Pergamon

0042-6989(94)00176-6

Vision Res. Vol. 35, No. 3, pp. 389–412, 1995
Copyright © 1995 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0042-6989/95 $7.00 + 0.00

Measurement and Modeling of Depth Cue Combination: in Defense of Weak Fusion

MICHAEL S. LANDY,*† LAURENCE T. MALONEY,* ELIZABETH B. JOHNSTON,*‡ MARK YOUNG*

Received 11 June 1993; in revised form 20 June 1994

Various visual cues provide information about depth and shape in a scene. When several of these cues are simultaneously available in a single location in the scene, the visual system attempts to combine them. In this paper, we discuss three key issues relevant to the experimental analysis of depth cue combination in human vision: cue promotion, dynamic weighting of cues, and robustness of cue combination. We review recent psychophysical studies of human depth cue combination in light of these issues. We organize the discussion and review as the development of a model of the depth cue combination process termed *modified weak fusion* (MWF). We relate the MWF framework to Bayesian theories of cue combination. We argue that the MWF model is consistent with previous experimental results and is a parsimonious summary of these results. While the MWF model is motivated by normative considerations, it is primarily intended to guide experimental analysis of depth cue combination in human vision. We describe experimental methods, analogous to *perturbation analysis*, that permit us to analyze depth cue combination in novel ways. In particular these methods allow us to investigate the key issues we have raised. We summarize recent experimental tests of the MWF framework that use these methods.

Depth Multiple cues Sensor fusion

The human visual system extracts information about depth and object shape from a variety of cues. There are cues resulting from object rotation (the kinetic depth effect or KDE) (Wallach & O'Connell, 1953) and from observer motion (motion parallax§) (von Helmholtz, 1910/1925). Two eyes and overlapped visual fields permit measurement of binocular disparity (Wheatstone, 1838) and allow for vergence cues to depth. The geometry of perspective provides a number of cues including texture density, texture element foreshortening and size (Cutting & Millard, 1984) and perspective cues from linear image elements. Other cues include occlusion, smooth shading, specularities (highlights) on glossy curved surfaces, blur, accommodation, and so on (see Gibson, 1950; Kaufman, 1974, for reviews).

Outside of the laboratory (and sometimes inside it as well), the visual system has available to it multiple sources of information about depth and shape at each location in the scene. Information from multiple cues is combined to provide the viewer with a unified estimate (and percept) of depth and shape, although the combination process can fail, leading to multistable percepts.

To illustrate the depth cue combination process that we envisage, imagine that we are viewing the simple scene depicted in Fig. 1 (a bowl of lemons on a table), and that we are moving. Many of the cues to depth are available under these circumstances: motion parallax, binocular stereopsis, texture, highlights, etc. Each cue is signaling depth and shape information about the same scene. Any inconsistency in the information about depth and shape provided by two cues is due either to stochastic error in the initial information available to the visual system, or to erroneous assumptions or calculations made in processing depth information (as when false stereo correspondences are chosen). The information about depth and shape available from any one cue may be inaccurate due to stochastic or processing error. A more accurate overall estimate may be obtained by combining the separate estimates.

MODELS OF DEPTH CUE COMBINATION

The Weak Observer

A simple way to combine multiple depth estimates is to first attempt to compute separate estimates of depth

*Psychology Department and Center for Neural Science, New York University, 6 Washington Place, Room 961, New York, NY 10003, U.S.A. [*Email* landy@nyu.edu].

†To whom all correspondence should be addressed.

‡Present address: Psychology Department, Sarah Lawrence College, Bronxville, NY 10708, U.S.A.

§In this article the term *motion parallax* is reserved for the case of depth estimation given differential image velocities generated by observer motion. As suggested by Braunstein *et al.* (1986, p. 220), this is as distinguished from *velocity gradients*, which result from translations of an object relative to the observer along a path perpendicular to the line of sight (and using polar perspective) and the KDE, which results from rotations of an object.

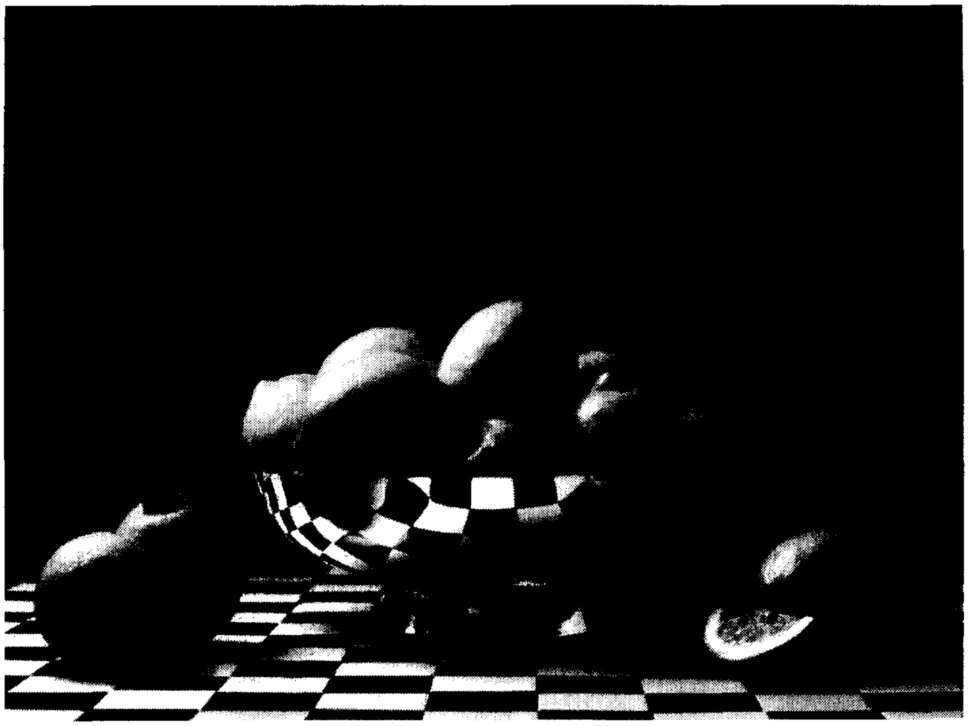


FIGURE 1. A photograph of a scene involving multiple cues to depth and shape. (This is a poor quality halftone of a 4 × 5 color photograph taken by Corinne Colen. Reprinted with permission.)

("depth maps") based on each depth cue considered in isolation, and then to average the separate depth estimates from each cue (the depth maps) to obtain an overall depth map for the scene. We will call this rule of combination the *Weak Observer*. The Weak Observer is illustrated in Fig. 2(A). It has the advantages that it is modular (the depth maps are computed independently) and that the rule of combination (averaging) is very simple. If we accepted the Weak Observer as a model of biological shape and depth perception, then we could take advantage of the modular structure by studying each depth cue in isolation. Similarly, the design of a Weak Observer algorithm for machine vision could begin with the design of isolated modules corresponding to different cues. One prediction of the Weak Observer is that interactions between depth estimates from different modules are limited to those attributable to sharing a common retinal input.

There are several problems, minor and major, with the Weak Observer. The foremost is that it doesn't really make sense. The information available from different depth cues is qualitatively different. A cue such as motion parallax can be used to estimate a depth map measured in physical units of depth (e.g. meters). A cue such as texture provides only relative depth information, that is ratios between the depths of different points in a scene. The outputs of the various modules in Fig. 2(A) cannot be meaningfully averaged. Before averaging depth maps based on distinct cues, we must change them to common units, a process we term *promotion*.

Even if we succeeded in promoting the depth maps to common units, we must still recognize that the resulting information available from each cue varies in reliability across the scene. In Fig. 1, for example, texture is only a reliable cue in the regions containing the lemons and the table surface. Depth information obtained from texture cues in other regions of the scene, such as the reflection of the table's surface texture in the bowl, will be inaccurate and should be given no weight. The Weak Observer (or any modular scheme) should change the weights assigned to different cues to reflect the reliability of the cues. If the results of independent depth calculations are widely discrepant the cue combination process should be *robust*, degrading more gracefully than the simple averaging rule allows. Further, if we were to change the viewing conditions slightly by forcing the observer to remain still, we would want to alter the weights in the average to reflect the absence of the motion parallax cue which is no longer available. These considerations suggest that the weighted averages of depth cues should be *dynamic*, changing within and between scenes, based upon the estimated reliability of the cues.

The Strong Observer

Figure 2(B) represents an extreme alternative to the Weak Observer that we term the Strong Observer. The Strong Observer does not divide the computation of depth into separate modules corresponding to different depth cues. Nakayama and Shimojo (1992), for example, recently proposed a depth processing model in which the scene interpretation S_i is chosen that maximizes the probability (likelihood) $P[I|S_i]$ of the image I. They propose that the observer, in effect, determines the most probable three-dimensional interpretation of the scene given the current retinal data. There is no need, in their scheme, to modularize the computation of depth

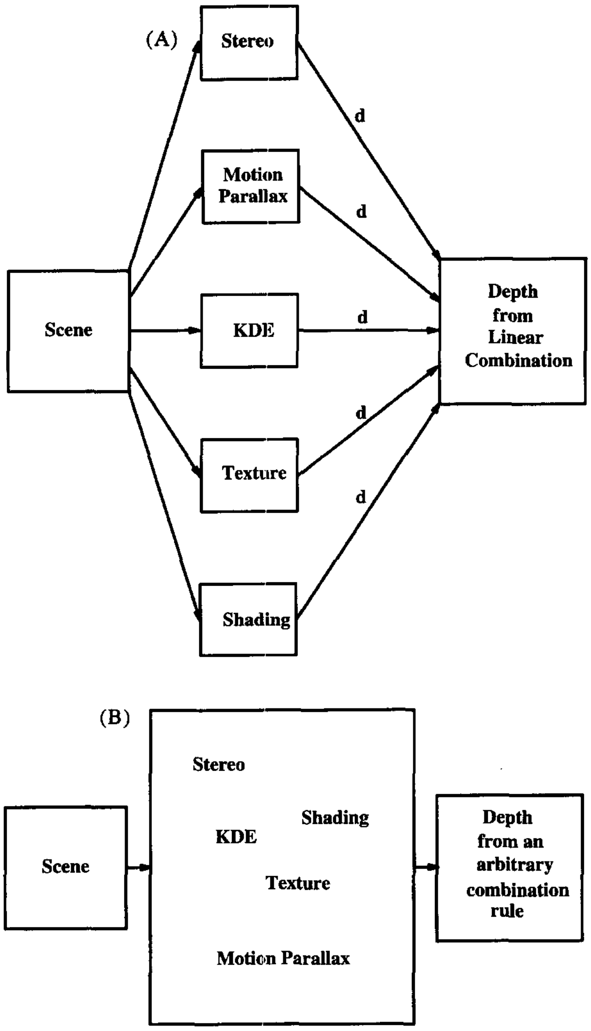


FIGURE 2. (A) Weak fusion. Each depth cue is processed independently. These estimates are combined linearly. (B) Strong fusion. Depth modules may interact, and the combination rule is not necessarily linear. If the interactions and combination rule are not constrained, the model can be arbitrarily complex and is no longer testable.

(although many of the benefits of their approach would carry over to an analogous model that did divide early depth and shape processing up by cue type). We will return to their proposal in the section on Bayesian and likelihood approaches below.

The Strong Observer is not (necessarily) modular, and it is not clear that there is any meaningful definition of "depth cue in isolation" for such an observer. With respect to the Strong Observer, the traditional depth cues discussed in the preceding section are artificial constructs of the experimenter. Interactions between depth cues are to be expected simply because the Strong Observer model is not organized in modules corresponding to depth cues.

The Weak and Strong Observers fall at the two ends of a continuum of possible models of depth and shape processing. Clark and Yuille (1990) distinguish *weak fusion* and *strong fusion** approaches to depth cue combination, which are analogous to the Weak and Strong Observers discussed above. Models that emphasize modularity tend toward the weak end of the spectrum, models that emphasize interactive, holistic processing tend toward the strong end. For strong models, distinctions among traditional cue types are de-emphasized or eliminated. What constitutes a distinct depth cue, then, is not given in advance, and must be developed and tested as part of a model of depth cue combination.

The Modified Weak Observer

Here we develop an alternative depth cue combination model designed to overcome the difficulties inherent in the extreme Weak and Strong Observers. The *modified weak fusion* (MWF) model is intended to be as modular as possible, consistent with the normative guidelines raised above and in the following discussion. We believe that it represents a useful guide to the design and interpretation of depth and shape experiments.

Cue promotion and calibration

Different depth cues provide markedly different *kinds* of information. For example, given knowledge of self-motion, the retinal motion induced by self-motion (motion parallax) is an *absolute* cue to the depth of stationary objects. That is, the depth derived from parallax information is specified completely by retinal velocity:†

$$\text{depth}_p = f_p(\text{velocity}), \qquad (1)$$

where depth_p is the distance from the observer to the object. Similarly, given knowledge of the interocular separation and gaze angles, binocular disparity is potentially an absolute cue to depth.‡ That is, any given pair of retinal locations which correspond to the same feature in the environment may be used, together with the gaze information, to compute the precise distance (e.g. in meters) to that three-dimensional location. On the other hand, without the information as to viewing distance (to the fixation point), a given disparity does not specify a fixed amount of depth. Rather, for a given disparity the depth derived from stereo disparity scales with the square of the unknown viewing distance parameter d:

$$\text{depth}_s = d + d^2 f_s(\text{disparity}). \qquad (2)$$

Here, f_s is the result of the correspondence computation (and perhaps some further rescaling) and provides relative depth values, but can not be interpreted as absolute depth (in meters) until scaled by the viewing distance.

*The term *fusion* here refers to the process of combining information from multiple sources, not the process of fusion in stereoscopic vision.

†Assuming that the absolute distance to the surface is specified by egomotion information (see Ono, Rivest & Ono, 1986).

‡We use the term *depth* both to denote distance from the observer to an object (absolute depth) and the difference in distance from the observer to each of two different objects (relative depth). In much of the literature the term depth is reserved for the latter concept, whereas the term *distance* is used for the former.

In the following, we will speak of a "depth map" as if it consisted of an array of measured distances across the visual field. Alternatively (and more plausibly), we could represent depth by means of the parameters of a model of piece-wise smooth surfaces and textures in a scene. The parameter settings are, then, the "depth representation" from which a "depth map" could be generated (see, e.g. Grimson, 1981). For our purposes in this article, the precise nature of the depth representation is of little consequence. We will make some assumptions concerning the depth representation when we present the perturbation analysis method below.

