CHAPTER 1

Ideal-Observer Models of Cue Integration

Michael S. Landy, Martin S. Banks, and David C. Knill

When an organism estimates a property of the environment so as to make a decision ("Do I flee or do I fight?") or plan an action ("How do I grab that salt shaker without tipping my wine glass along the way?"), there are typically multiple sources of information (signals or "cues") that are useful. These may include different features of the input from one sense, such as vision, where a variety of cues—texture, motion, binocular disparity, and so forth—aid the estimation of the three-dimensional (3D) layout of the environment and shapes of objects within it. Information may also derive from multiple senses such as visual and haptic information about object size, or visual and auditory cues about the location of a sound. In most cases, the organism can make more accurate estimates of environmental properties or more beneficial decisions by integrating these multiple sources of information. In this chapter, we review models of cue integration and discuss benefits and possible pitfalls in applying these ideas to models of behavior.

Consider the problem of estimating the 3D orientation (i.e., slant and tilt) of a smooth surface (Hillis, Ernst, Banks, & Landy, 2002; Hillis, Watt, Landy, & Banks, 2004; Knill & Saunders, 2003; Rosas, Wagemans, Ernst, & Wichmann, 2005). An estimate of surface orientation is useful for guiding a variety of actions, ranging from reaching for and grasping an object (Knill, 2005) to judging whether one can safely walk or crawl down an incline (Adolph, 1997). Errors in the estimate may lead to failures of execution of the motor plan (and a fumbled grasp) or incorrect motor decisions (and a risky descent possibly leading to a fall). Thus, estimation accuracy can be very important, so the observer should use all sources of information effectively.

The sensory information available to an observer may come in the form of multiple visual cues (the pattern of binocular disparity, linear perspective and foreshortening, shading, etc.) as well as haptic cues (feeling the surface with the hand, testing the slope with a foot). If one of the cues always provided the observer with a perfect estimate, there would be no need to incorporate information from other cues. But cues are often imperfectly related to environmental properties because of variability in the mapping between the cue value and a given property and because of errors in the nervous system's measurement of the cue value. Thus, measured cue values will vary somewhat unpredictably across viewing conditions and scenes. For example, stereopsis provides more accurate estimates of surface orientation for near than for far surfaces. This is due to the geometry underlying binocular disparity: A small amount of measurement error translates into a larger depth error at long distances than at short ones. In addition, estimates may be based on assumptions about the scene and will be flawed if those assumptions are invalid. For example, the use of texture perspective cues is generally based on the assumption that texture is homogeneously distributed across the surface, so estimates based on this assumption will be incorrect if the texture itself varies across the surface. For example, viewing a frontoparallel photograph of a slanted, textured surface could yield the erroneous estimate that the photograph is slanted. Unlike stereopsis, the reliability of texture perspective as a cue to surface orientation does not diminish with viewing distance.

Because of this uncertain relationship between a cue measurement and the environmental property to be estimated, the observer can generally improve the reliability of an estimate of an environmental property by combining multiple cues in a rational fashion. The combination rule needs to take into account the uncertainties associated with the individual cues, and those depend on many factors.

Along with the benefit of improving the reliability of perceptual estimates, there is also a clear benefit of knowing how uncertain the final estimate is and how to make decisions given that uncertainty. Consider, for example, estimating the distance to a precipitous drop-off (Maloney, 2002). An observer can estimate that distance most reliably by using all available cues, but knowing the uncertainty of that estimate can be crucial for guiding future behavior. If the future task is to toss a ball as close as possible to the drop-off, one would use the most likely distance estimate to plan the toss; the plan would be unaffected by the uncertainty of the distance estimate. If, however, the future task is to walk blindfolded toward the drop-off, the decision of how far to meander toward the drop would most certainly be influenced by the uncertainty of the distance estimate.

Much of the research in this area has focused on the question of whether cue integration is optimal. This focus has been fruitful for a variety of reasons. First, to determine whether the nervous system is performing optimally requires a clear, quantitative specification of the task, the stimulus, and the relationship between the stimulus and the specified environmental property. As Gibson (1966) argued, it forces the researcher to investigate and define the information available for the task. As Marr (1982) put it, it forces one to construct a quantitative, predictive account of perceptual performance. Second, for tasks that have been important for survival, it seems quite plausible that the organism has evolved mechanisms that utilize the available information optimally. Therefore, the hypothesis that sensory information is used optimally in tasks that are important to the organism is a reasonable starting point. Indeed, given the efficacy of natural selection and developmental learning mechanisms, it seems unlikely to us that the nervous system would perform suboptimally in an important task with stimuli that are good exemplars of the natural environment (as opposed to impoverished or unusual stimuli that are only encountered in the laboratory). Third, using optimality as a starting point, the observation of suboptimal behavior can be particularly informative. It can indicate flaws in our characterization of the perceptual problem posed to or solved by the observer; for example, it could indicate that the perceptual system is optimized for tasks other than one we have studied or that the assumptions made in our formulation of an ideal-observer model fail to capture the problem posed to observers in naturalistic situations. Of course, there remains the possibility that we have characterized the sensory information and the task correctly, but the nervous system simply has not developed the mechanisms for performing optimally (e.g., Domini & Braunstein, 1998; Todd, 2004). We expect that such occurrences are rare, but emerging scientific investigations will ultimately determine this.

In this way, "ideal-observer" analysis is a critical step in the iterative scientific process of studying perceptual computations. At perhaps a deeper level, ideal-observer models help us to understand the computational structure of what are generally complex problems posed to observers. This can in turn lead to understanding complex behavior patterns by relating them to the features of the problems from which they arise (e.g., statistics of natural environments or noise characteristics of sensory systems). Ideal-observer models provide a framework for constructing quantitative, predictive accounts of perceptual performance at Marr's computational level for describing the brain (Marr, 1982).

Several studies have found that humans combine sensory signals in an optimal fashion, taking into account the variation of cue reliability with viewing conditions, and resulting in estimates with maximum reliability (e.g., Alais & Burr, 2004; Ernst & Banks, 2002; Hillis et al., 2004; Knill & Saunders, 2003; Landy & Kojima, 2001; Tassinari, Hudson, & Landy, 2006). These results suggest that human observers are optimal for a wide variety of perceptual and sensorimotor tasks.

This chapter is intended to provide a general introduction to the field of cue combination from the perspective of optimal cue integration. We work through a number of qualitatively different problems, and we hope thereby to illustrate how building ideal observers helps formulate the scientific questions that need to be answered before we can understand how the brain solves these problems. We begin with a simple example of integration leading to a linear model of cue integration. This is followed by a summary of a general approach to optimality: Bayesian estimation and decision theory. We then review situations in which realistic generative models of sensory data lead to nonlinear ideal-observer models. Subsequent sections review empirical studies of cue combination and issues they raise, as well as open questions in the field.

LINEAR MODELS FOR MAXIMUM RELIABILITY

There is a wide variety of approaches to cue integration. The specific approach depends on the assumptions the modeler makes about the sources of uncertainty in sensory signals as well as what the observer is trying to optimize. Quantitative empirical evidence can then determine whether those assumptions are valid.

The simplest such models result in linear cue integration. For the case of Gaussian noise, linear cue integration is optimal for an observer who tries to maximize the precision (i.e., minimize the variance) of the estimate made based on the cues. Suppose you have samples x_i, $i = 1, \cdots, n$, of n independent, Gaussian random variables X_i that share a common mean μ and have variances σ_i^2. The minimum-variance unbiased estimator of μ is a weighted average

$$\hat{x} = \sum_{i=1}^{n} w_i x_i, \tag{1.1}$$

where the weight w_i of cue i is proportional to that cue's *reliability* r_i (defined as its inverse variance, $r_i = 1/\sigma_i^2$):

$$w_i = \frac{r_i}{\sum_{j=1}^{n} r_j} \tag{1.2}$$

(Cochran, 1937). The reliability r of this integrated estimate is

$$r = \sum_{i=1}^{n} r_i. \tag{1.3}$$

As a result, the variance of the integrated estimate is generally lower than the variance of the individual estimates and never worse than the least variable of them. Thus, if an observer has access to unbiased estimates of a particular world property from each cue, and the cues are Gaussian distributed and conditionally independent (meaning that for a given value of the world property being estimated, errors in the estimates derived from each cue are independent), the minimum-variance estimate is a weighted average of the individual estimates from each cue (Landy, Maloney, Johnston, & Young, 1995; Maloney & Landy, 1989).

To form this estimate, an observer needs to represent and compute with estimates of cue uncertainty. The estimates could be implicit in the neural population code derived from the sensory features associated with a cue or might be explicitly computed, for example, by measuring the stability of each cue's estimates over repeated views of the scene. They could also be assessed online by using ancillary information (viewing distance, amount of self-motion, etc.) that impacts cue reliability (Landy et al., 1995). Estimates of reliability need not be explicitly represented by the nervous system, but they might be implicit in the form of the neural population code (Ma, Beck, Latham, & Pouget, 2006).

If the variability in different cue estimates is correlated, the minimum-variance unbiased estimator will not necessarily be a weighted average. For some distributions, it is a nonlinear function of the individual estimates; for others, including the Gaussian, it is still a weighted average, but the weights take into account the covariance of the cues (Oruç, Maloney, & Landy, 2003).

BAYESIAN ESTIMATION AND DECISION MAKING

The linear model has dominated cue-integration research and has provided important insights into human perceptual and sensorimotor processing. However, most perceptual and sensorimotor problems encountered by an observer in the natural world cannot be accurately characterized by the linear model. In many cases, the linear model provides a reasonable "local" approximation to the ideal observer. In those cases, the complexity of the problem is reduced in the laboratory setting and this reduction is assumed to be known by the observer. To characterize the complex problems presented to observers in the real world, a more general computational framework is needed. Bayesian decision theory provides such a framework.

