LETTER Communicated by Steven Nowlan

Modeling the Combination of Motion, Stereo, and Vergence Angle Cues to Visual Depth

I. Fine
Center for Visual Science, University of Rochester, Rochester, NY 14627, U.S.A.

Robert A. Jacobs
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, U.S.A.

Three models of visual cue combination were simulated: a weak fusion model, a modified weak model, and a strong model. Their relative strengths and weaknesses are evaluated on the basis of their performances on the tasks of judging the depth and shape of an ellipse. The models differ in the amount of interaction that they permit among the cues of stereo, motion, and vergence angle. Results suggest that the constrained nonlinear interaction of the modified weak model allows better performance than either the linear interaction of the weak model or the unconstrained nonlinear interaction of the strong model. Further examination of the modified weak model revealed that its weighting of motion and stereo cues was dependent on the task, the viewing distance, and, to a lesser degree, the noise model. Although the dependencies were sensible from a computational viewpoint, they were sometimes inconsistent with psychophysical experimental data. In a second set of experiments, the modified weak model was given contradictory motion and stereo information. One cue was informative in the sense that it indicated an ellipse, while the other cue indicated a flat surface. The modified weak model rapidly reweighted its use of stereo and motion cues as a function of each cue's informativeness. Overall, the simulation results suggest that relative to the weak and strong models, the modified weak fusion model is a good candidate model of the combination of motion, stereo, and vergence angle cues, although the results also highlight areas in which this model needs modification or further elaboration.

1 Introduction

Recent years have seen a proliferation of new theoretical models of visual cue combination, especially in the domain of depth perception. This proliferation is due partly to a poor understanding of existing models and partly to a lack of comparative studies revealing the relative strengths and weaknesses of competing models. This article studies how multiple visual

 © 1999 Massachusetts Institute of Technology

cues may be combined to provide information about the three-dimensional structure of the environment.

Depth cue interactions have been extensively studied from a psychophysical and computational perspective (e.g., Rogers & Collett, 1989; Blake, Bülthoff, & Sheinberg, 1993; Nawrot & Blake, 1993; Tittle, Todd, Perotti, & Norman, 1995; Turner, Braunstein, & Anderson, 1997). Various models have been proposed to characterize these interactions (e.g., Bruno & Cutting, 1988; Bülthoff & Mallot, 1988; Clark & Yuille, 1990; Landy, Maloney, & Young, 1991). Landy, Maloney, Johnston, and Young (1995; see also Clark & Yuille, 1990) have defined three classes of models for combining visual cues for depth. Strong fusion models estimate depth by combining the information from different cues in an unrestricted manner. Weak fusion models compute a separate estimate of depth based on each depth cue considered in isolation. These estimates are then linearly averaged to yield a composite estimate of depth. The linear coefficients that weight the different cues are proportional to the cues' reliability.

Landy et al. (1995) proposed that aspects of the interactive properties of strong models and the modular properties of weak models can be combined in modified weak fusion models. Such models allow constrained nonlinear interactions, such as cue promotion and reweighting, between different cues. Most cues are incapable of providing absolute depth information when considered in isolation; for example, occlusion provides only order information, and motion parallax provides only shape information. However, once a number of missing parameters are specified, these cues become capable of providing absolute depth information. Cue promotion is the determination of these missing parameter values through the use of other depth cues. For example, motion parallax is an absolute depth cue if the viewing distance is known. There are a number of ways that this missing parameter could be specified, such as by means of the vergence angle or through the intersection of constraints using stereo disparities as well as motion parallax. According to Landy et al. (1995), this nonlinear stage, in which information from different cues is combined to promote any cue until it is capable of providing an absolute depth map, is followed by a linear stage, in which a weighted average is taken of the depth estimates of the different cues.

The results of some psychophysical experiments support relatively weak models, allowing little interaction between different cues for depth. Increases in the number of depth cues available in a stimulus display lead to increases in the amount of depth perceived and also to improvement in the consistency and accuracy of depth judgments (Bruno & Cutting, 1988; Bülthoff & Mallot, 1988; Dosher, Sperling, & Wurst, 1986; Landy et al., 1991). Bruno and Cutting (1988), for example, varied in a factorial design the availability of four depth cues (occlusion, relative size, height in the visual field, and motion perspective). Data from direct and indirect scaling tasks were consistent with observers' using a nearly linear additive procedure analogous to a weak fusion model.

It is clear, however, that the visual system is capable of using more complex rules of cue integration than simple linear averaging. Cue vetoing, a nonlinear combination rule whereby depth estimates are based on the cue in a visual scene ranked highest in a hierarchical ordering, has been observed with a number of visual cues. In the Ames room illusion, for example, perspective and other cues appear to veto "familiar size" (i.e., the adults in the far corners of the room are about equally tall). Turner et al. (1997) placed motion parallax and binocular disparity in conflict with each other in a surface detection task, with one cue signaling a surface and the other cue signaling points scattered randomly within a volume. Binocular disparity was weighted far more heavily, approaching a veto rule, than motion information, regardless of which cue was informative about the surface and despite the two cues being equally reliable when used in isolation.

The results of other experiments support strong fusion models with nonlinear combination rules more powerful than simple cue vetoing. Rogers and Collett (1989) found that when binocular disparity and motion parallax are placed in conflict in a shape judgment task, observers judged shape in accordance with disparity information, as in the Turner et al. (1997) experiment. However, fairly strong interaction between motion and stereo was implied by the percept of nonrigid motion. Nonrigid motion was also reported by observers in the Turner et al. experiments in trials where disparity and motion information were in conflict. Rather than simply vetoing the motion cue, the disparity information appeared to affect interpretation of the motion cue. A number of studies examining the interaction between stereoscopic depth displays and the kinetic depth effect (KDE) also seem to point toward a relatively strong model of depth cue combination (Nawrot & Blake, 1989, 1991, 1993). Retinal disparity can be used to disambiguate depth relations in otherwise ambiguous KDE displays, and adaptation and perceptual priming have been shown to transfer between stereoscopic and kinetic depth displays.

In summary, the current state of the literature suggests that the degree of interaction between cues may depend on the cues, the experimental conditions, and the task. One formidable possibility is that the visual system uses a bag of tricks to calculate depth, which would be difficult to model formally. However, most depth cues bear an orderly and lawful, albeit complicated, relationship to three-dimensional space. Given that, it is likely that human cue combination in depth perception is more orderly than implied by the expression "bag of tricks" and should be amenable to being modeled by some form of fully specified nonlinear model.

One difficulty in evaluating different models for depth cue combination is that strong and modified weak models are nonlinear and therefore difficult to analyze quantitatively. Computer simulations are a particularly useful way of examining visual cue combination when used as a complement to experimental investigations. They allow competing models to be evaluated quickly under a variety of conditions in a manner that permits detailed,

quantitative comparisons among different models. These comparisons can often reveal hidden or underspecified properties of qualitatively described theoretical models.

We present the results of simulations of three models for the combination of stereo, motion, and vergence angle cues for depth. The models were instances of a strong fusion model, a weak fusion model, and a modified weak fusion model. Investigators who advocate each of these three classes of models have omitted important details that are necessary if these models are to be specified fully and implemented. For example, investigators have failed to characterize the noise that corrupts the various visual signals that are used as inputs to the models. Consequently, when implementing the models, we have had to supply details that were not supplied by the theorists who originally proposed the models. In all cases, we have attempted to make sensible and straightforward choices, avoiding exotic, or at least less obvious, implementations of these models.

The goal of experiment 1 was to compare the performances of the three models so as to evaluate their relative plausibility as models of cue combination for both object depth and object shape perception. A variety of noise conditions such as flat noise and Weber noise were simulated because the noise model was expected to have a significant effect on performance. The goal of experiment 2 was to explore the modified weak fusion model more closely. In the case of depth perception, an important part of good cue combination is the ability to learn which cues are informative under which circumstances and to weight them accordingly. Using a pretrained model, we set either motion or stereo to always indicate a flat surface, while the other cue continued to indicate an ellipse. The cue indicating an ellipse was informative in the sense that the training feedback was always correlated with this cue; the cue indicating a flat surface was uninformative. The modified weak model successfully learned to reweight motion and stereo cues as a function of their informativeness. Overall, the simulations reported in this article suggest that the modified weak fusion model is a good model of the combination of motion and stereo signals relative to weak and strong fusion models. However, the results also highlight areas in which the modified weak fusion model needs modification or further elaboration.

2 Stimulus

The simulated stimulus was a two-dimensional ellipse whose width varied along the frontoparallel plane and whose depth varied along the line of sight (see Figure 1, panel A). Sixteen different ellipses were presented to each model; the width and depth of each ellipse varied independently and took values between 12 and 48 cm. The ellipse was positioned at one of eight viewing distances from the simulated observer, ranging between 72 and 408 cm. (Details of the stimulus are in appendix A.)

