Bayesian models of sensory cue integration

David C. Knill and Jeffrey A. Saunders [1]
Center for Visual Sciences
University of Rochester
274 Meliora Hall
Rochester, NY 14627

Submitted to Vision Research

[1]E-mail address:knill@cvs.rochester.edu

1 Introduction

Our senses provide a number of independent cues to the three-dimensional layout of objects and scenes. Vision, for example, contains cues from stereo, motion, texture, shading, etc.. Each of these cues provides uncertain information about a scene; however, this apparent ambiguity is mitigated by several factors. First, under normal conditions, multiple cues are available to an observer. By efficiently integrating information from all available cues, the brain can derive more accurate and robust estimates of three-dimensional geometry (i.e. positions, orientations, and shapes in three-dimensional space)[1]. Second, objects in our environment have strong statistical regularities that make cues more informative than would be the case in an unstructured environment. Prior knowledge of these regularities allows the brain to maximize its use of the information provided by sensory cues.

Bayesian probability theory provides a normative framework for modeling how an observer should combine information from multiple cues and from prior knowledge about objects in the world to make perceptual inferences[2]. It also provides a framework for developing predictive theories of how human sensory systems make perceptual inferences about the world from sensory data, predictions that can be tested psychophysically. The goal of this chapter is to introduce the basic conceptual elements of Bayesian theories of perception and to illustrate a number of ways that we can use psychophysics to test predictions of Bayesian theories.

1.1 Basics

To illustrate the basic structure of Bayesian computations, consider the problem of integrating multiple sensory cues about some property of a scene. Figure 1 illustrates the Bayesian formulation of one such problem estimating the position of an object, X, from visual and auditory cues, V and A. The goal of an optimal, Bayesian observer would be to compute the conditional density function, p(X — V,A). Using Bayes rule, this is given by

$$p(X \mid V, A) = \frac{p(V, A \mid X)p(X)}{p(V, A)} \tag{1}$$

where p(V, A — X) specifies the relative likelihood of sensing the given data for different values of X and p(X) is the prior probability of different values of X. Since the noise sources in auditory and visual mechanisms are statistically independent, we can decompose the likelihood function into the product of likelihood functions associated with the visual and auditory cues, respectively:

$$p(V, A \mid X) = p(V \mid X)p(A \mid X) \tag{2}$$

p(V — X) and p(A — X) fully represent the information provided by the visual and auditory data about the targets position. The posterior density function is therefore proportional to the product of three functions, the likelihood functions associated with each cue and the prior density function representing the relative probability of the target being at any given position. An optimal estimator could pick the peak of the posterior density function, the mean of the function, or any of a number of other choices, depending on the cost associated with making different types of errors[3, 4].