Priors, Likelihoods, and Posteriors

In the Bayesian framework, the information provided by sensory information to estimate a scene property or make a decision related to that property is represented by a "posterior" probability distribution

$$P(s|d) = \frac{P(d|s)P(s)}{P(d)}, \tag{1.4}$$

where s represents the scene property or properties of interest (possibly multidimensional) and d is a vector of sensory data. In this formulation, it is important to delineate what is known by the observer from what is unknown. The data, d, are given to and therefore known by the observer. The scene properties, s, are unknown. The probability distribution, $P(s|d)$, represents the probabilities of different values of s being "true," given the observed data. If the distribution is narrowly concentrated around one value of s, it represents reliable data; if broad, it represents unreliable data. If it is narrow in one dimension and broad in others, it reflects a situation in which the information provided by d reliably determines s along the narrow dimension but does not along the other dimensions.

Bayes' rule (Eq. 1.4) shows how to compute the posterior distribution from prior knowledge about the statistics of s—represented by the prior distribution $P(s)$ (that is, which values of s are more likely in the environment than others)—and knowledge about how likely scenes with different values of s are to give rise to the observed data d, which is represented by the likelihood function $P(d|s)$. Because d is given, the likelihood is a function of the conditioning variable and does not behave like a probability distribution (i.e., it need not integrate to one), and hence it is often notated as $L(s|d)$. The third term in the denominator, $P(d)$, is a constant, normalizing term (so that the full expression integrates to one), and it can generally be ignored in formulating an estimation procedure. From a computational point of view, if one has a good "generative" model for how the data are generated by different scenes (e.g., the geometry of disparity and the noise associated with measuring disparity) and a good model of the statistics of scenes, one can use Bayes' rule to compute the posterior distribution, $P(s|d)$, and hence derive a full representation of the information provided by some observed sensory data about scene properties of interest.

Gain/Loss Functions, Estimation, and Decision Making

Having computed the posterior distribution, the Bayesian decision maker next chooses a course of action. For that action to be optimal, one must have a definition of optimality. That is, one must have a loss function that defines the consequences of the decision maker's action. Optimality is defined as making decisions that minimize expected loss. For Bayesian estimation, the observer's task is to choose an estimate, and often the loss is defined as a function of estimation error (the difference between the chosen estimate and the true value of the parameter in the world). In other situations, the observer makes a discrete decision (e.g., categorizing the stimulus as signal or noise) or forms a motor plan (e.g., a planned trajectory to grasp an observed object).

Bayesian decision theory prescribes the optimal choice of action based on several ingredients. First, one needs a model of the environment: that is, a set of possible states of the world or *scenes* and a prior distribution across them (random variable S with prior distribution $P(s)$). This world leads to noisy sensory data d conditioned on a particular state of the world (with distribution $P(d|s)$). The task of the observer is to choose an optimal action $a(d)$, which might be an estimate of a scene parameter, a button press in an experiment, or a plan for movement in a visuomotor task. For experimental tasks or estimation, the action is the final output of the decision-making task upon which gain or loss is based. In other situations, like visuomotor control, the outcome itself—the executed movement—may be stochastic. So we distinguish the outcome of the plan (e.g., the movement trajectory) t as distinct from the selected action $a(d)$ (with distribution $P(t|a(d))$). The final ingredient is typically called the loss function, although we also use the negative of loss, or gain $g(t, s)$. Note that g is a function only of the actual scene s and actual outcome of the decision t. An optimal choice of action is one that maximizes expected gain

$$\begin{aligned} a_{\text{opt}} &= \arg\max_a EG(a), \text{ where} \\ EG(a) &= \iiint g(t, s) P(t|a(d)) P(d|s) \\ &\qquad P(s)\, dt\, dd\, ds. \end{aligned} \tag{1.5}$$

It is worth reviewing some special cases of this general method. For estimation, the final output is the estimate itself ($t \equiv a(d)$). If the prior distribution is uniform over the domain of interest ($P(s) = c$, a constant) and the gain function only rewards perfectly correct estimates (a delta function centered on the correct value of the parameter), then Eq. 1.5 results in maximum-likelihood estimation: that is, choosing the mode of $P(d|s)$ over possible scenes s. If the prior distribution is not uniform, the optimal method is maximum a posteriori (MAP) estimation, that is, choosing the mode of the posterior $P(s|d)$. If the gain function is not a delta function, but treats estimation errors symmetrically (i.e., is a function of $|\hat{s} - s|$, where $\hat{s}$ is the estimate and s is the true value in the scene), the optimal estimation procedure corresponds to first convolving the gain function with the posterior distribution, and then choosing the estimate corresponding to the peak of that function. For example, the oft-used squared-error loss function leads the optimal observer to use a minimum-variance criterion and hence the mean of the posterior as the estimate.

Bayesian Decision Theory and Cue Integration

The most straightforward application of Bayesian decision theory to cue integration involves the case in which the sensory data associated with each cue are conditionally independent. In that case, we can write the likelihood function for all of the data as the product of likelihood functions for the data associated with each cue,

$$P(d_1, \cdots, d_n|s) = \prod_{i=1}^{n} P(d_i|s), \tag{1.6}$$

where d_i is a data vector representing the sensory data associated with cue i (e.g., disparity for the stereo cue) and s is the scene variable being estimated. Combining Eqs. 1.4 and 1.6, we have

$$P(s|d_1, \cdots, d_n) \propto P(s) \prod_{i=1}^{n} P(d_i|s), \tag{1.7}$$

where we have dropped the constant denominator term for simplicity.

If the individual likelihood functions and the prior distribution are Gaussian, with variances σ_i^2 and σ_{prior}^2, then the posterior distribution will be Gaussian with mean and variance identical to the minimum-variance estimate; that is, for Gaussian distributions, the MAP estimate and the mean of the posterior both yield a linear estimation procedure identical to that of the minimum-variance unbiased estimator expressed in Eqs. 1.1–1.3. If the prior distribution is flat or significantly broader than the likelihood function, the posterior is simply the product of individual cue likelihoods and the mode and mean correspond to the maximum-likelihood estimate of s. If the Gaussian assumption holds, but the data associated with the different cues are not conditionally independent, the MAP estimate will remain linear, but the cue weights have to take into account the covariance structure of the data, resulting in the same weighted linear combination as the minimum-variance, unbiased estimate (Oruç et al., 2003).

While a linear system can characterize the optimal estimator when the estimates are Gaussian distributed and conditionally independent, the Bayesian formulation offers an equally simple, but much more general formulation. In essence, it replaces the averaging of estimates with the combining of information as represented by multiplying likelihood functions and priors. It also replaces the notion of perceptual estimates as point representations (single, specific values) with a notion of perceptual estimates as probability distributions. This allows one to separate information (as represented by the posterior distribution) from the task, as represented by a gain function.

Figure 1.1 illustrates Bayesian integration in two simple cases: estimation of a scalar variable (size) from a pair of cues (visual and haptic) with Gaussian likelihood functions, and estimation of a two-dimensional variable (slant and tilt) from a pair of cues, one of which is decidedly non-Gaussian (skew symmetry). While the latter may appear more complex than the former, the ideal observer operates similarly by multiplying likelihood functions and priors.

NONLINEAR MODELS: GENERATIVE MODELS AND HIDDEN VARIABLES

We now turn to conditions under which optimal cue integration is not linear. We will describe three qualitatively different features of cue-integration problems that make the linear model inappropriate. One such situation is when the information provided by two cues interacts because each cue disambiguates scene or viewing variables that the other cue requires to determine a scene property. Another problem is that the information provided by many cues, particularly visual depth cues, depends on what prior assumptions about the world hold true and cues can interact by jointly determining the appropriate world model. A special case of this is a situation in which an observer has to decide whether different sensory cues should or should not be combined into one estimate at all.

Cue Disambiguation

The raw sensory data from different cues are often incommensurate in the sense that they specify a scene property in different coordinate frames of reference. For example, auditory cues provide information about the location of a sound source in head-centered coordinates, whereas visual cues provide information in retinal coordinates. To apply a linear scheme for combining these sources of information, one would first need to use an estimate of gaze direction relative to the head to convert visual position estimates to head-centered coordinates or auditory position estimates to retinal coordinates, so that the two location estimates are in the same coordinates. Similarly, visual depth estimates based on relative motion should theoretically be scaled by an estimate of the viewing distance to provide an estimate of metric depth (i.e., an estimate in centimeters). On the other hand, depth derived from disparity needs to be scaled approximately by the square of the viewing distance to put it in the same units. Landy et al. (1995) called this preliminary conversion into common units *promotion.*

A normative Bayesian model finesses the problem in a very elegant way. Figure 1.2 illustrates the structure of the computations as applied to disparity and velocity cues to relative depth. The key observation is that the generative model for the sensory data associated with both the disparity and velocity cues depends not only on the scene property being estimated (relative depth) but also on the viewing distance to the fixated point. We will refer to viewing

