this amounts to a halving of the viewing distance. In practice there may be other constraints placed upon the value for viewing distance used to scale image data. Such constraints come from the convergence angle of the eyes, the subject's knowledge that the viewing distance has not changed between consecutive experimental runs, and the existence of a variety of objects in the subject's field of view which can provide familiar size information. It may be difficult for the observer to override entirely these other cues to viewing distance. One observation consistent with this view is that the surfaces do not appear to change drastically in size when the different stereo/motion ratios are presented. Changes in convergence angle are known to produce both changes in perceived size, and perceived depth/height from stereo alone (Cumming *et al.*, 1991), as expected from a change in the estimate of viewing distance. Manipulating the

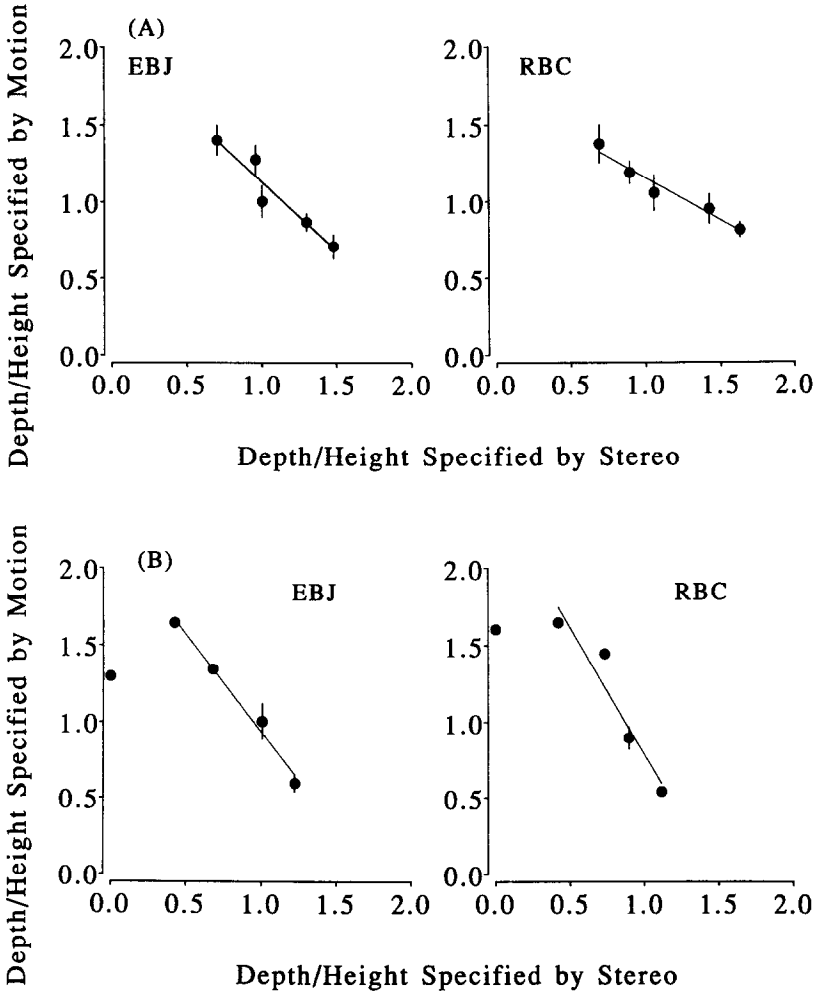

FIGURE 8. Incongruent combinations of stereo and motion, with two frame motion. The data are plotted in terms of the depth/height portrayed by stereo and the depth/height portrayed by motion for the cylinders which were judged to have equal apparent depth and height. (A) 200 cm viewing distance. The r^2 values are EBJ, 0.907; RBC, 0.956. (B) 50 cm viewing distance. The lines shown are regression line fits to the data (excluding the 0 point for the 50 cm data). The r^2 values are EBJ, 0.982; RBC, 0.926. The error bars show standard deviations. Comparison with Fig. 6 shows that stereo has a greater effect when the motion cue is weakened (the absolute slopes of the regression lines are larger here than in Fig. 6).

convergence position in stereo/motion experiments should help clarify the role of explicit distance estimates on shape perception.

