International Journal of Computer Vision 40(1), 73–91, 2000

How Optimal Depth Cue Integration Depends on the Task

PAUL R. SCHRATER AND DANIEL KERSTEN
Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 E. River Dr., Minneapolis, MN 55455, USA

Received November 1, 1999; Revised July 12, 2000

Abstract. Bayesian parameter estimation can be used to generate statistically optimal solutions to the problem of cue integration. However, the complexity and dimensionality of these solutions is frequently prohibitive. In this paper, we show how the complexity and performance characteristics of the optimal estimator for a task depend strongly on the detailed formulation of the task, including the choice of representation for the scene variables. In particular, some representations lead to simpler inference algorithms than others. We illustrate the problem of cue integration for the perception of depth from two highly disparate cues, cast shadow position and image size, and show how the complexity and performance of the depth estimators depends on the specific representation (choice) of depth parameter. From the analysis we predict human performance on a simple depth discrimination task from the optimal cue integration in each depth representation. We find that the cue-integration strategy used by human subjects can be described as near-optimal using a particular choice of depth representation.

Keywords: task, data fusion, cue integration, optimal estimation, Bayes nets, Bayesian inference, depth estimation, depth from shadows, depth from image size

1. Introduction

Cue integration (data fusion) is the process by which we combine different kinds of image measurements (e.g. edges, optic flow, color, etc.) to estimate quantities of interest in the scene (e.g. shape or depth or reflectance). Human visual perception uses well over a dozen different cues to depth, including binocular and motion parallax, pictorial cues, and the so-called physiological or proprioceptive cues (cf. (Cutting and Vishton, 1996)). For this many cues, cue integration becomes a complex and potentially computationally intractable problem.

Given a variable to be estimated like depth, cue integration is simplest when each cue provides a unique and mutually compatible estimate of depth. However, such simple cases are rare, and cue integration typically involves solving a number of problems. For instance, different cues provide qualitatively different kinds of information about depth. Some cues provide information about depth in different coordinate frames. Almost all cues are ambiguous (e.g. unique up to an affine transformation) unless other interacting scene variables are known. And for a given task, some of the cues are more reliable than others.

One solution to all these problems is to do Bayesian cue integration, which uses all the information at hand, both measured and due to prior knowledge, to form the statistically optimal estimator. However, while optimal strategies best integrate the information, there is no guarantee that the resulting strategies are simple enough to practically implement.

In this paper, we show how the complexity and performance characteristics of the optimal estimator for a task depend strongly on the detailed formulation of the task. This dependence exceeds the level of detail generally provided by an informal description of the task. For example, an informal description of a task might be to discriminate the depths of two objects. We show that the optimal estimator will depend on exactly what

frame of reference we use. For instance, we can compute the distance of one object relative to the other, or relative to the observer, or relative to some fixed point in world coordinates. The purpose of this paper is to describe this dependence and its applications to understanding cue integration strategies used by the human observer.

In the next section we describe the problems of cue integration in greater detail.

1.1. Cue Integration

Optimal cue integration strategies use all of the available information to provide the statistically best estimates of the scene variables of interest. Because of this, in the presence of multiple cues or scene variables these estimators suffer from problems of complexity and dimensionality. Here, *estimation complexity* is used to indicate the number and difficulty of the computations that result from having to consider variables jointly in an estimate (e.g. evaluating functions of N variables, searching for maxima over N variables, etc.). Optimal cue integration can quickly become intractable as the number of cues and variables grows.

The most common solution to this problem is to try to make separate estimates of scene variables from each relevant cue. Because implementations of this solution can be described as forming modules for computing each scene variable estimate, systems that make separate estimates of scene variables are called modular. Fully modular systems result when each scene variable has a separate estimate from each image measurement that is relevant (see Fig. 1).

The potential problems for cue integration created by an unjustified commitment to a particular modular structure have been addressed elsewhere (Clark and Yuille, 1990; Landy et al., 1995). Briefly, the problems can be summarized as: unjustified prior assumptions on related scene variables, incompatible estimates of scene variables, and difficulties combining different estimates (the fusion problem). To illustrate, consider the fusion problem. Given that we have several estimates for an unknown quantity x, what do we do with them? In order of simplicity, we could: discard the worst estimates as outliers; take a linear combination (often termed *weak fusion*); take linear combinations modified by prior knowledge or other constraints; or, we could cook up more complicated functions of the estimates potentially incorporating prior knowledge or other constraints.

Under particular conditions each of these fusion methods is optimal, but many situations arise in which it is sub-optimal to form separate estimates at all. An important instance is when there are image measurements that depend on several scene variables. In this case, optimal estimation may need to consider all the image measurements and scene variables together or *cooperatively*. For instance, any image measurement can be created by different combinations of surface geometry and reflectance, hence any estimate of surface geometry must take into account the reflectance, either through jointly estimating the quantities or by assuming prior values for reflectance (Knill and Kersten, 1991). Modular schemes typically assume prior values, but without a statistical justification.

In contrast, Bayesian (optimal) inference insures consistent inferences and optimal integration of cues based on the confidence in the estimates, using fusion rules that fall out of the inference. In addition, it affords the ability to specify precise and consistent prior information that can frequently be estimated offline.

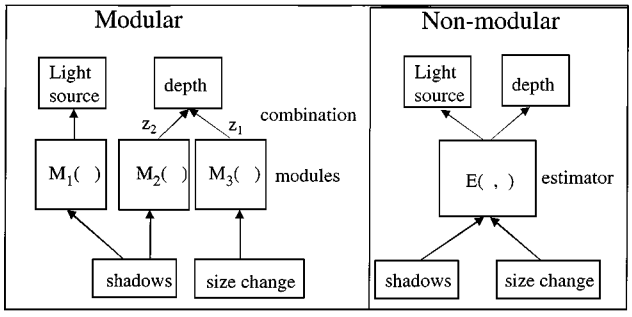

Figure 1. Modular vs. non-modular visual systems. Image measurements of a cast shadow position and the image size of an object are related to the depth. The non-modular system for cue integration has been called "strong fusion".

Clearly, the advantages of using optimal inference are manifold. In this paper, we will discuss how the computations for Bayesian cue integration can be simplified by restricting a visual system to performing specific tasks. Further, we show that the degree of modularity (and resulting estimation complexity) obtained depends on both the task and more surprisingly on the specific representation of the scene variables. These effects occur because changes in task and representation can strongly modify the statistical dependence between variables.

An outline of the remainder of the paper is as follows. We briefly review optimal inference. We will then show how a specific task affects the statistical dependence between variables, and analyze the conditions under which optimal inference systems can still be modular. We then show how the modularity of the inference is a function of the representation chosen for the scene variables, and how the representation can have an sizable impact on the performance of the estimator. We then illustrate these results with an extended example of a cue integration problem in which the optimal estimator undergoes strong changes in modularity and performance with changes in the depth representation. These results form theoretical predictions that are compared to a human psychophysical cue integration task.

2. Optimal Inference and Task Dependency

We begin with a brief exposition of Bayesian (optimal) inference.