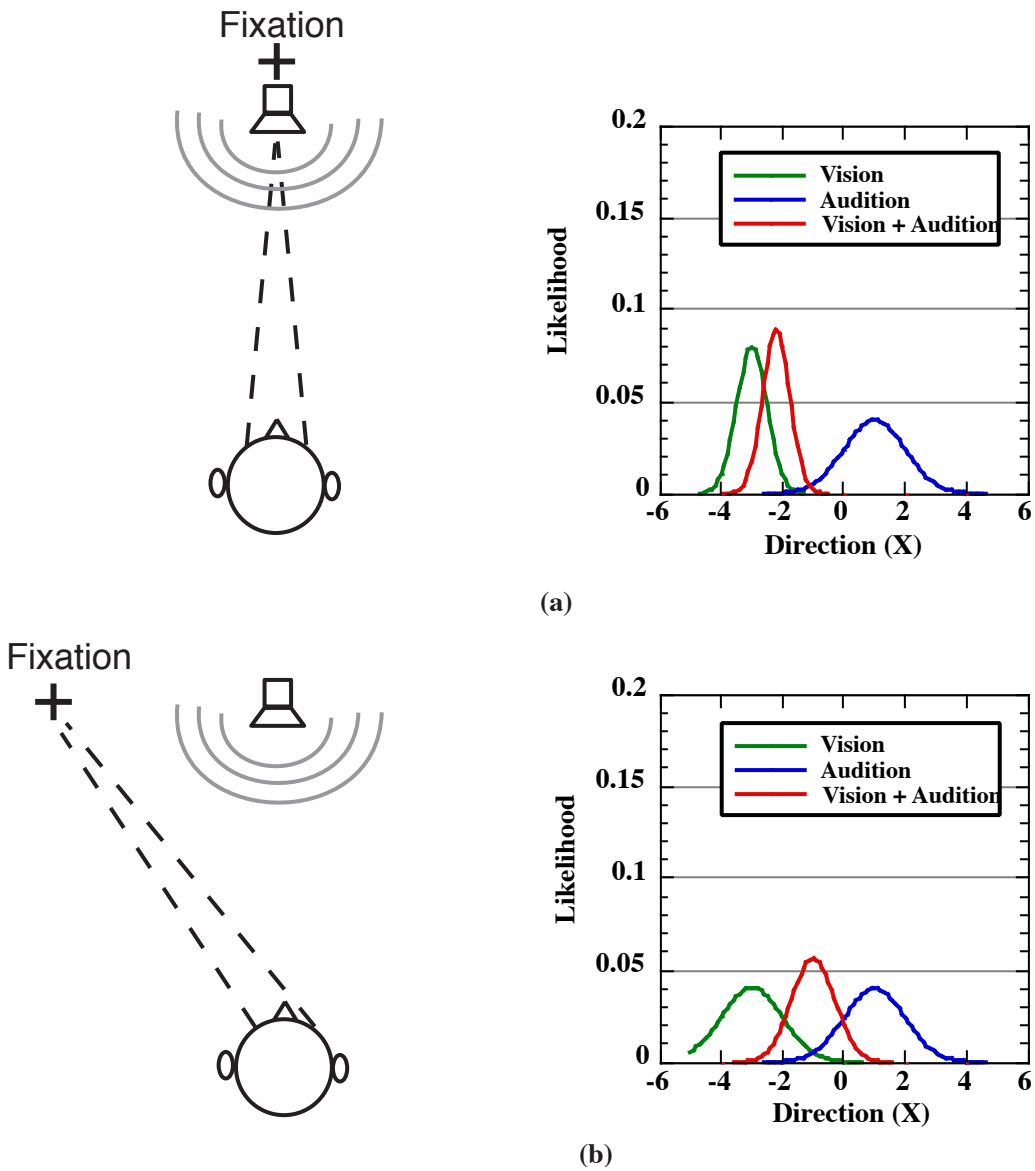


Figure 1: Two examples in which auditory and visual cues provide conflicting information about a targets direction. The conflict is apparent in the difference in means of the likelihood functions associated with each cue, though the functions overlap. Such conflicts are always present due to noise in the sensory systems. In order to optimally integrate visual and auditory information, a multimodal area must take into account the uncertainty associated with each cue. (a) When the vision cue is most reliable, the peak of the posterior distribution is shifted toward the direction suggested by the vision cue. (b) When the reliabilities of the cues is more similar, for example, when the stimulus is in the far periphery, the peak is shifted toward the direction suggested by the auditory cue. When both likelihood functions are Gaussian, the most likely direction of the is given by a weighted sum of the most likely directions given the vision and auditory cues individually, $\hat{X}_{A,V} = w_V \hat{X}_A + w_A \hat{X}_V$. The weights are inversely proportional to the variances of the likelihood functions.

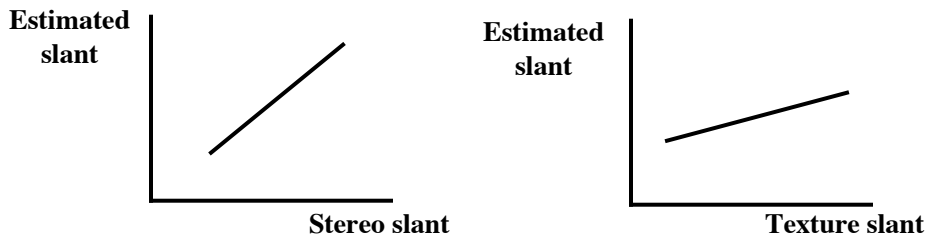


Figure 2: Perceived surface slant as a function of the slant suggested by one of the cues in a cue conflict stimulus. The example shown here illustrates a case in which a subject gives more weight to stereoscopic cues than to texture cues. (a) The slant suggested by texture information is fixed. (b) The slant suggested by stereo disparities is fixed. In practice, what we have labeled as perceived slant would be the value of a corresponding psychophysical measure, such as the point of subjective equality between a cue-conflict and a cue-consistent stimulus.

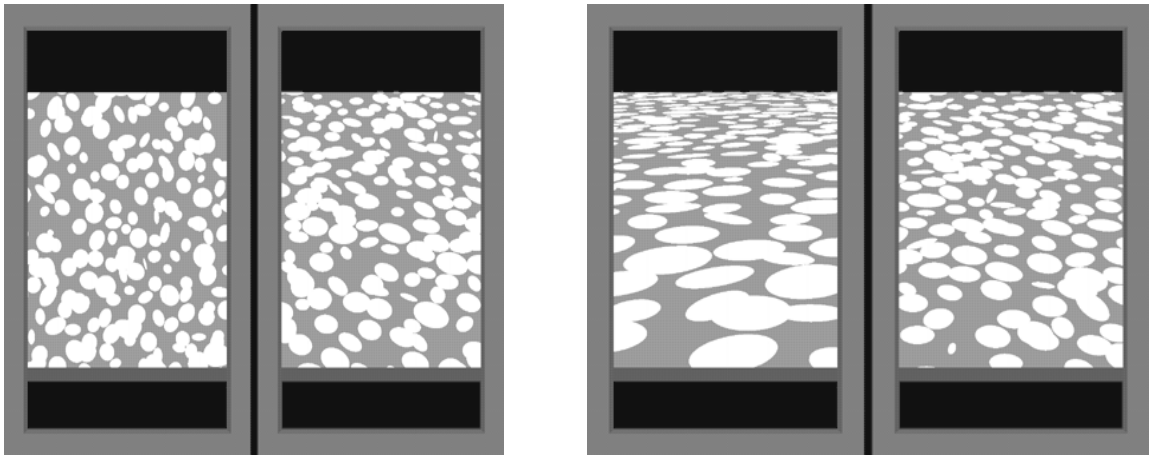
For our purposes, the point of the example is that an optimal integrator must take into account the relative uncertainty of each cue when deriving an integrated estimate. When one cue is less certain than another the integrated estimate should be biased toward the more reliable cue. Assuming that a system can accurately compute and represent likelihood functions, the calculation embodied in equations (1) and (2) and implicitly enforces this behavior (see figure 1). While other estimation schemes can show the same performance as an optimal Bayesian observer (e.g. a weighted sum of estimates independently derived from each cue), computing with likelihood functions provides the most direct means available to automatically account for the large range of differences in cue uncertainty that an observer is likely to face.