The kinetic depth effect provides a different set of constraints on object shape. The series of retinal images produced by a rotating object is identical to that produced by the rotation of an object which is moved away from the observer by a given factor, whose size is increased by that same factor, and which is rotated at the same angular velocity. In addition, KDE displays are subject to depth reversals. Thus, the depth portrayed by KDE depends on two parameters, the fixation distance d and the sign of the perceived rotation direction ϕ:

$$\text{depth}_k = d(1 + \phi f_k(\text{velocity})), \tag{3}$$

where $\phi = \pm 1$. Shading is a cue which provides an indication of the surface normal at each location and is often referred to as a shape cue (as opposed to a depth cue). But, by integrating this surface normal over space (Koenderink, van Doorn & Kappers, 1992), shading may be shown to provide the same form of information as the KDE. (In all of these formulations the specification of some details is suppressed including retinal location and self-motion parameters.) Thus, it is meaningful to average the data computed independently from shading and KDE, assuming that the two cues share the same distance and ambiguous reversal parameters d and ϕ.

The depth cue of occlusion provides an entirely different type of information. At an occluding contour the only information provided by the assertion of occlusion is that the depth on one side of the border is greater than on the other side. Nothing is implied by this cue about the amount of depth difference, nor does it specify anything about depth values away from the boundary. Finally, it has been suggested that at certain types of accretion/deletion boundaries there is a sensation of depth which is ambiguous: the two sides of the contour are merely perceived to be at *different* depths, but it is ambiguous which side is closer to the observer.

Thus, depth cues provide qualitatively different information. These qualitative differences must be taken into account by any rule of combination. For example, it is possible to combine depth$_s$ and depth$_k$ by averaging, but it would be nonsensical to perform a similar calculation using f_s(disparity) and f_k(velocity). As an examination of equations (2) and (3) demonstrate, f_k(velocity) is unitless, while f_s(disparity) is in units of inverse distance. The resulting average would defy interpretation.

This discussion is remininscent of measurement theoretic notions of "meaningfulness" and of ratio, interval and ordinal scales (Krantz, Luce, Suppes & Tversky, 1971; Roberts, 1979; Stevens, 1959). Depth cues such as KDE do not provide depth estimates corresponding to any of these scale types (due to the reversal ambiguity). Therefore, rather than talking about "scale types", we prefer to consider most depth cues as sources of absolute depth information once a number of parameters are specified (Maloney & Landy, 1989). The output of the KDE depth computation is taken to be a depth-map-with-two-parameters, the output of the disparity computation, a depth-map-with-one-parameter, and the output of the motion parallax computation, a depth map.

By specifying the missing parameters for a given cue, it is thereby *promoted* to the status of an absolute depth cue. Cues must be promoted to be on equal footing before the values obtained from them are commensurate. The notion of depth-map-with-parameter and promotion are both present in Schopenhauer (1847/1974):

> "... with the same visual angle an object may be small and near or large and distant. Only when its size is already known to us in another way are we able to know its distance ... Insofar as we have before us an uninterrupted succession of visibly connected objects, we are certainly able to judge distance from this gradual convergence of all lines and hence from linear perspective. Yet we cannot do this from the mere visual angle by itself, but the understanding must summon to its aid another datum which acts, so to speak, as a commentary to the visual angle ... Now there are essentially four such data ... (pp. 95–97)".

He then lists accommodation, vergence, atmospheric perspective, and linear perspective as candidate "second data" for the promotion of visual angle. These ideas appear in different forms in more recent work as well (e.g. Gogel, 1977). There is some correspondence between our notions of an absolute cue and a cue requiring further parameters and Gogel's absolute and relative cues. However, our distinction is a formal one related to the information content available using a cue, and Gogel's is an empirical definition based on the percepts engei.dered by various reduced-cue displays.

A single cue from one view of a scene cannot be used to promote itself, and thus interaction between different depth cues is inevitable. This sharing of information must occur if two qualitatively different depth cues are to contribute to the depth percept at a given location. We do not know how depth promotion is carried out in the human visual system [although we have begun to elucidate its mechanisms (Johnston, Cumming & Landy, 1994)], and we consider this an important area for future study. Here, we suggest a number of ways in which depth cue promotion could occur. This interaction can take a simple form. For example, if an absolute depth cue (such as motion parallax) is available at the same location as a depth cue which has one missing parameter, the viewer can assume that the two cues are indicating the same absolute depth value and hence solve for the missing parameter. If motion parallax and stereo disparity are available in a large number of image locations, one can obtain a more stable estimate of the viewing distance by using the value of d which minimizes the inconsistency between depth from disparity and depth from motion parallax (Maloney & Landy, 1989):

$$\min_{d} \sum_{x,y} \left([d + d^2 f_s(\text{disparity};x,y)] - f_p(\text{velocity};x,y) \right)^2, \quad (4)$$

thus promoting the stereo cue. In some cases, two cues can promote one another as long as they scale in a different manner with respect to a given parameter. Stereo disparity and the KDE both scale with distance, but stereo scales differently than KDE. If both cues are available at a number of locations, the observer can set the missing parameters in a similar manner by minimizing:

$$\min_{d,\phi} \sum_{x,y} \left([d + d^2 f_s(\text{disparity};x,y)] - [d(1 + \phi f_k(\text{velocity};x,y))] \right)^2. \quad (5)$$

On the other hand, an ordinal cue such as occlusion only provides assertions about depth order. Theoretically, it cannot be used to estimate depth *per se*, but can be used to disambiguate other cues [e.g. to specify ϕ for the KDE as has been found with human observers (Braunstein, Andersen & Riefer, 1982; Proffitt, Bertenthal & Roberts, 1984)]. Of course, the viewer is free to mistakenly take occlusion to indicate a fixed amount of depth and then combine it with other cues. The data of Bruno and Cutting (1988) suggest that observers do so when reporting perceived depth of simple frontoparallel rectangular surfaces using numerical rating scales. However, this performance is not totally counter to the information content of occlusion displays. For example, if the amount of surface that is occluded is known, this can constrain the depth difference at the occluding contour, although this also requires knowledge of the slant of the rear surface. Our argument is not that observers *must* use the information in depth displays in a sensible, meaningful manner. Rather, we suggest that this approach is normative, and further suggest that the normative approach (or some approximation) should be the null hypothesis, with departures from optimality as useful indications of compromises made in human depth perception. This is in the spirit of Geisler's (1989) suggestion (so far applied principally to visual detection and discrimination tasks) that experimenters first compare human performance with that of an ideal observer to discover what information is lost at various stages of visual processing.

The depth promotion process followed by a combination of multiple depth cues allows a visual system to use the best qualities of several cues. For example, consider an observer moving slowly around a scene in which there is fast object motion. Because the observer's motion is slow, depth estimates from motion parallax may be noisy and unreliable. However, they result in a complete depth map. Figure 3(A) illustrates depth estimates from motion parallax from a slice across a scene. The true depths are piecewise-constant, but the estimates are quite noisy. The fast-moving objects result in a depth

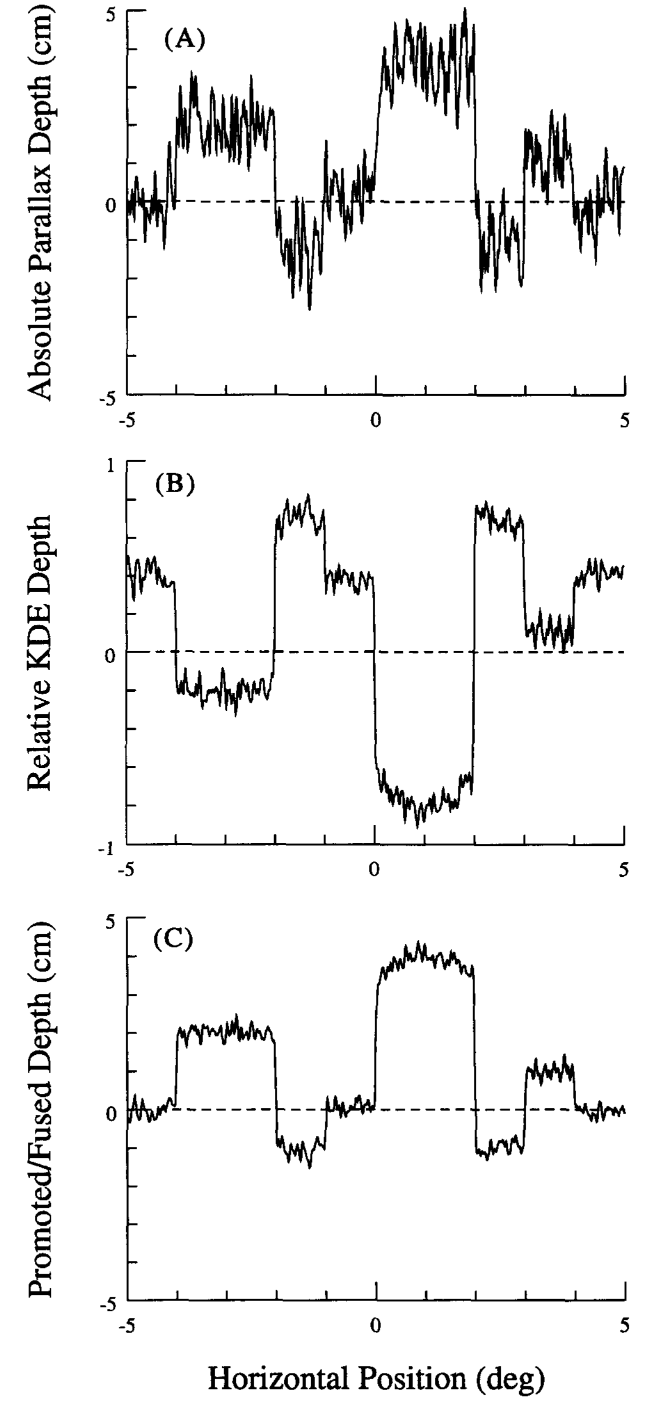


FIGURE 3. (A) Depth estimates in a horizontal slice across a scene made using an inaccurate, absolute cue (parallax with slow observer motion). (B) Depth estimates across the same image locations made using an accurate, relative cue (KDE, fast object motion). Note the depth reversal. (C) The two cues are compared to promote the KDE estimates and are then combined. The fused depth is accurate and on an absolute scale. See text for further details.

map from KDE which is accurate, but provides only relative depth values which are also subject to depth reversal. These estimates are illustrated in Fig. 3(B) in which both decreased noise and a mistaken depth reversal are visible. The missing parameters for KDE depth (fixation distance and reversal) may be estimated by least-squares regression of the KDE depth against the motion parallax depth. Then, the parallax and promoted KDE estimates can be combined [using a weighted average with greater weight given to the more reliable KDE cue (see section entitled Reliability, ancillarity, availability below)]. Figure 3(C) shows the KDE cue promoted and fused with the parallax cue. Under these circumstances, promotion by a weak absolute cue has permitted an accurate cue-with-two-parameters to be used to provide accurate estimates of depth, something it could not do in isolation.

A cue can be used to promote itself if two sets of data are gathered using that cue. If the only available cue is stereo disparity, then there is a large class of shapes which are possible (parameterized by the unknown viewing distance d). With only a single glance at the scene (and no or restricted alternative cues) there is no reason why the promotion of the cue need be carried out with the correct value of the viewing distance, and there is strong evidence that, in fact, stereopsis often utilizes an incorrect value of the viewing distance parameter (Johnston, 1991; also see Foley, 1980; Gogel, 1960). However, if two stereo views are available (e.g. for a rigid, rotating object), then promotion should be possible. This is essentially the argument made by Richards (1985), and we see some evidence for it (Johnston *et al.*, 1994, discussed in section entitled Tests of the framework below). Although each stereo view is consistent with a one-parameter class of objects, there will only be one shape in common between the classes of solutions resulting from each of the two views. This is analogous to the notion that depth from KDE for only two views results in a three parameter class of solutions [up to depth scaling which is the "affine-equivalent" class referred to by Todd and Bressan (1990) and Bennett, Hoffman, Nicola and Prakash (1989), translation in depth, and depth reversal]. Once a third view is available, there are only two degrees of freedom left [as in equation (3)]. Of course, if the amount of rotation is small and the data are unreliable, then the observer will be forced to form a "soft intersection" by choosing only those elements of the two sets of solutions which are insignificantly different from one another. Still, with further rotation the solution should become more constrained.

To summarize, a given depth cue can be modeled as indicating the absolute depth to each feature in the scene once a number of scaling parameters are specified. This cue promotion is required before cues which are parameterized differently may be combined, and thus some interaction between cues is mandatory. Promotion may be accomplished by combining cues with different scaling behavior or by looking at multiple instances of a single cue. We have only suggested a number of ways in which cues may be promoted, and encourage further research to determine the methods used by human observers. Once cues have been promoted they may be combined using very simple weak fusion techniques. Note that cue promotion is not the same thing as cue calibration. Cue promotion is an instantaneous reaction to a second source of information used to promote the first cue. "Calibration", in biological vision, usually describes a slow learning process which requires time to provide accurate estimates from a given cue (e.g. to cope with changes in the data over time such as those caused by aging optics, growth of the eye, etc.). "Calibration". in computer vision, typically refers to the process of estimating the camera's characteristics. With either definition of calibration, it should be a simple matter to identify cases of cue promotion as being distinct from calibration. At the same time, there is clearly evidence that the promotion process itself is based both on current scene information as well as recent observation history. For example, the disambiguation of KDE is affected by a stereo preview, which does not suffer from a reflection ambiguity (Dosher, Sperling & Wurst, 1986), and by prior adaptation to dynamic stereo displays (Nawrot & Blake, 1989, 1991).

Robustness

The theory of robust statistics (Hampel, 1974; Huber, 1981) provides a framework for estimation under uncertainty, and in particular provides estimation techniques which are resistant to outlier observations. Statistical decision procedures are typically based on assumptions concerning the probability distributions that generate the data available to the decision procedure. If the distributions are *perfectly* known, there are standard methods for computing optimal statistical estimators. One can imagine an estimator that is optimal if certain distributional assumptions are satisfied (e.g. independence, normality), but whose performance rapidly degenerates when those assumptions are not quite satisfied. A general definition of robustness is that "small changes" in the underlying distribution of the data produce "small changes" in the distribution of the estimator (Rey, 1980, Chap. 3).

The 5% trimmed mean is an example of a robust estimator. The 5%-trimmed mean consists of an average where the most discrepant 5% of the observations have been removed from the sample. If the data of interest are normally distributed, the trimmed mean produces an estimate of the population mean somewhat inferior to the ordinary, untrimmed arithmetic mean of the data (the trimmed mean is more variable). In fact, the untrimmed mean is the best estimator (best unbiased estimator with minimum variance) *if* the data are independent, identically-distributed normal random variables. When the true distribution is not precisely the normal distribution, and prone to producing outliers that are a considerable distance from the population mean, the performance of the untrimmed mean worsens more rapidly than that of the 5%-weighted mean until the latter is preferable. The trimmed mean is *robust* with respect to failures of the distributional assumptions.

If we are not certain about the underlying distribution of the data, a robust estimator may be preferable to a nonrobust optimal estimator. By using a 5%-trimmed mean instead of the untrimmed mean, we are, in effect, buying insurance against failures of the distributional assumptions we have made. The cost of the insurance is the reduced performance if the data are, in fact, independent, identically-distributed normal random variables. We suggest that cue combination would benefit from the use of robust statistical methods.