Probabilistic approaches to scene estimation require the specification of $p(S, I)$, the joint probability distribution on the vector of scene attribute variables S and image measurement variables I. This joint distribution contains all the required information for making optimal inferences and doing optimal encoding of the image information. For the problem of inferring scene descriptions from image measurements, we use Bayes' rule to write the posterior probability as:

$$p(S \mid I) = \frac{p(S, I)}{p(I)} = \frac{p(I \mid S)p(S)}{\int_S p(I \mid S)p(S)\, dS} \qquad (1)$$

Optimal inference uses $p(S \mid I)$, but the form of the estimators for S depends on the task.

2.1. *Defining Tasks*

Intuitively, tasks are the actions that agents perform within particular contexts. Each task implicitly or explicitly places a set of demands on a visual system through the visual inference of scene attributes required for successful completion of the task. Similarly, the cost associated with a failure to complete the task induces a cost function on successful visual inference. Thus, the first component of a task based visual system is a specification of the cost of inaccurate estimates of scene properties. For instance, for a reaching task, the shape and position of the object relative to the observer are important, but the object's spectral reflectance (color) typically is not. The second component is the specification of the context in which the task is performed. In terms of decision theory, the context can be modeled by the prior term $p(S)$. For example, a reaching agent can frequently assume that objects are stationary.

Bayesian decision theory provides a precise language to model the costs of errors determined by the choice of visual task (Yuille and Bulthoff, 1996; Brainard and Freeman, 1997). The *risk* $R(\hat{S}; I)$ of guessing $\hat{S}$ when the image measurement is I is defined as the expected loss:

$$R(\hat{S}; I) = \int_S L(\hat{S}, S)p(S \mid I)\, dS, \qquad (2)$$

with respect to the posterior probability, $p(S \mid I)$. The best interpretation of the image can then be made by finding the $\hat{S}$ which minimizes the risk function. One possible loss function is a delta function $-\delta(\hat{S} - S)$. In this case the risk becomes $R(\hat{S}; I) = -p(\hat{S} \mid I)$, and then the best strategy is to pick the most likely interpretation. This is called *Maximum a posteriori estimation* (MAP). A second kind of loss function assumes that costs are constant over all guesses of a variable. This is equivalent to marginalization of the posterior with respect to that variable.

2.2. *How a Task Determines the Inference Computations*

In a Bayesian decision theory framework, tasks specify two things: a cost function and a prior distribution that models knowledge about scene attributes in the context of the task. We will consider how each affects the inference computation in turn.

2.2.1. Effect of the Cost Function. Scene variables can be classified as being relevant or irrelevant to the specific task at hand. An relevant variable S_r needs to be estimated precisely, and an irrelevant one S_{ir} imprecisely or not at all. As we will show below, the irrelevant variables can further be broken down into those that do or do not influence the estimate of the relevant variables. Imagine we have three scene variables, S_r, S_g, and S_{ind} that determine image formation. S_r is the relevant scene variable to estimate precisely for the task at hand. S_g influences our estimate of S_r, and S_{ind} does not. We will now show how this classification emerges from the cost function and the independence structure of the posterior distribution.

The costs assigned to incorrect estimates are set by the task. In general, narrow loss functions will be assigned to the scene variables that need to be estimated, and broad or nearly constant loss functions will be assigned to the remaining scene variables. Thus the cost function naturally divides the scene variables into two groups, relevant S_r, and irrelevant S_{ir}. Formally, $\{S_{ir}\} = \{S \mid L(\hat{S}, S) = c\}$ and $\{S_r\} = \{S \mid L(\hat{S}, S) \neq c\}$.

The optimal decision involves computing the expected loss with respect to the posterior. As noted above, the constant loss function is equivalent to marginalization of $p(S \mid I)$ over all of the irrelevant variables. If we perform this marginalization offline, then we can base our estimator only on the reduced posterior $p(S_r \mid I) = \int_{S_{ir}} p(S_r, S_{ir} \mid I) dS_{ir}$.

In this marginalization integral, some of the irrelevant variables will have an effect on shaping the reduced posterior, while others will have no effect. This impact will depend on the degree of dependence between the relevant and irrelevant variables. In particular, if some irrelevant variable is independent of S_r, then the joint distribution factors, and hence the relevant posterior factors out of the marginalization integral i.e.,

$$p(S \mid I) = p(S_r, S_{ir} \mid I) = p(S_r, S_g \mid I) p(S_{ind} \mid I)$$

so that

$$\int_{S_{ir}} p(S_r, S_{ir} \mid I)\, dS_{ir} = \int_{S_g} p(S_r, S_g \mid I)\, dS_g \times \int_{S_{ind}} p(S_{ind} \mid I)\, dS_{ind}$$

where $\int_{S_{ind}} p(S_{ind} \mid I)\, dS_{ind} = 1$. Because of this factorization, scene variables that are independent of S_r can be safely ignored. In addition, image measurements that depend solely on these variables may also be ignored. What we have done is to factor the irrelevant variables into two groups, those that won't influence the inference due to independence, and those scene variables S_g, that are not estimated but nevertheless affect the inference through marginalization. These variables are typically termed "nuisance" variables or more recently "generic" variables (Freeman, 1994), from which stems the 'g' subscript. There are two extreme cases of how the nuisance variables affect the inference depending on whether the prior distribution approaches a delta function (i.e. the nuisance variable has a known value) or it approaches a uniform distribution. We treat both these cases in turn.

2.2.2. Effect of Known Nuisance Variables. In most cases, the nature of a task supplies prior information that is equivalent to fixing or restricting the values of some of the scene variables. For instance, if an observer's task is to identify objects on an assembly line, then a number of relevant variables are typically fixed, such as the viewing direction and distance, and the light source distance and direction. Restricting the task domain to rigid bodies allow the observer to treat object geometry as time invariant. Note that most constraints used to regularize vision problems can be expressed as fixing a set of scene variables. For instance, in a world of polynomial surfaces, the constraint that the task only involves flat surfaces in the world, can be rephrased as all non-linear polynomial coefficients are fixed at zero.

If the prior probability distribution on a nuisance variable approaches a delta function $\delta(S_g - S_{g_0})$, marginalizing across the variable is equivalent to conditioning on the fixed value of that variable: $\int p(S \mid S_g) p(S_g)\, dS_g = \int p(S \mid S_g) \delta(S_g - S_{g_0})\, dS_g = p(S \mid S_{g_0})$.

Conditioning can have a substantial impact on the statistical dependence between the remaining variables, which can sometimes increase, and sometimes decrease or eliminate the dependence.[1] As an example of the latter, consider a simple scene consisting of a light source, a planar background, and a single object that casts a shadow on the background (see Section 4). Assume the observer of this scene is asked to estimate the light source direction in one of two conditions: the depth of the object is unknown, or the depth of the object is known (see Fig. 2). Further assume the observer has two cues available: the cast shadow position,

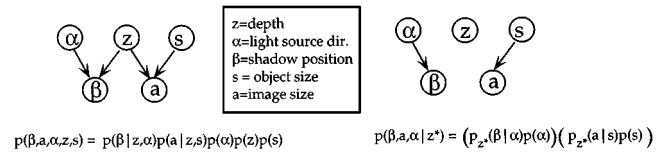

Figure 2. Example showing how the set of required variables varies as a function of the task specifications. Left: Depth uncertainty. Right: Depth specified ($z = z^*$).

and the image size of the object. Depending on the condition, either both cues or only the shadow cue are relevant. When the observer knows the depth of the object, the image size cue is irrelevant to estimating light source direction, because it only provides information about depth. In the presence of depth uncertainty, both cues are relevant, because the image size cue provides information about depth that can be used to disambiguate the shadow cue, and because a given shadow position can be produced by a family of object depths and light source positions.