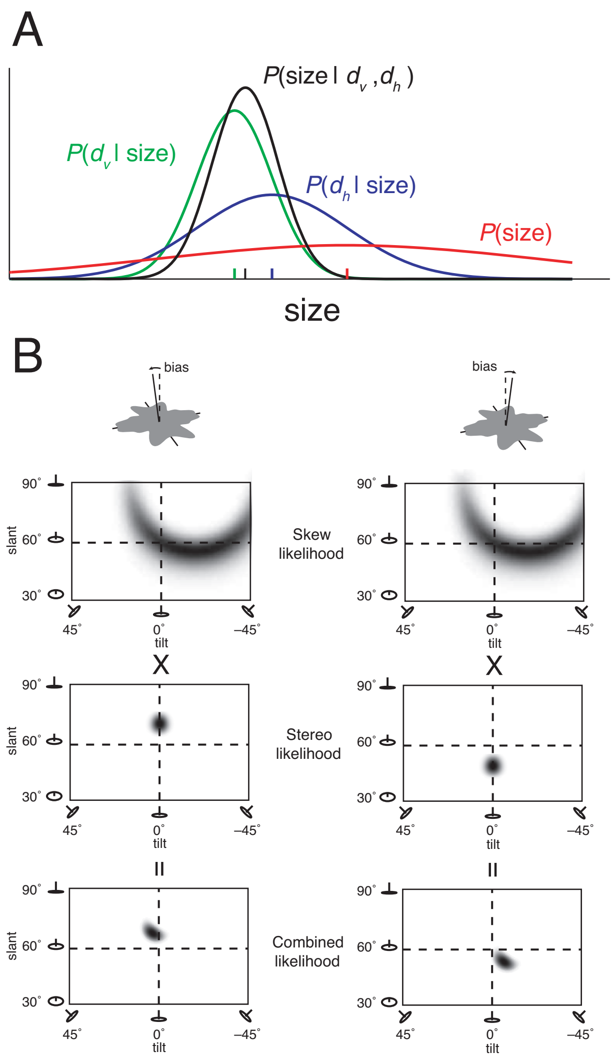

Figure 1.1 Bayesian integration of sensory cues. (A) Two cues to object size, visual and haptic, each have Gaussian likelihoods (as in Ernst & Banks, 2002). The resulting joint likelihood is Gaussian with mean and variance as predicted by Eqs. 1.1–1.3. (B) Two visual cues to surface orientation are provided: skew symmetry (a figural cue) and stereo disparity (as in Saunders & Knill, 2001). Surface orientation is parameterized as slant and tilt angles. 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. Assuming that visual measurements of the orientations of these angles in the image are corrupted by Gaussian noise, one can compute a likelihood function for three-dimensional (3D) surface orientation from skew. The result, as shown here, is highly non-Gaussian. The shape of the likelihood function is highly dependent on the spin of the figure around its 3D surface normal. Top row of graphs: skew likelihood for the figure shown at the top. Middle row: two stereo likelihoods centered on larger (left) and smaller (right) values of slant. Bottom row: When combined with stereoscopic information from binocular disparities, assuming the prior on surface orientation is flat. This leads to the prediction that perceptual biases will depend on the spin of the figure. It also leads to the somewhat counterintuitive prediction illustrated here that changing the slant suggested by stereo disparities should change the perceived tilt of symmetric figures. This is exactly the pattern of behavior shown by subjects.

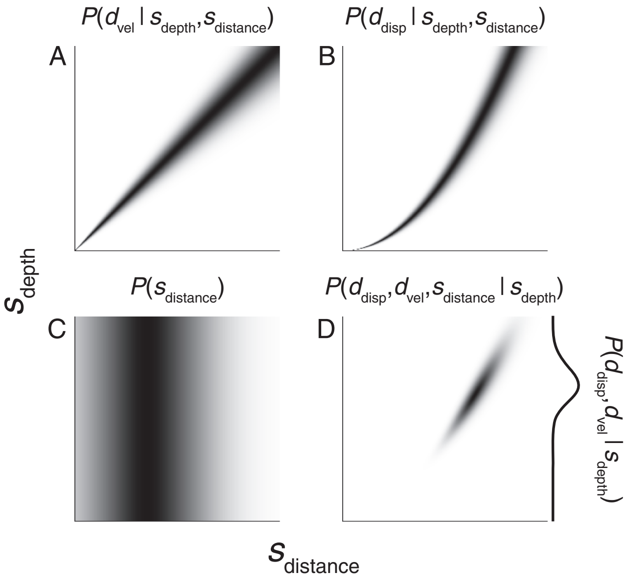

Figure 1.2 Cue disambiguation. (A) Likelihood function for the motion cue as a function of depth and viewing distance $P(d_{vel}|s_{depth}, s_{distance})$. The depth implied by a given retinal velocity is proportional to the viewing distance. (B) Likelihood function for the disparity cue $P(d_{disp}|s_{depth}, s_{distance})$. The depth implied by a given retinal disparity is approximately proportional to the square of the viewing distance. (C) A Gaussian prior on viewing distance. (D) Combined likelihood function $P(d_{disp}, d_{vel}, s_{distance}|s_{depth})$. The right-hand side of the plot illustrates the likelihood for depth alone, $P(d_{disp}, d_{vel}|s_{depth})$, integrating out the unknown distance.

distance as a hidden variable in the problem (statisticians refer to these kinds of variables as nuisance parameters; for more discussion of the role of hidden variables and marginalization in visual cue integration, see Knill, 2003). The generative model for both the relative-disparity and relative-velocity measurements requires that both relative depth and viewing distance be specified. This allows one to compute likelihood functions for both cues, d_{vel} and d_{disp}, $P(d_{vel}|s_{depth}, s_{distance})$ and $P(d_{disp}|s_{depth}, s_{distance})$ (assuming one knows the noise characteristics of the disparity and motion sensors). Assuming that the noises in the two sensor systems are independent, we can write the likelihood function for the two cues as the product of the likelihood functions for the individual cues,

$$\begin{aligned} &P(d_{disp} d_{vel}|s_{depth}, s_{distance}) \\ &= P(d_{disp}|s_{depth}, s_{distance}) \\ &\quad \times P(d_{vel}|s_{depth}, s_{distance}). \end{aligned} \tag{1.8}$$

This is not quite what we want, however. What we want is the likelihood function for depth alone, $P(d_{disp}, d_{vel}|s_{depth})$. This is derived by integrating ("marginalizing") over the hidden variable, $s_{distance}$:

$$\begin{aligned} &P(d_{disp} d_{vel}|s_{depth}) \\ &= \int P(d_{disp}, d_{vel}, s_{distance}|s_{depth}) ds_{distance} \\ &= \int P(d_{disp}, d_{vel}|s_{distance}, s_{depth}) \\ &\quad \times P(s_{distance}) ds_{distance} \\ &= \int P(d_{disp}|s_{depth}, s_{distance}) \\ &\quad \times P(d_{vel}|s_{depth}, s_{distance}) P(s_{distance}) ds_{distance}. \end{aligned} \tag{1.9}$$

(Note that the second step required that depth and distance be independent.) The important thing to note is that the joint likelihood for s_{depth} is not the product of the individual cue likelihoods, $P(d_{\text{disp}}|s_{\text{depth}})$ and $P(d_{\text{vel}}|s_{\text{depth}})$. Rather, we had to expand the representational space for the scene to include viewing distance, express both likelihood functions in that space, multiply the likelihoods in the expanded space and then integrate over the hidden variable to obtain a final likelihood. If we had a nonuniform prior on relative depth, we would then multiply the likelihood function by the prior and normalize to obtain the posterior distribution. As illustrated in Figure 1.2, both cues are consistent with a large range of relative depths (depending on the viewing distance assumed), but because the cues depend differently on viewing distance, when combined they can disambiguate both relative depth and viewing distance (Richards, 1985).

An alternative to this approach would be to estimate the viewing distance from ancillary information (e.g., vergence signals from the oculomotor system). With these parameters fixed, optimal cue integration will again be linear. However, this approach is almost certainly suboptimal because it ignores the noise in the ancillary signals. The optimal approach is to incorporate the information from ancillary signals in the same Bayesian formulation. In this case, extraretinal vergence signals specify a likelihood function in depth-distance space that is simply stretched out along the depth dimension (because those signals say nothing about relative depth) (much like the prior in the lower-left panel of Fig. 1.2). In this way, vergence signals disambiguate viewing distance only in so much as the noise in the signals allows. If that noise is high, the disambiguating effects of the nonlinear interaction between the relative disparity and relative motion signals will dominate the perceptual estimate.

Robust Estimation and Mixture Priors

One might ask how a normative system should behave when cues suggest very different values for some scene property. Consider a case in which disparity indicates a frontoparallel surface, but the texture pattern in the image suggests a surface slanted away from frontoparallel by 60°. A linear system would choose some intermediate slant as its best estimate, but if the relative reliabilities of the two cues (i.e., the inverse of the variances of the associated likelihood functions) were similar, this estimate would be at a slant (say 30°) that is wildly inconsistent with both cues.

On the face of it, this appears like a standard problem in robust statistics. For example, the mean of a set of samples can be influenced strongly by a single outlier, and robust, nonlinear statistical methods, such as the trimmed mean, are intended to alleviate such problems (Hampel, 1974; Huber, 1981). The trimmed mean and related methods reduce the weight of a given data point as the value of that data point becomes increasingly discrepant from the bulk of the sample. The application of robust statistical methods to cue integration is difficult, however, because one is usually dealing with a small number of cues rather than a large sample, so it is often unclear which cue should be treated as the discrepant outlier.

A discrepant cue may result from a particularly noisy sample, but it may also indicate that the estimate from that cue was fallacious due to a mistaken assumption (Landy et al., 1995). The second problem is more common than the first. The observation that outliers may arise from fallacious assumptions suggests a reconceptualization of the outlier problem. Consider the case of depth perception. All pictorial depth cues rely on prior assumptions about objects in the world (texture relies on homogeneity, linear perspective on parallelism, relative motion on rigidity, etc.). A notable and very simple example is that provided by the compression cue (Fig. 1.3A). The visual system interprets figures that are very compressed in one direction as being slanted in 3D in that direction. For example, the visual system uses the aspect ratio of ellipses in the retinal image as a cue to the 3D slant of a figure, so much so that it gives nearly equal weight to that cue and disparity in a variety of viewing conditions (Hillis et al., 2004; Knill & Saunders, 2003). Of course, the aspect ratio of an ellipse on the retina is only useful if one can assume that the figure from which it projects is a circle. This is usually a reasonable assumption because most ellipses in an image are circular in the world. When disparities suggest a slant differing by only a small amount from that suggested by the compression cue, it makes sense to combine the two cues linearly. When disparities suggest a very different slant, however, the discrepancy provides evidence that one is viewing a noncircular ellipse. In this situation, an observer should down-weight the compression cue or even ignore it.