2 Psychophysical tests of Bayesian cue integration

2.1 The linear case

When several independent sensory cues are available to an observer for estimating some scene property like depth, we can write the average integrated percept as a function of the average value estimated from each cue individually, $z = f(z_1, z_2)$. Within a small neighborhood of one value of z, we approximate $f()$ as a linear function, so that we can write $z = w_1 z_1 + w_2 z_2 + k$. The weights w_1 and w_2 provide a measure of the relative contribution of the two cues to the integrated percept. A standard psychophysical procedure for measuring a set of cue weights is to find the point of subjective equality between a cue-consistent stimulus with a stimulus in which the cues are made to conflict by a small amount. If we fix the value of one cue and vary the value of the other cue used to create the cue conflict stimulus, we can derive a set of functions like those in figure 2. The relative slopes of the two curves provides a measure of the relative weights, w_1/w_2. Were we to apply this method to an optimal integrator, we would find that the relative weights satisfy the relationship, $w_1/w_2. = \sigma_2^2/\sigma_1^2$, where σ_1^2 and σ_2^2 are the variances (uncertainty) in the estimates derived from each cue individually.

Discrimination thresholds, the difference in the value of z needed by an observer to correctly discriminate stimuli at some fiduciary level of performance (e.g. 75% of the time), provide the standard psychophysical measure of uncertainty. For a gaussian model of uncertainty, thresholds



(a) (b)

Figure 3: (a) Two textured surfaces rendered at slants of 0° and 40°. The difference in slant is barely discernible. (b) Two textured surfaces rendered at slants of 70° and 60°. The difference is clearly discernible. The example illustrates the fact that texture information is a much more reliable indicator of 3D surface slant for surfaces at high slants than for surfaces at low slants.

are proportional the standard deviation of internal perceptual representations. A psychophysical test of Bayes optimality, then, would be to

Step 1: Measure discrimination thresholds T_i for stimuli containing only one or the other cue being investigated. These are approximately proportional to σ_i.

Step 2: Measure cue weights using the procedure described above.

Step 3: Test the predicted relationship, $w_1/w_2. = T_2^2/T_1^2$.

Knill and Saunders[6] applied this logic to the problem of integrating binocular depth cues with texture information for estimating planar surface slant. Figure 3 shows images of a number of textured, flat surfaces that are slanted away form the line of sight. The figure illustrates the fact that texture information is more unreliable at low slants than at high. The uncertainty in binocular cues (e.g. disparity), however, does not change as rapidly as a function of slant. This observation leads to the prediction that observers will give more weight to texture cues at high slants than low slants. Figure 4a shows a plot of the average discrimination thresholds for stimuli containing only one or the other cue. Fig. 4b shows a plot of the predicted texture cue weights derived from these thresholds (normalized so that the texture and binocular weights sum to 1) along with the average weights measured for all subjects. The measured weights closely follows those predicted from the thresholds. What appears as a slight under-weighting of texture information may reflect the fact that the stimuli used to isolate binocular cues for estimating single-cue thresholds (random dot patterns) were not equivalent to those used to measure cue weights (randomly tiled patterns). The disparity information in the latter may well have been more reliable than in the former, making the binocular cues more reliable in the combined cue stimuli used to measure cue weights.