Experiments 1 and 3 demonstrate that the combination of stereo and motion cannot be considered simply a summation of the depths specified by each cue in isolation. There is no single relative weighting of stereo and KDE that produces the combined percept observed. (Although assigning a weight of zero to stereo would account for the data of Expt 1, it is incompatible with the data of the other three experiments.) Thus, as a result of presenting stereo and KDE information simultaneously there is some modification of the individual depth or depth/height estimates produced by stereo and KDE. However, it is not possible to account for the data in Expts 2 and 4 simply by assuming that viewing distance is calculated from the combination of stereo and motion. Schemes that do this predict that only object size, not object shape, should change as a result of incongruent combinations of stereo and motion. The data suggest that there is some interaction of stereo and motion that alleviates the stereo scaling problem, but that stereo and KDE then yield separate depth estimates.

TABLE 2. Experiment 4—weights assigned to motion and stereo, when the motion consists of only two frames

Subject	Distance	KDE	Stereo
EBJ	200	0.57	0.43
RBC	200	0.67	0.33
EBJ	50	0.42	0.58
RBC	50	0.32	0.68

Calculated assuming that the weights sum to 1, and that the perceived depths from motion and stereo are veridical.

TABLE 3. Weight given to stereo as a function of number of motion frames averaged over subjects

Distance	Two frame	Many frame
200	0.38	0.23
50	0.63	0.55

If we accept that there are independent depth estimates from stereo and KDE (however they are modified by the simultaneous presence of the two cues) we can then consider the linearity of the interaction and the relative weighting of the cues. The results of Expt 2 demonstrate that a linear combination rule is appropriate to describe the interaction of stereo and KDE, at least over small degrees of perturbation of the relative depths specified by the two cues. At the close viewing distance used (50 cm), the relative weighting of stereo and motion was approximately equal. The relative weighting of stereo was approximately halved at the far viewing distance (200 cm). These results are very different from those of Rogers and Collett (1989), who found a very low weight assigned to motion. However, there is no real discrepancy between our findings and theirs. In their study the variation in image velocities produced a sensation of rotation of the whole object. In this way it was possible to resolve the motion flow field with the disparity field without requiring changes in perceived depth. Our observation that the relative weighting of stereo and motion depends upon viewing distance, and the difference between our results and those of Rogers and Collett, show that the weighting of the cues depends to some extent upon the viewing conditions.

To investigate this issue further we examined incongruent combinations in which the KDE information was weakened by presenting only two frames of motion. The combination of stereo and two frame KDE proved to be veridical, although neither the stereo nor the KDE percept was veridical alone. This appears to be an instance of promotion, where the estimate of the viewing distance parameter must be obtained from a combination of the information provided by the two cues. Although the results with two frame motion are most dramatic, it should also be stressed that the same promotion type of interaction was observed with many frame motion—stereo appeared to be scaled veridically in the presence of a motion cue, at viewing distances where stereo alone was far from veridical.

Experiment 4 addressed the issue of whether the weighting of stereo and motion can be altered by reducing the effectiveness of the motion information. With only two frames, motion alone provided inadequate information to solve the task. In spite of this, the basic form of the interaction was similar, suggesting that the scaling difficulties were overcome by the cue combination, and then separate depth estimates were combined linearly. These data also showed clearly that perceived shape is not dominated by the shape *perceived* from motion alone. Depth from two frame motion alone was underestimated, but the combination of two frame motion and stereo was veridical.

The other motivation for the two frame motion studies was to investigate the claim that human subjects extract only affine structure from kinetic depth. Todd and Bressan (1990) and Todd and Norman (1991) postulate that structure can be extracted from motion only up to the particular affine transformation of stretching along the line of sight. One line of evidence they provide to support the affine theory is their finding that perceived depth from two frame motion is not significantly different from either four or eight frame motion. In their studies the two frames chosen were not symmetrical about the midline. This is an important difference between the work of Todd *et al.* and Expts 3 and 4 above. When a circular cylinder in the fronto-parallel plane is stretched along the line of sight, it becomes an elliptical cylinder. Although this is also true of a cylinder which is rotated around a vertical axis, the angle formed between the occluding contour and the face is altered by this affine stretch. Therefore it is possible that using only two frames of motion that were not symmetrical about the midline, subjects would be able to perform the ACC task veridically. It is theoretically possible that this explains the observation that subjects were able to extract veridical metric structure from many-frame motion, while the depth/height ratio was significantly underestimated from two frame motion alone. Thus, our results are quite compatible with those of Todd *et al.*

