A statistical framework for robust fusion of depth information

Laurence T. Maloney
Michael S. Landy

Department of Psychology
and Center for Neural Science
New York University

ABSTRACT

We describe a simple statistical framework intended as a model of how depth estimates derived from consistent depth cues are combined in biological vision. We assume that the rule of combination is linear, and that the weights assigned to estimates in the linear combination are variable. These weight values corresponding to different depth cues are determined by *ancillary measures*, information concerning the likely validity of different depth cues in a particular scene. The parameters of the framework may be estimated psychophysically by procedures described. The conditions under which the framework may be regarded as normative are discussed.

1. Introduction

Independent Depth Modules. Until recently, research in computational vision has concentrated on the study of single depth cues in isolation. Much of this research has derived inspiration directly from empirical studies in biological vision (See Kaufman, 1974, Chaps. 7, 8), and, conversely, some of the resulting analytic and computational work has proven directly relevant to the study of biological vision. Cues studied computationally include kinetic depth (Hoffman, 1982; Koenderink and van Doorn, 1986; Landy, 1987; Longuet-Higgins and Prazdny, 1980; Ullman, 1979, 1983, 1984), disparity (Dev, 1975; Landy and Hummel, 1986b; Marr and Poggio, 1976 and 1979; Mayhew and Frisby, 1980; Prazdny, 1975; Sperling, 1970), texture gradients (Witkin, 1981), surface contours (Stevens, 1981), accommodation (Pentland, 1985), occlusion (Haynes and Jain, 1988), and others.

In this paper we examine the question of how multiple depth estimates are combined into a single estimate of depth at each location in a scene (the *depth fusion problem*). We analyze depth fusion as a problem in statistical decision theory and derive an optimal statistical depth fusion computation. At the same time, we are interested in the depth fusion rule used by the human visual system. We suggest that the optimal robust estimation procedure is also used in human vision. In addition, we show that this theory is testable and that its parameters can be estimated psychophysically.

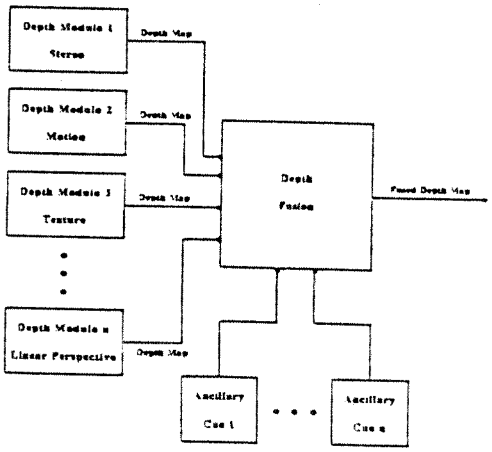

Figure 1: Depth cues and depth fusion.

Fig. 1 illustrates a working model of the visual process that emerges from previous psychophysical and computational research. Multiple independent visual modules each make use of a distinct depth cue, and the resulting pieces of information are 'fused' into a single depth estimate at each point in the scene. Examined in detail, this model fails to capture biological depth phenomena in several respects. The assumption that different depth cues are processed independently is suspect, as is the assumption that a unique depth estimate is always obtained at each location.

However, particularly if we choose appropriate experimental conditions, the performance of each module may be studied psychophysically, or analyzed computationally in isolation from the remaining modules. Consequently, we assume

Assumption 1: Multiple depth modules process distinct depth cues independently.

The Rule of Combination. Implicit in both the biological and computational approaches is the existence of a *rule of combination* which reduces the many pieces of information relevant to depth to a single estimate of the three-dimensional layout of a scene at each point. The problem of specifying this rule of combination can be treated *descriptively* by measuring the performance of 'The Psychophysical Observer,' or *normatively* as the problem of developing a model of optimal visual processing as constrained by whatever visual information is actually available ('The Ideal Observer'; See Geisler, 1988).