2.2.3. Effect of Marginalizing Unknown Nuisance Variables, S_g. Problems frequently occur in which the image data given the scene variables can be expressed as

$$I = f(x, S) + \nu(x, S)$$

where the function f expresses the deterministic imaging equations, $S_g = x$ represents the nuisance variable, S represents the remaining set of scene variables and $\nu(x, S)$ is a term representing the imaging noise, which is frequently a constant or a slowly-varying function of x and S.

If, in addition to being slow-varying, the measurement noise distribution has zero mean, and the likelihood admits a quadratic approximation,[2] then we can make explicit the effect of the nuisance variable on the resulting distribution.

In Appendix B, we show that the likelihood can be written

$$p(I \mid x, S) \simeq g(x, S) \exp\left(\frac{-\frac{n}{2}(I - f(x, S))^2}{\sigma(x, S)^2}\right)$$

where the imaging noise $\frac{1}{\sigma(x,S)^2} = \frac{\partial^2 k(I \mid x,S)}{\partial I^2}|_{f(x,S)}$, and $g(x, S) = \exp(-\frac{n}{2}k(f(x, S) \mid x, S))$, and $k(I \mid x, S) = \frac{-\log(p(I \mid x,S))}{n}$.

We also show that the marginal integral across x can be approximated as:

$$\int p(x, S)\, p(I \mid x, S)\, dx \approx \frac{\sqrt{2\pi\sigma(\hat{x}(I, S), S)^2}}{|\partial f(\hat{x}(I, S), S)/\partial x|}\, p(\hat{x}(I, S), S) g(\hat{x}(I, S), S) \tag{3}$$

where $\hat{x}(I, S) = \arg\max_x \exp(-\frac{n}{2}(I - f(x, S))^2/\sigma(x, S)^2)$.

By assuming that the imaging noise $\sigma(\hat{x}(I, S), S)^2 \approx \sigma^2$ (i.e. is nearly a constant), this last expression shows that the likelihood after marginalization is dominated by values of S where $|\partial f(\hat{x}(I, S), S)/\partial x|$ is close to zero and disappears where $|\partial f(\hat{x}(I, S), S)/\partial x|$ is large. The important point here is that the posterior can be dominated by the rate of change of the imaging function with the nuisance variable x, which will be important when we consider the effects of changes of representation on the inference in a later section. This result and derivation is similar to Freeman (1994). The key differences are that Freeman simply assumes a gaussian posterior and centers the approximation around a fixed value of x, rather than around the maximum $\hat{x}(I, S)$.

In general, the variables that the distribution is marginalized over have a big impact on the likelihood, and hence on the statistical properties of the estimator. As a practical consequence, the uncertainty we have on the nuisance variables has a greater impact on the estimator than the particular properties of the measurement noise distribution when the measurement noise distribution is slowly varying across changes in scene variables.

We have shown how the dependence between variables is a function of the task. In the next section we show how the degree of modularity and estimation complexity of an optimal inference is a function of both

the statistical dependence between scene variables and the choice of representation of the scene variables.

3. Dependence of Optimal Cue Integration on Task and Representation

In this section, we discuss three ways in which task specification affects the computational architecture for an optimal cue integration problem. We show, given the task: 1) how the degree of modularity and resulting estimation complexity of the optimal estimator is a function of conditional independence between image cues; 2) that a particular task, such as discriminating depth, does not unambiguously specify the representation of the scene variable, and furthermore; 3) the choice of variable representation results in differences in the complexity of the cue integration and the performance of the optimal estimators.

The third point in particular will bring us to our central idea, illustrated with an example in Section 4, that the choice of scene variable representation for a decision can determine the modularity and performance of the optimal estimator.

3.1. *Conditional Independence Determines Modularity*

Having discussed how statistical dependence between variables depends on a task, we can now show how this statistical dependence determines how variables interact in performing optimal inference, which has consequences for data fusion. The key element in optimal inference is the posterior, which we can rewrite both in terms of the joint distribution and in terms of likelihoods and priors:

$$p(S_{\mathrm{r}}, S_{\mathrm{g}} \mid I) = \frac{p(I, S_{\mathrm{r}}, S_{\mathrm{g}})}{p(I)} = \frac{p(I \mid S_{\mathrm{r}}, S_{\mathrm{g}})\, p(S_{\mathrm{r}}, S_{\mathrm{g}})}{p(I)} \qquad (4)$$

The probabilistic structure of the joint probability distribution $p(I, S_{\mathrm{r}}, S_{\mathrm{g}})$ can be represented by a Bayes Net (Pearl, 1988; Jensen, 1966), which is simply a graphical model which expresses the conditional independence between the variables. Using labels to represent variables and arrows to represent conditioning (with $a \rightarrow b$ indicating b is conditioned on a[3]), independence can be represented by the absence of connections between variables. Using these graphical models we can determine the interactions between variables by inspection.

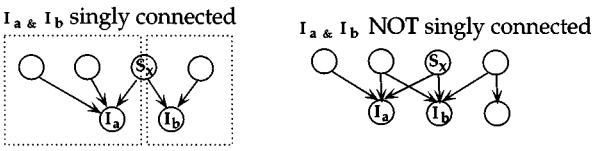

Figure 3. Whether independent data measures are singly connected to the estimated variable S_x determines whether or not estimation modules can be created for S_x. *Left* example of Bayesian modularity. Boxes show how the variables can be split to form two modules. *Right* example of a non-modular estimation.

For instance if two sets of variables are completely independent, then the graphs of the variables are disjoint.

Because modularity is the ability to use different image cues to produce independent estimates of the variable S_x, what determines modularity in a Bayesian inference is whether or not the data are conditionally independent given S_x. When this is true, we can produce separate likelihood functions for S_x which can be combined by multiplication (i.e. $p(I_a, I_b \mid S_x) = p(I_a \mid S_x)\, p(I_b \mid S_x)$), a property we will call *Bayesian modularity*. Graphically, this requirement is equivalent to the different image measurements being singly connected to the variable of interest. Figure 3 shows examples of a singly connected net and a non-singly connected net. The non-singly connected net corresponds to the case in which the data cues depend on more than one scene variable, which is exactly the case that calls for cooperative computation.

3.2. *Dependence of Cue Integration on Representations of Scene Variables*

In moving from a general task description to a specific implementation, there can be a choice with regard to the exact scene variables used to do the inference. For example, a task which involves inferring the relative distance of objects from the observer can estimate any function of the distances which do not change the relative depth ordering. However, this choice does make a difference in terms of the properties of the estimator, in particular in its performance and Bayesian modularity.

