Combining Visual Cues to Depth and Shape: A Comparison of Three Models

Ione Fine
Center for Visual Science
University of Rochester
Rochester, NY 14627
fine@cvs.rochester.edu

Robert A. Jacobs
Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
robbie@bcs.rochester.edu

Abstract

Performance in estimating the depth and shape of an ellipse on the basis of stereo, motion, and vergence angle information was compared for three models of visual depth cue combination. The three models were a weak model (strict modularity, with no interaction between motion and stereo cues), a modified weak model (restricted interaction allowed between motion and stereo cues), and a strong model (unconstrained interaction between all visual cues). Results are that the modified weak model performed best overall indicating that its structure, which contains both modular and interactive features, has advantages over both the extreme modular organization of the weak model and the extreme interactive organization of the strong model. In addition, the different weighting of motion and stereo cues by the modified weak model in the depth and shape judgment tasks provides a motivation for multiple visual representations of three-dimensional space.

Introduction

Recent years have seen a proliferation of new models of visual cue combination, especially in the domain of depth perception. This proliferation is due to a poor understanding of existing models, and to a lack of comparative studies that reveal the strengths and weaknesses of competing models. This paper studies how multiple visual cues may be combined to provide information about the three-dimensional structure of the environment. We are particularly concerned with two related computational issues.

The first issue concerns the relationship between representations of three-dimensional space and the task that an observer performs. There is often an implicit assumption in the literature that people use a single representation of space. Such a view has been put forward explicitly by Gogel (1990), and is often taken as a default simplification by other investigators.

Psychophysical and physiological evidence suggests, however, that different tasks may involve the use of different spatial representations. Philbeck and Loomis (1997) found that observers were capable of accurately estimating the egocentric viewing distance to a point when asked to walk to it blindfolded, but showed systematic biases that were dependent upon viewing distance when asked to verbally estimate the depth-to-width ratio of a pair of perpendicular sticks. They suggested that different representations of 3-D space were involved in the two judgments. Gross and Graziano (1994) found neural maps in the parietal cortex of monkeys that were centered on different body parts, such as the arm or eye. They argued that different tasks place different demands on the monkey's sensorimotor system. These demands are met through the use of multiple neural maps representing objects in space using coordinate frames centered on different body parts.

This paper examines a second motivation for multiple representations of three-dimensional space, besides that suggested by Graziano and Gross. We examined differences in the weighting of motion and stereo in object shape and object depth judgment tasks, and concluded that the need to weight the two depth cues differently for the two tasks provides a motivation for having separate representations for the shape and depth of objects. It is sensible to use different combinations of motion and stereo cues for shape and depth judgments. Motion signals provide a cue to shape that do not need to be scaled with viewing distance, whereas stereo signals do need to be appropriately scaled (see below). Consequently, motion signals should be weighted more heavily than stereo signals for all viewing distances when performing a shape judgment task. In contrast, both motion and stereo signals need to be scaled with viewing distance when judging the depth of an object. Therefore stereo should be weighted more heavily when making depth judgments than shape judgments. The fact that motion provides a scale-invariant cue for shape suggests that shape judgments should be easier to make than depth judgments.

A second issue addressed in this paper is the question of how representations of 3-D space are constructed from the many visual depth cues that are available. We focus on the trade-off between modularity and fusion. Investigators have long been aware that modular systems have several advantages over non-modular systems. First, some components of modular systems operate relatively independently of other components, and so modular designs tend to have fewer parameters than non-modular designs. Their more parsimonious organization makes them more constrained, and simpler to understand. Second, modular systems often learn faster and recover from damage more quickly. Because the parameters associated with one module are relatively decoupled from the parameters associated with other modules, changes in one portion of the system do not necessitate changes in other portions. Lastly, some theorists

have argued that modular designs make more efficient use of neural hardware than non-modular designs (e.g., Barlow, 1986; Cowey, 1981). For these reasons, many researchers have favored cue combination models in which modularity is preserved. These models typically possess a separate processing module for each visual cue. Information gained from each cue processed in isolation is only combined in the last stage of the system.