A second reason to use robust methods is to be less sensitive to fallible processing in early vision (Schunck, 1989; Sinha & Schunck, 1992). For example, if disparity processing results in an incorrect choice of correspondence between the two eyes, this may introduce extreme outlier estimates of depth.

The robustness of shape and depth perception can be experimentally tested. Suppose that we are viewing a simulated scene, and that all but one of the cues available at a location in the scene signal the same estimate of depth. The remaining cue is manipulated so that it signals depth information that is more or less discrepant with the other cues, and suppose that we could measure the effect of this cue on the final estimate of depth. Figure 4 plots two hypothetical outcomes. The degree of discrepancy, positive or negative, is plotted on the x-axis, the influence of the discrepant cue is plotted on the y-axis. The straight line is the influence curve of a nonrobust estimator (e.g. averaging all cues with fixed weights). As the discrepancy increases in magnitude, the effect of the discrepancy upon the depth estimate increases linearly. The second curve is the *influence curve* of a robust estimator (due to Hampel, 1974; see Huber, 1981; Rey, 1980). As the discrepancy of the odd cue increases from zero, it is treated like the other cues and should affect the depth percept linearly. As the discrepancy increases beyond the range present in normal scenes, robust statistical considerations should come to play and the discrepant cue should then have less and less of an effect on the fused percept. An alternative definition of robustness is that an estimator is robust precisely when the influence curve of any datum is bounded (Hampel, 1974). We describe in a later section how the influence curve of a cue associated with the depth combination rule may be empirically determined.

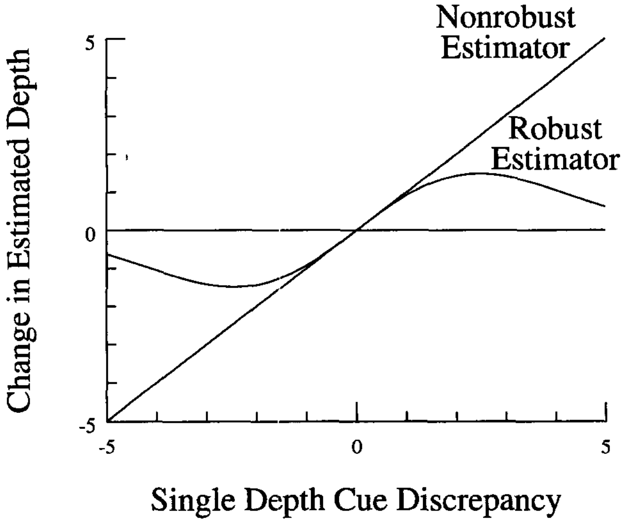


FIGURE 4. A representative influence curve for a robust statistical estimator. As a single cue specifies an increasingly discrepant amount of depth relative to other cues, at first overall perceived depth changes accordingly, but eventually the discrepant cue is downweighted. A nonrobust estimation procedure such as a simple average would predict the linear function.

We find it useful to think of robust statistics in terms of a "reality check". If the discrepancies in the scene you are viewing are typical of those to which you are accustomed in viewing natural scenes, then accept the information that is offered. If, however, the discrepancy between individual depth estimates is outside the limits typical of real scenes then you are forced to do something else. This can be to simply pick one cue (the "veto" of Bülthoff & Mallot, 1988), actively seek more information, and so on.

It is important in empirically studying visual response to depth cues that the scene used as stimulus be veridical. When cues are put in conflict it should be no surprise that they interact in complex ways [as with the famous Ames Room demonstrations (see Ittelson, 1952)]. Normative considerations of robustness predict such interactions. This is not to say that one shouldn't study the system outside of its normal operating region, but rather that the results be interpreted with caution.

Reliability, ancillarity, availability

Our treatment of the issue of combining cues with different degrees of reliability is also motivated by normative statistical considerations. If n identically distributed random normal samples are drawn then the minimal variance unbiased estimator of the mean μ of the underlying distribution is the sample mean (as noted above):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{6}$$

If these samples come from different distributions with the identical mean μ but with different known variances σ_i^2, then the minimal variance unbiased estimate is again an average, but in this case a weighted average is preferred where each sample is weighted by its inverse variance:

$$\hat{x} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}. \tag{7}$$

[see, e.g. Bove (1990) for an application to range sensor fusion, and Searle, Braida, Davis and Colburn (1976) for an application to the combination of spatial auditory cues]. Thus, the normative computation for depth fusion for depth sensors perturbed by zero mean independent Gaussian error is a weighted average. The weighted average is a natural consequence of treating depth estimates as noisy and using an optimal, signal detection theoretic framework. This assumption of cue weights inversely related to variability was used by Taylor (1962) in a theory of figural aftereffects.

The optimal computation requires the observer to know the variance of each sample. Natural viewing conditions include a wide range of scene content, and this has an impact on the *availability* of the various depth cues. If you close one eye then stereo disparity information is no longer available and any depth estimates which are derived from a depth-from-stereo computation are surely nonsensical. Yet, when you close one eye the world does not suddenly become flat. We take this as indicating that the weights used to average depth cues are malleable. An estimate of cue variance should be based on the objective "availability" of cues in the scene and on the quality of depth information available from the cue, e.g. if a region of a scene contains no texture elements or if the texture elements are sparse, the weight assigned to texture in that region of the scene should drop if more reliable cues are available.

The measurement of individual cues is made more or less reliable based on scene content as well. For example, stereo depth measurement scales with the square of the viewing distance. Small errors in the measurement of stereo disparity are amplified into large errors in depth for large viewing distances. At large viewing distances the stereo cue becomes less reliable and its estimate should be given less weight.* A low level of image contrast will surely decrease the reliability of all depth cues, but it need not affect their reliability to the same degree, and the observer should take this into account.

Recent behavioral work suggests that animal visual systems are sensitive to the reliability of depth cues. For example, Ellard, Goodale and Timney (1984) found that the Mongolian gerbil, when forced to jump from one platform to another, used both a looming and a motion parallax cue. When the looming cue was reduced (by shortening the platform), the animals apparently sought more parallax information by making more and larger head movements. Similarly, Goodale, Ellard and Booth (1990) found that these animals also combined image size and motion parallax cues adaptively. This goes a step beyond the MWF framework to suggest that observers also estimate the total reliability of the combined depth estimate, and will seek more sources of information if that reliability is insufficient for the task at hand.

How is it possible for the observer to estimate the reliability of individual cues? Various measures of the degree of egomotion (e.g. from the vestibular system) provide one piece of information for estimating the reliability of motion parallax information. A measure of the absolute viewing distance (from other depth cues or vergence angle information) constrains the reliability of depth from stereo disparity. Various measures of the spatial frequency content of a scene can constrain the reliability of a shape-from-texture estimate. In each case, side information which is not necessarily relevant to the actual estimation of depth, termed an *ancillary measure*, is used to estimate or constrain the reliability of a depth cue. Ancillary statistics are statistics that are conditionally sufficient and reduce the estimation variability of a parameter, but which are independent of the value of the parameter being estimated (Cox & Hinkley, 1974; Kendall & Stuart, 1979). By analogy, ancillary measures do not necessarily allow for estimation of depth by themselves, but do provide information about the performance of other depth estimators. If ancillary information changes as the content of a scene changes, the observer should take note and vary the weights of individual cues accordingly. Some ancillary information (such as estimates of the fixation distance as ancillary information for stereopsis) will constrain the reliability estimate of that cue everywhere in the scene. In other cases (such as estimates of local texture content), the ancillary information constrains the reliability estimate only in a local region of the scene. Thus, relative estimates of cue reliability can vary across the scene. Again, if an experimenter is not aware of this possibility, then a change in weights based on cue availability or observer estimated cue reliability will be mistaken as a counterexample to weak fusion.

The coexistence of mechanisms for robust estimation and for weighting cues based on reliability has important consequences. For example, a popular stimulus manipulation is to add noise to stimuli as a means of estimating internal noise and observer efficiency (e.g. Pelli, 1981). With tasks involving multiple cues, this same technique can be used to make one cue less reliable than another and hence alter the weights used for the two cues (Young, Landy & Maloney, 1993; discussed below). However, if a cue becomes sufficiently unreliable, not only will its weight be lowered, but it will also occasionally produce strongly discrepant estimates, which will be further discounted by mechanisms for robustness.

Summary of the MWF argument

We propose a model for depth fusion which at its heart uses a weak fusional method: weighted averaging. However, several wrinkles are added to the basic weak fusion scheme motivated by the theoretical considerations covered above, leading to a scheme which we term *modified weak fusion* (MWF).

In the case of multiple cues to depth, we would like to argue for a finer set of distinctions than simply weak versus strong fusion. The issue is more than simply one of definition. Referring again to Fig. 2, if no constraints are placed on the interactions between two cues [as in Fig. 2(B)], then by calling an interaction strong fusion we are, in fact, not constraining the form of the interaction at all. The resulting theory is not falsifiable. Instead, we suggest that a middle ground be struck which to a first approximation is the simplest form of weak fusion: a

*There is some evidence that this also applies to motion parallax (Ono et al., 1986).

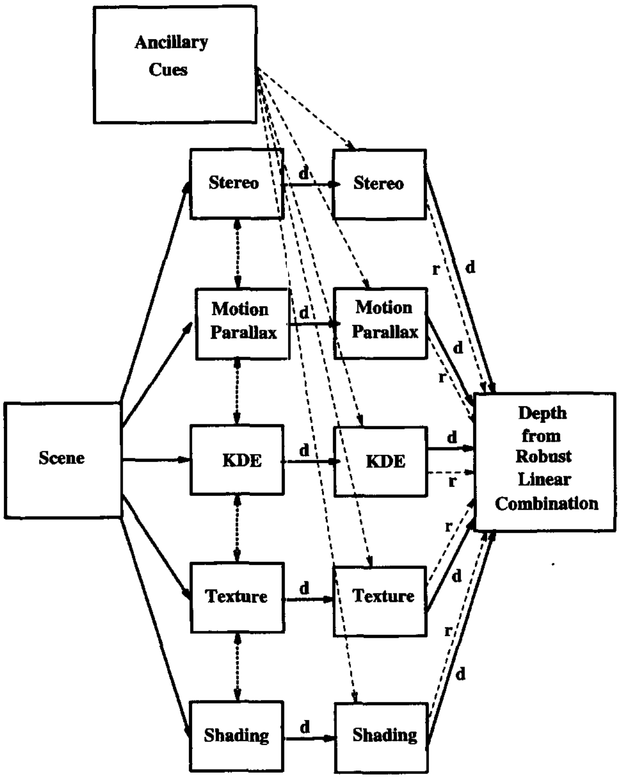


FIGURE 5. Modified weak fusion. Cues interact solely for the purposes of cue promotion. Each depth cue produces both a depth map and, using ancillary cues, a reliability map. The depth combination rule is linear (for small perturbations) and robust (for large perturbations). (Compare with Fig. 2.)

linear combination of the separate cues. However, for the reasons discussed above, we allow for interactions to occur between separate depth modules, but only of a highly constrained sort. These are justified on statistical grounds, and yet result in a theory which can be tested.

Interactions between different cues are required to promote all cues to be absolute depth cues (or at a minimum to be commensurate). These interactions are illustrated on the left in Fig. 5. Each independent cue provides other cues with their current depth map (which may not yet be promoted) *solely for purposes of depth promotion*. A second stage is required for each cue to amalgamate information from within the scene as viewed by that cue and from ancillary measures so that each depth cue pathway may estimate its own reliability. The order of these two stages is of no consequence since even unreliable cues may participate in interactions required for cue promotion (as in Fig. 3). Finally, each cue provides a depth map and a map of estimated reliabilities (scene content and hence cue reliability can vary from position to position within the image) which are input to the final fusion stage. The fusion computation involves a weighted average where the weights take into account both the estimated reliabilities of each cue and the discrepancies between cues.

Modified weak fusion is not simply weak fusion: interactions are allowed between depth cues for the purposes of promotion. This is a necessity to make the cues commensurate; weak fusion without interactions was never a real possibility. At the same time, however, MWF is a very specific and simple form of strong fusion. The interactions between cues are highly constrained and are all required by the problem at hand. When one cue leads to an estimate which is discrepant from several other cues it is important to minimize the consequences of the one suspect cue, and this should not be taken as an indication of unmitigated strong fusion. When cues interact to achieve promotion, that is a prerequisite to meaningful cue combination. Finally, depth combination rules are likely to be dynamic in the sense of taking into account the context of a scene. If the viewing conditions are changed in a way which impinges on cue reliability, the observer will note this fact (through ancillary measures) and reweight the cues accordingly.

Throughout the remainder of the paper we will contrast the terms "modified weak fusion" and "strong fusion". *Henceforth, by "strong fusion" we mean any model of depth combination that permits interactions between cues that is different from the MWF model.* In an experimental test of the MWF as a hypothesis concerning human vision, "strong fusion" is the alternative hypothesis. Evidence of impermissible interactions would force one to reject the MWF hypothesis.

It is important to motivate why we preserve the degree of modularity present in the weak fusion model. It is certainly possible to imagine a depth computation that processes multiple depth cues in a single, interactive computation (strong fusion), and several have been proposed in the computational literature. However, if the observer computes depth from stereo and motion in combination, then she/he will be in trouble if a scene contains disparities but no motion, or contains motion but no disparities. One should also consider the case of ancillary cues which signal the reliability of one or another depth cue. If an ancillary measure for the absolute value of viewing distance is available for gauging the reliability of stereo, and another measure of the general quantity of shearing motion in the scene is available for gauging the reliability of the KDE, over time it is in the best interests of the observer to learn to break up the depth problem into smaller, separate component pathways which may be controlled by these separate ancillary cues. Modularity is a defensible position against the factorial structure of the world. Sloman and Rumelhart (1992) have recently developed an adaptive network model capable of reorganizing itself into effectively independent networks each of which takes input from a subset of the possible inputs. It is at least plausible that such an adaptive network could in principle be able to reorganize itself into modules corresponding to distinct depth cues and to identify and use ancillary cues appropriate for each module.

BAYESIAN ESTIMATION

The Bayesian framework provides a powerful, general technique for selecting optimum depth and shape estimation rules when sufficient information is known about the prior distributions of cues (see the excellent review by Yuille & Bülthoff, 1995). It would be possible, in many respects, to approximate a MWF observer using Bayesian methods. In this section we examine the Bayesian framework and its relation to the key issues of promotion, dynamic reweighting, robustness, and the linearity of cue combination. We conclude that the Bayesian framework provides an elegant method for the representation of what we have called "depth maps with parameters" and promotion. The remaining issues can be treated within the Bayesian framework only with some difficulty.