The normative fusion problem is a specific instance of the problem of normative aggregation of information (see Landy and Hummel, 1986a, for a review of some approaches to the problem of evidence aggregation), and is fundamentally a problem of statistical estimation. Hager (1988), in particular, casts the depth fusion problem within the framework of Bayesian statistical decision theory (Berger, 1980; Ferguson, 1967). Kak and Chen (1987) includes a selection of other recent work (See also Allen, 1985; Durrant-Whyte, 1986; Krotkov, 1987). Rules of combination derived from normative considerations are often complex, and computationally intensive (see Hager, 1988). For our purposes, though, their primary drawback is that they are difficult to evaluate as descriptive models of the psychophysical observer because of their generality.

From a descriptive standpoint, there has been a long history of psychological research on combination of cues. Much of that work concerned situations in which cues were put into conflict. For example, the Ames room (see Gregory, 1970) provides a stimulus in which the cue of perspective and other 2-D heuristics (e.g., the assumption that nearly parallel lines in 2-D must have been parallel in 3-D) are put in conflict with the other available depth cues (familiar size, linear perspective). The Ames chair (see Gregory, 1970) is another such example where visual heuristics are put in conflict with other cues. Hans Wallach (e.g., Wallach and Karsh, 1963) has also studied a number of such examples where one cue leads to a nonveridical percept and yet is so compelling that it vetoes other cues which, if used, would have led to the veridical percept. It is also well known that a real model of the Necker cube, for which the 2-D cues are ambiguous as to depth sign, can actually perceptually reverse despite other available cues such as disparity. A viewer in the perceptually reversed state will perceive a non-rigid object if s/he moves about the object because of the motion parallax cue based on the nonveridical reversed depth values. In each of these cases we are dealing with phenomena in which depth cues are severely in conflict and some subset of the cues is ignored entirely (see the discussion on robustness below). These experiments are not well suited to determining how depth cues are combined when there is a possibility of using the information from more than one cue.

More recently there has been some work on quantifying the way in which depth cues are combined. Dosher, Sperling, and Wurst (1986), following Schwartz and Sperling (1983), studied the combination of linear perspective, stereo disparity, and what they term proximity luminance covariance (or PLC, where brighter areas appear closer). The stimuli were rotating Necker cubes, and the subjects specified whether the cube was perceived rotating front-to-the-right or front-to-the-left. It turned out that a simple additive cues model was able to account for the data entirely.

Bülthoff and Mallot (1987) combined a number of depth cues (shading, contour, edge disparity, and what they term intensity-based stereo). For a given cue combination, the perceived shape of an ellipsoidal surface was measured by allowing subjects to adjust a stereo probe spot to appear to lie on the surface of the ellipsoid. This allowed them to measure the perceived surface (relative to the stereo spot's perceived depth), and thus probe how each cue had been calibrated internally by the subject relative to disparity. We have used the same probe technique, although in a forced choice paradigm, to measure stereo depth interpolation (Würger and Landy, 1989). Other papers on depth cue combination include Cutting and Millard (1984), Epstein (1968), Gillam (1968), Proffitt, Bertenthal, and Roberts (1984), Stevens and Brookes (1987 and 1988), Todd and Akerstrom (1987), Todd and Mingolla (1983), and Youngs (1976).

In summary, when depth cues are in conflict, the depth processing of biological visual systems is complex and difficult to characterize. When depth cues are consonant, the rule of combination for depth estimates is consistent with a weighted linear rule of combination (Dosher et al., 1986). Bruno and Cutting (1988) also can not reject an additive rule. Consequently, we limit our scope to models where depth cues are consonant and assume,

Assumption 2: The rule of combination for depth estimates obtained from different depth cues is linear.

Hager (1988) summarizes the preconditions that permit a linear rule of combination to be normative from the viewpoint of statistical decision theory.