Figure 1.3 illustrates how these observations are incorporated into a Bayesian model (see Chapter 9 and Knill, 2007b, for details). The generative model for the aspect ratio of an ellipse in the image depends on both the 3D slant of a surface and the aspect ratio of the ellipse in the world. The aspect ratio of the ellipse in the world is a hidden variable and must be integrated out to derive the likelihood for slant. The prior distribution on ellipse aspect ratios plays a critical role here. The true prior is

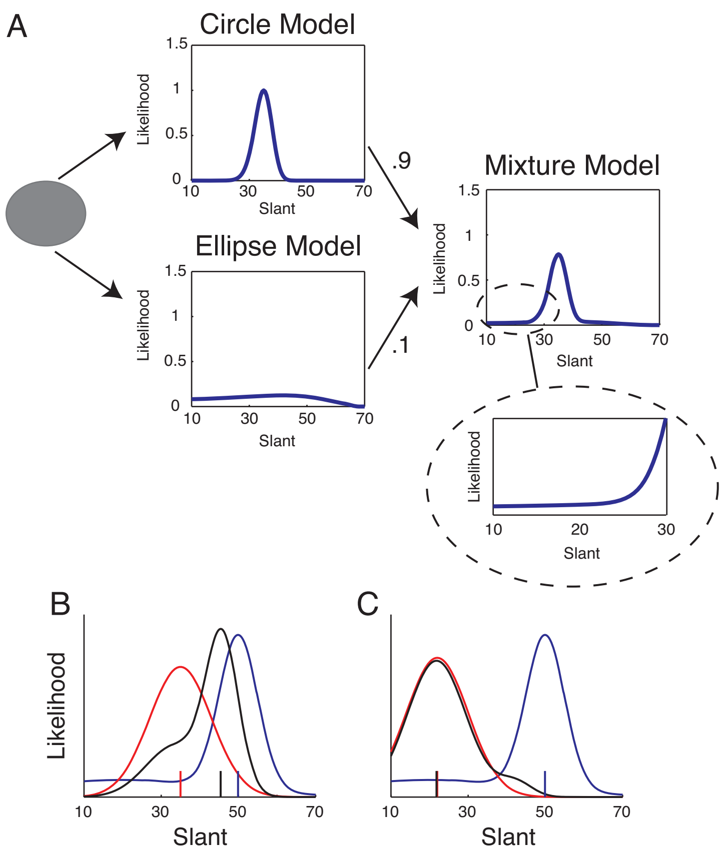

Figure 1.3 Bayesian model of slant from texture (Knill, 2003). (A) Given the shape of an ellipse in the retinal image, the likelihood function for slant is a mixture of likelihood functions derived from different prior models on the aspect ratios of ellipses in the world. The illustrated likelihoods were derived by assuming that noise associated with sensory measurements of aspect ratio has a standard deviation of 0.03, that the prior distribution of aspect ratios of randomly shaped ellipses in the world has a standard deviation of 0.25, and that 90% of ellipses in the world are circles. The mixture of narrow and broad likelihood functions creates a likelihood function with long tails, as shown in the blow-up. (B) Combination of a long-tailed likelihood function from compression (blue) and a Gaussian likelihood function from disparity (red) yields a joint likelihood function that peaks between the two individual estimates when the individual estimates are similar, much as with the linear, Gaussian model of cue integration. (C) When the cue conflict is increased, the heavy tail of the compression likelihood results in a joint likelihood that peaks at the disparity estimate, effectively vetoing the compression cue.

a mixture of distributions, each corresponding to different categories of shapes in the world. A simple first-order model is that the prior distribution is a mixture of a delta function at one (i.e., all of the probability massed at one) representing circles and a broader distribution over other possible aspect ratios representing randomly shaped ellipses. In Figure 1.3, the width of the likelihood for the circle model is due to sensory noise in measurement of ellipse aspect ratio. The result is a likelihood function that is a mixture of two likelihood functions—one derived for circles, in which the uncertainty in slant is caused only by noise in sensory measurements of shape on the retina, and one derived for randomly shaped ellipses, in which the uncertainty in slant is a combination of sensory noise and the variance in aspect ratios of ellipses in the world. The result is a likelihood function for the compression cue that is peaked at the slant consistent with a circle interpretation of the measured aspect ratio but has broad tails (Fig. 1.3A).

The likelihood function for both the compression cue and the disparity cue results from multiplying the likelihood function for disparity (which presumably does not have broad tails; but see Girshick and Banks, 2009, for evidence that the disparity likelihood also has broad tails) with the likelihood function for the compression cue. The resulting likelihood function peaks at a point either between the peaks of the two cue likelihood functions when they are close to one another (small cue conflicts, Fig. 1.3B) or very near the peak of the disparity likelihood function when they are not close (large cue conflicts, Fig. 1.3C). The latter condition appears behaviorally as a down-weighting or vetoing of the compression cue. Thus, multiplying likelihood functions can result in a form of model selection, thereby determining which prior constraint is used to interpret a cue. Similar behavior can be predicted for many different depth cues because they also derive their informativeness from a mixture of prior constraints that hold for different categories of objects. This behavior of integrating with small cue conflicts and vetoing with large ones is a form of model switching and has been observed with disparity-perspective conflict stimuli (Girshick & Banks, 2009; Knill, 2007b) and with auditory-visual conflict stimuli (Wallace et al., 2004).

Causal Inference

The development of the linear cue-combination model is based on the assumption that the individual cues are all estimating the same feature of the world (e.g., the depth or location of the same object). However, the observer may not know for sure that the cues derive from the same source in the world. The observer has to first infer whether the scene that gave rise to the sensory input consists of one or two sources (i.e., one or two objects) before determining whether the sources should be integrated. That is, the observer is faced with inferring the structure of the scene, not merely producing a single estimate.

Consider the problem of interpreting auditory and visual location cues. When presented with both a visual and auditory stimulus, an observer should take into account the possibility that the two signals come from different sources in the world. If they come from one source, it is sensible to integrate them. If they come from different sources, integration would be counterproductive. As we mentioned in the previous section, behavior consistent with model switching has been observed in auditory-visual integration experiments (Wallace et al., 2004). Specifically, when auditory and visual stimuli are presented in nearby locations, subjects' estimates of the auditory stimulus are pulled toward the visual stimulus (the ventriloquist effect). When they are presented far apart, the auditory and visual signals appear to be separate sources in the world and do not affect one another.

Recent work has approached this *causal-inference* problem using Bayesian inference of structural models (see Chapters 2, 3, 4, and 13). These models typically begin with a representation of the causal structure of the sensory input in the form of a Bayes net (Pearl, 1988). For example, Körding and colleagues (2007) used a structural model to analyze data on auditory-visual cue interactions in location judgments. The structural model (Fig. 1.4) is a probabilistic description of a generative model of the scene. According to this model, the generation of auditory and visual signals can be thought of as a two-step process. First, a weighted coin flip determines whether the scene consists of one cause (with probability p_{common}, left-hand branch) or separate causes for the auditory and visual stimuli (right-hand branch). If there is one cause, the location of that cause, x, is then determined (as a random sample from a prior distribution of locations), and the source at that location then gives rise independently to visual and auditory signals. If there are two causes, each has its own independently chosen location, giving rise to unrelated signals.

An observer has to invert the generative model and infer the locations of the visual and auditory sources (and whether they are one and the same). While there are numerous, mathematically equivalent ways to formulate the ideal observer, the formulation that is consistent with the others in this chapter is one in which an observer computes a posterior distribution on both the auditory and visual locations, x_a and x_v. The prior in this case is a mixture of a delta function along the diagonal in $x_a - x_v$ space (corresponding to situations in which the auditory and visual signals derive from the same source) and a broad distribution over the entire space (corresponding to situations in which the locations are independent). In this formulation, the prior distribution has broad tails, but the result is similar. If the two signals indicate locations near one another, the posterior is peaked at a point on the diagonal corresponding to a position between the two. If they are far apart, it peaks at the same point as the likelihood function. The joint likelihood function for the location of the visual and auditory sources can be described by the same sort of mixture model used in the earlier slant example (for further discussion of causal and mixture models, see Chapters 2, 3, 4, 12, and 13).

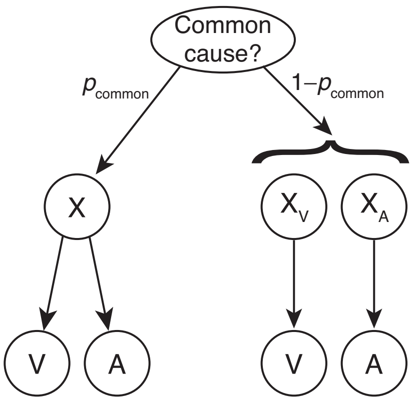

Figure 1.4 A causal-inference model of the ventriloquist effect. The stimulus either comes from a common source or from two independent sources (governed by probability p_{common}). If there is a common source, the auditory and visual cues both depend on that common source's location. If not, each cue depends upon an independent location.

Conclusions (Theory)

The previous theoretical discussion has a number of important take-home messages. First, Bayesian decision theory provides a completely general normative framework for cue integration. A linear approximation can characterize the average behavior of an optimal integrator in limited circumstances, but many realistic problems require the full machinery of Bayesian inference. This implies that the same framework can be used to build models of human performance, for example, by constructing and testing model priors that are incorporated into human perceptual mechanisms or by modeling the tasks human perceptual systems are designed to solve using appropriate gain functions. Second, the representational framework used to model specific problems depends critically on the structure of the information available and the observer's task. In the aforementioned examples, appropriate representational primitives include "average" cue estimates (in linear models), additive mixtures of likelihood functions, and graphical models. Finally, constructing normative models of cue integration serves to highlight the qualitative structure of specific problems. This implies that normative models can suggest the appropriate scientific questions that need to be answered to understand, at a computational level, how the brain solves specific problems. In some cases, this may mean that appropriate questions revolve around the weights that observers use to integrate cues. In others, they may revolve around the mixture components of the priors people use. In still others, they center on the causal structure assumed by observers in their models of the generative process that gives rise to sensory data.