The Bayesian approach

Ferguson (1967) and Berger (1985) describe the Bayesian approach to statistical decision theory in detail. What follows is an outline of the Bayesian approach. Let I denote the "image data" available for depth and shape. Let $\{S_n|n = 1, \cdots, N\}$ be the set of all possible scenes. For notational convenience we will assume there are only a finite number of scenes, and also a finite number of possible images $I_1, \cdots, I_m$. The image I is one of these images, and given I, we wish to decide which S_n is the scene we are viewing. We assume that we know the probability, $P[I|S_n]$, that the observed data I were generated by any given scene S_n. We assume we also know the probability that we encounter any particular scene, $P[S_n]$, and the probability that we would observe the image I, denoted $P[I]$.*

Applying Bayes' Theorem, we compute

$$P[S_n|I] = \frac{P[I|S_n]P[S_n]}{P[I]}, \tag{8}$$

the *posterior probability distribution* that the scene we are viewing is S_n given the observed data I. Last of all, we must reduce the distribution $P[S_n|I], n = 1, \ldots, N$ to an estimate of the scene we are viewing, $S_{\hat{n}}$. One commonly-used method is to select the index n that maximizes $P[S_n|I]$, the maximimum *a posteriori* estimator (MAP) (see Yuille & Bulthoff, 1995). The MAP rule picks the scene that is most likely, given what we know. The MAP rule is an example of a decision rule: it takes the image I and computes an index $\hat{n}$ that serves as our estimate of the true scene S_t.

Bayesian decision theory provides a way to choose a decision rule that takes into account the consequences of our decisions. Those consequences are summarized as a loss function, $L(t, \hat{n})$, which specifies how much we lose if we decide that the current scene is $S_{\hat{n}}$ when, in fact, S_t is what is really out there. Since $\hat{n} = d(I)$, we can represent the loss for any decision rule as, $L(t, d(I))$. This quantity is a random variable, since I is. To better measure how risky a particular decision rule is, we compute the *expected loss* or *risk*,

$$R(t, d) = \sum_{j=1}^{m} L(t, d(I_j))P[I_j|S_t]. \tag{9}$$

This quantity is readily interpretable. Given that S_t is the true scene, each image I_j has probability $P[I_j|S_t]$ of occurring. When image I_j occurs, the decision rule d incurs loss $L(t, d(I_j))$. Equation (9) is the average loss we would incur if S_t were the true scene and we chose to use decision rule d.

Now, we could have two candidate decision rules d_1 and d_2 where d_1 has very low risk if S_1 is the true scene, and very high risk if S_2 is the true scene, and vice versa. We assume the rules are equally risky on all the remaining scenes. Of course, we don't know which scene is the true scene, and so it's not clear which of these two rules to use. If, however, we knew that S_1 occurred very rarely, while S_2 is something we encounter often, we would likely favor the decision rule that has lower risk with respect to S_2.

We know the probability of each scene and, consequently, we can compute the expected risk (termed the *Bayes' risk*) given $P[S_n]$:

$$r(d) = \sum_{i=1}^{n} R(i, d)P[S_i]. \tag{10}$$

If we combine equations (9) and (10),

$$r(d) = \sum_{i=1}^{n} \sum_{j=1}^{m} L(i, d(I_j))P[I_j|S_i]P[S_i], \tag{11}$$

and use the identity $P[I_j|S_i]P[S_i] = P[S_i|I_j]P[I_j]$ (a variant on Bayes' Theorem), and interchange the order of summation, then

$$r(d) = \sum_{j=1}^{m} \left[\sum_{i=1}^{n} L(i, d(I_j))P[S_i|I_j] \right] P[I_j]. \tag{12}$$

The term in brackets is the loss we expect should we observe image I_j,

$$\sum_{i=1}^{n} L(i, d(I_j))P[S_i|I_j]. \tag{13}$$

If we minimize this term for each I_j, we minimize the Bayes' risk in equation (12) as well. The Bayes' decision rule is to choose $d(I_j)$ to minimize this term. Note that equation (13) is the expected loss with respect to the posterior probability distribution. The Bayes' decision rule is simply, "once I is observed, choose the estimate $\hat{n} = d(I)$ that minimizes the expected posterior risk". If we do know the necessary prior distributions and conditional distributions, and we are, in fact, attempting to minimize the expected loss, the resulting Bayes' estimation rule is optimal.†

If the loss increases as the square of the discrepancy between $\hat{n}$ and t ("square error loss"), then the Bayes' estimation rule will be the mean of the distribution $P[S_n|I]$. If the loss increases as the absolute value of the discrepancy between $\hat{n}$ and t, then the Bayes' estimation rule will be the median of the distribution $P[S_n|I]$ ("absolute error loss"). Such choices of loss

*The prior distribution $P[I]$ can be computed from the prior distribution $P[S_n]$ and the conditional distribution $P[I|S_n]$.

†If we use continuous parameters to describe the scene (rather than the discrete parameter n) then it may be that there is no optimal rule. There are then, however, rules that approach optimal performance as close as desired (Ferguson, 1967).

functions would only make sense if scenes were indexed appropriately.

If I is simply the contents of the retina, then the Bayes' estimation rule is, in general, a form of strong fusion. Nothing about the computations above requires that we identify "cues to depth" or segregate information into modules by cue. Yuille and Bülthoff (1995) use the term *strong coupling* for such an application of the Bayes' framework.

If the organization of a visual system were modular, perhaps because of computational limitations, we could still apply the Bayes' method to each of the modules. Suppose that depth information is segregated by cue type. M_1 might be a disparity map, M_2 might be a retinal velocity map, etc. Each of these maps is derived from the image data I and we will process the different $M_k, k = 1, \cdots, p$ separately.

The output of each module m would be a posterior distribution $P[S_n|I_m]$. The posterior distribution can be regarded as a *likelihood* function, a measure of evidence that each scene S_n is the true scene (Edwards, 1972). Following one of the weak coupling methods of Yuille and Bülthoff, we could combine these likelihoods across modules by multiplying them, to compute an overall likelihood function,

$$L[S_n|I] = \prod_{k=1}^{p} P[S_n|M_k]. \tag{14}$$

The likelihood function L is not in general a probability distribution and is not to be thought of as a probability distribution. It is a numerical measure of the evidence in favor of choosing each scene S_n as the estimate. There is no reason to expect the depth maps $M_k, k = 1, \ldots, p$ to be independent random variables. The multiplication in equation (14) is simply an atheoretical but plausible way of combining evidence across several sources.

Bayesian approaches are not, therefore, inherently weak or strong. They permit incorporation of prior knowledge and explicit loss functions into the estimation process. Nakayama and Shimojo (1992) develop a Bayesian-like procedure for estimating shape and depth, which, by design, does not use prior information about the relative frequency with which scenes are encountered. They argue that many classes of scenes (notably those exhibiting transparency) occur so infrequently that it is difficult to see how any reliable estimate of the probability $P[S_j]$ could be obtained. They also note that observers have little difficulty perceiving very unusual scenes without distortion. In the absence of useful prior information, they choose a "uniform prior", setting $P[S_j]$ to a constant. Their estimation is then the scene j that maximizes $P[S_j|I] = CP[I|S_j]$ where C is a constant, that is, the MAP rule. The MAP rule with a uniform prior also maximizes $P[I|S_j]$ which, in this case, is also the maximum likelihood rule, "choose j to maximize $P[I|S_j]$". Such rules are asymptotically optimal in many respects, and have other desirable properties (Kendall & Stuart, 1979, Chap. 18). Indeed, the Bayesian estimation rule is asymptotically equivalent to the maximum likelihood rule (Kendall & Stuart, 1979, p. 104).

Linear combination of cues

Yuille and Bülthoff (1995) demonstrate that some forms of weak coupling approximate a weighted linear rule of combination when cues are not too discrepant. Suppose that we are interested in estimating the depth D in a particular direction in the visual field and that there are two depth cues M_1 and M_2 available with posterior distributions $P[D|M_1]$ and $P[D|M_2]$. The MAP estimate of depth from cue 1 alone is the point D_1 where $P[D|I_1]$ reaches its maximum. The MAP estimate of depth from cue 2 alone is the point D_2 where $P[D|I_2]$ reaches its maximum. The likelihood function obtained by the weak-coupling method discussed above is $L[D] = P[D|I_1]P[D|I_2]$. Suppose it reaches its maximum at D_{12}. Yuille and Bülthoff show that, if the estimates from the two cues are not too discrepant, the Bayes' estimator approximates the linear combination rule,

$$D_{12} = \frac{w_1 D_1 + w_2 D_2}{w_1 + w_2}, \tag{15}$$

where

$$w_i = -\frac{d^2 \log P_i[D|M_i]}{dD^2}(D_i), i = 1, 2. \tag{16}$$

They also show that the weights w_i are positive. This rule of combination approximates a linear rule of combination with positive weights summing to 1.

If the posterior distributions are normal with means D_i and variance σ_i^2, then

$$w_i = \frac{1}{\sigma_i^2}.$$

The weights are inversely proportional to the variances of the two distributions, precisely as in our example above. The weak-coupling Bayesian estimate here coincides with the maximum likelihood and least-squares estimates.

Robustness

Note that the weights computed in equation (16) depend on the second derivative of the posterior distribution at its maximum. Empirical estimates of first and second derivatives are notoriously unstable (Anderssen & Bloomfield, 1974), and the assumption that the distribution is perfectly Gaussian in a very small neighborhood of the maximum is not empirically verifiable. It would be possible to make small changes in the shape of the assumed distribution in the vicinity of the maximum that would lead to very different weight estimates in equation (16), while samples from the two distributions would be empirically indistinguishable. In brief, if we use equation (16) to estimate the weights, the estimator above is not robust.

Bayesian estimators are chosen to be optimal when there is perfect knowledge of certain prior and conditional distributions. There is no reason to expect that the optimal estimator will remain approximately optimal when the true distributions are slightly different from our assumed distributions (Huber, 1981). The performance of a robust estimator, in contrast, would not be greatly affected by small failures of distributional assumptions. Nothing about the Bayesian framework requires that an estimator be robust. Berger (1985) discusses attempts to make Bayesian estimators more robust with respect to failures in the distributional assumptions.

Dynamic reweighting and ancillary cues

The MWF framework concentrates attention on the weights used to combine estimates of depth and shape derived from different cues, and on the influence of ancillary cues on these weights. Recall that ancillary cues are cues (such as vestibular input) which convey information concerning the reliability of depth cues. We will discuss below how these weights can be measured experimentally. With the weak coupling Bayesian rule of combination discussed above, the reliability of cue type m is encoded by the posterior distribution $P[S_n|I_m]$. (Recall that the posterior distributions are then multiplied together to produce a likelihood function.) Any ancillary cue not available in the input to module m is unable to affect the reliability assigned to the cue. There is no obvious reason why the inputs to each module should contain this information. The restricted role assigned to ancillary cues in the framework seems to have no natural expression in the weak coupling Bayesian model discussed here. Of course, if it were demonstrated empirically that the weights assigned to a depth cue were determined by the input I_m to its module, then this objection to the weak coupling model would not be relevant.

A strong coupling Bayesian method could easily incorporate the notion of ancillarity. We see no way, however, to reject such a model of depth and shape processing, and consequently consider it to be of little value in guiding experimentation.

Promotion

Recall that certain depth cues could not be used, in isolation, to compute depth estimates. We termed the output of the hypothesized module corresponding to such a depth cue to be a "depth map with parameters". Once the missing parameters in such a representation were estimated, the depth map with parameters could be promoted to a depth map and combined with other depth maps.

Suppose, for example, that the quantities ϕ_i are the distances (in meters) to n identical objects in the scene. We parameterize them in a depth map by,

$$\theta_i = \frac{\phi_i}{\phi_n}, i = 1, \ldots, n-1.$$

That is, θ_i is the (known) ratio of the depth of object i to object n and we let θ_n be ϕ_n, the unknown distance to the nth object. The output of the relative size module is a depth map with parameter θ_n. Any estimate of θ_n permits us to promote the depth map: the parameters $\theta_1, \cdots, \theta_n$ all known specifies the distances to the n identical objects.

Within the Bayesian framework such a depth map with parameters could be readily expressed by posterior distributions whose marginal distributions with respect to the missing parameters were uniform.* Returning to the example above, let us assume there are only two objects in the scene, that θ_1 is the ratio of the first object's distance to that of the second, and θ_2 is the second object's distance. The output of the relative size module provides information about the value of θ_1, but no information about θ_2. The resulting bivariate probability density function is

$$f(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\theta_1^2}{2\sigma^2}\right)\frac{1}{\theta_{\max}}\chi_{[0,\theta_{\max}]}(\theta_2), \quad (17)$$

where $\theta_{\max}$ is a large constant representing the greatest possible distance that θ_2 can be, $\chi_{[0,\theta_{\max}]}(\theta_2)$ is a function that is 1 if θ_2 is in the interval $[0, \theta_{\max}]$ and otherwise 0, and σ is a parameter computed by the module. If σ is small, the resulting distribution along the θ_1 axis is sharply peaked, indicating that the module has a sharp estimate of θ_1 and conversely. All we know about θ_2, though, is that it is somewhere between its physical limits, that is we know nothing about it we did not know before the module completed its computations.

Suppose that we had a second module which is able to accurately estimate θ_2, but cannot estimate θ_1. Then we might represent its posterior by

$$g(\theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\theta_2^2}{2\tau^2}\right)\frac{1}{\theta_{\max}}\chi_{[0,\theta_{\max}]}(\theta_1), \quad (18)$$

If τ is small, the module is signaling an accurate estimate of θ_2. Suppose τ and σ are both very small. Then the first module has a good estimate of θ_1 and the second module has a good estimate of θ_2. Multiplying the posterior distributions of the two modules together (and suppressing the χ functions for clarity),

$$L(\theta_1, \theta_2) = g(\theta_1, \theta_2) f(\theta_1, \theta_2)$$
$$= \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\theta_1^2}{2\sigma^2} - \frac{\theta_2^2}{2\tau^2}\right), \quad (19)$$

a sharply peaked function. We have, in effect, promoted both modules and combined them in one step. In this way, depth maps with parameters are easily expressed by posterior distributions that are uniform with respect to the missing parameters.

In summary, the Bayesian approach is of potentially great value. It does not lend itself, however, to expressing the kinds of issues (and hypotheses) we address here. Reformulating the MWF observer as a weak coupling Bayesian observer would obscure rather than enhance the issues we seek to resolve.

*If θ_n can take on any positive value, then it does not have a uniform prior as no constant function on the positive reals can be a probability density function. A uniform distribution on $[0, \theta_{\max}]$ with $\theta_{\max}$ chosen to be very large is, for practical purposes, sufficient. We take this as our uniform prior in the sequel. See the discussion on approximate Bayes' rules in Ferguson (1967) for methods of approximating Bayes' rules in such cases.

DEPTH CUE COMBINATION RESEARCH AND THE MWF FRAMEWORK

The issues discussed above suggest the possibility of reinterpreting previous empirical work on depth cue combination in the light of the MWF model. Some previous work has been consistent with the following model (which is assumed implicitly in several papers): (l) the depth system is divided into modules; (2) estimates of depth from these modules are combined into one numerical estimate at each spatial location; and (3) the rule of combination is insensitive to the magnitude of deviations. This model is essentially the Weak Observer introduced above.