Ancillary Measures. It is plausible that any normative rule of combination for depth cues would (should) change, depending on the physical characteristics of the scene and the viewing conditions. It is plausible, for example, that the amount and location of texture in a scene would affect the weight given to depth information derived from texture gradients. The speed and direction of egomotion, available from the vestibular system, should influence the weight given to depth derived from motion parallax, and the angular rotation of an object, the weight given to kinetic depth information associated with it. *In general, the quality of information available from a given source will depend on specific characteristics of a scene which can in some cases be independently measured.* Side information that is relevant to assessing the quality of depth information available from various depth modules is represented by the *ancillary measure* inputs to the fusion box in Fig. 1.

Ancillary statistics are conditionally sufficient statistics that serve to reduce the variability of estimate of a parameter, but whose distribution is independent of the values of the parameter being estimated (See Cox and Hinkley, 1974, pp. 33-35, 220-221; Kendall and Stuart, 1979, pp. 232-234), By analogy, ancillary measures do not in themselves provide information about depth in a scene, but instead provide information concerning the likely performance of different depth modules.

We make

Assumption 3: The weights assigned to depth estimates from different depth modules may vary from scene to scene and from location to location within a scene.

Assumption 4: The weights assigned to depth estimates from different depth modules are determined by ancillary measures.

Within statistical decision theory, it is normative to condition maximum likelihood estimates of parameters on ancillary statistics ('The Conditionality Principle,' Cox and Hinkley, 1974, pp. 38-39). Where ancillary measures can be treated as ancillary statistics, their use is normative. Bayesian methods automatically make proper use of ancillary statistics (Cox and Hinkley, 1974, p. 369). We do not, however, restrict ancillary measures to be ancillary statistics.

Within this theoretical framework, we can develop a simple psychophysical method that makes it possible to measure, for human observers, the weights assigned to the various depth cues in a particular environment. Consequently, we can test the proposed model against actual visual performance, and test the various assumptions concerning the use of candidate ancillary measures and dynamic reweighting of depth information sources. In the next section, we describe in greater detail the computational model of the Ideal Depth Observer outlined above. In the section following we describe the psychophysical methods used to test it as a framework for biological depth fusion of consonant cues.

2. Theory: An Ideal Depth Observer.

Framework. Consider an abstract, binocular visual system which is sensitive to any of the following depth cues: motion parallax, kinetic depth information, retinal disparity, texture gradients, and linear perspective. This visual system contains five modules which independently derive depth information from each of the five sources. Our goal is to combine the output of these five modules to estimate depth in the scene. More specifically, the output of the visual system is intended to be a depth map of surface points in a scene.

There are several ways to denote a depth map and corresponding ways to supply coordinates to the three-dimensional layout of the scene (see Roth-Tabak and Jain, 1988). We choose to use a spherical coordinate system centered on a fixed point near the sensors in the visual system. The distance to an object in the direction (θ,ϕ) is then denoted $D(\theta,\phi)$. Since the visual system has two simultaneous views of the scene and multiple sequential views, all from different vantage points, a depth map which records only a single distance in a given direction is not really adequate to the data available. A proper representation would be a three-dimensional model of the environment. For our purposes, however, the simple depth-map representation $D(\theta,\phi)$ will prove to be sufficient. The argument here is in principle extendable to an improved three-dimensional representation of the environment which is viewpoint-independent and allows for observer motion.

Cues and Cue Class. The five depth cues chosen fall into three different classes depending on the kind of depth information available from the cue. The first class comprises motion parallax and retinal disparity. Either the motion parallax cue or the retinal disparity cue alone provide enough information in principal to compute a complete depth map of the scene. Let $D_d(x,y)$ denote the estimate of depth at location (x,y) available through the disparity cue. Let $D_m(x,y)$ denote the estimate of depth available through the motion parallax cue. There are actually two depth maps derivable from motion parallax, one for the left, and one for the right eye. For the sake of presentation, we ignore one for now. The confounding of information between disparity and simultaneous parallax estimates from two eyes is one problem, however, that we must address in the psychophysical studies outlined below.