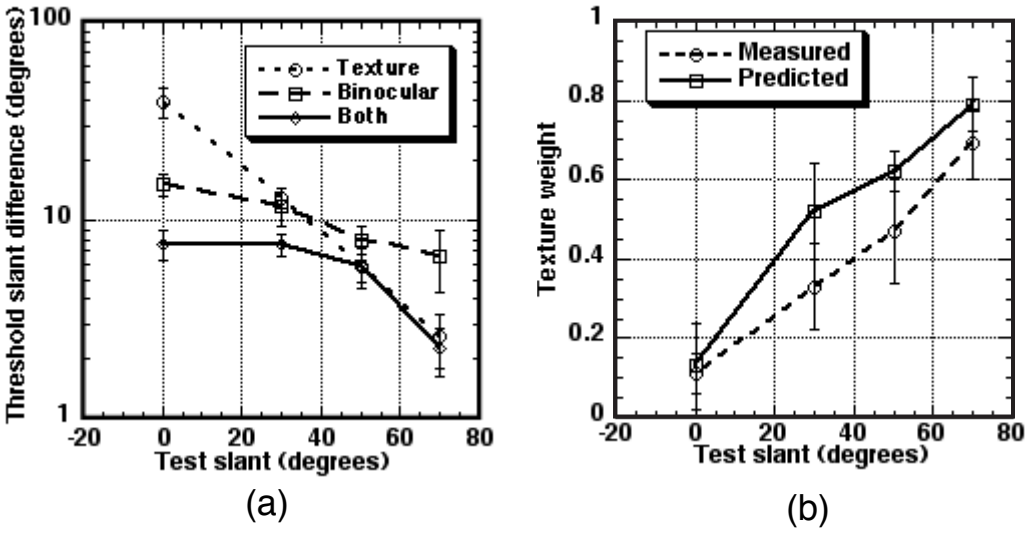


Figure 4: (a) Average slant discrimination thresholds when only texture information was available (monocular viewing of a texture similar to those in figure 3), when only stereo disparities were available (stimuli contained small random dots) or when both cues were available (binocular viewing of textures like those in figure 3. (b) Average texture cue weights measured using the cue perturbation technique described in the text. Stereo and texture cue weights were normalized to sum to 1, so that a texture weight of 0.5 would reflect equal weighting of the two cues.

Studies of human cue integration, both within modality (e.g., stereo and texture)[5, 6] and across modality (e.g., sight and touch or sight and sound)[7, 8, 9, 10] consistently find cue weights that vary in the manner predicted by Bayesian theory. While these results could be accounted for by a deterministic system that adjusts cue weights as a function of viewing parameters and stimulus properties that covary with cue uncertainty, representing and computing with probability distributions (as illustrated in figure 1 is considerably more flexible and can accommodate novel stimulus changes that alter cue uncertainty.

2.2 A nonlinear case

When the likelihood functions associated with one or another cue are not Gaussian, simple linear mechanisms do not suffice to support optimal Bayesian calculations. Non-gaussian likelihood functions arise even when the sensory noise is Gaussian as a result of the non-linear mapping from sensory feature space to the parameter space being estimated. In these cases, computations on density functions (or likelihood functions) are necessary to achieve optimality[11].

Skew symmetry is an example of a 3D cue that has a highly non-gaussian likelihood function[12]. Mirror symmetric planar patterns project to approximately skew-symmetric figure sin the image plane (under orthographic projection, the approximation is exact) (see figure 5). The information provided by skew can be parameterized by the angles of a figure's projected symmetry axes. When humans view images of symmetric figures slanted away in depth, even with correct stereoscopic viewing, they see the figures to have orientations that are slightly biased from the true orientation.

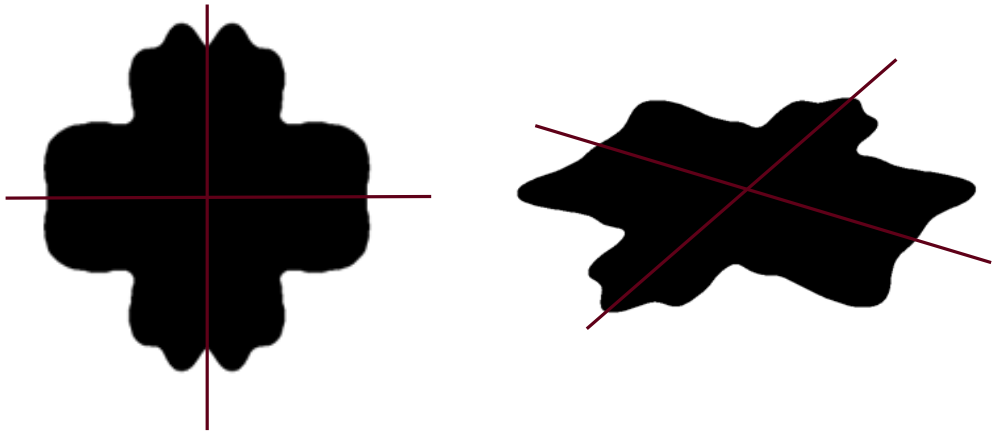


Figure 5: Skew symmetric figures appear as figures slanted in depth because the brain assumes that the figures are projected from bilaterally symmetric figures in the world. The information provided by skew symmetry is given by the angle between the projected symmetry axes of a figure, shown here as solid lines superimposed on the figure.

These biases vary with the "spin" of a figure (its orientation around the normal to the surface). These spin-dependent biases are well accounted for by a Bayesian model that optimally combines skew symmetry information (represented by a highly non-Gaussian likelihood function in figure 6) with stereoscopic information about 3D surface orientation. Figure 7 shows subjects' data along with model predictions.