We simulated a point traveling around the perimeter of the ellipse at

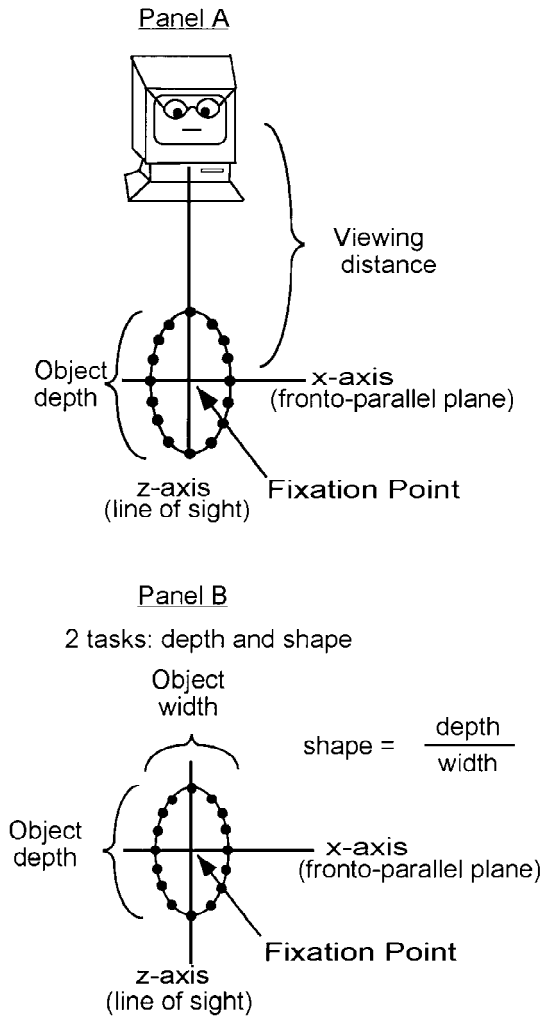

Figure 1: (Panel A) Illustration of the simulated stimulus. (Panel B) Illustration of the object shape task and the object depth task.

a constant velocity, rather like a train traveling around a track, instead of modeling the ellipse itself rotating. This was a different stimulus from that used by Johnston, Cumming, and Landy (1994) in their psychophysical experiments and is a less realistic stimulus than theirs, although it does produce a reliable impression of depth in human observers when extended in height (Perotti, Todd, Lappin, & Phillips, 1998; Jacobs & Fine, 1998). This

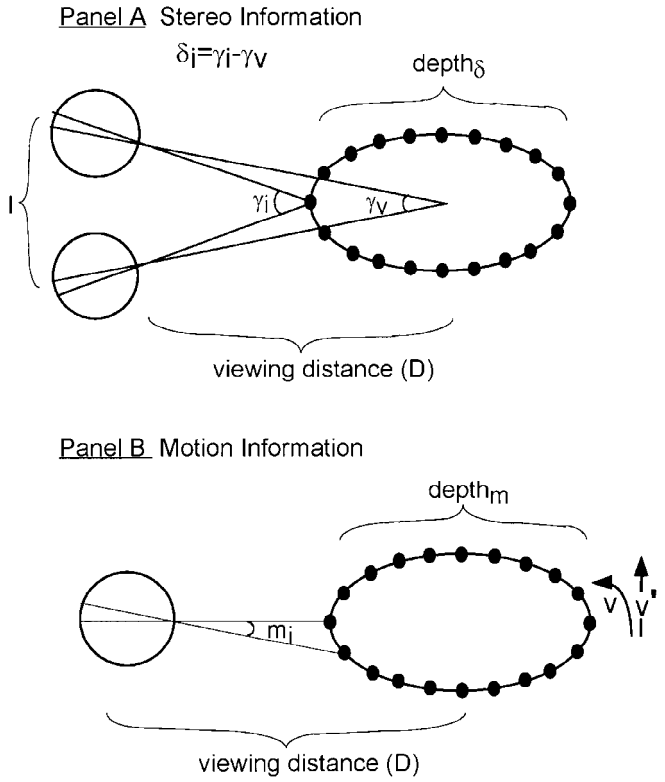

Figure 2: (Panel A) Illustration of the simulated stereo signal. (Panel B) Illustration of the simulated motion signal.

stimulus has the advantage that it avoids artifactual depth cues resulting from changes in retinal angle subtended by the ellipse over time. For each of 20 time slices of the point traveling around the perimeter of the ellipse, three sources of information were given to the simulated observers: stereo disparity, retinal motion, and vergence angle.

Stereo information consisted of the stereo disparity angle subtended by the point on the ellipse at each moment in time (see Figure 2, panel A). It was assumed that the simulated observer always fixated the center of the ellipse. Let the vergence angle γ_v be the angle between the lines connecting the fixation point and the centers of the left and right retinas. Let the angle γ_i be the angle between the lines connecting the location of the point on the ellipse at time step i and the images of this point on the left and right retinas. The stereo disparity at time step i, denoted δ_i, is equal to $\gamma_i - \gamma_v$.

Motion information consisted of the monocular retinal velocity of the

point at each moment in time expressed in degrees of retinal angle (see Figure 2, panel B). We assumed a cyclopean eye. The retinal velocity at time step i is the angle m_i between the lines connecting the aperture of the eye and the locations of the point on the ellipse at time steps $i - 1$ and i. The velocity of the point traveling around the ellipse was a function of the perimeter of the ellipse; the point traveled more slowly for ellipses with small perimeters and more quickly for ellipses with large perimeters. By choosing the point's velocity to be dependent on the perimeter of the ellipse, we removed artifactual depth and shape cues based on the overall magnitudes of the retinal velocities, and also prevented knowledge of the retinal velocities from being used as a cue from which viewing distance could be inferred.

The vergence angle (γ_v) of an observer fixated on the center of the ellipse was the third source of information given to the simulated observers. This angle was directly related to the viewing distance (D) through the equation

$$\gamma_v = 2 \tan^{-1}\left(\frac{I}{2D}\right), \tag{2.1}$$

where I is the interocular distance. We chose to use the vergence angle as one of a number of cues that observers appear to use to estimate viewing distance. There are a large number of cues for viewing distance, and viewing distance estimates appear to increase and grow more accurate as the number of cues increases. Bradshaw, Glennerster, and Rogers (1996) found that horizontal disparities were scaled by an estimate of the egocentric viewing distance that was approximately an additive function of vertical disparities and vergence angle. However, depth constancy was far from complete in their study, unlike those done with more naturalistic viewing conditions (Glennerster, Rogers, & Bradshaw, 1993; Durgin, Proffitt, Olsen, & Reinke, 1995), suggesting that other cues besides vergence angle and vertical disparities also provide viewing distance information.

Three noise conditions were examined: a Weber noise condition, a flat noise condition, and a velocity-uncertainty noise condition. In the Weber noise condition, motion, stereo, and the vergence angle were corrupted by additive gaussian noise whose distribution had a mean of zero and a standard deviation proportional to the signal magnitude (i.e., proportional to the disparity angle, the retinal motion, and the vergence angle).

In the flat noise condition, motion and stereo cues were corrupted by additive gaussian noise with mean zero and a constant variance, while the vergence angle was corrupted by Weber noise as in the Weber noise condition. Once again motion uncertainty was modeled as uncertainty about the retinal velocities.

An alternative way to model motion noise is as uncertainty about the *velocity of the moving point on the ellipse* rather than uncertainty about the *retinal* velocity. In the velocity-uncertainty condition, noise in the motion

cue was modeled as uncertainty about the velocity of the moving point on the ellipse. In this velocity-uncertainty condition, stereo and vergence angle signals were corrupted by noise with the same distribution as in the Weber condition, while the motion signals were corrupted by adding zero-mean gaussian noise to the velocities of the point traveling around the ellipse.

Weber noise was added to the vergence angle signal in all noise conditions because a Weber noise model is a conservative one, due to the vergence angle's being inversely related to viewing distance. In addition, a fourth condition was considered as a control. In this no-noise control condition, noise was not added to any of the cues. This condition was used to check that it was added noise that limited performance of the models. In all noise models, motion and stereo noise levels were set at values chosen to make stereo a slightly more reliable cue for judging the depth of an ellipse. These noise levels are consistent with psychophysical data (e.g., Rogers & Graham, 1982). (Table 1 in appendix A contains the equations used for the noise models.)

3 Tasks

The depth of an ellipse is the distance from the point on the ellipse closest to the observer to the point farthest away; its width is the distance from the left-most point to the right-most point (see Figure 1, panel B). The shape of an ellipse is defined as the ratio of the ellipse's depth to its width. This ratio is sometimes referred to as the *form ratio*. Cues from which shape can be calculated independently of absolute depth, width, or viewing distance are known as *scale-invariant cues*. Cues from which shape cannot be computed independent of such information are known as *scale-dependent cues*.

Motion is a scale-invariant cue because both width and depth scale linearly with viewing distance (see Figure 3). For example, an object of 40 cm depth at a viewing distance of 240 cm produces the same retinal motion signal as an object of 20 cm depth at half that viewing distance. Because width from motion also scales linearly with viewing distance, shape can be directly computed without explicit knowledge of the viewing distance. However, motion alone provides only a shape cue; without information about the viewing distance, or the size or velocity of the object, there is no way of inferring object depth.