The second class of cues, which includes texture gradients and linear perspective, provides a depth map known up to a single unknown scaling factor. That is, we can estimate the ratio between any two depths assigned by texture gradients, but not the absolute distance unless we first pin down the unknown scale factor. We can represent the depth information obtainable from these two cues by $f_t D_t(x,y)$ (texture) and $f_p D_p(x,y)$ (linear perspective). where f_t and f_p are the unknown scale factors.

If, for example, an object of known size is available in the scene, then this cue (called 'the familiar size cue') can be used to assign the values of the scaling factors and produce actual depth estimates from texture and perspective cues across the scene. A second way to estimate f_t is to regress the values D_t against the values D_m and D_d. The regression coefficients γ_m and γ_d satisfy,

$$\min_{\gamma_m,\gamma_d} \sum_{x,y} \| D_t(x,y) - \gamma_m D_m(x,y) - \gamma_d D_d(x,y) \|^2 \qquad (1)$$

and $f_t = 1/(\gamma_m+\gamma_d)$ is the least-squares estimate of the unknown scaling factor.

The third class of cue, the kinetic depth effect (or KDE), is even more ambiguous than the previous class of cues. Like texture and linear perspective, KDE provides depth estimates up to an unknown multiplicative scale factor. But in addition, there is a sign ambiguity. KDE displays are at least two-way ambiguous, and the percept is multistable. By sign ambiguity we mean an ambiguity of the sign of depth relative to a fixed positive depth (generally on the axis of rotation of the object). The set of local KDE depth estimates $D_k(x,y)$ are known up to two scale factors γ_{d_1} and γ_{d_2}, and depth estimated from KDE alone is $\gamma_{d_1}(1 + \gamma_{d_2} D_k(x,y))$. γ_{d_1} is the usual positive scale factor; γ_{d_2} handles the ambiguity of sign and takes values of ±1.

Other depth cues provide different qualities of information. Occlusion does not allow one to estimate depth at all, but rather provides a list of assertions as to the depth ordering of points on either side of image contours and is used to disambiguate cues which have an inherent ambiguity of depth sign, such as KDE (Braunstein, Andersen, and Riefer, 1982; Proffitt et al., 1984). Shading can provide relative depth information but there can also be sign ambiguity, leading to occasional confounding of convex and concave surfaces. It requires the estimation of an additional parameter relating to the location of the illuminant.

The three classes of cues we have chosen to include in our model (motion parallax, disparity, KDE, texture, and linear perspective) represent (a) two absolute depth cues that provide actual depth estimates, (b) two relative depth cues that provide actual depth estimates only when a scale parameters are assigned values, and (c) a cue that is almost a relative depth cue except for an additional unknown parameter capturing sign ambiguity.

The kinds of information available from different depth cues are often loosely described in a terminology derived from the theory of measurement and scale type (Roberts, 1979; Stevens, 1959). Motion parallax information is 'on an absolute scale'; texture gradient information is 'on a ratio scale'. As noted above, common depth cues such as KDE and shading with their sign-ambiguities do not fit into this scheme. While we will use the language of scale types below for succinctness, we primarily view the different depth modules as *sources of absolute depth information possibly with missing parameter values.* The problem of combining such data types is naturally part of statistical decision theory.

The Rule of Combination. If we confine attention to motion parallax and retinal disparity, the two cues each of which provide a full depth map, a straightforward rule for combining the two estimates into a single depth estimate is

$$\begin{aligned} D(x,y) \;=\; & \alpha D_m(x,y) \\ & + (1-\alpha) D_d(x,y), \\ & 0 \le \alpha \le 1, \end{aligned} \tag{2}$$

the weighted average of the two estimates. How should α be chosen? Intuitively, no single, fixed value of α would seem adequate to all viewing conditions. If, for example, the visual system is momentarily not moving or moving very slowly, then D_m, derived from motion parallax, should be given little weight and α given a value near 0. Correspondingly, if the visual system is moving rapidly, then α should be increased. The value of α, then is a function of an objective parameter, the motion of the visual system.