THEORY MEETS DATA

Methodology

A variety of experimental techniques has been used to test theories of cue integration. Many researchers have used variants of the perturbation-analysis technique introduced by Young and colleagues (Landy et al., 1995; Maloney & Landy, 1989; Young, Landy, & Maloney, 1993) and later extended to intersensory cue combination (Ernst & Banks, 2002). Consider the combination of visual and haptic cues to size (Ernst & Banks, 2002). The visual stimuli are stereoscopic random-dot displays that depict a raised bar on a flat background (Fig. 1.5). The haptic stimuli are also a raised bar presented with force-feedback devices attached to the index finger and thumb. Four kinds of stimuli are used: visual-only (the stimulus is seen but not felt); haptic-only (felt but not seen); two-cue, consistent stimuli (seen and felt, and both cues depict the same size); and two-cue, inconsistent stimuli (in which the visual stimulus depicts one size x_v and the haptic stimulus indicates a different size x_h). Subjects are presented with two stimuli sequentially and indicate which was larger. For example, a subject is shown two visual-only stimuli that depict bars with heights x_v and $x_v + \Delta x_v$. The threshold value of Δx_v (the just-noticeable difference or JND) is used to estimate the underlying single-cue noise σ_v. An analogous single-cue experiment is used to estimate the haptic-cue noise σ_h. Interleaved with the visual-only and haptic-only trials, the two-cue stimuli are also presented. On such trials, subjects discriminate the perceived size of an inconsistent-cues stimulus in which the size depicted by haptics is perturbed from that depicted visually, $x_h = x_v + \Delta x$, as compared to a consistent-cues stimulus in which $x'_h = x'_v = x'$. The size x' of the consistent-cues stimulus is varied to find the point of subjective equality (PSE), that is, the pair of stimuli $((x_h, x_v)$ and $(x'_h, x'_v))$ that are perceived as being equal in size. Linear, weighted cue integration implies that x' is a linear function of Δx with slope w_h (the weight applied to the perturbed cue). The weight may be predicted independently from the estimates of the individual cue variances and Eq. 1.2.

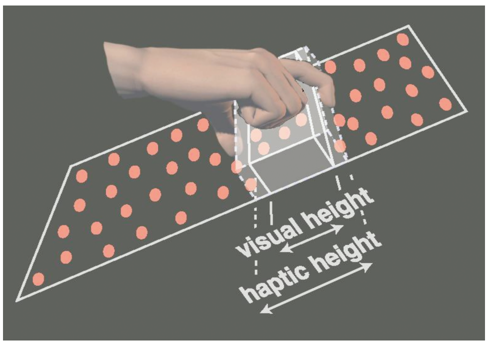

Figure 1.5 Multisensory stimulus used by Ernst and Banks (2002). A raised bar was presented visually as a random-dot stimulus with binocular disparity displaying a bar height x_v and, in inconsistent-cues stimuli, haptically with a height $x_h = x_v + \Delta x$.

There are a few issues with this method. First, one might argue that the artificial stimuli create cue conflicts that exceed those experienced under natural conditions and therefore that observers might use a different integration method than would be used in the natural environment. One can ask whether the results are similar across conflict sizes to determine whether this is a serious concern in a particular set of conditions.

Second, the method does not necessarily detect a situation in which the results have been affected by other unmodeled information such as another cue or a prior. Consider, for example, an experiment in which texture and motion cues to depth were manipulated, and the perturbation method was used to estimate the two cue weights (Young et al., 1993). The observers estimated depth by using texture and motion cues, but they may also have incorporated other cues such as blur and accommodation that specify flatness and/or a Bayesian prior favoring flatness (Watt, Akeley, Ernst, & Banks, 2005). As a result of using these other cues, observers should perceive all of the stimuli as flatter than specified by texture and motion, and therefore the texture and motion weights should sum to less than one ($w_t + w_m < 1$). This perceptual flattening would occur equally with both the consistent- and inconsistent-cues stimuli, and therefore would not affect points of subjective equality. In particular, the inconsistent-cues stimulus for which $\Delta x = 0$ is identical to the consistent-cues stimulus in which $x' = x_t = x_m$ and thus these two stimuli must be subjectively equivalent (except for measurement error). In this experimental design, the consistent-cues stimuli are used as a "yardstick" to measure the perceived depth of the inconsistent-cues stimulus. To uncover a bias in the percept, the yardstick must be qualitatively different from the inconsistent-cues stimulus. In the texture-motion case, for example, when the inconsistent-cues stimuli have reduced texture or motion reliability (by adding noise to texture shapes or velocities), but the consistent-cues stimuli do not have the added stimulus noise, the relative flattening of the noisy stimuli becomes apparent, and the separately measured weights sum to less than one (Young et al., 1993).

This experimental design is still useful if the observer incorporates a nonuniform prior into the computation of the environmental property of interest. For example, suppose the observer has a Gaussian prior on perceived depth centered on zero depth (i.e., a prior for flatness) and that the reliability r_i of each experimenter-manipulated cue i is unchanging across experimental conditions. The prior has the form of a probability distribution, but it is a fixed (nonstochastic) contributor to the computation. That is, as measured in stimulus units, the use of the prior will have no effect on the estimation of single-cue JNDs, nor on the estimation of relative cue weights. All percepts will be biased toward zero depth, but that will occur equally for the two discriminanda in each phase of the experiment and should not affect the results. Thus, the prior has no effect when the cue reliabilities $\{r_i\}$ are the same for the two stimuli being compared. The prior does have an effect when observers compare stimuli that differ in reliability: The stimulus with lower reliability displays a stronger bias toward the mean of the prior (Stocker & Simoncelli, 2006). This would occur in comparisons of single-cue to two-cue stimuli because cue integration typically increases reliability. Critically, it would also occur in conditions in which cue reliability depends on the resulting estimate so that reliability for each cue varies from trial to trial as, for example, occurs in the estimation of slant from texture (Knill, 1998).

Overview of Results

Many studies have supported optimal linear cue integration as a model of human perception for stimuli involving relatively small cue conflicts. By and large, these studies have confirmed the two main predictions of the model: With small cue conflicts, cue weights are proportional to cue reliability, and the reliability for stimuli with multiple cues is equal to the sum of individual cue reliabilities. Such studies have been carried out for combinations of visual cues to depth, slant, shape (Hillis et al., 2002; Hillis et al., 2004; Johnston, Cumming, & Landy, 1994; Knill & Saunders, 2003; Young et al., 1993), and location (Landy & Kojima, 2001). Multisensory studies have also been consistent with the model, including combinations of visual and haptic cues to size (Ernst & Banks, 2002; Gepshtein & Banks, 2003; Hillis et al., 2002) and visual and auditory cues to location (Alais & Burr, 2004). Some studies have found suboptimal choices of cue weights (Battaglia, Jacobs, & Aslin, 2003; Rosas et al., 2005; Rosas, Wichmann, & Wagemans, 2007).

Cue promotion is an issue for many cue-integration problems. Consider, for example, the visual estimation of depth. Stereo stimuli are misperceived in a manner that suggests that near viewing distances are overestimated, and far viewing distances underestimated, for the purposes of scaling depth from retinal disparity (Gogel, 1990; Johnston, 1991; Rogers & Bradshaw, 1995; Watt et al., 2005). This misscaling could be ameliorated by combining disparity and relative-motion cues to shape. However, the evidence for this particular cue interaction has been equivocal (Brenner &

Landy, 1999; Johnston et al., 1994; Landy & Brenner, 2001). It is important to note that people are essentially veridical at taking distance into account—little if any overestimation of near distances and little if any underestimation of far distances—when all cues to flatness are eliminated (Watt et al., 2005). In other words, failures to observe depth constancy may be due to the influence of unmodeled flatness cues such as blur and accommodation.

There is some evidence for robustness in intrasensory cue combination; that is, evidence that individual cues are down-weighted as they become too discrepant from estimates based on other cues. Most laboratory studies involve only two experimenter-manipulated cues with small conflicts, but some have looked at cue integration with large discrepancies. Bayesian estimation using a mixture prior can lead to robust behavior. For example, Knill (2007b; also see Chapter 9) has described two models for estimation of slant from texture, a more constrained and accurate model that assumes the texture is isotropic and a second that does not make this assumption. By assuming a mixture prior over scenes (between isotropic and nonisotropic surface textures), one can predict a smooth switch from the predictions of one model to the other as the presented surface texture becomes increasingly nonisotropic. Human performance appears to be consistent with the predictions of this mixture-prior model (Knill, 2007b). Recently, Girshick and Banks (2009) confirmed Knill's result that observers' percepts are intermediate between cue values for disparity and texture when the discrepancy between the two cues is small, and that percepts migrate toward one cue when the discrepancy is large. Like Knill, they found that the cue dictating the large-conflict percept was consistently disparity in some conditions. But unlike Knill, they also found other conditions in which the large-conflict percept was consistently dictated by texture. Girshick and Banks showed that their data were well predicted by a Bayesian model in which both the texture and disparity likelihoods had broader tails than Gaussians.

Empirical studies suggest that integration is impeded when the display indicates the two cues do not come from the same source. For example, optimal cue integration is found in combinations of visual and haptic cues to object size (Ernst & Banks, 2002), but if the haptic object is in a different location from the visual object, observers no longer integrate the two estimates (Gepshtein, Burge, Ernst, & Banks, 2005). Bayesian structural models have been successful at modeling phenomena like this, for example, in the ventriloquist effect (Körding et al., 2007).

Humans also appear to be optimal or nearly so in movement tasks involving experimenter-imposed rewards and penalties for movement outcome (Trommershäuser, Maloney, & Landy, 2003a, 2003b, 2008). Yet when analogous tasks are carried out involving visual estimation and integration of visual cues, many observers use suboptimal strategies (Landy, Goutcher, Trommershäuser, & Mamassian, 2007). As we argued earlier, observers should be more likely to approach optimal behavior in tasks that are important for survival. It seems reasonable that accurate visuomotor planning in risky environments is such a task and that the visual analog of these movement-planning tasks is not such a task. There remain many open questions in determining the limits of optimal behavior by humans in perceptual and visuomotor decision-making tasks with experimenter-imposed loss functions (i.e., decision making under risk).