Unfortunately, highly modular systems tend to be non-robust because visual cues considered in isolation can be highly ambiguous. Consequently, other researchers have favored combination schemes that encourage interactive processing of multiple cues so that information based on one cue may be used to disambiguate the interpretations of other cues. Interactive models allow the visual system to combine information based on a range of cues in order to eliminate, or at least ameliorate, the ambiguity inherent in the information carried by a single cue. However, such models are difficult to study; their highly nonlinear nature makes them unconstrained and difficult to analyze. The choice of a model often involves a compromise between the competing advantages of modularity and cue fusion. Ideally, one might wish to characterize visual processing using a model that has both interactive and modular features.

Landy, Maloney, Johnston, and Young (1995) defined three classes of models for combining visual cues for depth. *Strong* 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 cue considered in isolation. These estimates are then linearly averaged to yield a composite depth estimate. The coefficients of the linear weighting of the different cues are proportional to each cue's reliability. *Modified weak* fusion models combine aspects of interactive and modular processing. Constrained nonlinear interaction, described as cue promotion, is permitted between different cues. Most cues are incapable of providing absolute depth information when considered in isolation; for example, occlusion only provides depth order information, and motion parallax only provides shape information. However, once the necessary missing parameters are specified, these cues become capable of providing absolute depth information. Cue promotion allows the determination of these missing parameter values through the use of other cues. For example, motion parallax is an absolute depth cue only if the viewing distance is known. Two ways this missing information could be specified are through knowledge of the vergence angle, or through a combination of stereo and motion parallax information. In a modified weak model, nonlinear cue promotion, in which information from different cues is combined in order to promote each cue so that it can provide an absolute depth map, is followed by a linear stage in which a weighted average is taken of the depth map estimates of the different cues.

Landy et al. (1995) argued that the existing psychophysical literature was compatible with the class of modified weak fusion models. However, this is an extremely broad class of models, and it is hard to evaluate the compatibility of psychophysical data with this class of models without selecting a particular instance from this class, and studying it in detail. A motivation of this paper is to investigate fully implemented versions of weak, modified weak, and strong fusion models.

This paper reports 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, and were implemented through the use of artificial neural networks. The goal of the experiment 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, as the noise model might be expected to have a significant effect on performance. Results indicate that the shape task was significantly easier than the object depth task. The modified weak fusion model showed the best performance on the absolute depth task, and both the strong and the modified weak model performed equally well at the shape task. The superior performance of the modified weak model suggests that constrained nonlinear interaction provides a good model of depth cue combination, combining good performance with parsimonious design. It was also found that the relative weighting of motion and stereo was strongly affected by the task as well as by the viewing distance and, to a lesser degree, the noise condition.

General Methods

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. Sixteen different ellipses were presented; 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. We simulated a point traveling around the perimeter of the ellipse at a constant velocity instead of modeling the ellipse itself rotating. An advantage of this stimulus is that it avoided artifactual depth cues resulting from changes in retinal angle subtended by the ellipse over time. For each of twenty time slices of the point traveling around the perimeter of the ellipse three sources of information were given to the simulated observers: retinal motion, stereo disparity, and vergence angle. These quantities were computed based on the relevant geometric equations.

The viewing distance was the distance from the observer to the ellipse's center. The observer fixated the center of the ellipse and, therefore, the vergence angle was inversely related to the viewing distance. Stereo information consisted of the stereo disparity angle subtended by the point on the ellipse at each moment in time. Motion information consisted of the monocular retinal velocity, expressed in degrees of retinal angle, of the point at each moment in time. The velocity of the

point was proportional to the perimeter of the ellipse, thereby removing artifactual distance cues based on the overall magnitudes of the retinal velocities.

Three noise conditions were examined: a Weber noise condition, a flat noise condition, and a velocity uncertainty noise condition. In the Weber noise condition, the motion, stereo, and vergence angle cues were corrupted by additive Gaussian noise whose distribution had a mean of zero and whose variance was proportional to the magnitude of the signal (e.g., the disparity angle or the retinal velocity). In the flat noise condition, motion and stereo cues were corrupted by additive Gaussian noise with mean zero and constant variance. Note that in the Weber and flat noise conditions, motion uncertainty was modeled as uncertainty about the retinal velocity. An alternative is to consider noise as arising from uncertainty about the velocity of the moving point in the environment. In the velocity uncertainty condition, motion noise was modeled in this way, while stereo and vergence cues were corrupted by Weber noise. In both the flat noise model and the velocity uncertainty model (as well as the Weber noise model) vergence noise was modeled as being Weber noise. This was because a Weber noise model is a conservative one, due to the vergence angle 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 as a check to make sure that it was added noise that limited observers' performances. In all conditions, motion and stereo noise levels were equated to make the two cues approximately equally reliable for judging the depth of an ellipse. Approximately equally reliable motion and stereo cues is consistent with psychophysical data (Rogers and Graham, 1982; Turner, Braunstein, and Anderson, 1997).