We can generalize Eq. (2) by adding texture gradient and linear perspective cues to the rule of combination. Now the depth estimate is a weighted linear combination of information drawn from the first four depth cues.

$$\begin{aligned} D(x,y) \;=\; & \alpha_m D_m(x,y) \\ & + \alpha_d D_d(x,y) \\ & + \alpha_t \hat{f}_t D_t(x,y) \\ & + \alpha_p \hat{f}_p D_p(x,y), \\ 0 \le \; & \alpha_\bullet \le 1 \,, \quad \sum \alpha_\bullet = 1 \end{aligned} \tag{3}$$

The unknown scale factors f_t and f_p must be estimated separately as discussed above. The choice of weight α_t would seem to depend on the amount of textured surface visible in the scene, and the weight α_p on some measure of the number of perspective cues available. In any case, the choice of weights depends upon the physical characteristics of the visual system, the viewing conditions (e.g., the amount of egomotion), and the physical characteristics of the scene. A normative model could permit different weights in different regions of the scene (e.g., $\alpha_t(x,y)$).

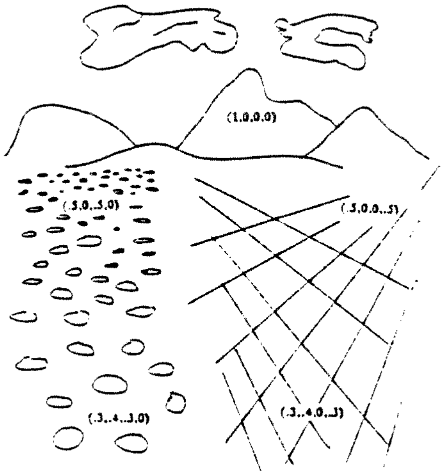

Figure 2: Weights assigned to different parts of a scene.

Figure 2 illustrates the behavior of a hypothetical fusion algorithm across different regions of a landscape. The four weights $(\alpha_m,\alpha_d,\alpha_t,\alpha_p)$, for various parts of the scene are shown. Beyond the range of binocular vision, in the absence of texture cues, and in the absence of linear perspective cues, the weights (1,0,0,0) are used. Where texture is present, α_t takes on a non-zero value. Throughout the scene, the values of the weights reflect the local reliability of the respective depth cues.

Ancillary Measures. To complete the model of the Ideal Depth Observer, we must specify a rule that selects the weights $(\alpha_m,\alpha_d,\alpha_t,\alpha_p)$. Ancillary measures (mentioned above) represent measurements of characteristics of a scene that are not primarily depth information but information relevant to predicting the performance of the various modules. Examples of ancillary measures (and what they might affect) are degree and duration of motion (motion parallax), whether one or two eyes is in use (disparity), whether a region falls within the monocular or binocular region of visual space (disparity), various measures of the amount of texture in the scene (texture gradients), and the number of triplets of line segments whose extensions coincide (linear perspective).

Given a set of ancillary measures, the rule that selects the weights $(\alpha_m,\alpha_d,\alpha_t,\alpha_p)$ is some mapping from the space of ancillary measures to the space of vectors (actually the unit simplex in 4D) $(\alpha_m,\alpha_d,\alpha_t,\alpha_p)$. In many instances, we can 'guess' that specific ancillary measures will prove important. Degree of egomotion is likely relevant to weighting the worth of motion parallax information. But, in general, we do not know how to choose the 'best' ancillary measures.