In contrast to motion, stereo is not a scale-invariant cue. Although the width of an object indicated by retinal stereo disparities scales linearly, the depth of an object indicated by a given retinal signal scales with the square of the viewing distance (see Figure 3). The same disparity retinal signal indicates an object of 20 cm depth at a viewing distance of approximately 172 cm or an object of 40 cm depth at a viewing distance of 240 cm. Stereo disparities are therefore scale dependent; there is no way of inferring shape information independent of the viewing distance. Although stereo disparities are occasionally described as absolute depth cues, it is necessary to have

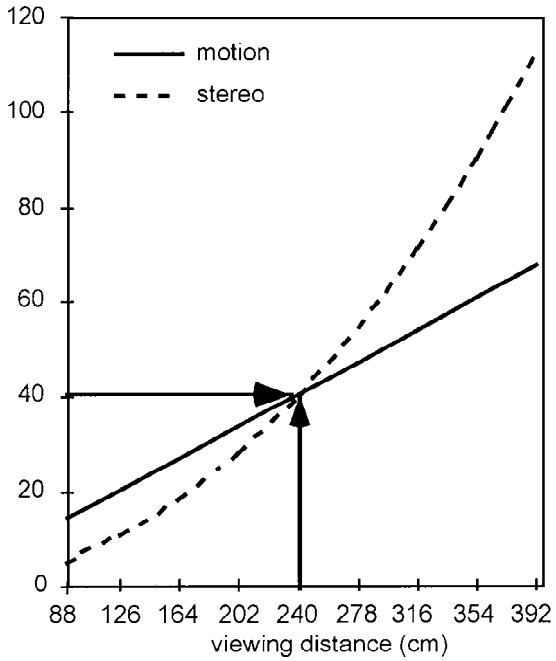

Figure 3: Scaling of motion and stereo retinal signals with distance from the observer.

an estimate of the vergence angle or the viewing distance to obtain either object depth or shape information from stereo information. This need to scale disparities by the viewing distance is referred to as the *stereo scaling problem*. As would be expected from the geometry, both Johnston (1991) and Durgin et al. (1995) have found evidence that depth estimates mediated by stereo disparities were scaled by the viewing distance estimate. In addition, Trotter, Celebrini, Stricanne, Thorpe, and Imbert (1992) found that responses of V1 cells were modulated by changes in the viewing distance.

Differences in the geometrical information provided by the scale-invariant cue of motion and the scale-dependent stereo cue motivated us to examine both an object depth task and an object shape task.

4 Models of Cue Combination

A series of nonlinear artificial neural networks trained using the backpropagation optimization algorithm were used to simulate the different observers. Each network performed a regression, possibly nonlinear, that mapped inputs to outputs. In this study, any reasonable regression procedure could

be used. In contrast to researchers who use neural networks for the purposes of biological modeling, our simulations were intended as a functional study of cue combination. Neural networks were used because they have a number of convenient computational properties. They show comparatively fast learning and good generalization on a wide variety of tasks (Chauvin & Rumelhart, 1995). Their theoretical foundations are also becoming increasingly better understood (e.g., Chauvin & Rumelhart, 1995; Smolensky, Mozer, & Rumelhart, 1996). In addition, they are efficient and easy to implement. Their parameter values can be estimated using a gradient-descent procedure in which the relevant derivatives are computed using an implementation of the chain rule known as the *backpropagation algorithm* (Rumelhart, Hinton, & Williams, 1986). The recursive nature of this algorithm makes neural networks efficient to run on relatively large-scale tasks and easy to program.

The instances of the strong fusion, weak fusion, and modified weak fusion models used in our simulations are illustrated in Figure 4. Each box in the panels represents an independent network, and the labeled lines represent the flow of information between the networks. With one exception, noted below, the networks have a generic form (an input layer fully connected to a hidden layer, which is fully connected to an output layer; the hidden units of the networks use the logistic activation function, and the output units use a linear activation function; the networks are trained to minimize the sum of squared-error objective function). The inputs to the networks were linearly scaled to fall in the interval between -1 and 1 (stereo disparities and retinal velocities) or between 0 and 1 (vergence angle); the desired outputs were scaled to fall in the interval between 0 and 1. Each network of each model was trained independently for 3000 epochs, and the networks were trained in their logical order (e.g., if the output of network A is an input to network B, then network A was trained before B). At the end of training, network performances had reached asymptote. In general, the simulations showed virtually no overfitting, possibly due to the fact that the noisy input signals prevented the networks from memorizing the training data. The number of hidden units and the learning-rate parameter for each network were optimized under the Weber noise condition in the sense that networks with fewer or more hidden units or with a different learning rate showed equal or worse generalization performance. (Further details of the simulations are provided in appendix A.)

Figure 4 (panel A) illustrates the strong fusion model. The model consisted of two networks. The first network (labeled "viewing distance") received an estimate of the vergence angle (γ_v) as input and calculated an estimate of viewing distance (d_v). The second network (labeled "unconstrained interaction") received as input a set of 20 stereo disparities (δ_i, $i = 1, \ldots, 20$), a set of 20 retinal velocities (m_i, $i = 1, \ldots, 20$), and the viewing distance estimate produced by the preceding network. The output was an estimate of either the depth or the shape of the ellipse. Because this network contained

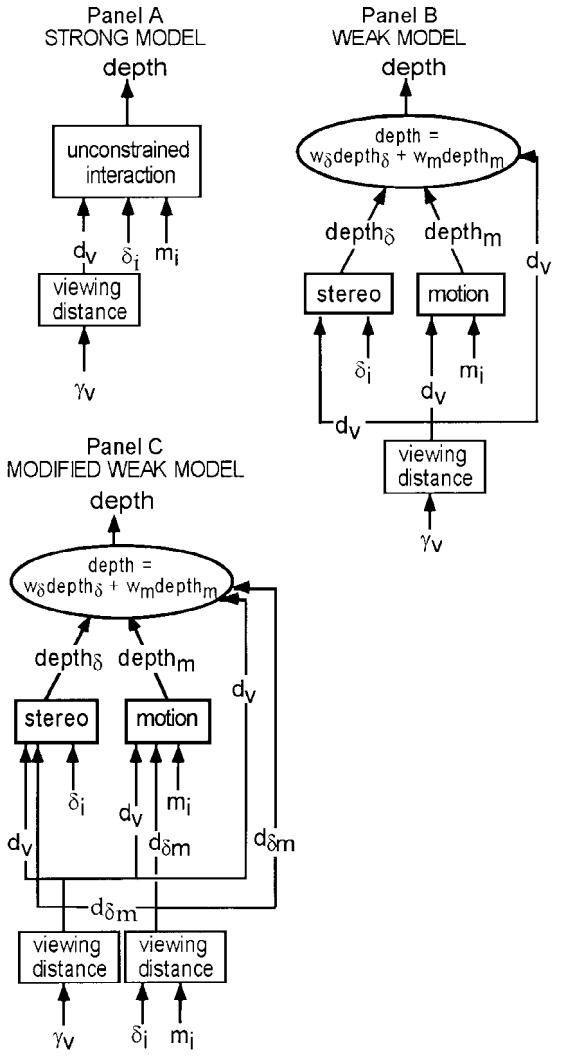

Figure 4: Instances of the strong fusion, weak fusion, and modified weak fusion models used in the simulations.

hidden units and was fully connected, the strong model was relatively unconstrained and could form high-order nonlinear combinations of stereo, motion, and vergence angle information.

The weak fusion model, shown in panel B, consisted of four underlying networks. The first network, like the first network in the strong model, re-

ceived as input the vergence angle (γ_v) and computed an estimate of the viewing distance (d_v). The stereo computation network used the viewing-distance estimate computed by the initial network (d_v) and the set of stereo disparities (δ_i) to estimate either the depth or the shape of the ellipse. The motion computation network used the viewing-distance estimate computed by the initial network (d_v) in conjunction with the set of 20 retinal velocities (m_i) to provide an independent estimate of ellipse depth or shape. The weighting network was given the estimate of viewing distance estimate (d_v) as input and then computed the linear coefficients (w_δ and w_m) used to average the outputs of the stereo ($depth_\delta$) and motion ($depth_m$) computation networks so as to produce the best final estimate of depth. For the object depth task, for example, the weighting network computed the weights w_δ and w_m as a function of the estimated viewing distance (d_v) using the equation

$$depth = (w_\delta \times depth_\delta) + (w_m \times depth_m), \tag{4.1}$$

where *depth* is the weak fusion model's final estimate of object depth, $depth_\delta$ is the output estimate of the underlying stereo computation network, $depth_m$ is the output estimate of the underlying motion computation network, and w_δ and w_m are, respectively, the weights used to average the output estimates of the stereo and motion networks. Whereas the other networks of the cue combination models have a generic form, the weighting network is nonstandard in the sense that its output unit is a sigma-pi unit (Rumelhart, Hinton, & McClelland, 1986). Specifically, the weighting network has four layers of units: an input layer, a hidden layer, a layer consisting of two units (the activations of these units are the values w_δ and w_m), and an output unit. The weights on the connections from the two units in the third layer to the output unit are set equal to the depth or shape estimates produced by the stereo computation network and motion computation network, respectively. Because the two units in the third layer use the logistic activation function, the weights w_δ and w_m are constrained to lie between zero and one; they are not constrained to sum to one.