3.2.1. Affects on Performance. First, the particular scene variables we estimate matter because Bayesian inference is not invariant to reparametrizations of continuous variables. Thus if we perform optimal inference on one variable, we cannot just transform the result to get optimal inference on the transformed variable. This is due to the fact that transforming

the variant x of probability distribution $dF = p(x)\,dx$ yields $dF = p(g(y))g'(y)\,dy$ where $x = g(y)$. Thus the transformation will not yield the same inferences unless $g(y)$ is linear. This causes, for instance, binomial and beta distributed densities which are identical in x space to be substantially different in $y = 1/x$ space (Edwards, 1992). While this fact has been used to critique Bayesian inference (Edwards, 1992), it also has the interpretation that the kind of information contained about a variable and its transform by one distribution is not the same as the information contained by another distribution.

This lack of invariance makes a difference in the performance characteristics of an estimator across reparametrizations of continuous variables. Using the same argument as above, one can show that measures of performance (e.g. Fisher Information) for the transformed random variable are not necessarily equal to the same value of the performance measure calculated using the original random variable.

3.2.2. Affects on Modularity/Complexity of the Inference. Second, the choice can determine how many nuisance variables must be considered. For instance, consider scene variables x and y such that x and y are statistically dependent given the image data, but $x + y$ and $x - y$ are independent. If we estimate x, then we must consider y a nuisance variable. However, if we estimate $x + y$, then nuisance variables disappear.

The complexity of the inference and the properties of the estimator depend on which variables we marginalize the distribution across. Changing representations can introduce new unknowns that require marginalization and change the dependence between the variable of interest and the nuisance variables. We will show examples of both of these effects in Section 4.

3.2.3. Affects on Marginalization. Third, the choice of representation for the nuisance variables can have a strong impact on the posterior and hence the properties of the estimator. Consider the ideal (noiseless) imaging equation $I = f(x, S)$, where S are the variables of interest and x is a nuisance variable. What happens to the posterior after marginalizing the nuisance variable if x is changed to $x = m(y)$? The integral over y becomes:

$$\int p(g(y), S)\, p(I \mid m(y), S) m'(y)\, dy \approx \frac{\sqrt{2\pi\sigma(m(\hat{y}), S)^2}}{|\partial f(m(\hat{y}), S)/\partial y|}\, p(m(\hat{y}), S) g(m(\hat{y}), S) \qquad (5)$$

However, $|\partial f(m(\hat{y}), S)/\partial y|$ can differ from $|\partial f(\hat{x}, S)/\partial x|$, and thus the posterior after marginalization will depend on the choice of representation.

What we have shown in this section is that the properties of optimal estimators are largely determined by: 1) the pattern of conditional dependence between variables; 2) the nuisance variables that the distribution is marginalized across; and 3) the representation chosen for the problem.

4. Estimating Depths from Image Size and Shadow Displacement

In this section we perform a detailed analysis of Bayesian cue integration for a simple problem, in order to generate predictions that can be tested with human psychophysical data. One of the challenges of testing models of human vision is that the experimenter does not have direct access to the variable representations on which a decision is being made. We show that different representations predict different patterns of cue integration for ideal observers, including which variables interact, and how cue integration is confidence-driven. Confidence-driven cue integration refers to strategies that weight a cue's contribution to the inference to match its ability to provide reliable estimates. Given these differences in ideal performance with changes in representation, we can begin to infer the scene variable representation used by the human observer by comparing human and ideal performance.

4.1. Theory

We illustrate the dependence of Bayesian cue integration on task demands and conditional independence with a simple scene due to Kersten et al. (l996). The scene consists of a flat central square, a flat checkerboard background and a light source. The square floats in front of the background, and the light source is positioned so that the square casts a shadow onto the background. The observer judges the depth of this square vs. the depth of another square (simulated to be physically identical in 3D) presented at a different time. The viewing distance, and the orientation of the square and background were kept fixed. In this simplified world the only cues to depth are the image size a of the square, and the position of the cast shadow β (measured by the visual angle subtended by the direction of gaze and the shadow position). An example stimulus is shown in Fig. 4.

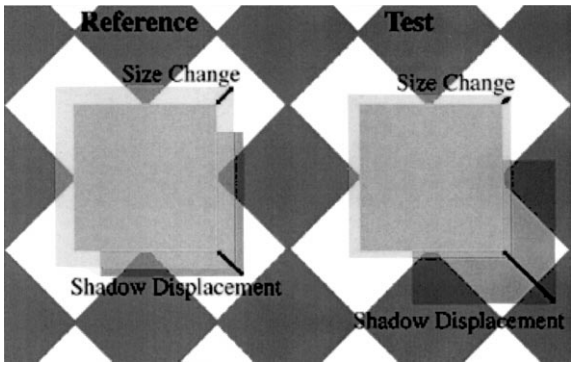

Figure 4. An illustration of the stimuli used in the experiment. Two movies depicting a square moving in depth are sequentially shown to the observer. The image size of the square becomes larger and the shadow moves away from the square with decreasing depth from the checkerboard background. The image on the left illustrates the reference condition in which the image size was maximal and the shadow displacement minimal. The right hand side shows the test condition which has variable image size and shadow displacements. Subjects judged whether the reference or test square moved further in depth at the end of the movie in a two-alternative forced-choice method.

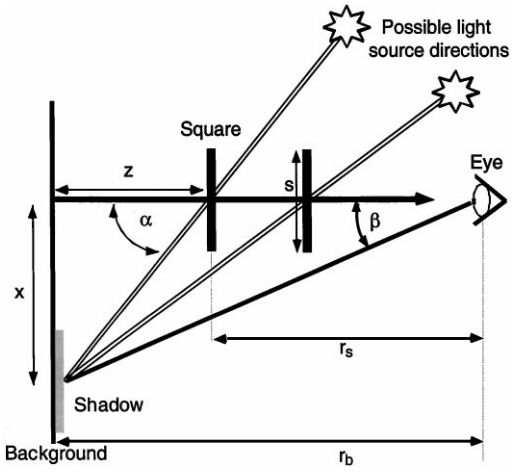

Figure 5. Diagram illustrating the problem of inferring depth from image size and cast shadow position in 1-D for the central square in front of a checkerboard background (see Fig. 4). There are three depth variables, distance to the background r_b, distance to the square r_s, and the distance of the square from the background z. The cast shadow position x depends both on the light source position α and z. We assume that the observer can measure the angle subtended by the shadow position β. The image size a (not shown) of the object depends on the physical 3D size of the square s and the viewing distance r_s.

These cues are substantially different. The image size is determined by the depth of the square from the observer and the physical size of the square. Image size information is most naturally used to estimate the *egocentric* distance to the square. On the other hand the shadow position is determined by variables in a different depth representation. Cast shadow position is determined by the *allocentric* distance of the square from the background and the position of the light source. Thus to integrate the shadow and image size data, we must convert one, or both of the variables into a common depth representation.

From the standpoint of traditional estimation, a strong case can be made not to integrate the cues. When we know that the sizes of the two squares are identical, then we can simply compare the likelihoods for depth given the image size. When the likelihoods are singly peaked, the optimal decision simplifies to comparing image sizes, and judging the larger one closer. Similarly for the shadow cue, assuming the light source direction is the same for both intervals, the square farther from the background can be decided on the basis of which shadow position is farther from the square. Thus it might seem more natural not to integrate the cues, and instead make separate judgments of depth from the cues.