A Bayesian integration model predicts that changing the 3D slant suggested by stereo disparities will lead to changes in the perceived tilt of a stereoscopically viewed symmetric figure (see figure 6). Subjects show exactly this behavior. The results would not be predicted by a deterministic scheme of weighting the estimates derived from each cue individually.

3 Psychophysical tests of Bayesian priors

Three-dimensional vision is well understood to be an ill-posed problem in the sense that multiple interpretations of a scene are generally consistent with a given set of image data. This is in part due to the inherent ambiguity of inverting the 3D to 2D perspective projection and in part due to noise in the image data. Despite this, our percepts of the 3D world are remarkably accurate and stable. The fact that our environment is highly structured makes this possible. Prior knowledge of statistical regularities in the environment allows the visual system to accurately estimate the 3-dimensional layout of surfaces in a scene even in images with seemingly impoverished information. Specific models of this type of knowledge, in the form of "a-priori" constraints, play a major role in computational theories of how the visual system estimates three-dimensional surface shape from a variety of cues. Examples include motion (rigidity[13]), surface contours (isotropy[14], symmetry[15], lines of curvature[16], geodesics[17]), shape from shading (lambertian reflectance, point light source[18]) and texture (homogeneity[19, 20], isotropy[21, 22]).

Building accurate Bayesian models of human perceptual performance requires that we formulate psychophysically testable predictions from models of the possible prior constraints that subjects

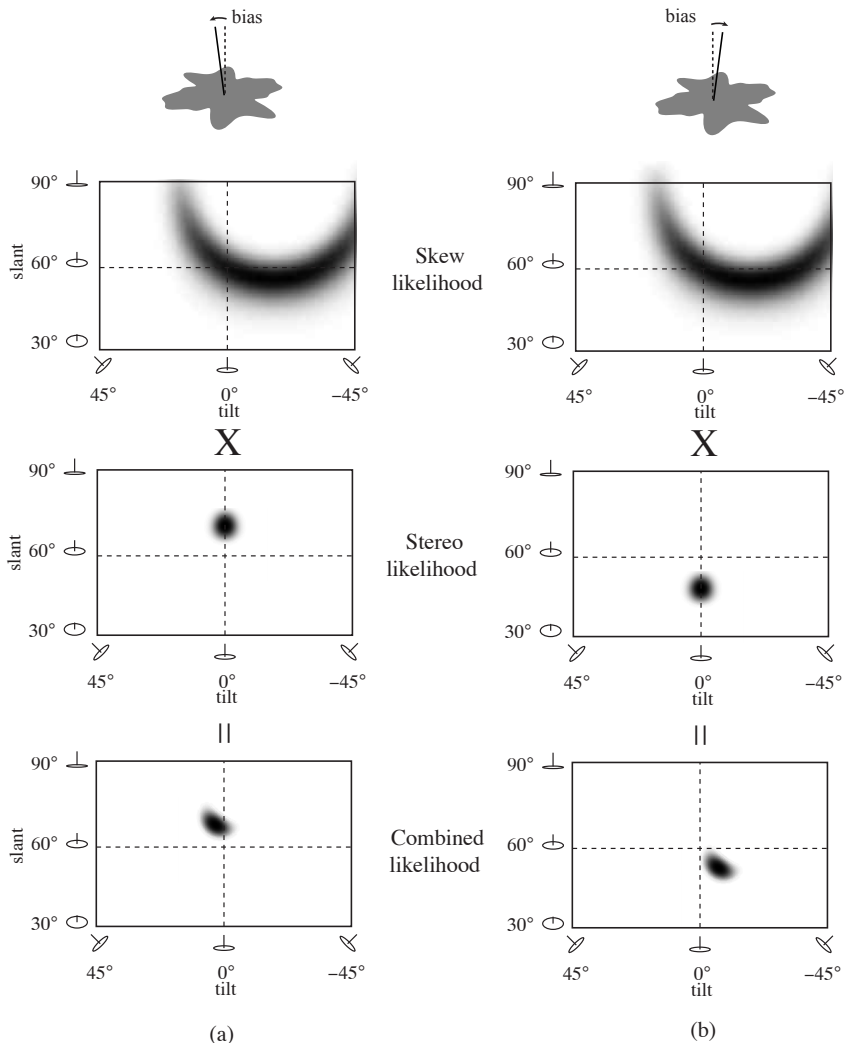


Figure 6: When viewed stereoscopically, slanted, symmetric figures appear tilted away from their true orientation. The bias is determined by the spin of the figure within its 3D plane. Assuming that visual measurements of the orientations of the skew symmetry axes in the image are corrupted by Gaussian noise, one can compute a likelihood function for 3D surface orientation from skew. The result, as shown here is highly non-Gaussian. When combined with stereoscopic information from binocular disparities, an optimal estimator multiplies the likelihood functions associated with skew and stereo to produce a posterior distribution for surface orientation, given both cues (assuming the prior on surface orientation is flat). (a) When binocular disparities suggest a slant greater than that from which a figure was projected, the posterior distribution is shifted away from the orientation suggested by the disparities in both slant and tilt, creating a biased percept of the figures tilt. (b) The same figure, when binocular disparities suggest a smaller slant, gives rise to a tilt bias in the opposite direction. This is exactly the pattern of behavior shown by subjects.