ISSUES AND CONCERNS

Realism and Unmodeled Cues

Perceptual systems evolved to perform useful tasks in the natural environment. Accordingly, these systems were designed to make accurate estimates of environmental properties in settings in which many cues are present and large cue conflicts are rare. The lack of realism and the dearth of sensory cues in the laboratory give the experimenter greater stimulus control, but they may place the perceiver in situations for which the nervous system is ill suited and therefore may perform suboptimally.

Bülthoff (1991) describes an experiment in which the perceived depth of a monocularly viewed display is gauged by comparison with a stereo display. Depth from texture alone, and depth from shading alone were both underestimated, but when the two pictorial cues were combined, depth was approximately veridical. The depth values appeared to sum rather than average in the two-cue display. Bülthoff and Yuille (1991) interpreted this as an example of "strong fusion" of cues (in contrast with the "weak fusion" of weighted averaging). However, these were impoverished displays and contained other visual cues (blur, accommodation, etc.) that indicated the display was flat. Bayesian cue integration predicts that the addition of cues to a display will have the effect of reducing the weight given to these cues to flatness (because increasing the amount of information about depth increases the reliability of that information), resulting in greater perceived depth than that with either of the experimenter-controlled cues alone.

There is now clear evidence that display cues to flatness can provide a substantial contribution to perceived depth. Buckley and Frisby (1993) observed a striking effect that illustrates the importance of considering unmodeled cues in general and specifically the role of cues from the computer display itself. Their observers viewed raised ridges presented as real objects or as computer-graphic images. In one experiment, the stimuli were stereograms viewed on a computer display. Disparity- and texture-specified depths were varied independently and observers indicated the amount of perceived depth. The data revealed clear effects of both cues. Disparity dominated when the texture-specified depth was large, and texture dominated when the texture depth was small. In the framework of the linear cue-combination model, the disparity and texture weights changed depending on the texture-specified depth.

Buckley and Frisby next asked whether the results would differ if the stimuli were real objects. They constructed 3D ridges consisting of a textured card wrapped onto a wooden form. Disparity-specified depth was varied by using forms of different shapes. Texture-specified depth was varied by distorting the texture pattern on the card so that the projected pattern created the desired texture depth once the card was bent onto the form. The results differed dramatically: Now the disparity-specified depth dominated the percept. Buckley and Frisby speculated that unmodeled focus cues—blur and accommodation—played an important role in the difference between the computer display and real results.

We can quantify their argument by translating it into the framework of the linear model. There are three depth cues in their experiments: disparity, texture, and focus cues; focus cues specify flatness on the computer-display images and the true shape on the real objects.

The real-ridge experiment is easier to interpret, so we start there. In the linear Gaussian model, perceived depth is based on the sum of all available depth cues, each weighted according to its reliability:

$$\begin{aligned}\hat{d} &= w_d d_d + w_t d_t + w_f d_f \\ w_d &+ w_t + w_f = 1,\end{aligned} \tag{1.10}$$

where the subscripts refer to the cues of disparity, texture, and focus. The depth specified by the focus cues was equal to the depth specified by disparity: $d_f = d_d$. Thus, Eq. 1.10 becomes:

$$\hat{d} = (w_d + w_f) d_d + w_t d_t. \tag{1.11}$$

The texture cue d_t had a constant value k for each curve in their data figure (their Fig. 3); therefore,

$$\hat{d} = (w_d + w_f) d_d + (1 - w_d - w_f) k. \tag{1.12}$$

For this reason, when perceived depth is plotted against disparity-specified depth (d_d), the slope corresponds to the sum of the weights given to the disparity and focus cues: $w_d + w_f$. The experimentally observed slope was $\sim$0.95. Thus, the texture weight w_t was small in the real-ridge experiment.

In the computer-display experiment, focus cues always signaled a flat surface ($D_f = 0$); therefore,

$$\hat{d} = w_d d_d + w_t d_t = w_d d_d + (1 - w_d - w_f) k. \tag{1.13}$$

Thus, the slope of the data in their figures was an estimate of the disparity weight w_d. The slope was always lower in the computer-display data than in the real data, and this probably reflects the influence of focus cues.

Frisby, Buckley, and Horsman (1995) further explored the cause of increased reliance on disparity cues with real as opposed to computer-displayed stimuli. Observers viewed the real-ridge and computer-display stimuli through pinholes, which greatly increased depth of focus, thereby rendering all distances roughly equally well focused. The computer-display data were unaffected by the pinholes, but the real-ridge data were significantly affected: Those data became similar to the computer-display data. This result makes good sense. Viewing through a pinhole renders the blur in the retinal image similar for a wide range of distances and causes the eye to adopt a fixed focal distance. This causes no change in the signals arising from stereograms on a flat display, so the computer-display results were unaffected. The increased depth of focus does cause a change in the signals arising from real 3D objects—focus cues now signal flatness as they did with computer-displayed images—so the real ridge results became similar to the computer-display results. The work of Frisby and colleagues, therefore, demonstrates a clear effect of focus cues. Tangentially, their work shows that using pinholes is not an adequate method for eliminating the influence of focus cues.

Computer-displayed images are far and away the most frequent means of presenting visual stimuli in depth-perception research. Very frequently, the potential influence of unmodeled cues is not considered and so, as we have seen in the earlier analysis, the interpretation of empirical observations can be suspect. One worries that many findings in the depth-perception literature have been misinterpreted and therefore that theories constructed to explain those findings are working toward the wrong goal.

Watt et al. (2005) and Hoffman, Girshick, Akeley, and Banks (2008) explicitly examined the role of focus cues in depth perception. They found that differences in the distance specified by disparity and the physical distance of the stimuli (which determines blur and the stimulus to accommodation) had a systematic effect on perceived distance and therefore had a consistent and predictable effect on the perception of the 3D shape of an object.

Estimation of Uncertainty

The standard experimental procedure for testing optimality includes measurements of the reliability of individual cues ($\sigma \propto$ JND). For some cue-integration problems, such as the combination of auditory, haptic, and/or visual cues to spatial location, this is relatively straightforward. However, for intramodal cue integration, difficulties arise in isolating a cue. And as we argued earlier in the analysis of the Buckley and Frisby (1993) study, this can lead to errors in interpretation. Consider, for example, the estimation of surface slant from the visual cues of surface texture and disparity. It is easy to isolate the texture cue by viewing the stimulus monocularly. In contrast, it is impossible to produce disparity without surface markings from which disparities can be estimated.

The best one can do in these situations is to generate stimuli in which the information provided by one of the two cues is demonstrably so unreliable as to be useless for performing the psychophysical task. For the stereo-texture example, Hillis et al. (2004) and Knill and Saunders (2003) did this by using sparse random-dot textures when measuring slant-from-disparity thresholds. While a random-dot texture generates strong disparity cues to shape, the texture cue provided by perspective distortions (changes in dot density) is so unreliable that its contribution is likely to be small (Hillis et al., 2004; Knill & Saunders, 2003). Hillis and colleagues (2004) showed this by examining cases in the two-cue experiment in which the texture weight was nearly zero; in those cases, a texture signal was present, but the percepts were dictated by the disparity signal. They found that those two-cue JNDs were the same as the disparity-only JNDs. The close correspondence supports the assumption that the disparity-alone discrimination thresholds provided an estimate of the appropriate reliability for the two-cue experiment. Knill and Saunders (2003) took a different approach. In their study, the random-dot textures used in the stimuli to measure slant-from-stereo thresholds were projected from the same slant as indicated by disparities and thus contained perspective distortions that could in theory be used to judge slant. They showed, however, that when discriminating the slants of these stimuli viewed monocularly, subjects were at chance regardless of the pedestal slant or the difference in slant between two stimuli.

Single-cue discrimination experiments are typically used to estimate the uncertainty associated with individual cues. Suppose that, in addition to the uncertainty inherent in the individual cues (which we model as additive noise sources N_1 and N_2), there is uncertainty due to additional noise late in the process, which we term *decision noise* (N_d, Fig. 1.6). Suppose further that this noise corrupts the estimate after the cues are combined but prior to any decisions involving this estimate. If this is the case, the single-cue experiments will provide estimates of the sum of the cue uncertainty and the uncertainty created by the added late noise (e.g., $\sigma_1^2 + \sigma_d^2$). The optimal cue weights are still those defined by Eq. 1.2 (based on the individual cue reliabilities, e.g., $r_1 = 1/\sigma_1^2$). By using the results of the single-cue discrimination experiments, the experimenter will estimate the single-cue reliabilities as, for example, $r_1' = 1/(\sigma_1^2 + \sigma_d^2)$. The resulting predictions of optimal cue weights based on Eq. 1.2 will be biased toward equality

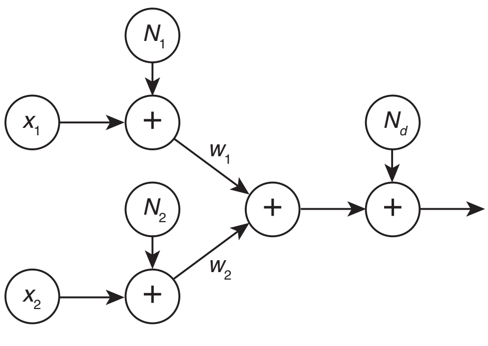

Figure 1.6 Illustration of a model that incorporates both early, single-cue noise terms (N_1 and N_2) as well as a late, postcombination, decision-noise term (N_d).

(weights of 0.5 for each cue in the slant/disparity experiment). Fortunately, decision noise affects PSEs and JNDs in a predictable way, and so one can gauge the degree to which decision noise affected the measurements. Both Knill and Saunders (2003) and Hillis and colleagues (2004) concluded the effects of decision noise were negligible.