In contrast, Bayesian inference requires choosing a common depth representation to integrate the cues. However, then the size of the square and the light source direction can no longer be neglected. We considered three possible common depth representations for the inference. Each of these depth representations leads to a different Bayes net and different properties of the optimal estimator. For each of the three representations, however, the best way to determine which square is closer is to compute MAP depth estimates for both intervals and choose the smaller (closer to the observer) value.

The geometric diagram in Fig. 5 illustrates the variables for the task. We will consider three ways of computing the depths: the relative distance of the square from the background $z_r = z/r_b$, the absolute distance of the square from the background z, and the absolute distance from the observer r_s. These estimates are illustrated in Fig. 6.

4.1.1. Representation 1: Estimating Relative Distance from Background (z_r). One way of judging the depths of the two squares is to compute the relative distance from the background. This leaves 4 unknowns, α, s, z, and r_b with only two data variables, the image size a and the shadow position β. If the observer estimates $z_r = z/r_b$, and represents the relative object size

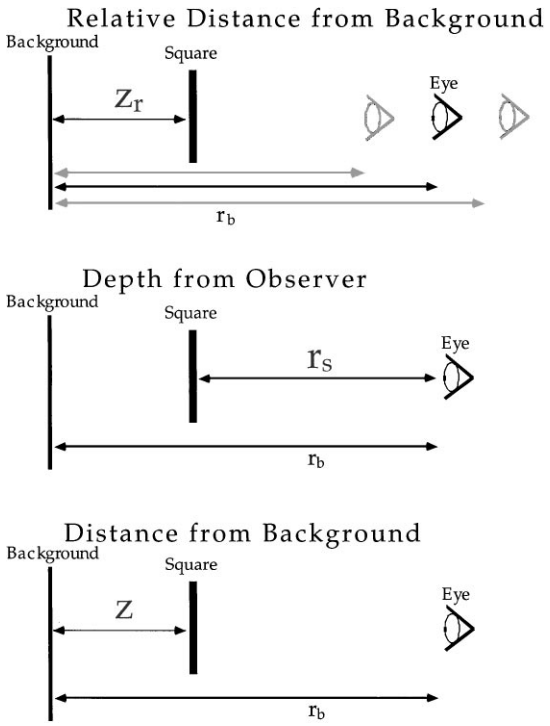

Figure 6. Diagram illustrating the depth variables to be estimated. The variable $z_r = z/r_b$ can't be shown directly, because it is an equivalence class of z and r_b distances.

$s_r = s/r_b$, then the resulting estimator does not need knowledge of the value of r_b. By computing with the scaled variables, we make our inferences more reliable because we have eliminated the uncertainty we might have in r_b.

While computing distance relative to an arbitrary background may seem contrived, the idea is similar to computing depth relative to the fixation distance frequently used in depth from stereo. From a psychological standpoint, object depth is often evaluated relative to a background context. There are situations, like sitting at one's desk, where a fixed object (the desk) is familiar enough for it to make sense to compute distances relative to it. In addition, many perceptual tasks do not require metric distance information (I can see that there is a pen on my desk without calculating the distances from myself to each of the objects).

In this representation the observer needs to estimate the relative distance $z_r = z/r_b$ of the square from the background checkerboard wall. Both the image size of the square and the shadow position are functions of z_r. The shadow position measurement β (in terms of visual angle),[4] is a function of z_r and light source position α:

$$\beta = \tan^{-1}(z_r \tan(\alpha)) + n_\beta \tag{6}$$

The term n_β models the noise in the measurement. For simplicity we take this to be a Gaussian random variable, so that β is Gaussian distributed. The likelihood function is given by:

$$p(\beta \mid z_r, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \times \exp\left(- \frac{(\beta - \tan^{-1}(z_r \tan(\alpha)))^2}{2\sigma_\beta^2} \right) \tag{7}$$

The image size a is given by:

$$a = \frac{s}{r_s} + n_a = \frac{s/r_b}{1 - z/r_b} + n_a = \frac{s_r}{1 - z_r} + n_a$$

where s_r is the actual size of the square relative to the distance to the background, and n_a is a term which models the noise in the measurement. We modeled the size measure noise as log normal. We believe this is a reasonable choice of noise distribution because both s_r and $1 - z_r$ are physically constrained to be positive and more importantly, to model existing data on human size and distance perception. Because human size and distance discrimination thresholds are well-fit by increasing power-laws, it is reasonable to assume that the variance of the measurement noise effectively scales up with the magnitude of the variable (Stevens, 1957), which is a key property of the log normal distribution. Given these assumptions, the likelihood for a is given by:

$$p(a \mid z_r, s_r) = \frac{1}{\sqrt{2\pi}\sigma_a a} \times \exp\left(- \frac{\left(\log(a) - \log\left(\frac{s_r}{1-z_r}\right)\right)^2}{2\sigma_a^2} \right) \tag{8}$$

We assume that the observer potentially has several measurements of shadow position and image size available. The set of measurements for β and a are represented using set function notation: $\{\beta\}, \{a\}$. To estimate z_r we compute $p(z_r \mid \{\beta\}, \{a\})$. Assuming that the repeated measurements of the image size a and the shadow position β are independent, $p(z_r \mid \{\beta\}, \{a\})$ can

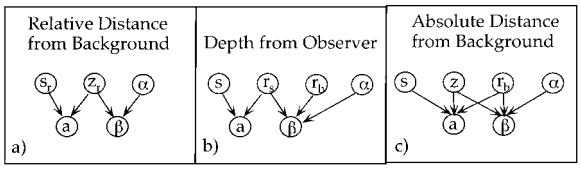

Figure 7. Bayes nets for the three depth representations. a) Bayes net for relative distance to the background. This task involves estimating object relations (world centered), and requires the least prior knowledge. b) Bayes net for distance to observer. Notice that the use of the shadow information requires integrating across two variables, hence the shadow cue should have more uncertainty for this task. c) Bayes net for metric depth from the background. Estimating the distance from the background, z, is complicated by the image size and shadow position measurements also being jointly dependent on the observer's distance to the background.

be written:

$$
\begin{aligned}
p(z_{\mathrm{r}} \mid \{\beta\}, \{a\}) &= \frac{p(\{\beta\} \mid z_{\mathrm{r}})\, p(\{a\} \mid z_{\mathrm{r}})\, p(z_{\mathrm{r}})}{p(\{\beta\}, \{a\})} \\
p(z_{\mathrm{r}} \mid \{\beta\}, \{a\}) &\propto p(\{\beta\} \mid z_{\mathrm{r}})\, p(\{a\} \mid z_{\mathrm{r}})\, p(z_{\mathrm{r}}) \\
&= \left(\int_{\alpha} \prod_{i=1}^{N} p(\beta_i \mid z_{\mathrm{r}}, \alpha)\, p(\alpha)\, d\alpha \right) \\
&\quad \times \left(\int_{s_{\mathrm{r}}} \prod_{i=1}^{N} p(a_i \mid z_{\mathrm{r}}, s_{\mathrm{r}})\, p(s_{\mathrm{r}})\, ds_{\mathrm{r}} \right) \\
&\quad \times p(z_{\mathrm{r}}),
\end{aligned}
$$

where N is the number of measurements. The Bayes net which corresponds to this inference is shown in Fig. 7(a). Note that this network is Bayes modular, which shows up in the factoring of the likelihoods above.