Tasks

The depth of an ellipse is the distance from the point on the ellipse closest to the observer to the point furthest away; the width is the distance from the leftmost point to the rightmost point. The shape of an ellipse, sometimes referred to as the form ratio, is defined as the ratio of the ellipse's depth to it's width. 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 independently 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. 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 only provides a shape cue; without information about the viewing distance, size, or velocity of the object there is no way of inferring the absolute depth of the object.

In contrast to motion, stereo is not a scale-invariant cue. Though the width of an object indicated by retinal stereo disparities scales linearly, the depth indicated by a given retinal signal scales with the square of the viewing distance. Stereo disparities are therefore scale-dependent cues; there is no way of inferring shape information independently of the viewing distance. Though stereo disparities are occasionally described as absolute depth cues, it is necessary to have an estimate of the vergence angle or the viewing distance to provide either object depth or shape information from disparities. This need to scale disparities by the viewing distance is referred to as the stereo scaling problem.

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.

Models of Cue Combination

A set of artificial neural networks trained using the back-propagation algorithm was used to simulate the different models. Each network performed a regression, possibly nonlinear, that mapped inputs to outputs. The instances of the strong fusion, weak fusion, and modified weak fusion models used in our simulations are illustrated in the three panels of Figure 1. Each box in these panels represents an independent network, and the labeled lines represent the flow of information between the networks.

Figure 1, Panel A illustrates the strong fusion model. This model consisted of two networks. The first network received an estimate of the vergence angle (γ_v) as input, and calculated an estimate of viewing distance (d_v). The second network received as input a set of twenty stereo disparities ($S_i, i = 1, \ldots, 20$), a set of twenty 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 (only the depth estimate is shown in the figure). The model was relatively unconstrained and could form high-order nonlinear combinations of stereo, motion, and vergence angle information.

The weak fusion model (Panel B) consisted of four networks. The first network, like the first network in the strong model, received 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 computed by the initial network (d_v) with the set of stereo disparities (S_i) to estimate either the depth or the shape of the ellipse. The motion computation network used the viewing distance computed by the initial network (d_v) in conjunction with the set of twenty retinal velocities (M_i) to provide an independent estimate of ellipse depth or shape. The weighting network was given the viewing distance computed by the initial network (d_v) as input, and it computed the linear coefficients used to average the stereo and motion components' outputs. For the object depth task, for example, the network computed the weights w_s and w_m as a function of the vergence angle in the equation

$$depth = (w_s \times depth_s) + (w_m \times depth_m)$$

where $depth$ is the weak fusion model's estimate of object depth, $depth_s$ is the output estimate of the underlying