Todd (personal communication) pointed out that veridical perception from a single stereo pair might be possible if the stimulus is not oriented face-forward, in the same way that the ACC task is possible for two frame motion when the rotation is not symmetrical about the midline. This might explain the veridical perception found in the experiment with two frame stereo and motion (since each stereopair portrayed a rotated cylinder). Since the rotation used here was so small (a maximum of 3.5 deg from face forward), this would not have been a strong cue. To control for any such effect, we repeated the stereo ACC experiment at a viewing distance of 200 cm for two subjects, with stimuli portraying cylinders rotated 3.5 deg from face forward. The rotation produced no change in the shape distortion shown in Fig. 7. Therefore, the veridical perception observed here in Expt 3 for two frame stereo viewing appears to be the result of the interaction of stereo and motion, not merely a consequence of the stereo views used.

The fact that combining two frame motion with stereo produced veridical perception has important implications: under these circumstances the human visual system can clearly recover metric structure (or at least shape to within a scale factor), rather than being limited to an affine set of shapes. This indicates that a metric representation is used in human vision, and prompts the question: why is metric shape not extracted in the experiments reported by Todd *et al.* on the KDE. One possible explanation (Eagle & Blake, 1994) is that it is necessary to have multiple frames that span a wide range of rotation in order to extract metric shape reliably (this need arises because human velocity discrimination thresholds are poor, and extracting metric shape from three frame motion requires the use of information about changes in velocity). Using three frame motion, Eagle and Blake found that a task requiring metric representation could be performed if the rotation angle was large enough. Using two frame motion spanning the same rotation range (i.e. removing the middle frame)

they found that performance on the metric task was poor. Thus it seems that Euclidean properties of two structures can be discriminated from KDE information when their changes in displacement can be discriminated.

The combination of stereo and two frame KDE proved to be veridical, although neither the stereo nor the KDE percept was veridical alone. This is an example of modified weak fusion involving promotion. Promotion falls into the category of strong fusion processes. Other investigators have reported findings which indicate strong fusion processes. For example, Bülthoff (1991) found that perceived depth from a combination of shading and texture exhibited superadditivity—depth from texture or shading alone was significantly underestimated but the combination was veridical. However, their results could be explained by assigning weight to a default "flatness" term when only one cue is present, but giving no weight to flatness when two consistent cues are present. This term can be thought of as a default tendency towards seeing no depth when the information is poor (Young *et al.*, 1993). This flatness explanation could be applied to the stereo and motion data at the 200 cm viewing distance (Fig. 7, left-hand panel), where depth is underestimated from both cues alone but veridical for the combination. However, it would not apply to the data collected at 50 cm (Fig. 7, right-hand panel), where depth from stereo is equal or greater than depth from the combination. Thus, promotion is required to explain the interaction.

The close geometrical relationship between the information provided by stereo, and that provided by motion, makes these two cues ideal candidates for cooperative interaction prior to obtaining depth estimates. Although we have found clear evidence for this, it is interesting that there is also a stage of weighted linear summation. The adaptability of the weights is illustrated by the changes in relative weighting which resulted from changes in viewing distance, and from changes in the number of motion frames presented. Depth cues are not all given equal weight, nor are they fixed in a strict hierarchy of strength of percept. The effectiveness of a given depth cue in a given stimulus depends upon the reliability of the information provided by that cue. Depth cue combination is a highly adaptable process, able to take account of the varying availability and accuracy of the individual cues.

REFERENCES

Andersen, R. A., Graziano, M., Snowden, R. J. & Treue, S. (1990). Cortical processing of depth from motion. Paper presented at the Rank Prize Symposium on Neural Representation of 3-D Space, Grasmere, Cumbria, England.

Bennett, B. M., Hoffman, D. D. & Prakash, C. (1989). Structure from two orthographic views of rigid motion. *Journal of the Optical Society of America A, 6*, 1052–1069.

Berry, R. N. (1948). Quantitative relations among vernier, real depth and stereoscopic acuities. *Journal of Experimental Psychology, 38*, 708–721.

Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *Journal of Physiology, London, 211*, 599–622.

Bruno, N. & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General, 117*, 161–170.

Bülthoff, H. (1991). Shape from X: Psychophysics and computation. In Landy, M. S. & Movshon, J. A. (Eds), *Computational models of visual processing* (pp. 305–330). Cambridge, Mass.: MIT Press.

Clark, J. J. & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Boston, Mass.: Kluwer Academic Press.