4.1.2. Representation 2: Estimating Depth to Square (r_{s}). As we interact with the world, there are instances when viewer-centered depth is required, such as navigating and reaching to objects. Thus, it is reasonable to consider a second task in which one estimates the distance, r_{s}, from the observer to the squares. The Bayes net for this inference is shown in Fig. 7(b). In this case, we must convert the shadow position cue's dependence on the allocentric distance z to the egocentric depth r_{s}. Using $r_{\mathrm{b}} = z + r_{\mathrm{s}}$, we can write the shadow position measurement as:

$$
\beta = \tan^{-1}\left(\left(1 - \frac{r_{\mathrm{s}}}{r_{\mathrm{b}}} \right) \tan(\alpha) \right) + n_{\beta} \tag{9}
$$

The likelihood function is given by:

$$
\begin{aligned}
&p(\beta \mid r_{\mathrm{s}}, r_{\mathrm{b}}, \alpha) \\
&\quad = \frac{1}{\sqrt{2\pi}\sigma_{\beta}} \exp\left(- \frac{\left(\beta - \tan^{-1}\left(\left(1 - \frac{r_{\mathrm{s}}}{r_{\mathrm{b}}} \right) \tan(\alpha) \right) \right)^2}{2\sigma_{\beta}^2} \right)
\end{aligned} \tag{10}
$$

The image size a is given by:

$$
a = \frac{s}{r_{\mathrm{s}}} + n_a
$$

Hence the likelihood for a is given by:

$$
p(a \mid r_{\mathrm{s}}, s) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp\left(- \frac{\left(\log(a) - \log\left(\frac{s}{r_{\mathrm{s}}} \right) \right)^2}{2\sigma_a^2} \right) \tag{11}
$$

To base the decision on r_{s}, we compute $p(r_{\mathrm{s}} \mid \{\beta\}, \{a\})$:

$$
\begin{aligned}
p(r_{\mathrm{s}} \mid \{\beta\}, \{a\}) &\propto p(\{\beta\} \mid r_{\mathrm{s}})\, p(\{a\} \mid r_{\mathrm{s}})\, p(r_{\mathrm{s}}) \\
&= \left(\int_{r_{\mathrm{b}}} \int_{\alpha} \prod_{i=1}^{N} p(\beta_i \mid r_{\mathrm{s}}, r_{\mathrm{b}}, \alpha)\, p(\alpha)\, p(r_{\mathrm{b}})\, d\alpha\, dr_{\mathrm{b}} \right) \\
&\quad \times \left(\int_{s} \prod_{i=1}^{N} p(a_i \mid r_{\mathrm{s}}, s)\, p(s)\, ds \right) p(r_{\mathrm{s}})
\end{aligned} \tag{12}
$$

Note that this inference is Bayes modular, and that inference with the shadow cue requires dealing with the additional unknown r_{b}. Thus, for this representation, the uncertainty in our depth from shadow estimates increases as compared with the relative distance representation (Representation 1).

4.1.3. Representation 3: Estimating Absolute Distance to Background (z). Finally, the observer could compute z, the absolute distance from the square to the background. This requires converting the image size cue's dependence on the egocentric depth r_{s} to the allocentric distance z. After conversion, the distance to the background r_{b} becomes a second unknown for both cues. The Bayes net that corresponds to this inference is shown in Fig. 7(c). The measurements can be written in terms of z as:

$$
\begin{aligned}
\beta &= \tan^{-1}(z\, \tan(\alpha)/r_{\mathrm{b}}) + n_{\beta} \\
a &= \frac{s}{r_{\mathrm{b}} - z} + n_{\mathrm{a}}.
\end{aligned} \tag{13}
$$

The likelihood functions are:

$$p(\beta \mid z, r_{\mathrm{b}}, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \times \exp\left(- \frac{\left(\beta - \tan^{-1}(z\,\tan(\alpha)/r_{\mathrm{b}})\right)^2}{2\sigma_\beta^2} \right) \quad (14)$$

$$p(a \mid z, r_{\mathrm{b}}, s) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp\left(- \frac{\left(\log(a) - \log\left(\frac{s}{r_{\mathrm{b}} - z}\right)\right)^2}{2\sigma_a^2} \right) \quad (15)$$

To estimate z we compute $p(z \mid \{\beta\}, \{a\})$:

$$p(z \mid \{\beta\}, \{a\}) \propto p(\{\beta\}, \{a\} \mid z)\, p(z) = \left(\int_{r_{\mathrm{b}}} \left(\int_\alpha \prod_{i=1}^{n} p(\beta_i \mid z, r_{\mathrm{b}}, \alpha)\, p(\alpha)\, d\alpha \right) \times \left(\int_{\mathrm{s}} \prod_{i=1}^{n} p(a_i \mid z, r_{\mathrm{b}}, s)\, p(s)\, ds \right) p(r_{\mathrm{b}})\, dr_{\mathrm{b}} \right) p(z) \quad (16)$$

Note that the posterior no longer factors into separate likelihoods for z, due to the joint marginalization across r_{b}. Thus, estimating absolute z is not Bayes modular. This has consequences for cue integration that we explore below.

Because the three depth estimators require different marginalizations, we expect that they will have different estimation and performance characteristics. We show that these expectations are correct below by deriving explicit formula for the depth estimates and the Cramer-Rao bound on the variance of these estimates.

4.1.4. MAP Estimates. Maximum a posteriori estimates of depth for each of the three representations were computed from the posterior distribution after marginalization. Marginalizations were approximated using Laplace's method, as described in Appendix A. The goodness of all of the approximations was checked by performing the marginalization integrals numerically. However, because these numerical integrations have to be performed for each value of the image data separately, we only performed these checks for the range of values used in the psychophysical experiments described below.

4.1.5. Representation 1: MAP Estimates for z_{r} (Relative z). To compute the MAP estimates for z_{r} from shadow position β and the image size of the square a, we first need to marginalize the likelihoods for β and a across the light source direction α and s_{r} respectively.

To marginalize $\prod_{i=1}^{n} p(\beta_i \mid z_{\mathrm{r}}, \alpha)$ across α, we assume a prior on light source direction $p(\alpha)$ that is uniform over $[-\pi/2, \pi/2]$ (i.e. $p(\alpha) = 1/\pi$).

First note that:

$$\prod_{i=1}^{n} p(\beta_i \mid z_{\mathrm{r}}, \alpha) = \frac{1}{\left(2\pi\sigma_\beta^2\right)^{\frac{n}{2}}} \times \exp\left(- \frac{\sum_{i=1}^{n} (\beta_i - \tan^{-1}(z_{\mathrm{r}} \tan(\alpha)))^2}{2\sigma_\beta^2} \right) = \frac{1}{\left(2\pi\sigma_\beta^2\right)^{\frac{n}{2}}} \exp\left(-n \frac{\hat{\beta}^2 - \hat{\beta}^2}{2\sigma_\beta^2} \right) \times \exp\left(-n \frac{(\hat{\beta} - \tan^{-1}(z_{\mathrm{r}} \tan(\alpha)))^2}{2\sigma_\beta^2} \right) \quad (17)$$

where $\hat{\beta}$ is the mean of the N sample βs, and $\hat{\beta}^2$ is the mean of the β_i^2.