Four of the five underlying networks of the modified weak fusion model (panel C of Figure 4) were nearly identical to the weak fusion model. It differed from the weak model in including one additional network that was used to model an instance of cue promotion. Johnston et al. (1994) found that the combination of stereo and motion cues helped human observers solve the stereo scaling problem when they were asked to choose which of a set of cylinders appeared circular. We modeled this combination of motion and stereo by including a network that mapped sets of stereo disparities (δ_i) and retinal velocities (m_i) to provide an additional estimate of the viewing distance ($d_{\delta m}$). Retinal velocities scale inversely with viewing distance, whereas stereo disparities scale inversely with the square of the viewing distance. Consequently there is only one object depth at one viewing distance that is consistent with both motion and stereo retinal signals (see Figure 3). By

combining motion and stereo disparity information, through this intersection of constraints, both object depth and viewing distance can be computed without the need for additional information, such as the vergence angle. In the modified weak model, limited nonlinear interaction between motion and stereo was allowed for the purpose of computing this additional estimate of the viewing distance ($d_{\delta m}$). This viewing-distance estimate was generally more accurate than the vergence-angle estimate (d_v) under the noise conditions studied. Under the Weber noise condition, for example, the correlation coefficient between the estimate of viewing distance d_v and the real viewing distance was 0.7821, while the correlation coefficient for $d_{\delta m}$ and the real viewing distance was 0.9166, corresponding to a root mean square (RMS) error nearly twice as large for d_v than $d_{\delta m}$. This improved stereo-motion viewing-distance estimate was used as an additional input to the motion, stereo, and weighting networks of the modified weak fusion model.

5 Experiment 1

The first experiment compared the performances of the different models (strong, weak, and modified weak models) on the two tasks (object shape and object depth tasks) under various noise conditions (Weber noise, flat noise, velocity-uncertainty noise, and no noise). Figures 5 and 6 show the results on the object shape task and object depth task, respectively. The two graphs in each figure show the models' performances in the Weber noise condition and in the no-noise condition. Performances in the flat and velocity-uncertainty noise conditions were very similar to those in the Weber noise condition and, thus, are not shown. The horizontal axis of each graph gives the model; the vertical axis gives the generalization performance at the end of training. The metric used to quantify generalization performance is the correlation between the actual output of a model and the target output (the real shape or depth of an ellipse) using a set of test patterns that differed from the patterns used during training. The error bars in the graphs give the standard error of the mean for 10 runs of each model.

None of the models we simulated had any difficulty in solving either the depth task or the shape task in the absence of noise, as shown by the comparatively good performance of each of the models in the no-noise control condition. Rather than lack of computational power, it was added noise that was the most significant factor limiting performance for each model. Good generalization performance was therefore based on the ability of each model to resolve ambiguity due to noise. This result highlights the seriousness of the problem mentioned above: that theorists proposing cue combination models have failed to specify noise conditions that are realistic and can be used to distinguish the relative strengths and weaknesses of competing cue combination models. In the absence of noise, widely different models all show good performance.

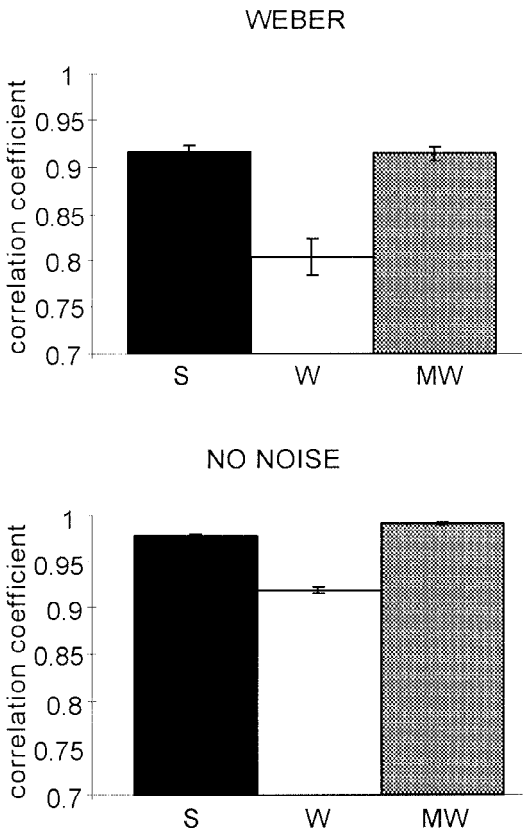

Figure 5: Generalization performances of the strong (S), weak (W), and modified weak (MW) models on the object shape task in the (top) Weber noise condition and (bottom) no-noise condition. Generalization performance was quantified as the correlation between a model's actual output and the target output using the set of test patterns. Standard error bars for 10 runs are shown.

The shape task was easier than the object depth task. As can be seen by comparing Figures 5 and 6, the generalization performances on the shape task were consistently better than those on the object depth task. Because the shape task was significantly easier for all three models, this result is unlikely to be due to a specific architectural property of a particular model. The results are also independent of the particular noise condition used. Shape is a scale-invariant property of objects, whereas object depth is susceptible to uncertainty in the viewing-distance estimate. It is the scale invariance of the shape task that makes it easier to solve.

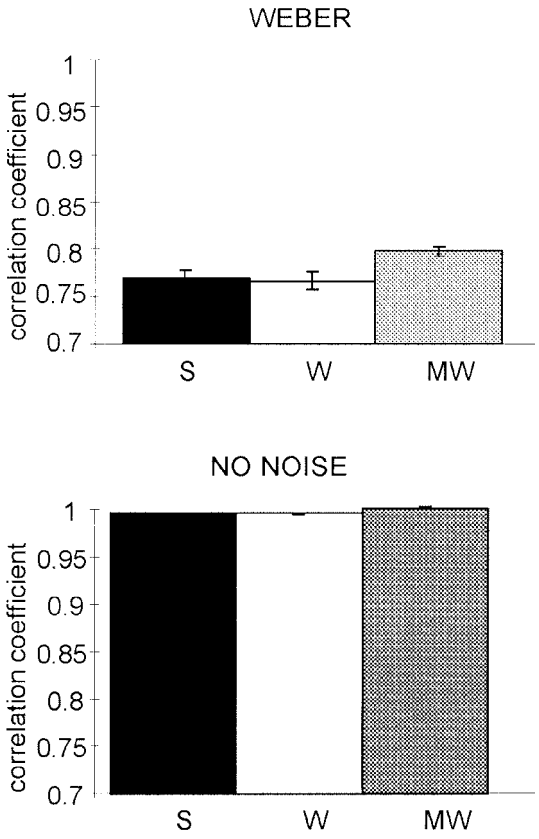

Figure 6: Generalization performances of the strong (S), weak (W), and modified weak (MW) models on the object depth task in the (top) Weber noise condition and (bottom) no-noise condition. Generalization performance was quantified as the correlation between a model's actual output and the target output using the set of test patterns. Standard error bars for 10 runs are shown.

The literature on visual perception often contains an implicit assumption that people use a single representation of three- dimensional space for all tasks (e.g., Gogel, 1990). Recent evidence suggests, however, that different tasks may involve the use of different spatial representations (e.g., Graziano & Gross, 1994). In particular, there are reasons to believe that observers have separate representations for the shape and depth of objects. The shape of objects is a useful cue for object recognition that is independent of distance-scaling effects, which provides a motive for representing shape independently of depth (Brenner, van Damme, & Smeets, 1997; Mishkin,

Ungerleider, & Macko, 1983). Our results show that the shape task is easier than the object depth task. Because object depth representations are necessarily susceptible to uncertainty in the viewing-distance estimate, making shape judgments dependent on object depth estimates would unnecessarily corrupt shape estimates. Separate representations could restrict the effects of uncertainty in viewing distance so that representations of scale- invariant properties are not needlessly corrupted.

Figures 5 and 6 also illustrate that the modified weak model showed the best performance in the object depth task and comparable performance to the strong model in the shape task. This was also the case in the flat and velocity-uncertainty noise conditions (not shown). This result is surprising because, in theory, the strong model should always be able to perform at least as well as the modified weak model due to the fact that it is less constrained. However, the strong model did not perform best; it seems that the complexity of the object depth task meant that the absence of built-in structure in the strong model allowed it to fall into relatively poor local minima of the error surface in the presence of noise during training. The addition of extra hidden units to the networks of the strong model did not remedy this problem.

In order to understand better the performances of the modified weak model relative to those of the strong model, we also simulated two variants of the strong model and one variant of the modified weak model. Recall that the strong model contains a network that maps the stereo and motion signals and the estimate of viewing distance based on the vergence angle (d_v) to estimates of object shape or object depth. In the first variant of the strong model, this network was also given as an input the estimate of viewing distance based on stereo and motion signals ($d_{\delta m}$). The generalization performances of this variant were nearly identical to those of the original strong model (the average correlation coefficients for the variant on the shape and depth tasks were 0.899 and 0.778; the corresponding values for the original strong model were 0.913 and 0.774).

In a second variant of the strong model, this network was given the viewing-distance estimate based on stereo and motion signals, but not the estimate based on the vergence angle (the first variant was given both of these estimates). This variant also did not perform better than the original strong model (its average correlation coefficients on the shape and depth tasks were 0.895 and 0.710). For the sake of completeness, we also simulated a variant of the modified weak model. In this variant, the networks of the model used the viewing-distance estimate based on stereo and motion signals, but not the estimate based on the vergence angle. This variant performed similar to the original modified weak model on the object shape task and worse than the original model on the depth task (the average correlation coefficients for the variant on the shape and depth tasks were 0.899 and 0.700; the corresponding values for the original modified weak model were 0.910 and 0.803). This outcome is surprising because the viewing-distance

estimate based on the stereo and motion signals, $d_{\delta m}$, is more accurate than the estimate based on the vergence angle, d_v. One probable explanation is that d_v, but not $d_{\delta m}$, is independent of noise in the stereo and motion cues, and this may be important for accurately estimating depth.