In some cases, previous researchers manipulated two or more depth cues, generally with large discrepancies between the depths signaled by the two cues, discovered that the cues interact, and concluded that the Weak Observer model is untenable. Since no depth model that permits promotion, is robust, or reweights cues normatively, can avoid some interactions or apparent interactions between depth cues, the finding of interactions between cues is not surprising. They are consistent with a modular organization such as the MWF framework that permits only limited interactions between modules when cues are not discrepant. In this section we review previous work in light of the issues raised before. Although some of this work appears to exclude a modular organization, we argue that the findings of interactions can be reinterpreted as involving issues of cue reliability, robustness, or promotion, that is, it is consistent with the Modified Weak Fusion framework.

Visual heuristics

An observer of an Ames room (see Ittelson, 1952) arrives at incorrect estimates of depth and size as a consequence of systematic distortions in linear and texture perspective cues. If the viewer sees two adults standing in the two far corners of the distorted room, one will appear far taller than the other. Thus, a weakened "familiar size" cue (i.e. "all adults are about equally tall") is abandoned in favor of the interpretation derived from the distorted perspective cues. The observer must choose between two sources of depth information and routinely chooses the wrong one.

The Ames chair may also be seen as creating cue conflict. The figure again involves an "accident of viewpoint". A three-dimensional scene is created which, when viewed from one special viewpoint, produces the same retinal image as a chair, even though the actual object is extremely distorted. The usual explanation is that even when conflicting cues are available (texture, shading, binocular disparity), the image may still appear chairlike because of visual heuristics which bar such accidents of viewpoint. Examples of such heuristics include: "lines which are nearly parallel in the image are parallel in 3D"; "lines which meet at a common vertex in the image also meet at a common vertex in 3D"; and "lines which meet or cross and form nearly a right angle in the image are, in fact, perpendicular in 3D". The first two heuristics involve a statement about the unlikely nature of such accidents of viewpoint, and have led to successful computer vision systems [Bennett, Hoffman & Prakash (1989) formalize and discuss such heuristics, and Lowe (1985) has built successful computer vision systems using them]. The latter is likely a cultural bias based on familiarity with human-made objects. The real three-dimensional version of the Penrose impossible triangle constructed by Gregory (1970) is another example of a misperception based on an accident of viewpoint.

Gregory (1970) discusses other cases in which misperceptions occur. A Necker cube has two possible interpretations (either lower or upper face forward). Yet, when a real three-dimensional wire-frame cube is constructed, viewers still can perceive the reversed interpretation. Thus, although one might expect binocular stereopsis (or even tactile cues!) to disambiguate the perspective interpretation of the cube, there are occasional failures. A view of the inside (concave side) of a facial mask often appears to be convex (see, e.g. Yellott, 1981). Again, this is a case where the ambiguity of one cue (shape-from-shading) which would normally be correctly disambiguated by a second cue (binocular disparity) is misperceived due to a third, more heuristic cue ("faces tend to be convex"). Cue promotion is necessary, but the cue promotion mechanism is error-prone.

Cue conflict and ambiguity

A large amount of research in cue combination concentrates on the cue conflict situation, or the similar case wherein an unambiguous cue might disambiguate a second, ambiguous cue (and if it fails to do so, results in cue conflict). Schwartz and Sperling (1983) found that linear perspective cues do not generally disambiguate the kinetic depth effect. Braunstein *et al.* (1982) and Proffitt *et al.* (1984) found that occlusion helps to disambiguate the kinetic depth effect. Prazdny (1986) described a random-dot cinematogram which portrays a flat object in front of a background which changes its two-dimensional shape consistent with a three-dimensional rotating wire object, and suggested that the appearance of three-dimensionality in these displays implies that the KDE effectively vetos the stereo disparity cue.* Stevens, Lees and Brookes (1991) found strong individual differences when binocular disparity and monocular curvature cues were in conflict. Rogers and Collett (1989) found that when subjects viewed displays including both motion parallax and stereo disparity, the appearance was determined by a combination of both cues in a manner which effectively maximized the

*We have constructed stimuli quite similar to those described by Prazdny (1986) and have shown them to a large number of people. Many observers state that they can perceive both the flatness of the stereo display *and* the depth depicted by the motion. This only serves to underscore our point that cases of extreme conflict are difficult to analyze and the results need not reflect what the visual system does under natural viewing conditions.

ɔnsistency between the percept and the information rovided by both cues. Others have investigated cue ɔnflict for combinations of kinetic depth and stereo isparity (Epstein, 1968; Wallach & Karsh, 1963), occluon and stereo disparity (Cavanagh, 1987), texture nd stereo disparity (Gillam, 1968; Stevens & Brookes, 987, 1988; Youngs, 1976), and occlusion and shading Ramachandran, 1988).

Many of these studies are interpreted according to a ıodel in which depth cues are arranged hierarchically, nd for any scene the highest ranking cue is used. This essentially a cue veto model. In the Ames displays the erspective cue and other visual heuristics veto cues of ımiliar size. The Prazdny (1986) cinematograms were sed to demonstrate that the KDE effectively vetoes ereo. However, these classes of results are also consistıt with the MWF framework. When cues are reasonbly consistent, they will combine in a linear fashion ınd this may even happen when they are inconsistent the reliabilities are sufficiently poor). When cues re reliable and inconsistent, robust statistical methods nply that new considerations come to play, and vetoing ehavior can be one of the results. The degree of cue ıconsistency and reliability determine the degree to hich results depart from simple depth averaging. n interesting idea (made by an anonymous reviewer) ıggests that the combination rule may also be endowed ith a memory. This could lead to hysteresis effects hereby averaging continues to occur as cues are ıade increasingly inconsistent, even beyond the point here averaging is used normally for the first glance at scene.

Other studies emphasize cue cooperation. This ıcludes cases where one cue is used to disambiguate ıother, such as occlusion disambiguating KDE 3raunstein *et al.*, 1982; Proffitt *et al.*, 1984), or the sparity of a highlight disambiguating shading (Blake & ülthoff, 1990, 1991). Disambiguation is simply a case of epth cue interaction in the service of cue promotion. he results of Rogers and Collett (1989) may also be en as a side-effect of cooperation between motion arallax and stereo for the purposes of cue promotion.

inearity and cue cooperation

Several recent studies have included conditions in hich multiple depth cues were consistent along with ixed-cues stimuli to learn more about cue interactions. ı several cases the results were consistent with a linear quasi-linear combination rule (Braunstein, 1968; runo & Cutting, 1988; Cutting & Millard, 1984; Dosher *al.*, 1986; Foley, 1977). The MWF framework predicts ıat when cues are perturbed by small amounts from a ɔnsistent-cues condition, linearity is to be expected. owever, in these studies departures from linearity were ten small with strongly inconsistent cues as well. In ırticular, the work of Dosher *et al.* (1986) involved ting a linear model to data collected over a wide range cue conditions from consistency to strong conflict ıd, surprisingly, found very strong support for a linear ɔmbination model. It is not clear why a linear model was successful over such a wide range of cue combinations, and why robustness considerations did not come into play. These displays involved four cues to depth: stereo, brightness as a cue to proximity, linear perspective, and kinetic depth. The subjects did not indicate the amount of perceived depth, but merely which KDE orientation was perceived (of the two possible rotations given the KDE reflection ambiguity). The linear model provided predictions for depth sign as a function of the portrayed stereo and brightness cues. It is possible that robustness considerations (i.e. nonlinearities) were not seen because of the paucity of cues. And, since perceived depth was not directly measured, it is also possible that perceived depth was not linear in the portrayed depths, even though depth sign was predictable from a linear combination (and a fixed, additive noise model).

Buckley and Frisby (1993) recently reported experiments involving combinations of stereo and texture cues using a magnitude estimation task. In most, but not all, of their conditions the results were consistent with a linear combination rule. In addition, the weights of the individual cues varied across display conditions. For example, stereo was given far higher weight when real objects were used than when using simulated video displays.

Shape measurement studies

A number of researchers have attempted to measure the amount of depth perceived at a given location in the visual field by asking the subject to compare the perceived depth of two different objects or features. This has been done to investigate the combination of depth information across space [interpolation (Würger & Landy, 1989)] to compare the amount of depth engendered by individual cues (Bülthoff & Mallot, 1988), and to examine cue interactions (Bülthoff, 1991; Bülthoff & Mallot, 1988; Johnston, Cumming & Parker, 1993; Johnston *et al.*, 1994; Landy, Maloney & Young, 1991; Parker, Johnston, Mansfield & Yang, 1991; Young *et al.*, 1993).

The first problem in studying cue interactions is to find a reliable experimental technique to measure perceived depth. Several methods have been used in the literature. Gogel (1976, 1977) presents a technique for measuring perceived depth indirectly by linking lateral object motion with lateral motions of the observer's head. When the observer perceives the object as stationary in exocentric space, lines from observer head positions through corresponding object locations intersect at a common point, which is the perceived depth of the object. This has been used successfully for analyzing depth in reduced-cue displays. It allows one to measure range to an object, but its applicability to measurements of relative depth or to more complex displays is questionable.

Bülthoff and Mallot (1988) presented subjects with a stimulus with one or more shape cues. The perceived shape was mapped by placing a binocular test spot at various locations on the shape and having the subject adjust the disparity of the test spot until it appeared to lie on the perceived three-dimensional surface of the stimulus shape. This technique is attractive since it is then possible to map out an entire perceived surface of arbitrary complexity. However, its interpretation rests on several assumptions. It requires that it be possible for the subject to compare the depth from a potentially multi-cue surface (which may or may not include disparity as one of the cues) to a single test spot with binocular disparity. Second, it assumes that this comparison is taking place at the level of "perceived depth" rather than by merely comparing disparities. Third, it assumes that the presence of the comparison dot does not alter the perceived surface shape. Finally, by using stereo alone to scale perceived depth, it assumes that stereo disparities are correctly calibrated [which is often false (Foley, 1980; Gogel, 1960; Johnston, 1991)]. Thus, cue promotion can stand in the way of a given experimental technique.

Another measurement technique involves comparisons of the stimulus shape to globally-defined norms. For example, Johnston (1991) describes a task called the *apparently circular cylinder* (ACC) wherein the observer chooses which among a series of cylinders varying in eccentricity appears to have a circular cross-section. This technique enabled the miscalibration of stereopsis to be measured which would not have been possible by reference to disparities of a comparison test spot. Finally, depth may be measured by comparison of a test stimulus in which individual cues are separately manipulated with a stimulus in which cues are congruent (e.g. Landy *et al.*, 1991). This allows one to measure the amount of perceived depth for a variety of shapes, and rests on the assumption that perceived depth will be veridical for a congruent cues stimulus as long as the number of cues and degree of realism in the simulation is sufficient.

Koenderink has championed the application of differential geometry both to the computational analysis of shape and the recovery of shape from various cues and to the perception of shape by human observers (Koenderink, 1990). Recently, this has led to empirical work to determine the shape representation scheme used by human observers. For shape measurement, his group has asked subjects to match surface normals using a perspective drawing of a plane and a normal to that plane (Koenderink *et al.*, 1992). [Previously, Mingolla and Todd (1986) had observers directly report the slant and tilt of surfaces, and Stevens and Brookes (1988) had observers adjust a stereo surface normal.] de Vries, Kappers and Koenderink (1993) asked observers to provide estimates of local surface *shape index* and *curvedness*, two indices of shape which are computed from the values of the two principal curvatures and taught to subjects. These are preliminary studies, but show promise as a means to studying shape representation.

The shape measurement studies have been quite useful. Bülthoff and Mallot (1988) found that several cues contributed to perceived shape including shading and two different forms of binocular stereopsis (both edge-based stereo and raw intensity-based stereo for stimuli which didn't include matchable edges or features). Later work (Blake & Bülthoff, 1990, 1991; Bülthoff, 1991) suggested that texture and specularities also contribute to perceived shape. In addition, they suggest a number of types of cue interaction: veto (where a "stronger" cue is selected over a weaker cue), accumulation (where cues simply summate), cooperation (where the cues interact in a nonlinear fashion), and disambiguation (where one cue helps resolve an ambiguity in a second cue). For example, a specularity (in particular, the disparity of a specularity) can disambiguate the sign of curvature derived from shading (a form of cue interaction in the service of promotion).

Some empirical results that have been interpreted as indicating strong fusion are actually consistent with MWF. Bülthoff (1991) discusses shape matching experiments where an ellipsoid with only pictorial cues (texture and/or shading) seen monocularly is adjusted to match the depth portrayed by a second ellipsoid for which the depth cue is stereo disparity. Using stereo disparity as the standard, he found that either texture or shading alone produced an underestimate of depth, but if both cues were present the resulting percept had a greater, more veridical amount of depth (actually, the depths from the two cues were approximately additive, rather than averaged in the combined percept). This was interpreted by Bülthoff and Yuille (1991) as indicating strong fusion of the texture and shading cues (which they term "strong coupling"). However, these display conditions affect both the availability of the cues of texture and shading and also the number of available cues. It is likely that additional cues were available which signaled a flat display (e.g. visibility of the edges of the monitor, accommodation, etc.). When fewer cues are available, objects can thus be perceived as having less depth than is actually portrayed (as in Todd & Reichel, 1989). By making additional shape cues available (texture *and* shading), the cues to flatness will receive less weight and the additivity of texture and shading is predictable. Displays with underestimated relative depth may be seen as having inadequate depth cues to support full depth. This provides a possible explanation for the sorts of displays which support the equidistance and specific distance tendencies (Gogel, 1965; Gogel & Tietz, 1973). The absence of a cue is not the same thing as the presence of a cue which indicates zero depth.

Tittle and Braunstein (1991, 1993) suggested that their data on stereo and KDE combination also imply a strong fusion scheme (cooperation between stereo and KDE to determine depth). However, the only interaction term found in their results may be summarized as: "motion may be helpful in solving the stereo correspondence problem". This rule applies to their demonstration, where the correspondence problem becomes too difficult for observers when disparities are large and the portrayed object is transparent (also described by Pong, Kenner & Otis, 1990). It also arises in Braunstein, Andersen, Rouse and Tittle (1986) and Rouse, Tittle and Braunstein (1989), where observers who fail tests of static stereopsis are indeed able to make stereo discriminations (to disambiguate KDE) when disparities are dynamic. In both cases, this is an interaction between motion and stereopsis, but not between depth-from-motion and depth-from-stereo. The only exception is that stereo helps to disambiguate KDE (Braunstein *et al.*, 1986), which is simply another example of cue interaction in the service of cue promotion.

In addition, Tittle and Braunstein (1991) found a difference between disparity alone and disparity plus KDE. They interpreted this as cooperation between the two cues. However, in the disparity-only case, there was stimulus motion (lateral translation under orthographic perspective), and this motion could be interpreted as a cue to flatness. Thus, the same argument suggests that these results are consistent with MWF. An appeal should not be made to strong fusion when a simpler, testable model is available. In fact, the second experiment of Tittle and Braunstein (1991) examines combinations of various amounts of stereo and KDE depth, and the results were completely consistent with a linear combination scheme.