Another important assumption of these methods is that subjects use the same information in the stimulus to make the single-cue judgments as the multiple-cue judgments. This can be a particular concern when the single-cue stimuli do not generate a compelling percept of whatever one is studying. In the depth domain, one has to be concerned about interpreting thresholds derived from monocular displays with limited depth information, particularly when presented on computer displays. One way around this is to use subjects' ability to discriminate changes in the sensory features (e.g., texture compression) that are the source of information in a cue and use an ideal-observer model to map the measured sensory uncertainty onto the consequent uncertainty in perceptual judgments from that cue. Good examples of this outside the cue-integration literature are work on how motion acuity limits structure-from-motion judgments (Eagle & Blake, 1995) and heading judgments form optic flow (Crowell & Banks, 1996), and how disparity acuity limits judgments of surface orientation at different eccentricities and distances from the horopter (Greenwald & Knill, 2009). In a visuomotor context, Saunders and Knill (Saunders & Knill, 2004) used estimates of position and motion acuity to parameterize an optimal feedback-control model and showed that the resulting model did a good job at predicting how subjects integrate position and motion information about the moving hand to control pointing movements. Knill (2007b) used psychophysical measures of aspect-ratio discrimination thresholds to parameterize a Bayesian model for integrating figural compression and disparity cues to slant, but he did not test optimality with the model.

Finally, it is important to note that the implications of optimal integration differ for displays in which a cue is missing (e.g., the focus cue when viewing a display through a pinhole) and displays in which the cue is present but is fixed. When two cues are present rather than just one, both contribute to perceived depth, but both of their uncertainties contribute to the uncertainty of the result. For example, Eq. 1.3 implies that the JNDs for depth should satisfy

$$\frac{1}{\mathrm{JND}_c^2} = \frac{1}{\mathrm{JND}_1^2} + \frac{1}{\mathrm{JND}_2^2}, \tag{1.14}$$

where JND_c is the threshold for discriminating consistent-cues, two-cue stimuli, and JND_1 and JND_2 are the individual uncertainties of the two cues (proportional to the standard deviation of estimates based on each cue), typically measured by isolating each cue.

Suppose an experimenter fails to isolate each cue when measuring the single-cue thresholds and, instead, single-cue thresholds are measured with both cues present in the stimulus, with one cue held fixed while discrimination threshold is measured for the other, variable cue. For such an experiment, the relationship between thresholds measured for each cue is different. Subjects' JNDs should instead satisfy[1]

$$\frac{1}{\mathrm{JND}_c} = \frac{1}{\mathrm{JND}'_1} + \frac{1}{\mathrm{JND}'_2}, \tag{1.15}$$

where JND'_1 and JND'_2 are the JNDs measured for each cue in the presence of the other, fixed cue. In fact, this relationship applies regardless of the weights that subjects give to the cues, optimal or not. It only depends on the linearity assumption.

Bradshaw and Rogers (1996) ran such an experiment, measuring the JNDs of the two constituent cues in the presence of the other cue, but the depth indicated by the second cue was fixed at zero (flat). That is, both cues' noise sources were involved. Bradshaw and Rogers interpreted the resulting improvement in JND for two-cue displays as indicative of a nonlinear interaction of the cues. But, their data were, in fact, reasonably consistent with the predictions of Eq. 1.15, that is, with the predictions of linear cue combination (optimal or not).

[1]To see this, assume that JNDs are defined to be the standard deviation of the noise that must be matched by the change in perceived depth to reach threshold. Thus, in the normal perturbation experiment in which single-cue JNDs are measured in isolation, $\mathrm{JND}_1 = \sigma_1$, $\mathrm{JND}_2 = \sigma_2$, and by Eq. 1.3, $\mathrm{JND}_c = \sigma_c = 1/\left[\left(1/\sigma_1^2\right) + \left(1/\sigma_2^2\right)\right]^{1/2}$ from which Eq. 1.14 follows. In the Bradshaw and Rogers (1996) experiment, when only cue 1 is manipulated and cue 2 is fixed, the weighted combination of the cues must overcome the combined noise, so that $w_1\Delta_1 + w_2\Delta_2 = w_1\mathrm{JND}'_1 + w_2(0) = \sigma_c$ and similarly for JND_2, where Δ_i is the difference in cue value for cue i between the two discriminanda at threshold. Because the weights are assumed to sum to 1, Eq. 1.15 follows.

Estimator Bias

In introducing the cue-combination models in the theory section of this chapter, we made the common assumption that perceptual estimates derived from different cues are unbiased; that is, we assumed that for any given value of a physical stimulus variable (e.g., depth or slant), the average perceptual estimate of that variable is equal to the true value. This assumption is generally incorporated in descriptions of optimal models because it simplifies exposition: Disregarding bias allows one to focus only on minimizing variance as an optimality criterion. However, it seems to us that it is generally impossible to determine whether sensory estimates derived from different cues prior to integration are unbiased relative to ground truth, largely because we only have access to the outputs of systems that operate on perceptual estimates (decision or motor processes) that may themselves introduce unknown biases. It is possible, however, to determine whether they are internally consistent, that is, whether their estimates agree with one another on average.

If the estimators in the optimal combination model (Eq. 1.1) are not internally calibrated, problems may arise. Consider presenting a 3D stimulus with a slant of 0° to the eye and hand. Let us say that vision and touch are equally reliable, but that vision is biased by 20° because the person is wearing spectacles. The internal inconsistency introduces a serious problem: If the stimulus is seen but not felt, its perceived slant will be 20°. If it is felt but not seen, the percept will be 0°. If it seen and felt, the percept will be 10° (assuming equal cue weights in Eq. 1.1). The internal inconsistency of the estimators has undermined one of the great achievements of perception: the ability to perceive a given environmental property as constant despite changes in the proximal stimuli used to estimate the property. Thus, it is clearly important for sensory estimators to maintain internal consistency with respect to one another (see Chapter 12).

There is a rich literature on how sensory estimators maintain internal consistency and external accuracy (Burian, 1943; Miles, 1948; Morrison, 1972; Ogle, 1950). The problem is referred to as sensory recalibration. Adams, Banks, and van Ee (2001) studied recalibration of estimates of slant from texture and slant from disparity by exposing people to a horizontally magnifying lens in front of one eye. The lens was worn continuously for 6 days. People were tested before, during, and after wearing the lens with three types of stimuli: slanted planes specified by texture and viewed monocularly, slanted planes specified by disparity and viewed binocularly, and two-cue, disparity-texture stimuli viewed binocularly. The introduction of the lens caused a change in the disparities measured at the two eyes such that a binocularly viewed plane that was previously perceived as frontoparallel was now perceived as slanted by $\sim 10°$. The apparent slant of monocularly viewed planes did not change. Thus, the introduction of the lens had created a conflict between the perceived slants for disparity- and texture-based stimuli even when they specified the same physical slant. Over the six days, observers adapted until frontoparallel planes, whether they were defined by texture alone, disparity alone, or both, were again perceived as frontoparallel. When the lens was removed, everyone experienced a negative after-effect: A disparity-defined frontoparallel plane appeared slanted in the opposite direction (and a texture-defined plane did not). The negative after-effect also went away in a few days as the observers adapted back to the original no-lens condition. These observations clearly show that the visual system maintains internal consistency between the perceived slants of disparity and texture stimuli even when the two cues are put into large conflict by optical manipulation. Because they maintain calibration with respect to one another, the visual system can achieve greater accuracy and precision by appropriate cue combination as described earlier in the chapter.

Girshick and Banks (2009) also obtained persuasive data that disparity and texture estimators maintain calibration relative to one another. They measured the slants of single-cue stimuli that matched the apparent slant of two-cue stimuli. Specifically, they measured the slant of a disparity-only stimulus that matched the perceived slant of a two-cue, disparity-texture conflict stimulus and they measured the slant of a texture-only stimulus that matched the perceived slant of the same disparity-texture conflict stimulus. The disparity-only stimulus was a sparse random-dot textured plane viewed binocularly and the texture-only stimulus was a Voronoi-textured plane viewed monocularly. On each trial, one interval contained a two-cue stimulus, and the other contained one of the two single-cue stimuli. Observers indicated the one containing the greater perceived slant. No feedback was provided. The slant of the single-cue stimulus was varied according to a staircase procedure to find the value that appeared the same as the two-cue stimulus. Figure 1.7 shows the results. Each data point represents the disparity- and texture-specified slants that yielded the same perceived slant as a particular two-cue, disparity-texture stimulus. Clearly, the disparity- and texture-specified slants were highly correlated and one was not biased relative to the other, showing that the disparity and texture estimators were calibrated relative to one another.

Variable Cue Weights

This discussion brings up one last point: Cue weights need not be constant, independent of the value of the parameter being estimated. Effective cue reliability can vary with conditions, including with changes in the parameter itself. This phenomenon has been observed, for example, with estimation of surface slant.

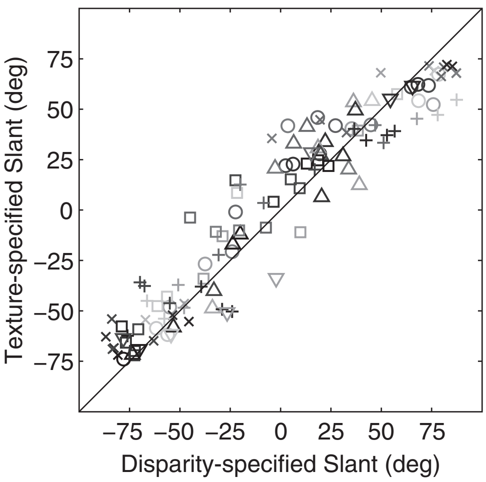

Figure 1.7 The slants of single-cue stimuli that matched the apparent slant of two-cue stimuli. The data are from Experiment 2 of Girshick and Banks (2009). On each trial, two stimuli were presented in sequence. One was a single-cue stimulus: either texture only or disparity only. The other was a two-cue stimulus: texture plus disparity. The two-cue, texture-disparity stimulus had various amounts of conflict between the slants specified by the two cues; some conflicts were as large as 75^o. After each trial, observers indicated which of the two stimuli had the greater perceived slant. Each data point represents the disparity- and texture-specified slants that yielded the same perceived slant as a particular two-cue, disparity-texture stimulus. Different symbols represent different conflicts between the texture- and disparity-specified slants and different observers. (Adapted from Girshick & Banks, 2009.)