It should be emphasized that no strong conclusions can be drawn concerning the superiority of the modified weak model over the strong model (or any of the variants of it that we simulated). We suggest, however, that the superior performances of the modified weak model provide evidence that the constraints imposed on it are at least not overly restrictive. Although nontrivial constraints are imposed on the modified weak model, they do not seem to impair its ability significantly to find a satisfactory solution to both the shape and depth tasks.

The modified weak model performed significantly better than the weak model. This is because constraints imposed on the weak model prevented any interaction between motion and stereo cues. In the case of the modified weak model, constrained interaction between motion and stereo signals provided a relatively accurate estimate of the viewing distance. This accurate source of information about the viewing distance gave the modified weak model a significant advantage over the weak model.

The relatively good performance of the modified weak model suggests that the modularity constraints imposed on it (the model contains separate stereo and motion depth computation networks) do not prevent it from finding a good solution. The architecture of the modified weak model provides an adequate compromise between modularity and the power to combine cues, thereby showing both good performance and parsimonious design. Stereo and motion information could interact in a constrained manner to provide an additional estimate of viewing distance, while the overall architecture remained essentially modular.

Although the comparative simulation results suggest that the modified weak fusion model is a good candidate model of the combination of motion and stereo cues, further simulation results with this model indicate behaviors that are sensible from a computational viewpoint but inconsistent with existing psychophysical data. In this sense, the simulation results show shortcomings of the modified weak model. We highlight these shortcomings in order to provide a fair evaluation of the strengths and weaknesses of this model and to encourage advocates of the model to consider modifications that may make the model's behavior more consistent with psychophysical results.

Figure 7 gives the weighting of motion and stereo as a function of viewing distance for the different tasks for the modified weak model in the Weber noise condition. The horizontal axis represents the viewing distance, and the vertical axis represents the weights assigned to motion and stereo (w_m and w_δ in equation 4.1). The weightings of motion and stereo cues in flat and velocity-uncertainty noise conditions were similar to the weightings in the Weber noise conditions and therefore are not shown. As might be expected,

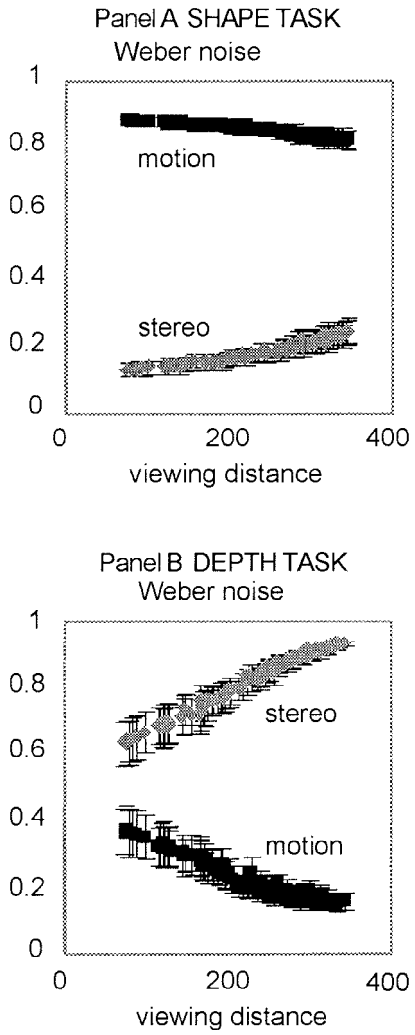

Figure 7: Weights assigned to motion and stereo information by the modified weak model as a function of viewing distance for the object shape and object depth tasks. Standard error bars for 10 runs are shown.

the weights added approximately to one over all distances for both depth and shape tasks in all noise conditions, although they were not constrained to do so.

In the case of the shape task (panel A of Figure 7), motion information was weighted far more heavily than stereo information for all three noise

conditions. This is consistent with the fact that retinal velocities provide a scale-invariant cue to shape. Because the motion cue to shape is not susceptible to noise in the viewing-distance estimate, it remains consistently the most reliable cue under all conditions tested. The weight assigned to stereo increased with viewing distance in all three noise conditions.

In the object depth task (panel B of Figure 7), the opposite results were found: stereo was weighted more heavily than motion for all three noise conditions. Again, the weight assigned to stereo increased with viewing distance for all three noise conditions.

That the weight assigned to stereo increased with distance is an unexpected finding because it is inconsistent with psychophysical data. The results differ from the psychophysical findings of Johnston et al. (1994), as well as those of several other investigators who found increased reliance on motion as the viewing distance increased (see Tittle et al., 1995, for a discussion). Johnston et al. (1994) explained this by arguing that motion is a more reliable cue at farther distances. The difference between the performance of the simulated modified weak model and that of the observers in Johnston et al.'s study is not easy to explain by assuming slightly different noise conditions for motion and stereo than the three we used. It is also not easy to explain by considering differences between the KDE displays used by Johnston et al. and the displays that we used. (Appendix B provides a lengthy discussion of these issues.) In short, analysis of the equations relating either motion or stereo information to estimates of object depth shows that for a point traveling around a fixed ellipse at a constant velocity, the depth estimates based on stereo become more accurate as the viewing distance increases relative to the depth estimates based on motion. Therefore, it is not the case that motion is providing more reliable information at greater viewing distances. One possible explanation of the difference between the simulation results reported here and the psychophysical data is that human observers have different biases in their estimates of viewing distance than those included in the modified weak model. The distance judgments of human observers tend to be biased toward viewing distances of approximately 1 meter; viewing distances less than this value tend to be overestimated, whereas viewing distances greater than this value tend to be underestimated. This phenomenon is known as the *specific distance tendency*. The study of Johnston et al., which reported that subjects relied more strongly on motion at farther viewing distances, used distances of 0.5 and 1.2 meters. It is likely that observers' estimates of viewing distance are more accurate at 1.2 meters than they are at 0.5 meter, and this may affect their relative use of motion and stereo. Our model, and the modified weak fusion model as outlined by Landy et al. (1995), does not include biases in viewing-distance estimates. Our simulation results suggest that advocates of this model may want to include such a mechanism in future versions.

As a final conclusion based on the results of experiment 1, we return to the issue of single versus multiple representations of visual space. Both the

modified weak model and the strong model performed better on the shape task than the depth task for all the noise conditions. The relative weighting of motion and stereo was significantly different for shape and depth tasks for all noise conditions and over a wide range of viewing distances. These differences between the shape and the depth task provide a source of motivation for having separate representations of object depth and object shape. Landy et al. (1995) proposed the existence of a depth map to which all cues were promoted. Our results motivate the additional existence of a shape map. Separate representations for the depth and shape of an object would permit independent cue weighting functions, allowing each judgment to be separately computed so as to minimize the effects of noise.

6 Experiment 2

Experiment 2 examined the ability of the modified weak model to compensate for changes in the relative usefulness of different cues. Landy et al. (1995) suggested that changes in the weights assigned to different cues for visual depth might serve to compensate nearly instantly for changes in their relative reliability. Young, Landy, and Maloney (1993) found that human observers altered the weights that they assigned to depth estimates based on texture and motion cues as a function of the cues' reliabilities. Turner et al. (1997) exposed observers to displays where either motion parallax or stereo disparity specified a three-dimensional sinusoidal corrugation in depth, while the other cue indicated random points scattered randomly within the volume. They found that performances on a depth judgment task improved when the observers were told whether motion or stereo was the informative cue. It is thought that this improvement in performance is due to a change in the relative degree to which observers relied on motion and stereo cues. Performance was better when the same cue was relevant for an entire block of trials than when the relevant cue changed on a trial-by-trial basis. This result suggests that a significant amount of cue reweighting might not occur instantaneously.

We began with a previously trained system that simulated the modified weak model. Either the stereo or the motion cue indicated an ellipse varying in width and depth; the other cue was set to indicate a flat surface on the fixation plane. The cue that indicated a flat surface was therefore uninformative as far as judging the depth or the shape of the ellipse was concerned. We examined the depth and shape estimates of the modified weak model when provided with this contradictory information from motion and stereo. We were interested in how the depth and shape estimates and the weights assigned to the different cues changed over time with additional training. We predicted that the weight assigned to the informative cue would increase at the expense of the weight assigned to the uninformative cue. We also predicted that the depth and shape estimates of the model would improve as the weight assigned to the informative cue increased. In the simulations

reported in this section, the weights assigned to stereo and motion were constrained to be nonnegative and to sum to one. In addition, we consider only the Weber noise condition (the results with the other noise conditions were qualitatively similar).

The weight assigned to the informative cue was examined over time for the object depth task. When motion was the informative cue, the weight assigned to motion (averaged over all test patterns) increased dramatically over about 300 pattern presentations. The opposite occurred when stereo was the informative cue, though the effect was less strong due to ceiling effects because the model had initially relied heavily on stereo information. Analogous results were found for the shape task. When stereo was the informative cue, the weight assigned to stereo significantly increased over time. The weight assigned to motion significantly increased when motion was the informative cue, although the model initially relied heavily on motion, and again, therefore, there were ceiling effects.