Ignoring the likelihood factors that exclude α, we need to compute the integral:

$$\int_{-\pi/2}^{\pi/2} \exp\left(-n \frac{(\hat{\beta} - \tan^{-1}(z_{\mathrm{r}} \tan(\alpha)))^2}{2\sigma_\beta^2} \right) p(\alpha)\, d\alpha \quad (18)$$

We use Laplace's method (see Appendix A), which involves computing a second order Taylor series expansion of the exponent $h(\hat{\beta} \mid \alpha, z_{\mathrm{r}}) = -\frac{(\hat{\beta} - \tan^{-1}(z_{\mathrm{r}} \tan(\alpha)))^2}{2\sigma_\beta^2}$, around $\alpha_m = \arg\max_\alpha h(\hat{\beta} \mid \alpha, z_{\mathrm{r}})$. Computing $\partial h(\hat{\beta} \mid \alpha, z_{\mathrm{r}})/\partial\alpha$ and setting it equal to zero, α_m can be solved for yielding: $\alpha_m(\hat{\beta}, z_{\mathrm{r}}) = \tan^{-1}(\tan(\hat{\beta})/z_{\mathrm{r}})$.

From the Appendix, the integral can be approximated:

$$\int_{-\pi/2}^{\pi/2} p(\alpha)\, p(\hat{\beta} \mid z_{\mathrm{r}}, \alpha) = p(\alpha_m(\hat{\beta}, z_{\mathrm{r}})) \exp(nh(\hat{\beta} \mid \alpha_m(\hat{\beta}, z_{\mathrm{r}}), z_{\mathrm{r}})) \times \sqrt{\frac{2\pi}{-n \frac{\partial^2 h(\hat{\beta} \mid \alpha_m(\hat{\beta}, z_{\mathrm{r}}), z_{\mathrm{r}})}{\partial \alpha^2}}}$$

Plugging in the expressions for α_m, $h(\hat{\beta} \mid \alpha, z_r)$, and $p(\alpha) = 1/\pi$ and simplifying yields:

$$\int_{-\pi/2}^{\pi/2} \prod_{i=1}^{n} p(\beta_i \mid z_r, \alpha)\, p(\alpha)\, d\alpha \simeq \frac{c z_r}{\left(z_r^2 \cos(\hat{\beta})^2 + \sin(\hat{\beta})^2\right)} \tag{19}$$

where $c = \frac{\sqrt{2}}{\pi (2\pi\sigma_\beta^2)^{\frac{n-1}{2}} \sqrt{n}}$.

Using this expression we find the maximum likelihood z_r occurs at:

$$\arg\max_{z_r}(p(\beta \mid z_r)) = \tan(\hat{\beta}). \tag{20}$$

This approximation is very good. Numerical evaluations of the integral showed that the approximate maximum likelihood estimates were within 2% of actual, and the mean Kullback-Leibler divergence between the approximate and the actual distributions was much less than one.

For the size change cue, we need to compute the integral:

$$\begin{aligned} p(\{a\} \mid z_r) &= \int_0^\infty \prod_{i=1}^{n} p(a_i \mid z_r, s_r)\, p(s_r)\, ds_r \\ &= \int_0^\infty \frac{p(s_r)}{\left(2\pi\sigma_a^2 \hat{a}^2\right)^{\frac{n}{2}}} \\ &\quad \times \exp\left(-n \frac{\left(\log(\hat{a}) - \log\left(\frac{s_r}{1-z_r}\right)\right)^2}{2\sigma_a^2}\right) ds_r \end{aligned}$$

where $\hat{a} = \prod_{i=1}^{n} a_i / n$ is the geometric mean of the n samples.

The integral is analytically tractable. If the prior is constant, it is easy to show that the integral evaluates to a constant times $\hat{a}(1 - z_r)$ and because $z_r \geq 0$, the maximum likelihood estimate of z_r is always zero. Thus, some knowledge of the relative size is crucial to compute the relative distance. We used a log normal prior on s_r:

$$p(s_r) = \frac{1}{\sqrt{2\pi}\sigma_{s_r} s_r} \exp\left(-\frac{\log(s_r/\mu_{s_r})^2}{2\sigma_{s_r}^2}\right).$$

To compute the integral, we combine the exponents for the prior and likelihood and use traditional methods of completing the square for $\log(s_r)$ and factoring off the part independent of s_r The result is:

$$\begin{aligned} p(\{a\} \mid z_r) &= \frac{1}{\sqrt{\pi\left(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2\right)\hat{a}}} \\ &\quad \times \exp\left(-\frac{\log(\hat{a}(1 - z_r)/\mu_{s_r})^2}{2\left(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2\right)}\right) \end{aligned} \tag{21}$$

where $\sigma_{\hat{a}}^2 = \sigma_a^2 / n$. The maximum z_r with respect to image size occurs at

$$\arg\max_{z_r}(p(\{a\} \mid z_r)) = 1 - \mu_{s_r}/\hat{a} \tag{22}$$

if $\mu_{s_r} < \hat{a}$ and at zero otherwise.

4.1.6. Representation 2: MAP Estimate for r_s. To find the optimal estimate of r_s from the shadow cue, we need to compute:

$$\int_{r_b} \int_{\alpha} p(\{\beta\} \mid r_s, \alpha, r_b)\, p(r_b)\, d\alpha\, p(r_b)\, dr_b$$

The first integral over α the same as performed in the last section, except that z_r is replaced with r_s/r_b. We also need to marginalize over r_b, the distance to the background. We assumed a log normal prior on r_b with parameters μ_{r_b} and σ_{r_b}. To marginalize, we used a modified method (see Appendix A, case 2) in which the Taylor series expansion is performed on the sum of the exponents of the prior and the likelihood around the location of the maximum of the combined prior and likelihood. We find:

$$\begin{aligned} &\int_{r_b} \int_{\alpha} p(\{\beta\} \mid r_s, \alpha, r_b)\, p(r_b)\, d\alpha\, p(r_b)\, dr_b \\ &\quad \simeq \int_{r_b} \frac{c\left(1 - \frac{r_s}{r_b}\right)}{\left(1 - \frac{r_s}{r_b}\right)^2 \cos(\hat{\beta})^2 + \sin(\hat{\beta})^2}\, p(r_b)\, dr_b \\ &\quad \simeq \frac{c(1 - r_s/\mu_{r_b})}{\left((1 - r_s/\mu_{r_b})^2 \cos(\hat{\beta})^2 + \sin(\hat{\beta})^2\right)} \end{aligned} \tag{23}$$

The maximum r_s occurs at:

$$\arg\max_{r_s}(p(\{\beta\} \mid r_s)) = \mu_{r_b}(1 - \tan(\hat{\beta})). \tag{24}$$