The JND for discrimination of surface slant from texture varies substantially with base slant (Knill, 1998) and the JND for slant from disparity varies with base slant as well (Hillis et al., 2004). Of course, JNDs can also vary with other stimulus parameters. For example, the reliability of slant estimates based on disparity varies with viewing distance (Hillis et al., 2004). As a result, one predicts changes in the relative weights of texture and disparity with changes in base slant (Hillis et al., 2004; Knill & Saunders, 2003) and distance (Hillis et al., 2004). For a large, slanted surface, one predicts changes in cue weights for different locations along the surface itself (Hillis et al., 2004).

The interesting point is that the optimal cue weights can change rapidly, from moment to moment or from location to location, sensitive to local conditions. These optimal weight settings are in response to changes in estimated cue reliability. This raises the question of how human observers estimate and represent cue reliability. One suggestion is that a neural population code can simultaneously encode both the estimate and its associated uncertainty (see Chapter 21 and Beck, Ma, Latham, & Pouget, 2007; Ma et al., 2006).

Simulation of the Observer

In the development of the linear model for a Bayesian observer, we pointed out that observers make measurements for each cue, then form the product of likelihood functions derived from these measurements and the prior distribution (Eq. 1.7). This results in the linear rule for Gaussian distributions. One can prove this by multiplying the likelihood functions corresponding to the expected cue measurements and the prior. But real observers do not have access to the expected cue measurement on any given trial. Rather, they have samples and must derive (and multiply) likelihood functions based on those samples. For symmetric likelihood functions like Gaussians, the predictions do not differ from those based on the expected measurements. However, for non-Gaussian likelihood functions or priors (e.g., mixture priors), one is forced to consider the variability of the cue estimates in formulating predictions.

In formulating Bayesian models incorporating both likelihoods and priors, one must confront the issue of where the prior comes from and how to estimate it. Three different approaches have been used in recent years. In one line of research, natural-image statistics are gathered and used to estimate a given prior distribution, and then human behavior is compared to the performance of a Bayesian ideal observer using that prior (see Chapter 11 and Elder & Goldberg, 2002; Fowlkes, Martin, & Malik, 2007; Geisler, Perry, Super, & Gallogly, 2001). An alternative approach is to ask what prior distribution is consistent with observers' behavior independent of whether it accurately reflects the statistics of the environment. The technique of Stocker and Simoncelli (2006) provides such an approach, by taking advantage of the differential effect of a prior on stimuli that differ in reliability. Others have fit parametric models of the prior distribution to psychophysical data (Knill, 2007b; Mamassian & Landy, 2001). Finally, Knill (2007a) has examined how the visual system adapts its internal prior to the current statistics of the environment.

OPEN QUESTIONS

The research on cue integration has been wide ranging, and it has led to interesting data and many successful models. Nevertheless, there is plenty of room for further progress. Here is a short list of interesting open questions:

1. How is cue reliability estimated and represented in the nervous system? Observers seem to be able to estimate cue reliability in novel environments, so this is presumably not learned in specific environments and then applied when one of the learned environments is encountered again. Clearly, cue reliability depends on many factors, and thus estimation of reliability is itself a problem of cue integration.
2. Are there general methods the perceptual system uses to determine when cues should be integrated and when, instead, they should be kept separate and attributed to different environmental causes? This problem can be cast as a statistical problem of causal inference.
3. How optimal is cue integration with respect to the information that is available in the environment? Scientists tend to classify environmental properties into distinct categories. The classical list of depth cues is an example. There are surely many other sources of depth information that observers' brains know about, but scientists' brains do not. A rigorous analysis of the linkages between information in natural scenes and human perceptual behavior should reveal previously unappreciated cues.
4. When human cue integration is demonstrably suboptimal, what design considerations does the suboptimality reflect? Are there examples in which the task and required mechanisms have been characterized correctly and the task is undeniably important to the organism, yet perception is nonetheless suboptimal?
5. There are now many examples in which Bayesian priors are invoked to explain aspects of human perception: a prior for slowness, light from above, shape convexity, and many more. Do these priors actually correspond to the probability distributions encountered in the natural environment?

REFERENCES

Adams, W. J., Banks, M. S., & van Ee, R. (2001). Adaptation to three-dimensional distortions in human vision. *Nature Neuroscience*, *4*, 1063–1064.

Adolph, K. E. (1997). Learning in the development of infant locomotion. *Monographs of the Society for Research in Child Development*, *62*, 1–158.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *20*, 1391–1397.

Beck, J., Ma, W. J., Latham, P. E., & Pouget, A. (2007). Probabilistic population codes and the exponential family of distributions. *Progress in Brain Research*, *165*, 509–519.

Bradshaw, M. F., & Rogers, B. J. (1996). The interaction of binocular disparity and motion parallax in the computation of depth. *Vision Research*, *36*, 3457–3468.

Brenner, E., & Landy, M. S. (1999). Interaction between the perceived shape of two objects. *Vision Research*, *39*, 3834–3848.

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 M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 305–330). Cambridge, MA: MIT Press.

Bülthoff, H. H., & Yuille, A. L. (1991). Shape-from-X: Psychophysics and computation. In P. S. Schencker (Ed.), *Sensor fusion III: 3-D perception and recognition, Proceedings of the SPIE* (Vol. 1383, pp. 235–246). Bellingham, WA: SPIE.

Burian, H. M. (1943). Influence of prolonged wearing of meridional size lenses on spatial localization. *Archives of Ophthalmology*, *30*, 645–666.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, *4*(Suppl.), 102–118.

Crowell, J. A., & Banks, M. S. (1996). Ideal observer for heading judgments. *Vision Research*, *36*, 471–490.

Domini, F., & Braunstein, M. L. (1998). Recovery of 3-D structure from motion is neither Euclidean nor affine. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1273–1295.

Eagle, R. A., & Blake, A. (1995). Two-dimensional constraints on three-dimensional structure from motion tasks. *Vision Research*, *35*, 2927–2941.

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, *2*, 324–353.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure-ground cues are valid for natural images. *Journal of Vision*, *7*(8):2, 1–9.

Frisby, J. P., Buckley, D., & Horsman, J. M. (1995). Integration of stereo, texture, and outline cues during pinhole viewing of real ridge-shaped objects and stereograms of ridges. *Perception*, *24*, 181–198.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711–724.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology*, *13*, 483–488.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, *5*, 1013–1023.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton-Mifflin.

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *Journal of Vision*, *9*(9):8, 1–20.

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

Greenwald, H. S., & Knill, D. C. (2009). Cue integration outside central fixation: A study of grasping in depth. *Journal of Vision*, *9*(2):11, 1–16.

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

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, *298*, 1627–1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, *4*, 967–992.

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, *8*(3):33, 1–30.

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

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.

Knill, D. C. (1998). Discrimination of planar surface slant from texture: Human and ideal observers compared. *Vision Research*, *38*, 1683–1711.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research*, *43*, 831–854.

Knill, D. C. (2005). Reaching for visual cues to depth: The brain combines depth cues differently for motor control and perception. *Journal of Vision*, *5*, 103–115.

Knill, D. C. (2007a). Learning Bayesian priors for depth perception. *Journal of Vision*, *7*(8):13, 1–20.

Knill, D. C. (2007b). Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision*, *7*(7):5, 1–24.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*, 2539–2558.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, *2*, e943.

Landy, M. S., & Brenner, E. (2001). Motion-disparity interaction and the scaling of stereoscopic disparity. In L. R. Harris & M. R. M. Jenkin (Eds.), *Vision and attention* (pp. 129–151). New York, NY: Springer-Verlag.

Landy, M. S., Goutcher, R., Trommershäuser, J., & Mamassian, P. (2007). Visual estimation under risk. *Journal of Vision*, *7*(6):4, 1–15.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *18*, 2307–2320.

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

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

Maloney, L. T. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world: Psychological and philosophical issues in perception* (pp. 145–189). New York, NY: Wiley.

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. In W. A. Pearlman (Ed.), *Visual communications and image processing IV. Proceedings of the SPIE* (Vol. 1199, pp. 1154–1163). Bellingham, WA: SPIE.

Mamassian, P., & Landy, M. S. (2001). Interaction of visual prior constraints. *Vision Research*, *41*, 2653–2668.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Miles, P. W. (1948). A comparison of aniseikonic test instruments and prolonged induction of artificial aniseikonia. *American Journal of Ophthalmology*, *31*, 687–696.

Morrison, L. (1972). Further studies on the adaptation to artificially-induced aniseikonia. *British Journal of Physiological Optics*, *27*, 84–101.

Ogle, K. N. (1950). *Researches in binocular vision.* Philadelphia, PA: Saunders.

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*, 2451–2468.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco, CA: Morgan Kaufmann.

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

Rogers, B. J., & Bradshaw, M. F. (1995). Disparity scaling and the perception of frontoparallel surfaces. *Perception*, *24*, 155–179.

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: Reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *22*, 801–809.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, *7*(6):3, 1–21.

Saunders, J. A., & Knill, D. C. (2001). Perception of 3D surface orientation from skew symmetry. *Vision Research*, *41*, 3163–3183.

Saunders, J. A., & Knill, D. C. (2004). Visual feedback control of hand movements. *Journal of Neuroscience*, *24*, 3223–3234.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*, 578–585.

Tassinari, H., Hudson, T. E., & Landy, M. S. (2006). Combining priors and noisy visual cues in a rapid pointing task. *Journal of Neuroscience*, *26*, 10154–10163.

Todd, J. T. (2004). The visual perception of 3D shape. *Trends in Cognitive Sciences*, *8*, 115–121.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003a). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *20*, 1419–1433.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003b). Statistical decision theory and trade-offs in the control of motor response. *Spatial Vision*, *16*, 255–275.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*, 291–297.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*, 252–258.

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, *5*, 834–862.

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.