Johnston (1991) was able to use the ACC technique to quantify the miscalibration of stereopsis, and suggested that all of her results could be explained by scaling disparity by a misestimate of the absolute viewing distance. Parker *et al.* (1991) used the ACC to study combinations of texture, shading and stereopsis and found that the addition of texture and stereopsis improved the depth estimates slightly. Johnston *et al.* (1993) investigated this question further for stereopsis and texture, and found that the results could be explained by a linear combination rule. Landy *et al.* (1991) came to the same conclusion for combinations of texture and object motion (KDE) using the shape comparison technique.

Summary: depth cue combination models

Although the combination of depth cues has been examined by many researchers, there have been relatively few models for depth cue combination which allow a theory of cue combination to be fit to empirical data. These models have either been linear or multiplicative. One linear model is described by Dosher *et al.* (1986) to fit their data on the disambiguation of KDE by stereo and brightness cues. The binary choice data are fit by assuming linear depth combination followed by a late noise process (i.e. Thurstonian Case V scaling). Bruno and Cutting (1988) fit depth rating judgments as four cues were manipulated (size, height, occlusion and motion), and found reasonable fits of a linear cue combination model. Massaro (1988) replied to this article by pointing out the difference between the absence of a cue (a static object has no KDE information) as opposed to a cue indicating no depth (two objects of equal size are an indication of equal depth). More importantly, Massaro fit the data of Bruno and Cutting using his fuzzy logical model of perception (FLMP), which is approximately a multiplicative model (with normalization), and found that this model fit the rating data better than the linear model more often than not. Cutting *et al.* reanalyzed their data and strengthened their case that a linear model provided an adequate fit, and demonstrated that the FLMP model was too general, being capable of fitting randomly generated data (Cutting & Bruno, 1988; Cutting, Bruno, Brady & Moore, 1992).

The MWF model suggests that depth combination is linear under restricted conditions, and under these conditions it reduces to the models of Bruno and Cutting (1988)* and with a suitable noise model, to that of Dosher *et al.* (1986). However, the MWF framework also indicates conditions under which nonlinearities are likely to occur as cues weights change due to changes in cue reliability, cue availability, or cue inconsistency or as cues interact to effect cue promotion. The MWF model is used to predict perceived depth, and hence is suitable for experiments in which depths of two stimuli are compared. The Bruno and Cutting (1988) data are based on depth judgments (from a rating scale), and hence carry the possibility of a nonlinear mapping from perceived depth to chosen rating. Obviously, this can make a linear combination rule look multiplicative (if the mapping to ratings is exponential) and can make a multiplicative combination rule look additive (if the mapping is logarithmic). We suggest that the appropriate datasets for contrasting alternative models of depth cue combination must therefore derive from tasks which allow an estimation of perceived depth at least on an interval scale using either depth comparisons (Landy *et al.*, 1991; Young *et al.*, 1993) or absolute shape judgments (Johnston *et al.*, 1993, 1994). As we shall indicate below, results from these studies are consistent with both the linearity of weighted depth averaging and the nonlinearity of promotion via cue interaction.

In summary, there is a plethora of results at this point concerning the combination of multiple cues to depth. Our aim here is to place these results in a common framework. We do this by first examining the task at hand from a theoretical point of view: what is the optimal combination rule, and under what conditions is it optimal? Having determined this, we find that the empirical results are reasonably consistent with this viewpoint. The key issue is to find a set of experiments which can be used to test the assumptions of the MWF paradigm. The next section outlines an experimental paradigm for testing MWF and determining its parameters. The experiments we have carried out

*Cutting *et al.* (1992) fit a model which they claim is an adaptation of MWF to their judgment task. It is a linear model with occlusion omitted from the fit (because we claim that occlusion is ordinal and therefore can only be used for disambiguation) and, needless to say, provides a poorer fit. However, it is not clear that degree of occlusion can not provide a cue to amount of depth in their displays (as noted previously). To distinguish the MWF model from a purely linear model requires more data (such as that of Johnston *et al.*, 1994). The two models cannot be distinguished using the data of Bruno and Cutting (1988).

so far have been consistent with the model we have presented.

METHODOLOGY: PERTURBATION ANALYSIS

At its heart, the MWF framework is quite simple: nondiscrepant depth cues are combined using a dynamic weighted average. Ancillary information concerning cue reliability determines the weights. If cues are discrepant, then the influence of discrepant cues should decrease with discrepancy (robustness). The MWF framework is the "maximally modular" model of biological vision. The only interactions between modules that are permitted are those needed for promotion. Hence the MWF framework predicts that interactions between cues will be observed experimentally, but it also sets out the conditions under which interactions will occur.

We propose here an experimental technique which permits testing the MWF framework by limiting experimental manipulations to small perturbations of depth portrayed by each cue. We term the method *perturbation analysis* in analogy to the engineering techniques of the same name.* We intend to operate in the range of the MWF framework where the rule of combination is a weighted average. We seek to measure each weight as a function of the objective reliability of the corresponding cue. Employing larger perturbations, we wish to test whether the visual system reacts to a single discrepant cue by reducing its weight.

It is useful to think of our approach by analogy to color discrimination research. Color vision researchers have often proposed models for color discrimination which consist of several linear mechanisms which share a common state of adaptation. Changes in the state of adaptation are seen as changing the parameters of each linear mechanism, but when adaptation is maintained at a constant state, the color mechanisms are essentially linear (Stiles, 1978). Small perturbations in the color of a test stimulus are required to preserve the identity of these linear color mechanisms to study their properties in isolation. Large perturbations affect the adaptational state and make the results difficult to interpret. Since we are interested in studying a system which adjusts its parameters based on cue availability and considerations of robustness, we are forced to use small perturbations to isolate the linear aspects of the depth combination rule.

The analogy to color vision might be extended a bit further. Recently there has been progress in modeling color constancy (see Maloney, 1992 for review). One interpretation of this work is that the state of color adaptation is the visual system's attempt to take into account the illuminant to best serve to associate a constant color descriptor for surface colors under varying illumination conditions. The theoretical results suggest that a reasonably large sample of colors must be visible at one time for a reliable calculation to occur. As such, displays with only a few colors (such as the test and field stimuli used universally in work on color discrimination and appearance) place the observer in a situation where a reliable estimate of surface color is not possible, and in which it is impossible to predict observer behavior based on these considerations. The color constancy mechanism, if indeed it exists, was designed to operate under natural, realistic viewing conditions (in which a rich array of colors are always visible), and is defeated by the typical experimental models. Simpler, more interpretable data result when a richer experimental model is employed which matches the requirements of the adaptational system (Brainard & Wandell, 1991). Similarly, the MWF framework suggests that results of depth cue combination experiments will be easier to interpret for displays which include several cues, appear realistic to the observer, and where the depths portrayed by the individual cues are reasonably consistent.

We next describe a psychophysical procedure (perturbation analysis) that permits us to measure the weights in the weighted average psychophysically. As above, we continue to speak of the depth map as if it were an array of distances. If instead it comprises parameterized surfaces, we only require that the rule of combination of parameterized surfaces derived from separate cues is an approximate averaging of the depth surfaces for small enough perturbations.

Suppose that we simulate a scene containing two cues. The cue information available concerning depth is discrepant: Cue 1 assigns a depth of d_1 to a specific point in the scene. Cue 2 assigns a depth of $d_2 = d_1 + \Delta\text{cue}$. This is called the *mixed-cues* stimulus. We also simulate a scene that is the same in every respect except that the point is assigned a single depth $d' = d'_1 = d'_2$ by both cues (the *consistent-cues* stimulus). We can adjust the value of d' (e.g. by a staircase method) until the subject considers the depth associated with the two cues in conflict d_1, d_2 to be the same as the depth associated with the two consistent cues d'_1, d'_2, as measured by indifference in a forced-choice task. We then express the depths as weighted averages of the depths assigned by different cues:

$$d' = \alpha_1 d_1 + \alpha_2 d_2 = \alpha_1 d_1 + \alpha_2(d_1 + \Delta\text{cue}). \quad (20)$$

Since only two cues are available, we assume that $\alpha_1 + \alpha_2 = 1$, and hence

$$\alpha_2 = \frac{d' - d_1}{\Delta\text{cue}}. \quad (21)$$

If we write $\Delta\text{depth} = d' - d_1$, then

$$\alpha_2 = \frac{\Delta\text{depth}}{\Delta\text{cue}}. \quad (22)$$

In words, the weight α_2 is the ratio between the change in estimated overall depth (Δdepth) and the amount the corresponding cue is perturbed (Δcue). In particular, the α_1, α_2 weights are readily estimated from psychophysical data. Further, equation (22) contains a straightforward

*An interesting alternative paradigm for estimating observer weights was proposed by Berg (1989) for auditory tasks involving multiple observations.

test of the entire model framework. The values estimated for α_2 must be between 0 and 1.

For a linear estimator, any size perturbations result in the same slope estimate α. For the robust estimator we propose, the estimated slope will vary for large enough perturbations Δcue. By choosing only small values of Δcue we avoid any effect of robust estimation procedures on the estimate of α.

The above technique for measuring weights α extends to any number of cues so long as only one of them is perturbed by Δcue at a time. If the consistent- and mixed-cues stimuli are drawn from the same set of stimuli, then it is clear that for a value of $\Delta\text{cue} = 0$, then d' must equal d_1, since in this case the consistent- and mixed-cues stimuli are truly the same. Thus, for this case, the paradigm constrains the estimated weights to sum to 1. However, we have carried out experiments in which the mixed-cues stimulus is drawn from a different set than the consistent-cues stimulus (Young *et al.*, 1993). In particular, the mixed-cues stimulus had depth cues which were intentionally made less reliable (described further in the next section). In this case, it is no longer necessarily true for $\Delta\text{cue} = 0$ that the measured d' must equal d_1. Hence, in this situation, the assumption that the weights α sum to 1 is then empirically verifiable by measuring the two weights separately.

TESTS OF THE FRAMEWORK

The perturbation analysis method allows one to test the assumption of a linear combination of cues. Consider our recent work using the perturbation analysis technique to examine combinations of two depth cues: motion and texture (Landy *et al.*, 1991; Young *et al.*, 1993). Subjects viewed displays which simulated the front half of a vertical elliptic cylinder (a stretched tin can) rotating back and forth about a horizontal axis. The cylinders were covered with a texture of dark spots on a lighter background.

Stimuli included both a textural cue and a motion cue to depth. The portrayed depth from each cue was varied independently as follows. First, an elliptic cylinder's surface was "carved" from a simulated slab of a three-dimensionally textured material with the depth extent determined by the depth to be portrayed by the texture cue. This isotropically textured surface was projected onto a cylinder with the depth extent to be portrayed by the motion cue, creating an anisotropically textured surface, and then rocked back and forth about a horizontal axis [Fig. 6(A)]. These surfaces were projected (using parallel projection throughout) onto the image plane [Fig. 6(B)].

The axis about which a cylinder rotated was one of the semiaxes of one of the cylinder's elliptic cross-sections, midway between the bottom and top of the display, perpendicular to the line of sight. The length of this axis (the *width axis*) corresponded to the width of the displayed surface, and was constant across surfaces in these experiments. The length of the other ellipse axis (the *depth axis*) was varied to give impressions of cylinders of variable depth extent. The texture compression in the projected image corresponded to a depth axis of d_t while the projected motion of texture elements corresponded to a depth axis of d_k.

In each trial, two stimuli were displayed side-by-side. In one of the two stimuli $d = d'_k = d'_t$ (the *consistent-cues* surface). In the other, the two depth values were generally inconsistent, $d_k \neq d_t$ (the *mixed-cues* surface). The subject's task was to indicate with a key press which surface appeared to extend further in depth. From these two-alternative forced-choice comparisons, we determined which surface with consistent depth cues appeared to have the same depth extent as a stimulus with inconsistent depth cues, thus empirically specifying the function f which yields the combined depth estimate for inconsistent motion and texture depth cues, $d = f(d_k, d_t)$.

By and large the results of these studies are consistent with the theoretical model outlined above. For example, Fig. 7(A) shows the psychometric functions for a series in which d_k for the mixed-cues stimulus was fixed throughout and d_t for the mixed-cues stimulus was varied across blocks of trials. It is clear that increasing the depth portrayed by texture in the mixed-cues stimulus (d_t) decreases the likelihood that a given consistent-cues surface will be perceived as having more depth. These data were fit with parallel cumulative normal distributions. The point of subjective equality was estimated from the 50% point of the function fitted, providing an estimate of $f(d_k, d_t)$. Several such measurements of $f(d_k, d_t)$ are plotted in Fig. 7(B) as a function of d_t. It is clear that the data fall on a straight line, consistent with the linear cue combination model. In this case, the slope suggests that the weight of the texture cue for these experimental conditions is 0.46.

In other experiments (Young *et al.*, 1993) it was found that with small perturbations of depth from a single cue, it is always possible to compute the weight for that cue. We found that the weight of the motion cue was reduced when its availability was reduced (smaller amounts of motion viewed). Additional experiments were run in which one or the other cue was reduced in reliability (by adding random variation in texture element shapes or in motion paths). It was found that lowering the reliability of a cue results in that cue receiving a lowered weight. In some cases, there is a compensatory increase in the weight of the other cue so that the sum of these weights was also close to one. However, there were also cases in which a degraded display resulted in the weights summing to a value significantly less than one (as compared to a consistent-cues stimulus with undegraded depth cues). Thus, in poor-quality displays the observer effectively seeks depth information from additional cues (flatness of the display, accommodation, etc.) all of which indicate less depth [or reverts to flat because of a specific distance or equidistance tendency (Gogel, 1965; Gogel & Tietz, 1973)].