Cornsweet, T. N. (1962). The staircase method in psychophysics. *American Journal of Psychology, 75*, 485–491.

Cumming, B. G., Johnston, E. B. & Parker, A. J. (1991). Vertical disparities and 3-D shape perception. *Nature (London), 349*, 411–413.

Cynader, M. & Regan, D. (1978) Neurons in cat parastriate cortex sensitive to the direction of motion in three-dimensional space. *Journal of Physiology, 274*, 549–569.

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

Eagle, R. A. & Blake, A. (1994). 2-D limits on 3-D structure-from-motion tasks. *Investigative Ophthalmology and Visual Science, 35*, 1277.

Foley, J. (1980). Binocular distance perception. *Psychological Review, 87*, 411–434.

Gogel, W. C. (1960). The perception of shape from binocular disparity cues. *Journal of Psychology, 50*, 179–192.

Gogel, W. C. (1972). Scalar perceptions with binocular cues of distance. *American Journal of Psychology, 85*, 477–497.

Graham, C. H. (1965). *Vision and visual perception*. New York: Wiley.

Graham, M. E. & Rogers, B. J. (1982). Simultaneous and successive contrast effects in the perception of depth from motion-parallax and stereoscopic information. *Perception, 11*, 247–262.

von Helmholtz, H. (1910). *Handbuch der physiologischen optik* (3rd edn). Hamburg: Voss. [Translated by Southall, J. P. C. (1962). *Physiological optics* (Vol. 3). New York: Dover.]

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research, 50*, 1351–1360.

Johnston, E. B., Cumming, B. G. & Parker, A. J. (1993). The integration of depth modules: Stereopsis and texture. *Vision Research, 33*, 813–826.

Judge, S. J. (1990). Vision: Knowing where you're going. *Nature (London), 348*, 115.

Landy, M. S., Maloney, L. T. & Young, M. J. (1991a). Psychophysical estimation of the human depth combination rule. In Schenker, P. S. (Ed.), *Sensor fusion III: 3-D perception and recognition. Proceedings of the SPIE, 1383*, 247–254.

Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. J. (1991b). In defense of weak fusion: Measurement and modeling of depth cue combination. Mathematical Studies in Perception and Cognition, 91-3, New York University.

Maloney, L. T. & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. In Pearlman, W. A. (Ed.), *Visual communications and image processing IV, Proceedings of the SPIE, 1199*, 1154–1163.

Mayhew, J. E. W. & Longuet-Higgins, H. C. (1982). A computational model of binocular depth perception. *Nature (London), 297*, 376–377.

McKee, S. P. (1981). A local mechanism for differential velocity detection. *Vision Research, 21*, 25–32.

McKee, S. P., Levi, D. M. & Bowne, S. F. (1990). The imprecision of stereopsis. *Vision Research, 30*, 1763–1780.

Nawrot, M. & Blake, R. (1991). The interplay between stereopsis and structure from motion. *Perception & Psychophysics, 49*, 230–244.

Orban, G. A., Lagae, L., Verri, A., Raiguel, S., Xiao, D., Maes, H. & Torre, V. (1992). First order analysis of optic flow in the monkey brain. *Proceedings of the National Academy of Sciences, U.S.A, 89*, 2595–2599.

Parker, A. J., Johnston, E. B., Mansfield, J. S. & Yang, Y. (1991). Stereo, surfaces and shape. In Landy, M. S. & Movshon, J. A. (Eds), *Computational models of visual processing* (pp. 359–381). Cambridge, Mass.: MIT Press.

Poggio, G. F. & Talbot, W. H. (1981). Mechanisms of static and dynamic stereopsis in foveal cortex of the rhesus monkey. *Journal of Physiology, 315*, 469–492.

Richards, W. (1985). Structure from stereo and motion. *Journal of the Optical Society of America A, 2*, 343–349.

Rogers, B. J. & Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature (London), 361*, 253–255.

Rogers, B. J. & Collett, T. S. (1989). The appearance of surfaces specified by motion parallax and binocular disparity. *Quarterly Journal of Experimental Psychology, 41A*, 697–717.

Rogers, B. J. & Graham, M. E. (1982). Similarities between motion parallax and stereopsis in human depth perception. *Vision Research, 22*, 261–270.

Roy, J. P., Komatsu, H. & Wurtz, R. H. (1992). Disparity sensitivity of neurons in monkey extrastriate area MST. *Journal of Neuroscience, 12*, 2478–2492.