Figure 8 shows the depth estimates of the modified weak model as a function of real depth. The horizontal axis gives the real depth of an ellipse; the vertical axis gives the depth estimate produced by the model. The fine solid line along the diagonal of each graph represents perfect depth constancy. The solid circles in the graphs represent the depth estimates of the model when both stereo and motion were informative cues providing information about the depth of the ellipse. When both cues were informative, there was a small, consistent tendency to overestimate the depth of "shallow" ellipses and underestimate the depth of "deep" ellipses. We believe that this is due to the use of a set of training patterns in which, on average, a pattern represented an ellipse of moderate depth. The model learned to bias its estimates toward this average value.

The bottom graph shows the depth estimates of the model when the motion cue indicated a flat surface. Data shown are averaged over all the test patterns and, thus, are averaged over the full range of viewing distances. We predicted that initially (before the model received additional training allowing it to compensate for the fact that one cue was uninformative) the ellipse would appear shallower when one cue indicated a flat surface. The solid triangles represent the initial depth estimates of the model. As predicted, the slope of the function relating the depth estimates to the real depths of the ellipses is comparatively flat; the model strikingly underestimated the depths of the ellipses. This result is consistent with the common finding of underestimation of depth by human observers in reduced cue conditions (e.g., Bülthoff & Mallot, 1988; Landy et al., 1991). The shaded squares represent the depth estimates of the model after additional training. The model learned to rely almost entirely on the stereo cue. This curve approaches the depth-estimate function of the model when both cues were informative (solid circles), although there is a slightly greater tendency to underestimate the depth of deep ellipses and overestimate the depth of shallow ellipses. The gray diamonds represent the depth estimate of the model

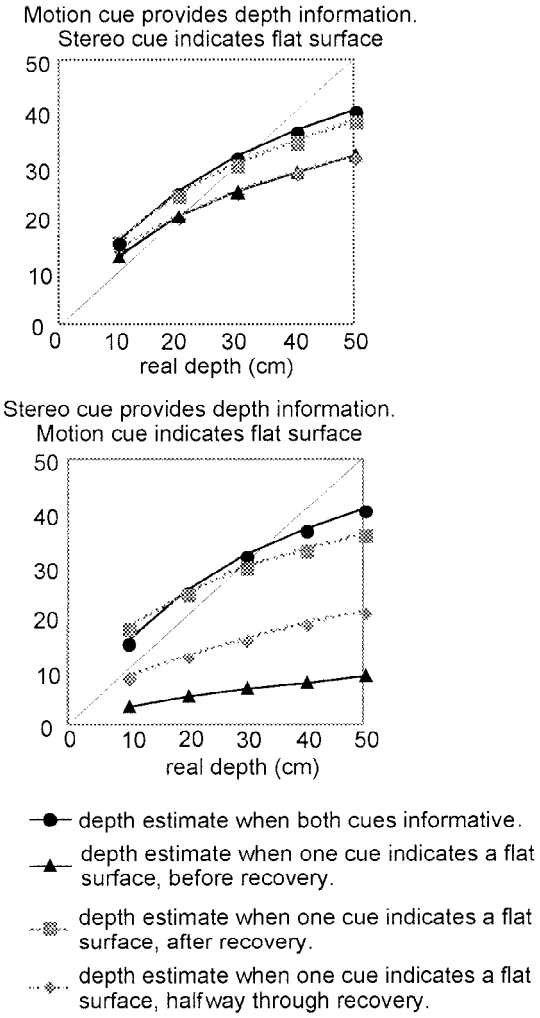

Figure 8: Depth estimates of the modified weak fusion model as a function of the real depth of an ellipse. (Top) The case when motion was the informative cue and the stereo cue indicated a flat surface. (Bottom) The case when stereo was the informative cue and motion indicated a flat surface. Standard error bars for 10 runs are smaller than the symbols.

halfway through the additional training period. As might be expected, the curve falls halfway between the initial depth estimates of the model and the estimates at the end of additional training. This improvement in performance over time is due to the fact that the model learned to reweight motion and stereo cues so as to rely more heavily on the informative cue. Qualitatively similar results were found when the stereo cue was uninformative (see the top graph of Figure 8).

Figure 9 shows the shape estimates of the modified weak model as a function of real shape. The horizontal axis gives the real depth-to-width ratio of an ellipse (normalized by the maximum depth-to-width ratio), and the vertical axis gives the depth-to-width ratio estimate of the model (also suitably normalized). The fine solid line along the diagonal represents perfect shape constancy. When stereo indicated a flat surface and motion was the informative cue (top graph), there was a small, consistent tendency to underestimate the depth of deep ellipses. These data resemble psychophysical performance in several studies on motion parallax that revealed a similar tendency by human observers to underestimate the depth of objects whose depth was greater than their width (Braunstein & Tittle, 1988; Caudek & Proffitt, 1993; Ono & Steinbach, 1990). Caudek and Proffitt (1993) speculated that observers were using a compactness assumption—an assumption that objects are about as deep as they are wide. Our simulation data, however, reveal that another possible cause is the reduced cue conditions used in the psychophysical experiments. It may be that subjects used a "flatness" assumption: observers interpret the absence of a visual cue to depth that normally appears in an environment as indicative of a lack of depth. In our simulations, underestimation of the depth of deep ellipses increased when either cue indicated a flat object (there is also an increase in the underestimation of the depth of shallow ellipses, though it is less easily noticed for these objects because of their small depths). Similarly, human observers may have interpreted the absence of expected cues, such as stereo or texture information, as indicative of a lack of depth, causing them to underestimate the depth of deeper ellipses.

The solid circles in Figure 9 represent the initial shape estimates of the model when both stereo and motion were informative cues; the shaded squares represent the model's shape estimates after additional training during which one cue was made uninformative. Overall, performance was better when both cues were informative, as would be expected. However, shape estimates for the very deepest ellipses were more accurate after recovery in the case when motion was the only informative cue than when both cues were informative (top graph). Performance for these deepest ellipses improved when the model was encouraged to use motion information alone rather than both motion and stereo information. Again, this is consistent with the fact that motion is a scale-invariant cue to object shape and stereo is not.

Although there have been relatively few studies of how human observers

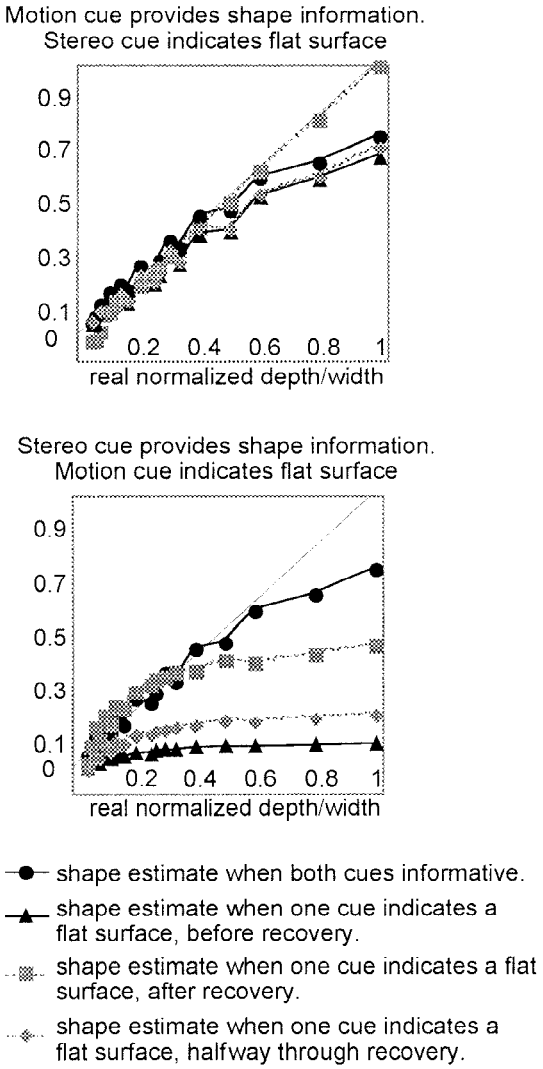

Figure 9: Shape estimates of the modified weak fusion model as a function of the real shape of the ellipse. (Top) The case when motion was the informative cue and the stereo cue indicated a flat surface. (Bottom) The case when stereo was the informative cue and motion indicated a flat surface. Standard error bars for 10 runs are smaller than the symbols.

compensate for reduced cue conditions, examination of the behavior of the modified weak fusion model reveals behavior that is qualitatively similar to psychophysical data in certain respects. For example, Turner et al. (1997) found that when human observers discriminated a surface from points scattered randomly within a volume, they were capable of good performance with motion or stereo information alone. However, when motion was the reliable cue, the presence of stereo as an unreliable cue impaired performance significantly. When the same cue was reliable through an entire block of trials, performance improved, suggesting that observers learned to reweight their relative reliance on motion and stereo over time. These experimental results are similar to the simulation results found using the modified weak model. The presence of a cue for "flatness" initially leads the model to underestimate both shape and depth in a manner that resembles psychophysical data collected under reduced cue conditions. The modified weak fusion model is capable of learning to reweight cues in order to use reliable cue information more extensively, similar to human observers. Because the modified weak model is broadly consistent with the limited amount of psychophysical data available, we tentatively conclude that the modified weak fusion model may provide a good model of how human observers learn to compensate for changes in cue informativeness.