A second series of experiments has been carried out using combinations of stereo disparity and motion (Johnston *et al.*, 1994). In these experiments the

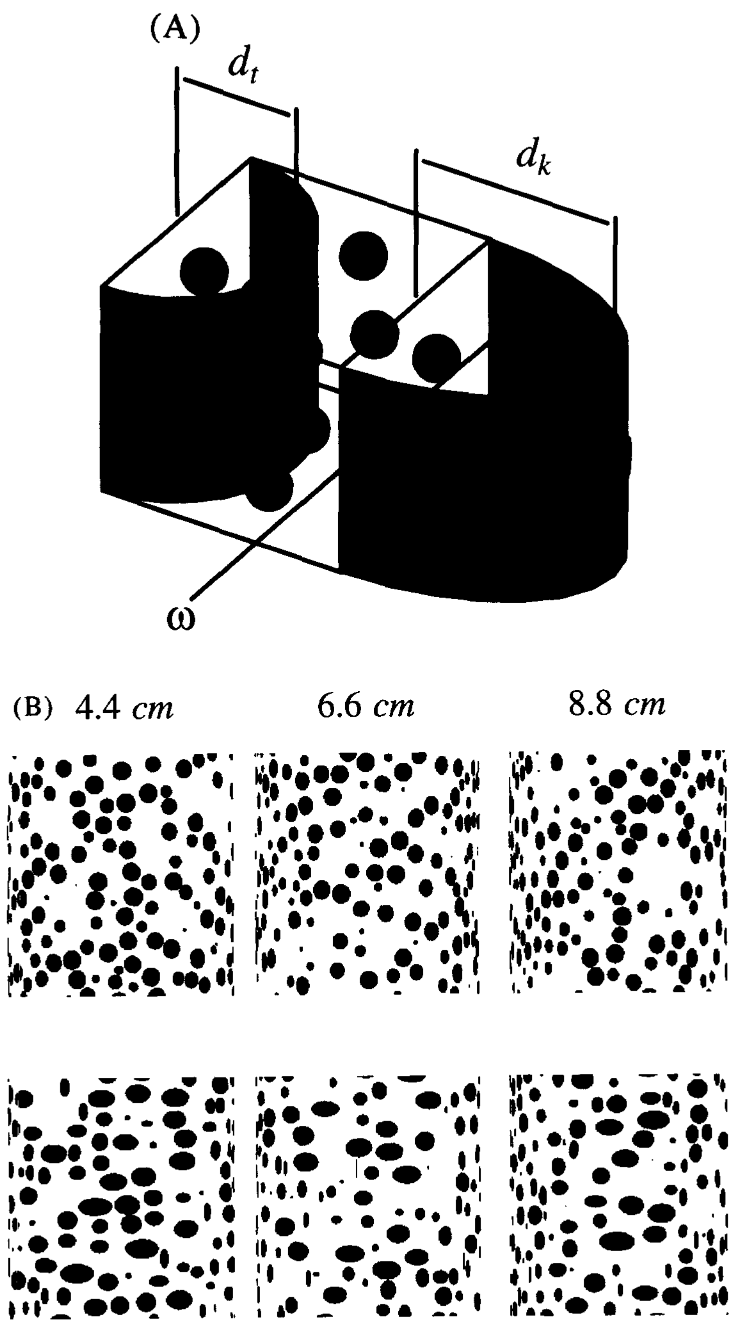
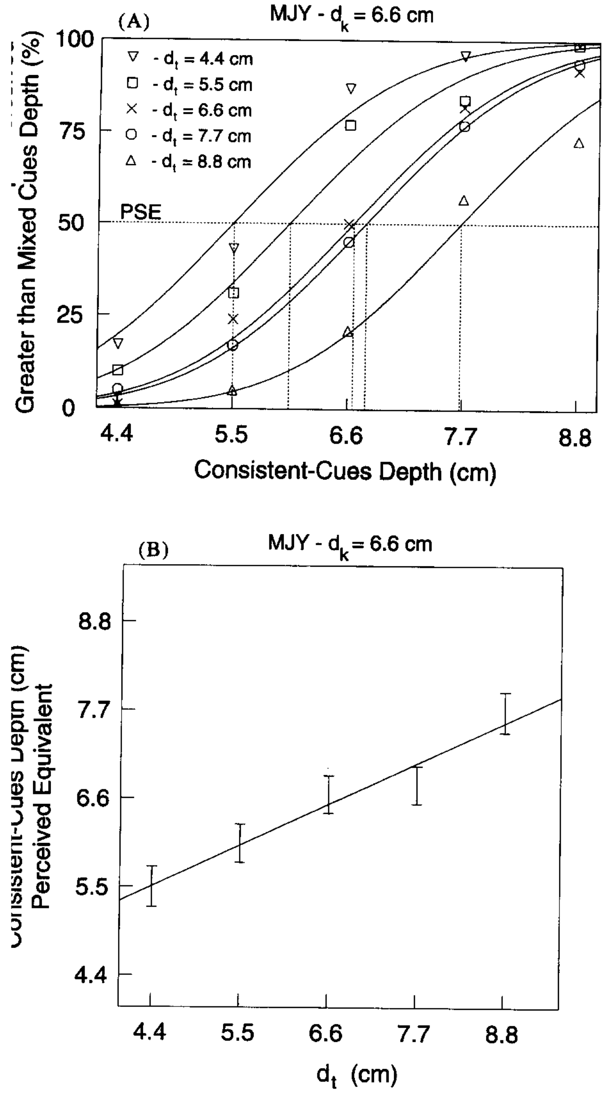


FIGURE 6. (A) A three-dimensional texture is carved to create an elliptic cylinder of a particular depth d_t. This surface is then projected to a new depth d_k and then rotated in depth. (B) The resulting surface is then projected onto the image plane using parallel projection. Images corresponding to three values of d_t are shown. The lower row of images result from adding noise to the texture foreshortening cue resulting in a lower reliability for shape-from-texture.

apparently circular cylinder task was used in which the observer specifies which among a series of elliptic cylinders varying in depth elongation appears to be circular in cross-section (Johnston, 1991). Figure 8(A) shows results of one experiment combining stereo and motion cues to depth using the ACC task. As depth portrayed by stereo is increased, the depth portrayed by motion must be correspondingly decreased for the stimulus to continue to appear circular. This tradeoff is linear, at least for most of the range of portrayed depths employed.

In this study the question of cue promotion was also examined. Stereo depth from disparity scales inversely with the square of the viewing distance, whereas depth from motion scales linearly with the viewing distance. Thus, in measuring depth from displays containing both of these cues, there is every reason to expect an interaction between the two cues for the purposes of cue promotion. Depth is computed from stereo disparities by a computation requiring knowledge of the absolute distance to the fixation point. Using the ACC task, Johnston (1991) discovered a strong nonveridicality in observer estimates of shape from stereo, and interpreted the results as a misinterpretation of stereo disparity as a result of using an incorrect value of the viewing distance to scale disparities into depth values. For short viewing



JRE 7. (A) Psychometric functions from a depth comparison ·iment involving texture and motion cues (Young *et al.*, 1993). curve represents a different value of d_t, and d_k is fixed throughout. ıbscissa shows the value of d for the comparison, consistent-cues ılus, and the ordinate shows the probability that the consistent-stimulus was perceived as having greater depth. As the discrepant of d_t is increased, the curves slide rightward. (B) The points of ctive equality estimated from (A) along with confidence limits. ıata are consistent with a linear trend. The slope of the curve provides an estimate of $\alpha_t = 0.46$.

ınces the viewing distance is overestimated, and for viewing distances it is underestimated. Thus, we ıosed that the different scaling properties of KDE motion would lead to an interaction between the cues (as in Richards, 1985). The results shown in 8(B) confirm this expectation. When stereo is the available cue, depth is misestimated. When motion e only available cue, depth is veridical. Note that a stimate of the viewing distance will not affect these on-only results; doubling the apparent viewing dis-e doubles the KDE depth estimate, but also doubles ıpparent size of the stimulus, resulting in an unged ACC. This is not the case for stereo. Thus, a o/motion interaction apparently helps to promote tereo depth cue, at which point depth combination is linear [as in Fig. 8(A)]. The data presented in Fig. 8(A) would be consistent with a mechanism for which KDE vetoes stereo, but other data collected at a shorter viewing distance are not consistent with this simple combination rule, providing support for the existence of a more sophisticated promotion mechanism. Of course, this is not the first time cue promotion has been found empirically. It is well known that the addition of occlusion information will partially promote a motion cue by disambiguating the sign of depth (Braunstein *et al.*, 1982; Proffitt *et al.*, 1984).

An interesting question is whether cues are ever fully promoted. An alternative has been suggested by Todd *et al.* (Todd & Bressan, 1990; Todd & Norman, 1991). They claim that the internal representation of depth from motion is not fully promoted, but rather is only computed up to a particular class of affine transfor-

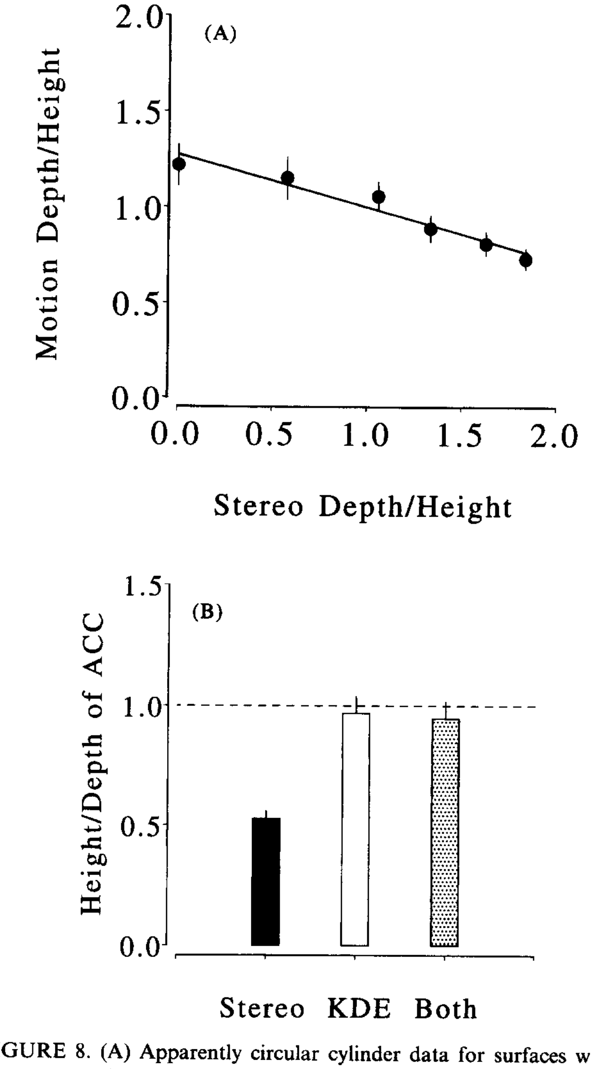


FIGURE 8. (A) Apparently circular cylinder data for surfaces with incongruent depths specified by kinetic depth and stereo (from Johnston *et al.*, 1994). The data points show the depth/height portrayed by stereo and by motion which resulted in a cylinder which appeared circular to the subjects. Viewing distance was 200 cm. (B) Portrayed height/depth for cylinders which appeared circular (data shown are an average over three subjects). For stereo alone, depth is underestimated at this viewing distance (i.e. an exaggerated amount of depth is required for the cylinder to appear circular, hence height/depth is < 1). For KDE alone and for the combination, perception is veridical.

mations (stretches along the line of sight). Further experiments by Johnston *et al.* (1994) bear on this point. They repeated the experiments discussed above using only two frames of KDE motion (with and without stereo). Depth from KDE alone was now underestimated by all subjects. The affine theory does not predict this difference between two frame and many frame KDE. Second, a combination of two frame KDE and stereo resulted in veridical perception. Hence, in these conditions it appeared that the stereo and motion cues were able to promote one another (as suggested by Richards, 1985). Finally, it was again the case that depth from stereo and (two frame) KDE were combined linearly by the observers. The important point is that if depth-from-motion is the only cue available and is only computed up to affine transformations along the line of sight, then certain judgments should be impossible (such as the ACC task or rigidity judgments). This is not the case. And, if it were the case, the MWF framework further suggests that promotion will occur when additional scene content is available to effect promotion.

To summarize, through a series of experiments we have repeatedly seen evidence of a linear combination rule for depth under restricted scene conditions. Cue weights change as cue reliabilities are manipulated. Cues interact to effect cue promotion. Thus, several aspects of the MWF model are seen in empirical data.

A final question of interest is whether, in fact, the rule of combination is robust. Using the perturbation analysis method, we may test whether a robust rule is used, as follows. Suppose that Cue 1 assigns a depth of d_1 to a specific point in the scene, and that Cue 2 assigns a (discrepant) depth of $d_2 = d_1 + \Delta\text{cue}$, where Δcue is initially taken to be small. We measure the weight assigned to Cue 1 by the perturbation analysis method as described above. Next we *increase* the size of Δcue and redo the measurement.

For a linear estimator, any size perturbations result in the same slope estimate α. For the robust estimator we propose, the estimated slope will vary with the size of the perturbations Δcue. In brief, the weight assigned to one or the other cue will decrease as the cues become more and more discrepant. Intuitively, the visual system chooses one cue over the other as the discrepancy between them grows and it becomes less and less plausible to merge them into a combined depth estimate.

Some of the results of Johnston *et al.* (1994) exhibit just such a pattern. The weights assigned to different cues do change as the size of the perturbation Δcue is varied. Their results are consistent with the proposal that the depth cue combination rule is robust. However, in experiments like these, where only two cues are merged at a time, we have no firm prediction as to which cue will be weighted less, which more. We can only predict that the weights will change as Δcue is varied, not the direction in which they will change. Other results of Johnston *et al.* (1994) and Young *et al.* (1993) do not show this change in weighting as a pair of cues become increasingly discrepant.

A stronger test of robustness would require stimuli containing three or more strong depth cues, one of which signals a depth discrepant with the depths signaled by the others. Now we predict that the weight assigned to the discrepant cue must decrease as the size of the discrepancy Δcue increases. Such measurements, in fact, permit us to directly measure the *influence curve* of the rule of combination for the discrepant cue (see Fig. 4). Suppose that $\alpha_{\Delta\text{cue}}$ is the weight assigned to the discrepant cue when the discrepancy is Δcue. Then the point $(\Delta\text{cue}, \alpha_{\Delta\text{cue}}\Delta\text{cue})$ lies on the influence curve. By varying Δcue, we may trace out the curve. Although this has not been done with depth judgements to our knowledge, Whitaker and MacVeigh (1992) measured the influence curve for the perceived spatial location of a cloud of dots as a function of the position of a single dot, resulting in data similar to the theoretical curve in Fig. 4.

SUMMARY AND CONCLUSIONS

Several different visual cues provide information about depth and shape in a scene. These include binocular stereopsis, object motion (the KDE), observer motion (motion parallax), texture perspective, and occlusion, among others. When several of these cues are simultaneously available in a single location in the scene, the visual system may attempt to combine them, may ignore some in favor of the others, or may attempt to represent the two estimates of depth as competing, bistable percepts. This problem of depth cue combination is an example of what is termed *the sensor fusion problem*. In this paper we argued that the combination of depth estimates is effected under normal viewing conditions by a particularly simple computation. Each depth cue is used to establish a depth map throughout the scene. The final estimate of depth at each location is a weighted average of the estimates derived from the individual cues.

We argue that the weights corresponding to different depth cues should vary from location to location within a scene, reflecting the quality of depth information available from each type of depth cue. In textureless regions of the scene, for example, little weight should be given to depth derived from texture perspective. The weight given to depth derived from motion parallax should reflect the observer's velocity. The weights used to form the final depth estimate are based on both scene content (an unavailable or unreliable cue is down-weighted) and also on the consistency of the estimates. An estimate (based on one cue) that is highly discrepant with estimates based on the remaining cues is down-weighted. The resultant rule of combination is consequently a robust statistical estimator.