Saito, H.-A., Yukie, M., Taneka, K., Hikosaka, K., Fukada, Y. & Iwai, E. (1986). Integration of image motion in the superior temporal visual sulcus of the macaque monkey. *Journal of Neuroscience, 6*, 145–157.

Sobel, E. C. & Collett, T. S. (1991). Does vertical disparity scale the perception of stereoscopic depth? *Proceedings of the Royal Society London B, 244*, 87–90.

Taneka, K., Hikosaka, K., Saito, H.-A., Yukie, M., Fukada, Y. & Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual area of the macaque monkey. *Journal of Neuroscience, 6*, 134–144.

Tittle, J. S. & Braunstein, M. L. (1991). Shape perception from binocular disparity and structure-from-motion. In Schenker, P. S. (Ed.), *Sensor fusion III: 3-D perception and recognition. Proceedings of the SPIE, 1383*, 225–234.

Todd, J. T. & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception & Psychophysics, 48*, 419–430.

Todd, J. T. & Norman, J. F. (1991). The visual perception of smoothly curved surfaces from minimal apparent motion sequences. *Perception & Psychophysics, 50*, 509–523.

Ullman, S. (1979). *The interpretation of visual motion*. Cambridge, Mass.: MIT Press.

Young, M., Landy, M. & Maloney, L. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research, 33*, 2685–2696.

Acknowledgements—This work was supported in part by grants EY08266 and EY06337 from the National Eye Institute. Bruce Cumming was supported by the MRC and the Oxford McDonnell-Pew Centre for Cognitive Neuroscience. We are grateful to Ron Cagenello, Larry Maloney, Andrew Parker, Mark Young and two anonymous reviewers for helpful comments.