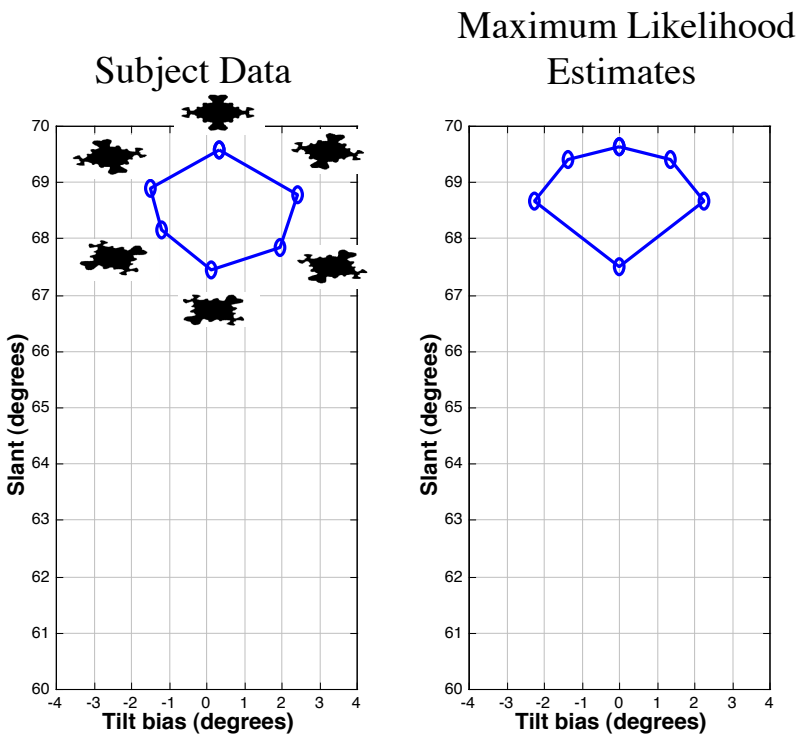


Figure 7: When viewed stereoscopically, slanted, symmetric figures appear tilted away from their true orientation. The bias is determined by the spin of the figure within its 3D plane. Shown here are average 3D orientation estimates from subjects viewing symmetric figures slanted away from the line of sight by 60°. Also shown are predictions of an optimal Bayesian integrator that assumes subjects overestimate slant-from-stereo (explaining the overall positive bias in slant judgments).

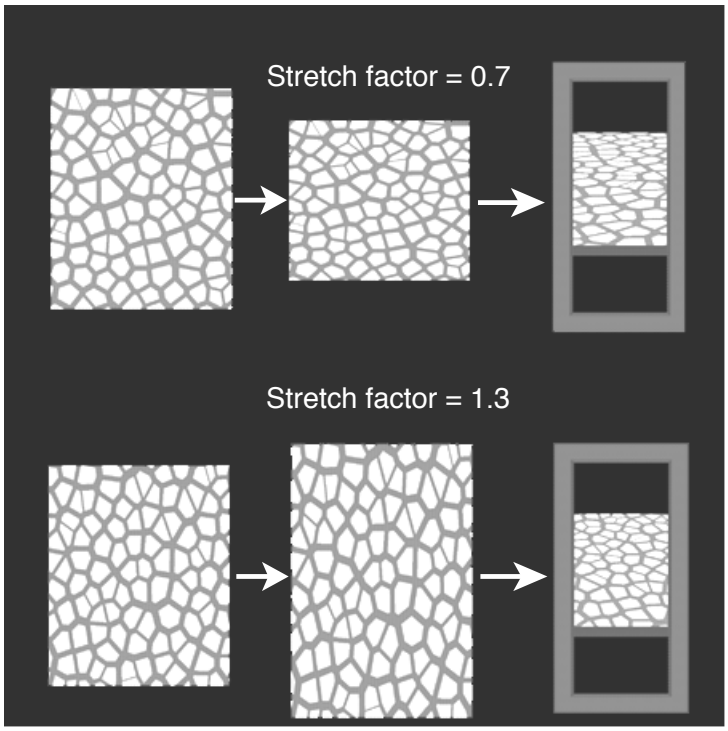


Figure 8: Stimuli for the experiments were created in three stages. First, a random, isotropic texture pattern was generated. This was then stretched by some amount in the vertical direction (here shown stretch factors of 0.7 and 1.3). The resulting texture was projected into the image at a slant of 65° and a vertical tilt. A subject that assumes surface textures are isotropic would underestimate the slant of the top stimulus and overestimate the slant of the bottom one.

might incorporate into their perceptual inferences. I will illustrate this using the example of texture isotropy. Texture patterns like those shown in figure 3 clearly provide information about 3D surface orientation; however, this can only be true if one has prior knowledge of the statistical structure of natural textures. Previous studies have shown that foreshortening information is a dominant cue for judgments of surface orientation and shape[23, 24, 25]. Since this cue relies on prior assumptions about the "shape" statistics of surface textures, knowing what assumptions human observers use is key to understanding how humans estimate surface orientation from texture.