We note that depth information available from single cues treated in isolation need not be commensurate: depth estimates derived from object motion, for example, are known only up to a scale parameter and a sign parameter, while estimates of depth obtained from (known) observer motion are absolute. We assume that different types of cues are processed as far as possible in isolation (*weak fusion*), but that the "missing parameters" for each depth cue must be filled in by comparison with depth estimates obtained using other cues. We term this process *promotion*. We assume that interactions between depth "modules" are limited to those needed to assess the weights assigned to different depth cues, and effect promotion. The resulting framework is termed *modified weak fusion*. One advantage of this framework is that it provides a realistic, falsifiable alternative to both weak and strong fusion models. A second advantage is that it makes explicit many of the key issues in depth fusion (promotion, dynamic weighting, ancillarity, robustness) that deserve experimental treatment and that should be considered in analyzing both the design and the outcome of depth fusion experiments.

We described novel methods, analogous to *perturbation analysis*, that permit us to measure the weights assigned to different depth cues, and summarized recent experimental tests of the MWF framework based on these methods. These methods are applicable to a wide range of fusion problems that do not involve depth vision or depth cues. For example, it might easily be applied to problems in spatial hearing involving the time-intensity tradeoff (Blauert, 1983). They have also been applied to visual localization tasks (Landy, 1993). They are particularly relevant to sensory and cognitive fusion problems where issues of promotion, dynamic weighting, robustness, and ancillarity arise.

The issue of robustness, in particular, has immediate implications for the study of depth vision and visual processing in general. If early visual processing is affected by the consistency of redundant visual information *per se*, then the results of those experiments which employ stimuli containing distorted, inconsistent, or impoverished depth cues must be interpreted with care. Particular care must be taken if the results of these experiments are to be compared to optimal performance models of any sort. Such experiments do measure human visual performance, but over a range where it would be unreasonable to expect that visual performance is optimized in any sense.

Consequently, it is important to understand what "reality checks", if any, are present in early visual processing and to create precisely controllable stimuli that are sufficiently *veridical* to pass these checks. If, for example, shading and specularity are not varied as independent variables in a given experiment, it would seem necessary to either make them consistent with some interpretation of illuminant and surface properties in a scene, or else to establish that inconsistent shading and specularity do not affect the experimental outcome. Experiments using veridical stimuli and perturbation analysis techniques have the potential to tell us about the range of human visual processing where the human visual system spends much of its time.

REFERENCES

Anderssen, R. S. & Bloomfield, P. (1974). Numerical differentiation procedures for non-exact data. *Numerische Mathematik, 22,* 157–182.

Bennett, B. M., Hoffman, D. D. & Prakash, C. (1989). *Observer mechanics; a formal theory of perception.* San Diego, Calif.: Academic Press.

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

Berg, B. (1989). Analysis of weights in multiple observation tasks. *Journal of the Acoustical Society of America, 86,* 1743–1746.

Berger, J. O. (1985). *Statistical decision theory; foundations, concepts, and methods* (2nd edn). New York: Springer.

Blake, A. & Bülthoff, H. H. (1990). Does the brain know the physics of specular reflection? *Nature, 343,* 165–168.

Blake, A. & Bülthoff, H. H. (1991). Shape from specularities: computation and psychophysics. *Philosophical Transactions of the Royal Society of London B, 331,* 237–252.

Blauert, J. (1983). *Spatial hearing.* Cambridge, Mass.: MIT Press.

Bove, V. M. Jr (1990). Probabilistic method for integrating multiple sources of range data. *Journal of the Optical Society of America A, 7,* 2193–2198.

Brainard, D. H. & Wandell, B. A. (1991). A bilinear model of the illuminant's effect on color appearance. In Landy, M. S. & Movshon, J. A. (Eds), *Computational models of visual processing* (pp. 171–186). Cambridge, Mass.: MIT Press.

Braunstein, M. L. (1968). Motion and texture as sources of slant information. *Journal of Experimental Psychology, 78,* 247–253.

Braunstein, M. L., Andersen, G. J. & Riefer, D. M. (1982). The use of occlusion to resolve ambiguity in parallel projections. *Perception & Psychophysics, 31,* 261–267.

Braunstein, M. L., Andersen, G. J., Rouse, M. W. & Tittle, J. S. (1986). Recovering viewer-centered depth from disparity, occlusion, and velocity gradients. *Perception & Psychophysics, 40,* 216–224.

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

Buckley, D. & Frisby, J. P. (1993). Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. *Vision Research, 33,* 919–933.

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

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

Bülthoff, H. H. & Yuille, A. L. (1991). Shape-from-X: Psychophysics and computation. In Schenker, P. S. (Ed.), *Sensor Fusion III: 3-D Perception and Recognition, Proceedings of the SPIE, 1383,* 235–246.

Cavanagh, P. (1987). Reconstructing the third dimension: Interactions between motion, binocular disparity, and shape. *Computer Vision, Graphics and Image Processing, 37,* 171–195.

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

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman & Hall.

Cutting, J. E. & Bruno, N. (1988). Additivity, subadditivity, and the use of visual information: A reply to Massaro (1988). *Journal of Experimental Psychology: General, 117,* 422–424.

Cutting, J. E. & Millard, R. T. (1984). Three gradients and the perception of flat and curved surfaces. *Journal of Experimental Psychology: General, 113,* 198–216.

Cutting, J. E., Bruno, N., Brady, N. P. & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121,* 364–381.

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.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Ellard, C. G., Goodale, M. A. & Timney, B. (1984). Distance estimation in the Mongolian gerbil: The role of dynamic depth cues. *Behavioural Brain Research, 14,* 29–39.

Epstein, W. (1968). Modification of the disparity-depth relationship as a result of exposure to conflicting cues. *American Journal of Psychology*, *81*, 189–197.

Ferguson, T. S. (1967). *Mathematical statistics; a decision theoretic approach*. New York: Academic Press.

Foley, J. M. (1977). Effect of distance information and range on two indices of visually perceived distance. *Perception*, *6*, 449–460.

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

Geisler, W. S. (1989). Sequential ideal observer analysis of visual discrimination. *Psychological Review*, *96*, 1–71.

Gibson, J. J. (1950). *Perception of the visual world*. Boston, Mass.: Houghton-Mifflin.

Gillam, B. J. (1968). Perception of slant when perspective and stereopsis conflict: Experiments with aniseikonic lenses. *Journal of Experimental Psychology*, *78*, 299–305.

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

Gogel, W .G. (1965). Equidistance tendency and its consequences. *Psychological Bulletin*, *64*, 153–163.

Gogel, W. G. (1976). An indirect method of measuring perceived distance from familiar size. *Perception & Psychophysics*, *20*, 419–429.

Gogel, W. G. (1977). The metric of visual space. In Epstein, W. (Ed.), *Stability and constancy in visual perception: Mechanisms & processes* (pp. 129–181). New York: Wiley.

Gogel, W. G. & Tietz, J. D. (1973). Absolute motion parallax and the specific distance tendency. *Perception & Psychophysics*, *13*, 284–292.

Goodale, M. A., Ellard, C. G. & Booth, L. (1990). The role of image size and retinal motion in the computation of absolute distance by the Mongolian gerbil (*Meriones unguiculatus*). *Vision Research*, *30*, 399–413.

Gregory, R. L. (1970). *The intelligent eye*. New York: McGraw-Hill.

Grimson, W. E. L. (1981). *From images to surfaces; a computational study of the early visual system*. Cambridge, Mass.: MIT Press.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383–393.

von Helmholtz, H. (1910/1925). Helmholtz's treatise on physiological optics. Translation of the 3rd edn (1910). New York: Dover.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Ittelson, W. H. (1952). *The Ames demonstrations in perception*. Princeton, N.J.: Princeton University Press.

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

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

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

Johnston, E. B., Landy, M. S., Cumming, B. G. & Maloney, L. T. (1991). Integration of stereo and motion shape cues. *Investigative Ophthalmology & Visual Science*, *32*, 1180.

Kaufman, L. (1974). *Sight and mind*. New York: Oxford University Press.

Kendall, M. & Stuart, A. (1979). *The advanced theory of statistics; Vol. 2: Inference and relationship* (4th edn). New York: Macmillan.

Koenderink, J. J. (1990). *Solid shape*. Cambridge, Mass.: MIT Press.

Koenderink, J. J., van Doorn, A. J. & Kappers, A. M. L. (1992). Surface perception in pictures. *Perception & Psychophysics*, *52*, 487–496.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement, Vol. I; additive and polynomial representations*. New York: Academic Press.

Landy, M. S. (1993). Combining multiple cues for texture edge localization. In Rogowitz, B. E. & Allebach, J. P. (Eds), *Human Vision, Visual Processing, and Digital Display IV*, *Proceedings of the SPIE*, *1913*, 506–517.

Landy, M. S. & Maloney, L. T. (1990). A statistical framework for combination of consonant depth cues. *Investigative Ophthalmology & Visual Science (Suppl)*, *31*, 173.

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

Lowe, D. G. (1982). *Perceptual organization and visual recognition*. Boston, Mass.: Kluwer.

Maloney, L. T. (1995). Color constancy and color perception: The linear models framework. In Meyer, D. E. & Kornblum, S. (Eds), *Attention & performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience—A Silver Jubilee* (pp. 59–78). Cambridge, Mass.: MIT Press.

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

Massaro, D. W. (1988). Ambiguity in perception and experimentation. *Journal of Experimental Psychology: General*, *117*, 417–421.

Mingolla, E. & Todd, J. T.(1986). Perception of solid shape from shading. *Biological Cybernetics*, *53*, 137–151.

Nakayama, K. & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, *257*, 1357–1363.

Nawrot, M. & Blake, R. (1989). Neural integration of information specifying structure from stereopsis and motion. *Science*, *244*, 716–718.

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

Ono, M. E., Rivest, J. & Ono, H. (1986). Depth perception as a function of motion parallax and absolute-distance information. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 331–337.

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

Pelli, D. G. (1981). Effects of visual noise. PhD thesis, University of Cambridge, Cambridge.

Pong, T.-C., Kenner, M. A. & Otis, J. (1990). Stereo and motion cues in preattentive vision processing—some experiments with random-dot stereographic image sequences. *Perception*, *19*, 161–170.

Prazdny, K. (1986). Three-dimensional structure from long-range apparent motion. *Perception*, *15*, 619–625.

Proffitt, D. R., Bertenthal, B. I. & Roberts, R. J. Jr (1984). The role of occlusion in reducing multistability in moving point-light displays. *Perception & Psychophysics*, *36*, 315–323.

Ramachandran, V. S. (1988). Perceiving shape from shading. *Scientific American*, *331*, 133–166.

Rey, W. J. J. (1980). *Introduction to robust and quasi-robust statistical methods*. Berlin: Springer.

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

Roberts, F. S. (1979). Measurement theory with applications to decisionmaking, utility, and the social sciences. In Rota, G.-R. (Ed.), *Encyclopedia of mathematics and its applications* (Vol. 7). Reading, Mass.: Addison-Wesley.

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

Rouse, M. W., Tittle, J. S. & Braunstein, M. L. (1989). Stereoscopic depth perception by static stereo-deficient observers in dynamic displays with constant and changing disparity. *Optometry and Vision Science*, *66*, 355–362.

Schopenhauer, A. (1847/1974). *On the fourfold root of the principle of sufficient reason* (Translation of 2nd edn, 1847, by Payne, E. F. J.). LaSalle, Ill.: Open Court.

Schunck, B. G. (1989). Robust estimation of image flow. In Schencker, P. S. (Ed.), *Sensor Fusion II: Human and Machine Strategies*, *Proceedings of the SPIE*, *1198*, 116–127.

Schwartz, B. J. & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society*, *21*, 456–458.

Searle, C. L., Braida, L. D., Davis, M. F. & Colburn, H. S. (1976). Model for auditory localization. *Journal of the Acoustical Society of America*, *60*, 1164–1175.

Sinha, S. S. & Schunck, B. G. (1992). A two stage algorithm for discontinuity-preserving surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14,* 36–55.

Sloman, D. & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In Healy, A., Kosslyn, S. & Shiffrin, R. (Eds), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, pp. 227–248). Hillsdale, N. J.: Erlbaum.

Stevens, K. A. & Brookes, A. (1987). Depth reconstruction in stereopsis. *Proceedings IEEE First International Conference on Computer Vision,* 682–686.

Stevens, K. A. & Brookes, A. (1988). Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research, 28,* 371–386.

Stevens, K. A., Lees, M. & Brookes, A. (1991). Combining binocular and monocular curvature features. *Perception, 20,* 425–440.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In Churchman, C. W. & Ratoosh, P. (Eds), *Measurement: Definitions and theories* (pp. 18–63). New York: Wiley.

Stiles, W. S. (1978). *Mechanisms of colour vision.* London: Academic Press.

Taylor, M. M. (1962). Figural after-effects: A psychophysical theory of the displacement effect. *Canadian Journal of Psychology, 16,* 247–277.

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

Tittle, J. S. & Braunstein, M. L. (1993). Recovery of 3-D shape from binocular disparity and structure from motion. *Perception & Psychophysics, 54,* 157–169.

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

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

Todd, J. T. & Reichel, F. D. (1989). Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review, 96,* 643–657.

de Vries, S. C., Kappers, A. M. L. & Koenderink, J. J. (1993). Shape from stereo: A systematic approach using quadratic surfaces. *Perception & Psychophysics, 53,* 71–80.

Wallach, H. & Karsh, E. B. (1963). The modification of stereoscopic depth-perception and the kinetic depth effect. *American Journal of Psychology, 76,* 429–435.

Wallach, H. & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45,* 205–217.

Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London, 33,* 371–394.

Whittaker, D. & MacVeigh, D. (1992). Sequential mapping of weighting functions for visual location. *Spatial Vision, 6,* 117–131.

Würger, S. M. & Landy, M. S. (1989). Depth interpolation with sparse disparity cues. *Perception, 18,* 39–54.

Yellott, J. I. Jr (1981). Binocular depth inversion. *Scientific American, 245,* 148–159.

Young, M. J., Landy, M. S. & Maloney, L. T. (1991). Depth from texture and motion. *Investigative Ophthalmology & Visual Science, 32,* 1180.

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.

Youngs, W. M. (1976). The influence of perspective and disparity cues on the perception of slant. *Vision Research, 16,* 79–82.

Yuille, A. L. & Bülthoff, H. H. (1995). Bayesian decision theory and psychophysics. In Knill, D. C. & Richards, W. (Eds), *Bayesian perspectives on visual perception.* Cambridge: Cambridge University Press. In press.

Acknowledgements—The work described in this paper was supported in part by grant EY08266 from the National Eye Institute, NIH, by an NRC-NASA Ames Research Associateship to Michael Landy, and by grant F49620-92-J-0187 from the Air Force Office of Scientific Research to Laurence Maloney. We would like to thank Misha Pavel, Dennis Proffitt and three anonymous reviewers for their helpful comments. We also thank Robert Picardi for technical assistance. Portions of this work have been presented at the annual meetings of the Association for Research on Vision and Ophthalmology (Sarasota, Fla, May 1990 and May 1991) and the SPIE (Philadelphia, Pa, November 1989; and Boston, Mass, November 1990) and published in abstract or proceedings form (Johnston, Landy, Cumming & Maloney, 1991; Landy & Maloney, 1990; Landy *et al.*, 1991; Maloney & Landy, 1989; Young, Landy & Maloney, 1991).