We can, however, ask, given a fixed set of ancillary measures, what is the normative rule that uses these measures to control the weights in the rule of combination? We emphasize that the desired Ideal Depth Observer is 'ideal' only with respect to a prespecified set of ancillary measures. The ancillary measures available to the human visual system are unknown. One goal of the psychophysical work proposed is to test whether likely ancillary measures are in fact used in human vision. We postpone discussion of this issue to the next section, 'The Psychophysical Observer.'

Given a set of ancillary measures for an abstract visual system, and given a set of depth cues, the problem of deriving the Ideal Observer is the problem of estimating the mapping from measures to weights. There are three methods that would permit us to estimate this map. The first is *analytic:* Derive the appropriate map that would satisfy an optimality criterion. The second uses simple learning algorithms and is connectionist in flavor. It involves simulating scenes, measuring simulated ancillary measures, and using a simple learning algorithm such as back-propagation to approximate the desired map. Such connectionist approaches represent a valuable addition to a purely analytic approach. For example, it would be of value to demonstrate that a particular ancillary measure is of little value in solving the fusion problem, or adds little predictive power to an existing set of ancillary measures. Such a result, while desirable, may not be analytically achievable. If a simple learning algorithm, however, consistently assigns negligible weights to a specific ancillary measure, we may be permitted to doubt its relevance to the problem at hand. At the very least, such a result would be a springboard for analytic results.

The third and last method is to model the human observer as utilizing the ideal depth combination rule. Having done this, we may *estimate* some parts of the Ideal Observer directly from human psychophysical data. We will describe this option further in the section on 'The Psychophysical Observer.' Psychophysical results play a strong role in guiding and inspiring our analytic and computational work.

3. Method: The Psychophysical Observer.

We assume that, in scenes where only the four depth cues of motion parallax, disparity, texture gradients, and linear perspective are present, the performance of the human observer can be modeled by Eq. (3). The addition of KDE will require two parameters, as discussed previously. We do not assume that performance is otherwise optimal and we do not assume that we know what factors affect the choice of weights for the human observer. In particular, we do not assume that we know what are the ancillary measures governing human performance. **We only assume that performance is describable as a linear combination of depth estimates derived from multiple cues** (possibly requiring the separate estimation of various free parameters). In other words, we will concentrate on conditions in which the depth cues are in (approximate) harmony, and the non-linearity of the robust estimator does not arise (see below).

We are interested in modeling human depth fusion using the ideal depth fusion model just outlined. In order for this to be a useful model, it must be possible to test the model's validity, and to estimate the model's parameters. We now show that it is possible to test whether or not the robust estimation procedure is used in human perception, and to derive the weights humans use for depth fusion psychophysically.

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 $d_2 = d_1 + \Delta\text{cue}$. 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. 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 consonant cues d'_1,d'_2 as measured by indifference in a forced-choice task. We conclude

$$\begin{aligned} d' &= \alpha_1 d_1 + \alpha_2 d_2 \\ &= \alpha_1 d_1 + \alpha_2(d_1 + \Delta\text{cue}) \end{aligned}$$

by specializing Eq. (3) for the case of two cues. Rearranging terms,

$$\alpha_2 = \frac{d' - d_1}{\Delta\text{cue}}. \qquad (4)$$

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

$$\alpha_2 = \frac{\Delta\text{depth}}{\Delta\text{cue}}. \qquad (5)$$

In words, the weight $\alpha_{\bullet}$ is the ratio between the change in estimated overall depth Δdepth and the amount the corresponding cue is perturbed Δcue. In particular, the $\alpha_{\bullet}$ weights are readily estimated from psychophysical data. Further, Eq. (5) contains a straightforward test of the entire model framework. The values estimated for α_2 must be between 0 and 1.