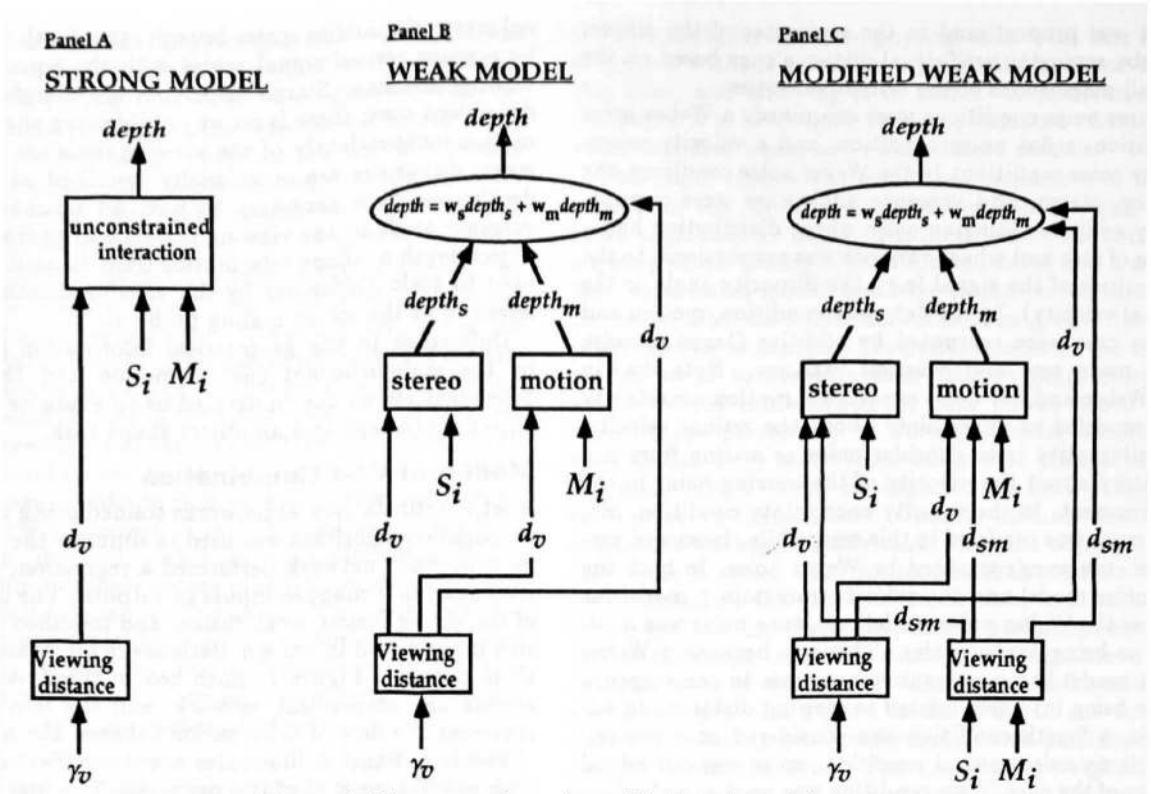

Figure 1: Strong, weak, and modified weak models.

stereo computation network, $depth_m$ is the output estimate of the underlying motion network, and w_s and w_m are the weights used to average the output depth estimates of the stereo and motion networks.

Four of the five underlying networks of the modified weak fusion model (Panel C) 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 (1991) found that the combination of stereo and motion cues helped solve the stereo scaling problem when human subjects 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 (S_i) and retinal velocities (M_i) to an additional estimate of the viewing distance (d_{sm}). As discussed above, 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. By combining motion and stereo disparity information, through this intersection of constraints, both object depth and viewing distance can be computed without any additional information such as the vergence angle. In the modified weak model, limited nonlinear interaction between motion and stereo was used to compute this additional estimate of the viewing distance, labeled d_{sm}. This viewing distance estimate was generally more accurate than the vergence angle estimate (d_v) under the noise conditions studied. This improved viewing distance estimate was used as an additional input to the motion, stereo, and weighting networks of the modified weak fusion model.

Experiment

Figures 2 and 3 show the performances of the weak, modified weak, and strong models on the object shape and object depth tasks, respectively. (Results reported in this paper are based on test patterns that were not used when training the models.) The four graphs in each figure correspond to the four noise conditions studied.

The first major result is that the models learned to perform the shape task better than the object depth task (compare Figures 2 and 3). The shape task was easier for all three models suggesting that it was not a specific property of a particular model that was responsible. This result was also independent of the noise model used. We believe that this result can be understood by noting that motion is a scale-invariant cue to object shape, but that there is no scale-invariant cue to object depth. This provides a motivation for separate shape and depth representations. 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.

The second main result is that the modified weak model showed the best performance in the object depth task