7 Summary

Recent years have seen a proliferation of new theoretical models of cue combination, especially in the domain of depth perception. This proliferation is partly due to a poor understanding of existing models and partly due to a lack of comparative studies revealing the relative strength and weaknesses of competing models. Three models of visual cue combination were simulated: a weak fusion model, a modified weak model, and a strong model. Experiment 1 compared the performances of the three models on a shape judgment task and an object depth task. The results suggest that the constrained nonlinear interaction of the modified weak model allows better performance than either the linear interaction of the weak model or the unconstrained nonlinear interaction of the strong model. It seems, therefore, that the modified weak fusion model represents a good compromise between the need for modularity and the need for cue interaction. Further examination of the modified weak model revealed that its relative weighting of motion and stereo cues was dependent on the task, the viewing distance, and, to a lesser degree, the noise model. Although the dependencies were sensible from a computational viewpoint, they were sometimes inconsistent with psychophysical experimental data. The fact that different weightings were used for different tasks suggests that it is sensible for human observers to use multiple representations of visual space.

Experiment 2 examined the ability of the modified weak model to compensate for changes in the relative usefulness of different cues. It was found

that the model is capable of learning to reweight cues in order to use reliable cue information more extensively, similar to human observers. Overall, the simulation results suggest that, relative to the weak and strong models, the modified weak fusion model is a good candidate model of the combination of motion, stereo, and vergence angle cues, although the results also highlight areas, such as the specification of noise models, in which this model needs modification or further elaboration.

Appendix A

This appendix provides details of the simulations that were not included in the main body of the text. The set of training patterns was based on ellipses varying between 10 and 50 cm in width and depth and viewing distances between 69 and 411 cm. The test data were based on ellipses varying between 12 and 48 cm in width and depth and viewing distances varying between 72 and 408 cm. Training patterns were presented randomly, and the network weights were updated after each pattern presentation using the backpropagation algorithm. Ten independent runs were simulated for each task for each model.

Three noise conditions were considered: Weber noise, flat noise, and velocity-uncertainty noise. The noise distributions were always gaussian with a mean of zero; the three conditions differed in terms of the variances of the noise distributions and the signals that were corrupted by noise. In the Weber and flat noise conditions, the stereo signals (δ_i, $i = 1, \ldots, 20$), motion signals (m_i, $i = 1, \ldots, 20$), and vergence angle signal (γ_v) were corrupted by noise; the variances of the noise differed in the different conditions. In the velocity-uncertainty condition, the stereo and vergence angle signals were corrupted by noise with the same distribution as in the Weber condition; the motion signals, however, were corrupted by adding zero-mean gaussian noise to the velocities (v_i, $i = 1, \ldots, 20$) of the point traveling around the ellipse. The equations characterizing the variances of each of these noise conditions are provided in Table 1.

The number of hidden units and the learning-rate parameter for each network were optimized under the Weber noise condition in the sense that networks with fewer or more hidden units or with a different learning rate showed equal or worse generalization performance. The network that mapped the vergence angle to an estimate of viewing distance had 1 input unit, 25 hidden units, and 1 output unit. The network in the strong model that mapped the estimate of viewing distance, the motion signal, and the stereo signal to an estimate of shape or object depth had 41 input units, 40 hidden units, and 1 output unit. In the weak model, the networks that mapped the estimate of viewing distance and either the motion or stereo signals to an estimate of shape or depth had 21 input units, 15 hidden units, and 1 output unit. The corresponding networks in the modified weak fusion model were identical except that they had 22 input units (the extra input

Table 1: Equations Characterizing the Variances of the Weber Noise, Flat Noise, and Velocity-Uncertainty Noise Conditions.

Weber	Flat	Velocity Uncertainty
$\sigma_{si}^2 = (k_\delta \delta_i)^2$	$\sigma_{si}^2 = (\frac{1}{2} k_\delta)^2$	$\sigma_{si}^2 = (k_\delta \delta_i)^2$
$\sigma_{mi}^2 = (k_m m_i)^2$	$\sigma_{mi}^2 = (\frac{1}{2} k_m)^2$	$\sigma_{vi}^2 = (k_m v)^2$
$\sigma_{\gamma_v}^2 = (k_{\gamma_v} \gamma_v)^2$	$\sigma_{\gamma_v}^2 = (k_{\gamma_v} \gamma_v)^2$	$\sigma_{\gamma_v}^2 = (k_{\gamma_v} \gamma_v)^2$

Note: δ_i denotes the stereo signals, m_i denotes the motion signals, v denotes the velocity of the point traveling around the ellipse, and γ_v denotes the vergence angle. The variance of the noise added to the ith stereo signal is denoted σ_{si}^2; the variance of the noise added to the ith motion signal is denoted σ_{mi}^2; the variance of the noise added to the ith velocity signal is denoted σ_{vi}^2; and the variance of the noise added to the vergence angle is denoted $\sigma_{\gamma_v}^2$. The constants k_δ, k_m, and k_{γ_v} were used to scale the variances. The coefficient of a half in the flat condition was used to equalize approximately the variance of the noise in flat and Weber noise conditions.

is the estimate of viewing distance based on motion and stereo signals). The network in the modified weak model that mapped motion and stereo signals to an estimate of viewing distance had 40 input units, 16 hidden units, and 1 output unit. The network in the weak model that computed the weights used to average the depth or shape estimates based on stereo or motion signals (w_δ and w_m in equation 2.1) had 1 input unit, a layer of 17 hidden units followed by a layer of 2 hidden units (the activations of these units were the weights w_δ and w_m), and 1 output unit. The corresponding network in the modified weak model was identical except that it had 2 input units.

Appendix B

Some researchers have claimed that motion is a more reliable cue to object depth than stereo at greater viewing distances (see Durgin et al., 1995; Johnston et al., 1994). This appendix analyzes the equations relating either motion or stereo information to object depth in order to show that for a point traveling around a fixed ellipse at a constant velocity, the depth estimates based on stereo become more accurate as the viewing distance increases relative to the depth estimates based on motion. Therefore, it is not the case that motion is providing relatively more reliable information at greater viewing distances.

For the sake of brevity, we consider only the flat noise condition (similar results are found using the Weber noise condition). The appendix first considers the variance of the object depth estimates when noise is added to

the stereo and motion signals but not to the vergence angle signal. Then it considers the case when all signals are corrupted by noise.

Consider object depth estimates based on stereo information first. Using the small angle approximation, it is the case that

$$depth_{\delta} \approx \frac{I}{\gamma_f} - \frac{I}{\gamma_n}, \tag{B.1}$$

where $depth_{\delta}$ is the object depth estimate and I is the interocular distance (using cm as the unit of measurement), and γ_f and γ_n are the angles subtended by the points on the ellipse farthest from and nearest to the observer (see Figure 2). The only variables in this equation that change with viewing distance are γ_f and γ_n. The dependencies of these quantities on the viewing distance are given by (again using the small angle approximation)

$$\gamma_f \approx \frac{I}{D + \frac{d}{2}} \tag{B.2}$$

and

$$\gamma_n \approx \frac{I}{D - \frac{d}{2}}, \tag{B.3}$$

where D is the viewing distance (in cm) and d is the depth of the ellipse (in cm).

Now consider object depth estimates based on motion information (using the small angle approximation):

$$depth_m \approx \frac{v'}{m_f} - \frac{v'}{m_n} \tag{B.4}$$

where $depth_m$ is the object depth estimate, v' is the component of the moving point's velocity (in cm per frame) that is parallel to the frontoparallel plane, and m_f and m_n are the retinal velocities (expressed in degrees of retinal angle per frame) when the point is at the locations on the ellipse farthest from and nearest to the observer. The only variables in this equation that change with viewing distance are m_f and m_n; the dependencies are given by

$$m_f = \frac{v'}{D + \frac{d}{2}} \tag{B.5}$$

and

$$m_n = \frac{v'}{D - \frac{d}{2}}. \tag{B.6}$$

Comparisons of equations B.1 and B.4, B.2 and B.5, and B.3 and B.6 indicate that object depth estimates from stereo information and from motion information scale similarly with viewing distance. Indeed, they scale identically except for a scaling factor.

When noise is added to the stereo and motion cues, it ought to be the case that the variances of the depth estimates based on stereo signals and on motion signals scale similarly with viewing distance. Consider the flat noise condition in which the noise added to the stereo and motion signals has a fixed distribution (for the moment, there is no noise added to the vergence angle). Using the fact that the disparity δ_i is equal to $\gamma_i - \gamma_v$, and the fact that in the flat noise condition zero-mean gaussian noise with variance σ_δ^2 is added to the disparity δ_i, equation B.1 can be rewritten as:

$$depth_\delta \approx \frac{I}{\gamma_v + (\delta_f \pm \sigma_\delta)} - \frac{I}{\gamma_v + (\delta_n \pm \sigma_\delta)} \tag{B.7}$$

$$\approx \frac{I}{\gamma_f \pm \sigma_\delta} - \frac{I}{\gamma_n \pm \sigma_\delta}. \tag{B.8}$$

For the motion cue, zero-mean gaussian noise with variance σ_m^2 is added to the retinal angle m_i. Equation B.4 can be rewritten as:

$$depth_m \approx \frac{v'}{m_f \pm \sigma_m} - \frac{v'}{m_n - \sigma_m}. \tag{B.9}$$

Inspection of equations B.8 and B.9 shows that the variances of the depth estimates based on stereo information and on motion information scale identically with viewing distance when the cues are corrupted by noise, except for a scaling factor.