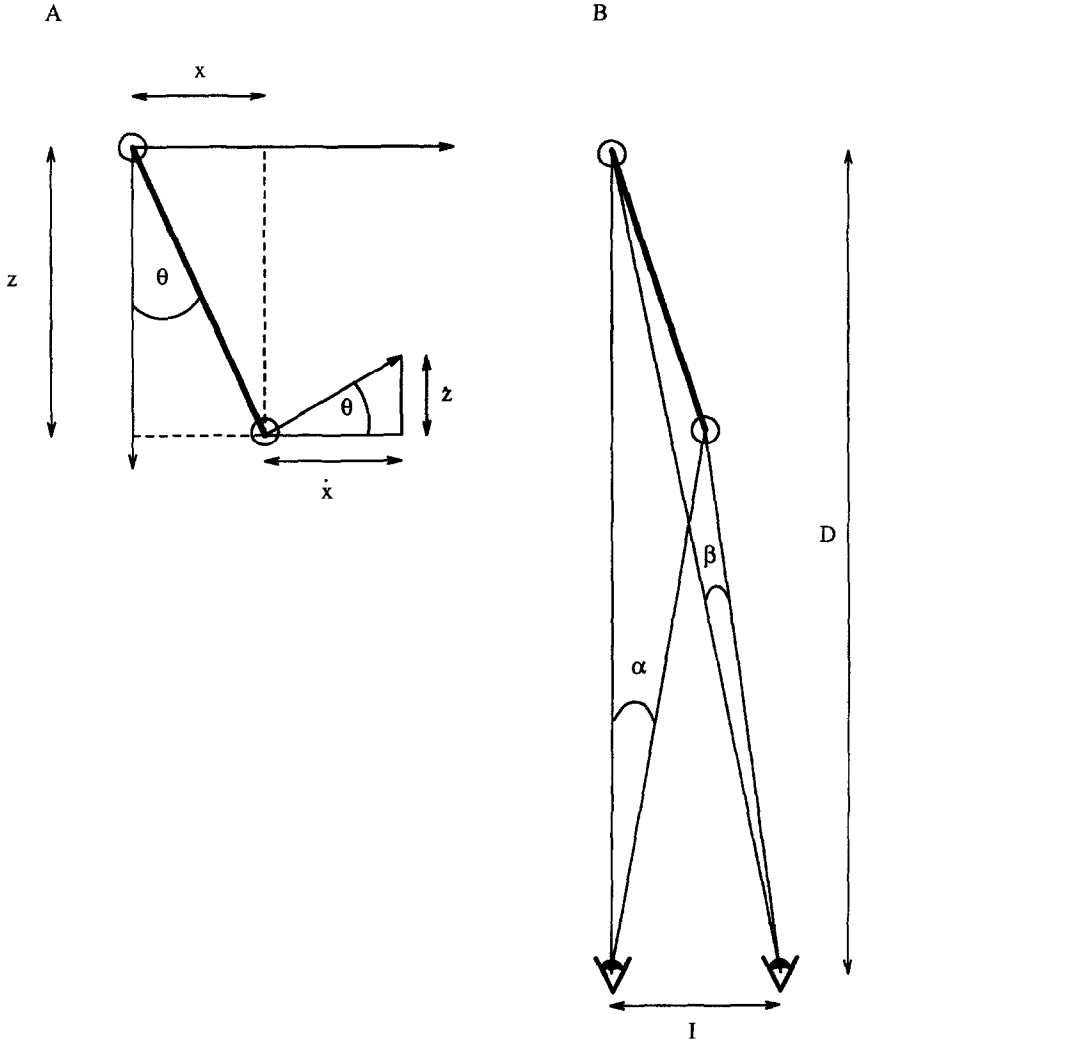

FIGURE A1. Schematic diagram of a stereo-motion integration scheme. (A) The heavy line joins two points in relative motion. The origin is coincident with the left point. The two angles θ are equal, but the sides of the triangles are measured by different visual parameters. This allows the viewing distance D, to be eliminated from equations for z and x (see text). (B) The disparity η is the difference in visual angle subtended by the two points, $\alpha - \beta$. D marks the viewing distance and I the interocular separation.

APPENDIX

Calculating Viewing Distance using Stereo-Motion Integration

Here we present a simple scheme for extracting a measure of viewing distance from a combination of stereo and KDE. We consider two points in relative motion, and define a coordinate frame whose origin is at one of the points (Fig. A1). By similar triangles

$$\tan\theta = \frac{x}{z}, \tag{A1}$$

$$\tan\theta = \frac{\dot{z}}{\dot{x}}, \tag{A2}$$

and thus,

$$\dot{x}x = \dot{z}z, \tag{A3}$$

where θ is the angle between the z-axis and the line joining the two points. The visual angle subtended by the two points at the left eye, α, at the right eye, β, and their relative disparity, $\eta = \alpha - \beta$, are calculated using the usual approximations (Graham, 1965),

$$x = \alpha D, \tag{A4}$$

and

$$z = \frac{D^2\eta}{I} \tag{A5}$$

where D is the unknown viewing distance and I is the interocular separation. Note that equation (A5) is a cruder approximation than equation (1), used when discussing the stimulus disparities, as this simplifies the equations that follow. Substituting equations (A4) and (A5) into equation (A3),

$$\alpha D^2\dot{\alpha} = \eta\dot{\eta}\frac{D^4}{I^2}, \tag{A6}$$

hence

$$D = I\sqrt{\frac{\dot{\alpha}\alpha}{\dot{\eta}\eta}}. \tag{A7}$$

Since I is known and α, $\dot{\alpha}$, η and $\dot{\eta}$ may be measured in the image, a measure of viewing distance can be obtained from two stereo views of a rotating point and the stereo scaling problem can thus be solved without recourse to any other sources of information about viewing distance. This equation for D can be substituted into equations (A4) or (A5) to give expressions for metric width or height:

$$x = \alpha I\sqrt{\frac{\dot{\alpha}\alpha}{\dot{\eta}\eta}}. \tag{A8}$$

The same can be done for metric depth:

$$z = I\frac{\dot{\alpha}\alpha}{\dot{\eta}}. \tag{A9}$$

The ratio of depth to width can also be computed:

$$z/x = \sqrt{\frac{\dot{\alpha}\eta}{\dot{\eta}\alpha}}. \tag{A10}$$

This is the three-dimensional shape measure required for our psychophysical task. Now, consider the consequences of manipulating the depth specified by stereo, while leaving the depth specified by motion unchanged. If all the disparity values are doubled, both η and $\dot{\eta}$ are doubled, so by equation (A7), D is halved. Equation (A10) shows that this produces *no change* in the value of depth/width or depth/height, simply a uniform scaling of the shape. Consequently, this model of stereo-motion integration, like that of Richards (1985), predicts that the perceived depth/height measured in the ACC task will not be altered by scaling the disparity field to specify another depth of cylinder while leaving the depth specified by motion unchanged.