By choosing only small values of the value Δcue we avoid any effect of robust estimation procedures on the estimate of α, and avoid creating multistable stimuli. Figure 3 illustrates what is called an 'influence function' in robust estimation (Hampel, 1974; Huber, 1981). The straight line with slope α is a plot of the influence of a discrepancy Δcue versus induced change Δdepth in estimated depth for a non-robust estimator. The second function plotted is the influence curve for a robust estimator. As Δcue increases in absolute value, the influence it exerts at first increases but then begins to decrease eventually to zero. For small enough values of Δcue, the robust estimator and the linear estimator coincide and α

summarizes the influence of Δcue over the range where cues are consistent. It is over this range that we seek to measure α. In brief, since we will not investigate the *trompe l'oeil* viewing conditions where the robust estimator and simple linear combination rule differ, the assumption of robustness does not affect the theoretical framework we employ. The assumption of robustness does impact the choice of psychophysical measure we use. It suggests that if the cues we select to combine are excessively discrepant, the human visual system may cease to operate in accordance with Eq. (3) and begin behaving as a robust estimator by lowering the weight of discrepant cues.

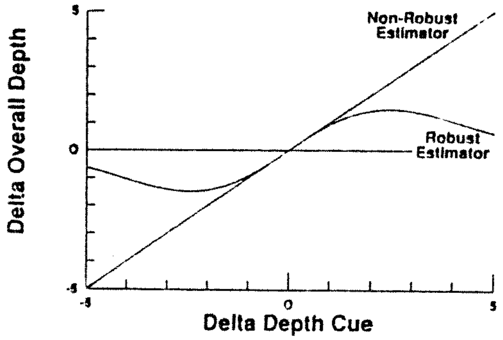

Figure 3: A robust influence function.

For the linear estimator, any size perturbations result in the same slope estimate α. For the robust estimator, the estimated slope will vary with the size of the perturbations Δcue. As Δcue increases, α will decrease. In the limit, very large perturbations are ignored (α is 0). One obvious test of the model proposed is to increase the range of discrepancies Δcue used and determine whether the measured slope α decreases. A more stringent test of the model is obtained by noting that, if we reverse the role of Cue 1 and Cue 2, we can independently measure α_1. We may then plot values of α_1 and α_2 across various conditions and determine whether they lie on the line $\alpha_1 + \alpha_2 = 1$.

The above technique for measuring $\alpha_\bullet$ extends to any number of cues so long as only one of them is perturbed by Δcue at a time. The constraint that the $\alpha_\bullet$'s sum to 1 is then an empirically verifiable consequence of the framework outlined in the previous section.

The theory and methods outlined here suggest that the values of the model weights $\alpha_\bullet$ are psychophysically measurable.

4. Conclusion

We have described a statistical framework for combining depth information from depth cues intended as a model of biological depth vision across a restricted range of experimental conditions. In particular, the choice of experimental conditions should guarantee that depth information from different cues is consistent, and depth estimates (up to a small number of missing parameters) based on distinct are computed independently of one another.

The depth information from different modules are first compared to estimate missing parameters and render the output of each module as an absolute depth map. The depth maps are then combined linearly with weights determined by ancillary measures. As described in the previous section, these weights can be estimated psychophysically.

5. Acknowledgements

This work was supported in part by grant EY08266 from the National Eye Institute.

6. References

Allen, P. K. (1985). *Object recognition using vision and touch.* Technical Report MS-CIS-85-60, GRASP LAB 65, Department of Computer and Information Science, University of Pennsylvania, Philadephia, Pennsylvania.

Berger, J. O. (1980). *Statistical Decision Theory; Foundations, Concepts, and Methods.* New York: Springer-Verlag.

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

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

Bülthoff, H. H., and Mallot, H. A. (1987). Interaction of different modules in depth perception. *Proceedings of the IEEE First International Conference on Computer Vision*, pp. 295-305.

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

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

Dev, P. (1975). Perception of depth surfaces in random-dot stereograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-2**, 333-340.

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

Durrant-Whyte, H. F. (1986). *Integration, coordination, and control of multi-sensor robot systems.* Technical Report MS-CIS-86-67, GRASP LAB 67, Department of Computer and Information Science, University of Pennsylvania, Philadephia, Pennsylvania.