The influences of noise on object depth estimates are not easy to ascertain by visual inspection of the relevant equations when noise is added to the vergence angle signal, as well as the stereo and motion signals. We have therefore conducted numerical analyses by plugging numbers into the equations and plotting the results. The equations used in these analyses are those in this appendix (though without the small angle approximation) and equation 2.1 in the main body of the text. We used a fixed ellipse with a point traveling around the ellipse at a constant velocity. The magnitude of the noise added to (or subtracted from) the motion signals was set equal to σ_{mi} (as defined in the flat noise condition; see appendix A); similarly, the magnitude of the noise used to corrupt the stereo signals was set equal to $\sigma_{\delta i}$, and the magnitude of the noise used to corrupt the vergence angle signal was set equal to σ_{γ_v}. Nine viewing distances were considered, spanning the range used in the simulations.

The results are shown in Figure 10. The horizonal axis of panel A gives the viewing distance in centimeters; the vertical axis gives the object-depth

estimate (the true object depth is 10 cm). Let $d_m^{\max}$ and $d_m^{\min}$ denote the largest and smallest depth estimates at a given viewing distance produced using combinations of noisy motion and vergence angle signals for a fixed amount of noise (for example, it may be that the largest depth estimate is produced when noise is added to the motion signals and subtracted from the vergence angle signal, whereas the smallest estimate is produced when noise is subtracted from the motion signals and added to the vergence angle signal). Similarly, let $d_\delta^{\max}$ and $d_\delta^{\min}$ be the largest and smallest depth estimates produced using combinations of noisy stereo and vergence angle signals. As is shown in panel A, with very short viewing distances (around 80 cm), depth estimates based on noisy motion and vergence angle signals are slightly more accurate than depth estimates based on noisy stereo and vergence angle signals. However, for all other viewing distances, depth estimates based on stereo signals are more accurate than those based on motion signals.

Define the motion and stereo errors at a given viewing distance, denoted ϵ_m and ϵ_δ, as follows:

$$\epsilon_m = \frac{1}{2}\left(|d_m^{\max} - d| + |d_m^{\min} - d|\right) \tag{B.10}$$

$$\epsilon_\delta = \frac{1}{2}\left(|d_\delta^{\max} - d| + |d_\delta^{\min} - d|\right), \tag{B.11}$$

where d is the true object depth. Define the accuracies of the depth estimates based on motion signals and on stereo signals as the reciprocals of the squared corresponding errors (ϵ_m^{-2} and ϵ_δ^{-2}). Finally, define the motion and stereo weights:

$$w_m = \frac{\epsilon_m^{-2}}{\epsilon_m^{-2} + \epsilon_\delta^{-2}} \tag{B.12}$$

$$w_\delta = \frac{\epsilon_\delta^{-2}}{\epsilon_m^{-2} + \epsilon_\delta^{-2}}. \tag{B.13}$$

In the case of the simulations, where the amount of noise is a random variable (rather than a single fixed value as in this appendix), we would expect the weights of the motion and stereo cues to be inversely related to their relative variances. The weights in equations B.12 and B.13, based on the relative accuracies of the depth estimates from motion and stereo signals for a fixed amount of noise, are shown in panel B. The horizontal axis gives the viewing distance; the vertical axis gives the weights. Consistent with the neural network simulation results (see Figure 7), the weight assigned to stereo increases with viewing distance, whereas the weight assigned to motion decreases.

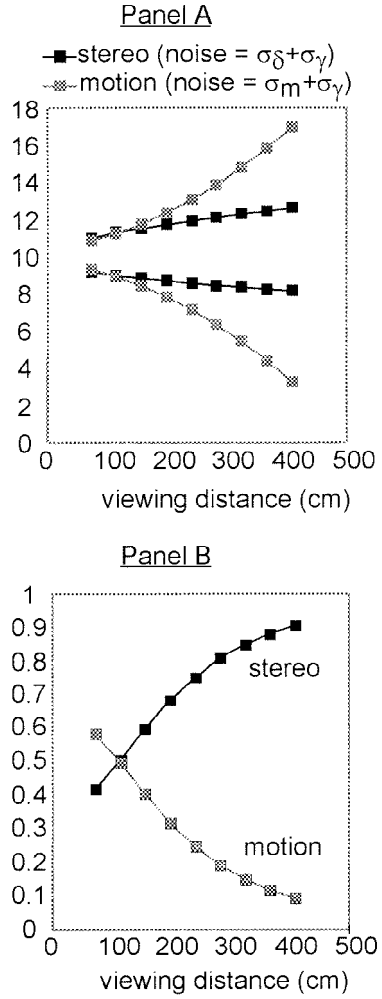

Figure 10: (Panel A) The upper and lower curves of shaded dots give the object depth estimates $d_m^{\max}$ and $d_m^{\min}$ produced when noise corrupts the motion and vergence angle signals; the upper and lower curves of solid dots give the depth estimates $d_\delta^{\max}$ and $d_\delta^{\min}$ produced when noise corrupts the stereo and vergence angle signals. (Panel B) The weights assigned to motion and stereo.

Acknowledgments

We thank R. Aslin for many useful discussions and for commenting on an earlier version of this article. This work was supported by NIH research grant R29-MH54770.

References

Blake, A., Bülthoff, H. H., & Sheinberg, D. (1993). Shape from texture: Ideal observers and human psychophysics. *Vision Research, 33*, 1723–1737.

Bradshaw, M. F., Glennerster, A., & Rogers, B. J. (1996). The effect of display size on disparity scaling from differential perspective and vergence cues. *Vision Research, 36*, 1255–1264.

Braunstein, M. L., & Tittle, J. S. (1988). The observer-relative velocity field as the basis for effective motion parallax. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 582–590.

Brenner, E., van Damme, W. J. M., & Smeets, J. B. J. (1997). Holding an object one is looking at: Kinesthetic information on the object's distance does not improve visual judgment of its size. *Perception and Psychophysics, 59*, 1153–1159.

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

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America, 5*, 1749–1758.

Caudek, C., & Proffitt, D. R. (1993). Depth perception in motion parallax and stereokinesis. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 32–47.

Chauvin, Y., & Rumelhart, D. E. (1995). *Backpropagation: Theory, architectures, and applications*. Hillsdale, NJ: Erlbaum.

Clark, J., & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Norwell, MA: Kluwer.

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

Durgin, F. H., Proffitt, D. R., Olsen., J. T., & Reinke, K. S. (1995). Comparing depth from motion with depth from binocular disparity. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 679–699.

Glennerster, A., Rogers, B. J., and Bradshaw, M. F. (1993). The constancy of depth and surface shape for stereoscopic surfaces under more naturalistic viewing conditions. *Perception, 22* (suppl.), 118.

Gogel, W. C. (1990). A theory of phenomenal geometry and its applications. *Perception and Psychophysics, 48*, 105–123.

Graziano, M. S. A., & Gross, C. G. (1994). Multiple representations of space in the brain. *Neuroscientist, 1*, 43–50.

Jacobs, R. A., & Fine, I. (1998). Integration of texture and motion cues to depth is adaptable. *Investigative Ophthalmology and Visual Science, 39*, S670.

Johnston, E. B. (1991). Systematic deviations of shape from stereopsis. *Vision Research, 31*, 1351–1360.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of motion and stereopsis cues. *Vision Research, 34*, 2259–2275.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35*, 389–412.

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

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414–417.

Nawrot, M., & Blake, R. (1989). On the perceptual identity of dynamic stereopsis and kinetic depth. *Science, 244*, 716–718.

Nawrot, M., & Blake, R. (1991). The interplay between stereopsis and structure from motion. *Perception and Psychophysics, 49*, 320–344.

Nawrot, M., & Blake, R. (1993). On the perceptual identity of dynamic stereopsis and kinetic depth. *Vision Research, 33*, 1561–1571.

Ono, H., & Steinbach, M. J. (1990). Monocular stereopsis with and without head movement. *Perception and Psychophysics, 48*, 179–197.

Perotti, V. J., Todd, J. T., Lappin, J. S., & Phillips, F. (1998). The perception of surface curvature from optical motion. *Perception and Psychophysics, 60*, 377–388.

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

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

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations*. Cambridge, MA: MIT Press.

Smolensky, P., Mozer, M. C., & Rumelhart, D. E. (1996). *Mathematical perspectives on neural networks*. Hillsdale, NJ: Erlbaum.

Tittle, J. S., Todd, J. T., Perotti, V. T., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology, 21*, 663–678.

Trotter, Y., Celebrini, S., Stricanne, B., Thorpe, S., & Imbert, M. (1992). Modulation of stereoscopic processing in primate area V1 by the viewing distance. *Science, 257*, 1279–1281.

Turner, J., Braunstein, M. L., & Anderson, G. J. (1997). The relationship between binocular disparity and motion parallax in surface detection. *Perception and Psychophysics, 59*, 370–380.

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.

Received August 26, 1997; accepted November 5, 1998.