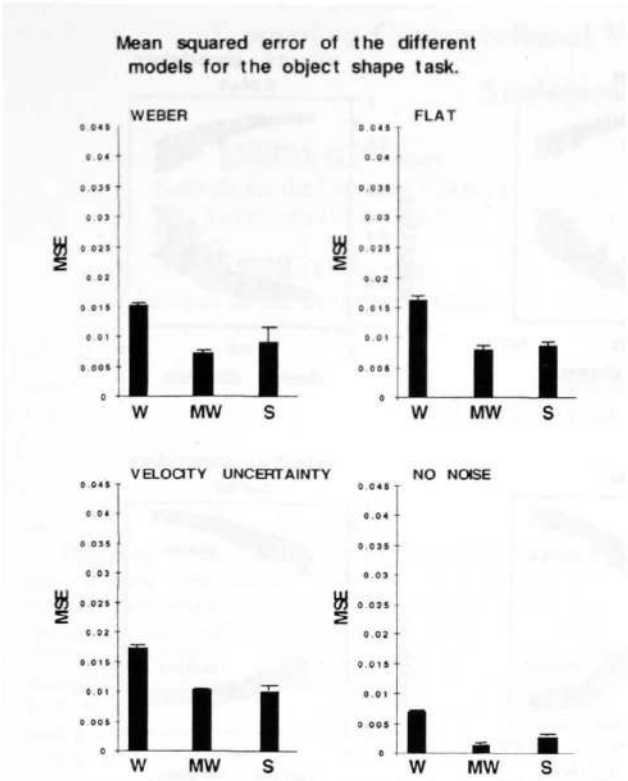

Figure 2: Performances of the weak (W), modified weak (MW), and strong (S) models on the object shape task.

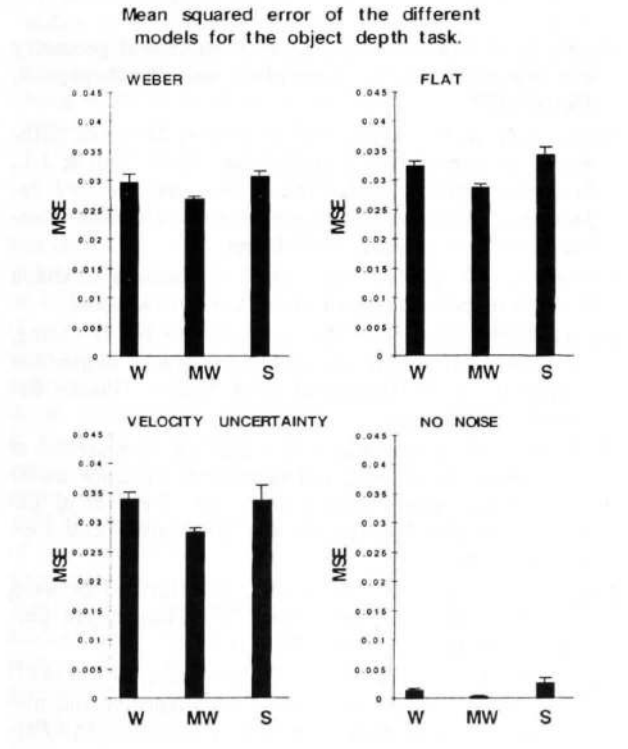

Figure 3: Performances of the weak (W), modified weak (MW), and strong (S) models on the object depth task.

and showed comparable performance to the strong model in the shape task. In theory the strong model should always be able to perform at least as well as the modified weak model, because it is less constrained. Interestingly, 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 frequently fall into relatively poor local minima of the error surface in the presence of noise during training. It should be emphasized that no strong conclusions can be drawn concerning the superiority of the modified weak model over the strong model. We suggest, however, that the good performance of the modified weak model provides evidence that the constraints imposed upon it are, at least, not overly restrictive. The modified weak model performed significantly better than the weak model in the object depth task. Constraints imposed upon 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 second source of information about the viewing distance reduced susceptibility to vergence angle noise, thereby giving the modified weak model a significant advantage over the weak model in the depth task. The superiority of the modified weak model suggests that the modularity constraints imposed upon it (the model contains separate stereo and motion depth computation networks) do not prevent it from finding a relatively good solution. The architecture of the modified weak model provides an adequate compromise between modularity and the power to combine cues.

Figure 4 indicates the weighting of motion and stereo as a function of viewing distance for both depth and shape tasks for the modified weak model. The horizontal axis represents the viewing distance, and the vertical axis represents the relative weighting assigned to motion and stereo (i.e. w_m and w_s).

In the case of the shape task (Panel A), motion information was weighted far more heavily than stereo information for all three noise conditions. This is consistent with the fact that retinal velocities, but not stereo disparities, provide a scale-invariant cue to shape. Interestingly, the weight assigned to stereo increased with viewing distance in all three noise conditions. In the object depth task (Panel B), 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 distance for all three noise models.

The relative weighting of motion and stereo was significantly different for shape and depth tasks. This difference provides 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 separate "shape map." Separate representations for the depth and shape of an object permit independent cue weighting functions, allowing depth and shape to be separately computed so as to minimize the effects of noise for each judgment.