A particularly strong constraint would be that surface textures are isotropic - that their statistical properties are invariant to orientation on a surface (they have no global orientation). Because isotropic textures have a specific average shape (circular), images of isotropic textures support much stronger inferences about surface geometry from the foreshortening cue than do images of anisotropic, homogeneous textures. In effect, when using an isotropic constraint, observers can use the local statistics of texture element shape (texture shape statistics) to make inferences about local surface orientation.

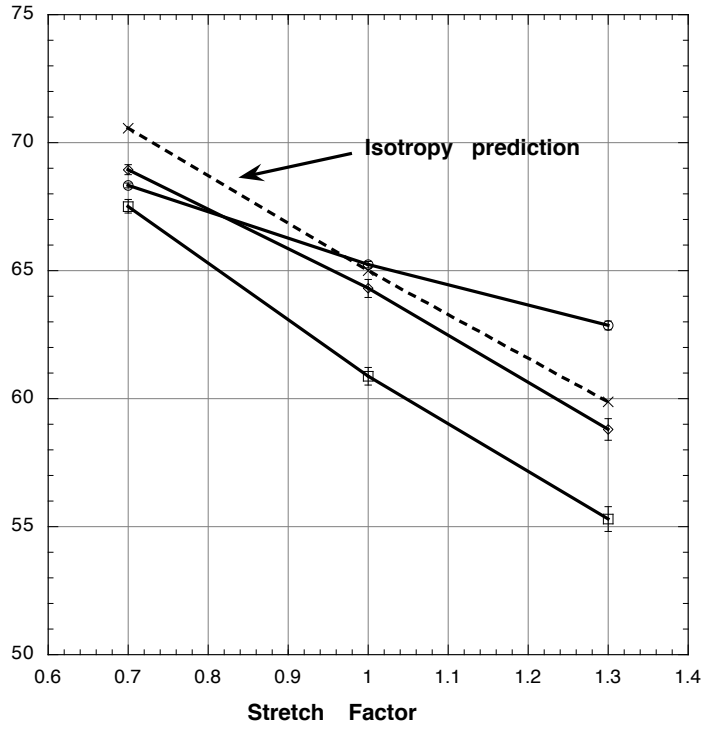


Figure 9: Plots of subjects' estimates of 3D surface slant as a function of the stretch factor used to create the surface textures prior to projecting them into the image. The dashed line shows the results predicted by the hypothesis that subjects assumed the surface textures were isotropic.

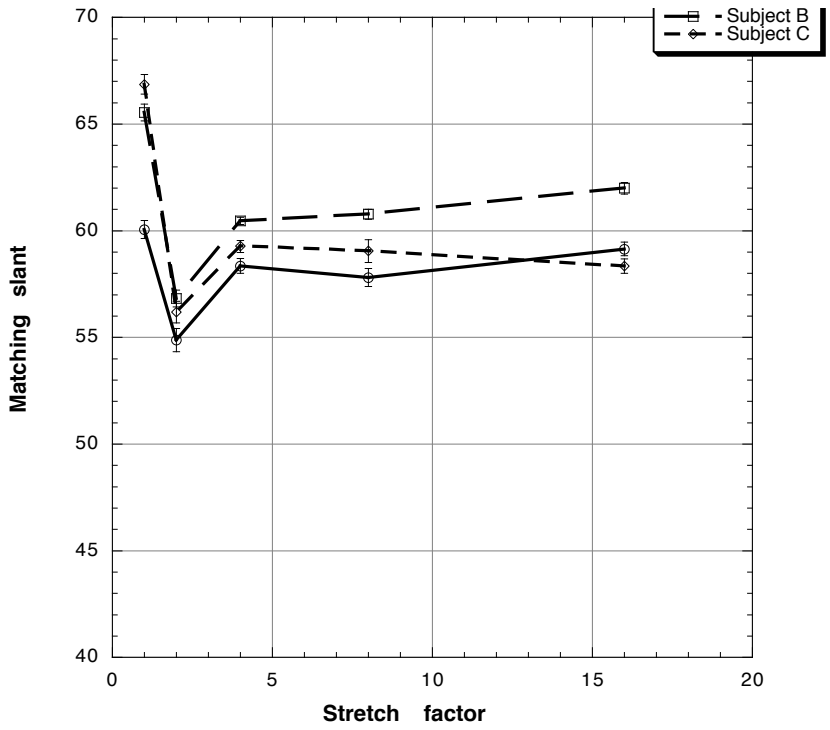


Figure 10: Plots of subjects' estimates of 3D surface slant as a function of the stretch factor used to create the surface textures prior to projecting them into the image. The stretch factors used here were considerably larger than in the previous experiment. Subjects show strong perceptual biases for images of surface textures stretched by a factor of 2 away from being isotropic, but the bias disappears for surface textures stretched by vary large amounts.