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.

Geisler, W. S. (1988). Sequential ideal-observer analysis of visual discrimination. *Review* , in press.

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

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

Hager, G. D. (1988). *Active reduction of uncertainty in multi-sensor systems.* Technical Report MS-CIS-88-47, GRASP LAB 147, Department of Computer and Information Science, University of Pennsylvania, Philadephia, Pennsylvania.

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

Haynes, S. M., and Jain, R. (1988). A qualitative approach for recovering depths in dynamic scenes. *Proc. of IEEE Workshop on Computer Vision*, Miami Beach, Nov.-Dec. 1987, 66-71.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America*, **72**, 888-892.

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

Kak, A., and Chen, S. (1987). *Spatial Reasoning and Multi-Sensor Fusion.* Proceedings of the 1987 Workshop, Sponsored by AAAI. Los Altos, CA: Morgan Kaufmann Publishers.

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

Kendall, M., and Stuart, A. (1979). *The Advanced Theory of Statistics; Volume 2: Inference and Relationship; 4th Edition.* New York: Macmillan.

Koenderink, J. J., and van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America, A*, **3**, 242-249.

Krotkov, E. P. (1987). *Exploratory visual sensing for determining spatial layout with an agile stereo camera system.* Technical Report MS-CIS-87-29, GRASP LAB 101, Department of Computer and Information Science, University of Pennsylvania, Philadephia, Pennsylvania.

Landy, M. S. (1987). Parallel model of the kinetic depth effect using local computations. *Journal of the Optical Society of America, A*, **4**, 864-877.

Landy, M. S., and Hummel, R. A. (1986a). A brief survey of knowledge aggregation methods. *Proceedings of the International Conference on Pattern Recognition*, Paris, France, October, 1986, 248-252.

Landy, M. S., and Hummel, R. A. (1986b). Multiresolution model of stereopsis. *Journal of the Optical Society of America A*, **3**, P88.

Longuet-Higgins, H. C., and Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London, Series B*, **208**, 385-397.

Marr, D., and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, **194**, 283-287.

Marr, D., and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London, Series B*, **204**, 301-328.

Mayhew, J. E. W., and Frisby, J. P. (1980). The computation of binocular edges. *Perception*, **9**, 69-86.

Pentland, A. P. (1985). A new sense for depth of field. *Proceedings, IEEE Joint Conference on Artificial Intelligence, 988-994.*

Prazdny, K. (1975). Detection of binocular disparities. *Biological Cybernetics*, **52**, 93-99.

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

Roberts, F. S. (1979). *Measurement Theory With Applications to Decisionmaking, Utility, and the Social Sciences*, G.-R. Rota (Ed.), Encyclopedia of Mathematics and its Applications, Volume 7. Reading, MA: Addison-Wesley.

Roth-Tabak, Y., and Jain, R. (1988), Building an environment model using depth information. Manuscript under review.

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

Sperling, G. (1970). Binocular vision: A physical and a neural theory. *The American Journal of Psychology*, **83**, 461-534.

Stevens, K. A. (1981). The visual interpretation of surface contours. *Artificial Intelligence*, **17**, 47-73.

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

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

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

Todd, J. T., and Akerstrom, R. (1987). Perception of 3-dimensional from from patterns of optic texture. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 242-255.

Todd, J. T., and Mingolla, E. (1983). Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 583-595.

Ullman, S. (1979). *The Interpretation of Visual Motion.* Cambridge, MA: MIT Press.

Ullman, S. (1983). Recent computational results in the interpretation of structure from motion. In A. Rosenfeld, B. Hope, and J. Beck (Eds.), *Human and Machine Vision*, 459-480. New York: Academic Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and non-rigid motion. *Perception* , **13**, 255-274.

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

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence*, **17**, 17-45.

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

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