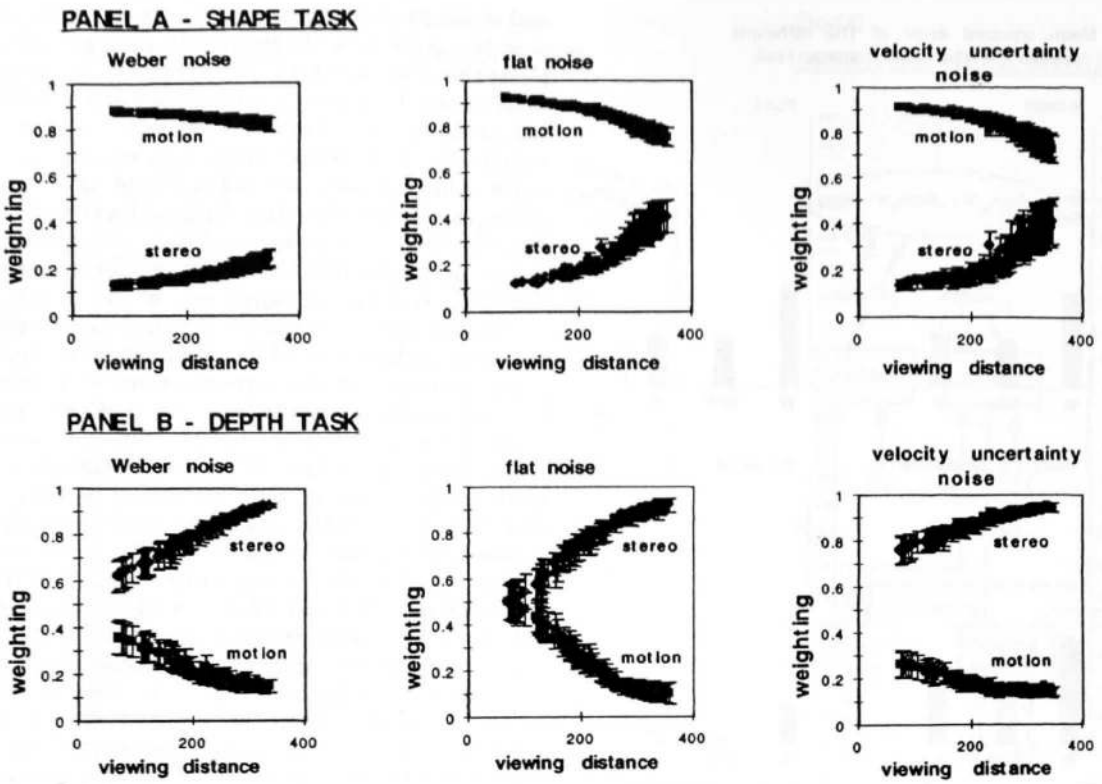

Figure 4: Weighting of motion and stereo as a function of viewing distance for both depth and shape tasks for the modified weak model.

Conclusion

We evaluated the overall performance of strong, weak, and modified weak models of depth cue integration and found that overall the modified weak fusion model outperformed the other two models. Performance in this domain therefore supports more general claims about the advantages of modified weak fusion as a compromise between modularity and fusion. Stereo and motion were weighted differently for shape and depth tasks suggesting the need for separate representations for shape and depth.

As can be seen from our results, neural nets provide a good means of quantitatively modeling highly nonlinear problems such as depth cue combination to reveal hidden or underspecified properties of qualitatively-described theoretical models.

Acknowledgments

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

References

Barlow, H.B. (1986) Why have multiple cortical areas? *Vision Research*, 26, 81-90.

Cowey, A. (1981) Why are there so many visual areas? In F.O. Schmidt, F.G. Warden, G. Adelman, & S.G. Dennis (Eds.), *The Organization of the Cerebral Cortex*. Cambridge, MA: MIT Press.

Gogel, W.C. (1990) A theory of phenomenal geometry and it's applications. *Perception and Psychophysics*, 48, 105-123.

Gross, C.G. & Graziano, M.S.A. (1994) Multiple pathways for representing visual space. In T. Inui & J.L. McClelland (Eds.), *Attention & Performance XVI: Information Integration in Perception and Communication*. Cambridge, MA: MIT Press.

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

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.

Philbeck, J.W. & Loomis, J.M. (1997) A comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

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

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