We tested whether humans assume that surface textures are isotropic by measuring the perceived 3D orientations of planar textures that were compressed or stretched by small amounts away from being isotropic (see figure 8)[11]. Were subjects to assume that image textures were projected from isotropic surface textures, these manipulations would lead to predictably biased estimates of surface orientation. Figure 9 shows the results of an experiment performed to test these predictions. Subjects judgments were biased in the manner predicted by the hypothesis that their visual systems assume surface textures are isotropic.

The human visual system does not, however, blindly impose the isotropy constraint. When the image data is inconsistent with the hypothesis that a surface texture is isotropic, subjects are able to "turn off" the isotropy constraint and the bias in their orientations estimates is diminished. This is shown by subjects performance when surface textures are compressed or stretched by large amounts prior to being projected into the image (figure10)

4 Conclusion

Bayesian probability provides a normative framework for combining sensory information from multiple cues and for combining sensory information with prior knowledge about the world. Exper-

iments to date that have quantitatively tested the predictions of Bayesian models of cue integration have largely supported the hypothesis that human observers are "Bayes' optimal" in their interpretation of image data.

References

[1] Landy, M. S., Maloney, L. T., Johnston, E. B. and Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, **35** (3), 389-412.

[2] Knill, D. C. and Richards, W. (eds.) (1996) *Perception as Bayesian Inference*, Cambridge U. Press, Cambridge, England.

[3] Yuille, A. and Bulthoff H. (1996) in *Perception as Bayesian Inference* Knill, D C and Richards, W (eds.), Cambridge University Press, Cambridge, England.

[4] Maloney, L. T. (2002) Statistical theory and biological vision, in *Perception and the physical world: Psychological and philosophical issues in perception*, Heyer, D. and Mausfeld, R., (eds.), Wiley, NY.

[5] Jacobs, R. A. (1999) Optimal integration of texture and motion cues to depth, *Vision Research*, **39**, 3621-3629.

[6] Knill, D. C. and Saunders, J. A. (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant?, *Vision Research*, 43 (24), 2539-58.

[7] van Beers, R. J., Sittig, A. C. and Denier van der Gon, J. J. (1999) Integration of proprioceptive and visual position information: An experimentally supported model, *J. Neurophysiology*, **81**, 1355-1364.

[8] Ernst, M. O. and Banks, M. S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion, *Nature*, **415** (6870): 429-433.

[9] Battaglia, P. W., Jacobs, R. A. and Aslin, R. N. (2003) Bayesian integration of visual and auditory signals for spatial localization, *Journal of the Optical Society of America A*, 20 (7), 1391-1397.

[10] Alais, D and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration, *Current Biology*, 14 (3), 257-262.

[11] Knill, D. C. (2003) Mixture models and the probabilistic structure of depth cues, *Vision Research*, 43 (7), 831-854.

[12] Saunders, J. and Knill, D. C. (2001) Perception of 3D surface orientation from skew symmetry, *Vision Research*, **41** (24), 3163 - 3185.

[13] Ullman (1979) The interpretation of structure form motion, *Proc. R. Soc. Lond. B*, **203**, 405 - 426.

[14] Brady, M. and Yuille, A. L. (1984) An extremum principle for shape from contour, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, No. 3, 288-301.

[15] Kanade, T. (1981) Recovery of the three-dimensional shape of an object from a single view, *Artificial Intelligence*, **17**, 409 - 460.

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

[17] Knill, D. C. (1992) Perception of surface contours and surface shape: from computation to psychophysics, *Journal of the Optical Society of America A*, **9** (4), 1449 - 1464.

[18] Ikeuchi, K. and Horn, B. K. P. (1981) Numerical shape from shading and occluding boundaries, *Artificial Intelligence*, **17**, 141 - 184.

[19] Garding, J. (1992) Shape from texture for smooth curved surfaces in pespective projection. *Journal of Mathematical Imaging and Vision*, **2** (4), 327-350.

[20] Malik, J. and Rosenholtz, R. (1995) Recovering surface curvature and orientation from texture distortion: A least squares algorithm and sensitivity analysis, *Proc. 3rd European Conf. on Computer Vision*, Volume 800 of *Lecture Notes in Computer Science*, 353-364, Springer-Verlag.

[21] Witkin, A. P. (1981). Recovering Surface Shape and Orientation from Texture. *Artificial Intelligence*, **17** (1), 17-45.

[22] Garding, J. (1995) Surface orientation and curvature from differential texture distortion, in *Proc. 5th International Conference on Computer Vision*, (Cambridge, MA), 733–739.

[23] Buckley, D., Frisby, J. and Blake, A. (1996) Does the human visual system implement an ideal observer theory of slant from texture? *Vision Research*, **36** (8), 1163-1176.

[24] Knill, D. C. (in press) Discriminating surface slant from texture: Comparing human and ideal observers, *Vision Research*.

[25] Knill, D. C. (1998) Ideal observer perturbation analysis reveals human strategies for inferring surface orientation from texture, *Vision Research*.