For the size change cue, marginalizing with respect to a log normal prior on s yields:

$$p(\hat{a} \mid r_s) = \frac{1}{\sqrt{\pi\left(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2\right)\hat{a}}} \exp\left(-\frac{\log(\hat{a} r_s/\mu_{s_r})^2}{2\left(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2\right)}\right). \tag{25}$$

The maximum r_s with respect to image size occurs at

$$\arg\max_{r_s}(p(\hat{a}\mid r_s)) = \frac{\mu_s}{\hat{a}}. \tag{26}$$

4.1.7. Representation 3: MAP Estimate for z. In optimal estimation of z we cannot consider the shadow cue and image size cues separately. To find the optimal estimate of z from the size and shadow cues, we need to compute:

$$\begin{aligned} p(\{\beta\},\{a\}\mid z) = &\int_{r_b}\left(\int_{\alpha}\prod_{i=1}^{n} p(\beta_i\mid z, r_b, \alpha)\,p(\alpha)\,d\alpha\right)\\ &\times\left(\int_{s}\prod_{i=1}^{n} p(a_i\mid z, r_b, s)\,p(s)\,ds\right)\\ &\times p(r_b)\,dr_b \end{aligned}$$

The integrals over s and α are identical to those above, with the appropriate change of variables. To compute the marginal across r_b, we assumed a log normal prior on r_b, and used Laplace's method case 2, in which the exponents of $p(\hat{\beta}\mid z, r_b)$ and $p(\hat{a}\mid z, r_b)$ are combined and the Taylor series expansion is performed around the joint maximum.

The resulting asymptotic approximation to the posterior is:

$$\begin{aligned} p(z\mid\{\beta\},\{a\}) \propto &\frac{\mu_s z\csc(\hat{\beta})\sec(\hat{\beta})}{\sqrt{2}\hat{a}\sqrt{\hat{a}^2z^2+\mu_s^2\left(\sigma_{\hat{a}}^2+\sigma_s^2\right)\tan(\hat{\beta})^2}}\\ &\times\exp\left(-\frac{(z-\frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))})^2}{2\frac{\hat{a}^2z^2+\mu_s^2(\sigma_{\hat{a}}^2+\sigma_s^2)\tan(\hat{\beta})^2}{\hat{a}^2(1-\tan(\hat{\beta}))^2}}\right)\\ &\times p(z) \end{aligned} \tag{27}$$

Neglecting the prior $p(z)$, the exact MAP estimator can be computed but is too complicated to present. However, for the values of the parameters used in the experiments, $\mu_s^2(\sigma_{\hat{a}}^2+\sigma_s^2)\tan(\hat{\beta})^2$ is small enough to be neglected. Then the maximum likelihood can be approximated by:

$$\arg\max_{z}(p(z\mid\{\beta\},\{a\})) \simeq \frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))} \tag{28}$$

when $\tan(\hat{\beta})\geq 1$, and 0 otherwise.

All the MAP estimates are summarized in Table 1.

4.1.8. Fisher Information. The Fisher Information contained in the observable data on the parameter x is given by:

$$\begin{aligned} \mathcal{I}(x) = -N\int_{data} &p(data\mid x)\\ &\times(\partial^2\log p(data\mid x)/\partial x^2)\,d(data) \end{aligned} \tag{29}$$

where the integral is over all possible values of the data. It measures information in the sense that $(1/N)\mathcal{I}^{-1}$ is a lower bound on the variance of any unbiased estimator (Rao, 1973). We can also interpret the Fisher Information as the variance parameter of the normal approximation to the likelihood function. Because unbiased maximum likelihood estimators are both asymptotically normal and achieve the lower bound with increasing sample size, the Fisher Information is also the asymptotic variance of the estimator (Tanner, 1996).

We computed the Fisher Information for each of the likelihood functions derived in the previous section. To compute the Fisher Information, we plugged the expressions for the likelihoods into Eq. (29), and computed the integral either analytically or using a Laplace approximation. For the likelihoods involving the shadow cue, the integrals are analytically tractable. However, the integrals over the likelihoods involving z

Table 1. Table of MAP estimates and Fisher information values for the three depth estimate representations. For the representations which admit modular estimates, the estimates are shown separately for the shadow and image size cues.

Task	Est from shadow	Est from size	Shadow Fisher info	Size Fisher info
Relative z	$z_r = \tan(\hat{\beta})$	$z_r = 1-\frac{\mu_{s_r}}{\hat{a}}$	$\frac{1}{\sqrt{2}\tan(\hat{\beta})^2}$	$\frac{2\hat{a}^2}{\mu_{s_r}^2(\sigma_{s_r}^2+\sigma_{\hat{a}}^2)}$
Dist. from obs.	$r_s = \mu_{r_b}(1-\tan(\beta))$	$r_s = \frac{\mu_s}{\hat{a}}$	$\frac{1}{\mu_{r_b}^2\tan(\hat{\beta})^2}$	$\frac{2\hat{a}^2}{\mu_s^2(\sigma_s^2+\sigma_{\hat{a}}^2)}$
Absolute z	$z = \frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))}$		$\frac{2\hat{a}^2(1-\tan(\hat{\beta}))^4}{\mu_s^2\tan(\hat{\beta})^2}$	

and the size change cue required approximation. The Laplace approximation, Case 1, (see Appendix A) was used in which the second derivative of the log likelihood played the role of the prior.

As an example of the approximate Fisher Information calculation, consider the likelihood $p(\{a\} \mid r_s)$ for a sample size of one. The second derivative of the log likelihood with respect to r_s is equal to $-\frac{1-\log(ar_s/\mu_s)}{(\sigma_a^2+\sigma_{s_r}^2)r_s^2}$, so that

$$\mathcal{I}(a \mid r_s) = \int_a p(a \mid r_s) \frac{1-\log(ar_s/\mu_s)}{(\sigma_a^2+\sigma_{s_r}^2)r_s^2}\, da \quad (30)$$

$$\mathcal{I}(a \mid r_s) = \frac{1}{\sqrt{\pi(\sigma_a^2+\sigma_{s_r}^2)}} \times \int_a \frac{1}{a}\exp\left(-\frac{\log(ar_s/\mu_{s_r})^2}{2(\sigma_a^2+\sigma_{s_r}^2)}\right) \times \frac{1-\log(ar_s/\mu_s)}{(\sigma_a^2+\sigma_{s_r}^2)r_s^2}\, da$$

Expanding the exponent in a Taylor series around the maximum likelihood $r_s^* = \frac{\mu_s}{a}$ and using Laplace's approximation we find:

$$\mathcal{I}(a \mid r_s) = \frac{2}{(\sigma_a^2+\sigma_{s_r}^2)r_s^2}$$

Evaluating the Fisher Information at the maximum likelihood yields:

$$\mathcal{I}(a \mid r_s^*) = \frac{2a^2}{(\sigma_a^2+\sigma_{s_r}^2)\mu_s^2}$$

The results of the other Fisher Information calculations are summarized in Table 1.

4.1.9. Cue Integration Using the Fisher Information. When independent likelihood functions for the depth variable can be derived (Bayesian modularity), the *minimum variance* estimator can be expressed in terms of the individual MAP estimates and the Fisher Information for each of the cues (Blake et al., 1996; Rao, 1973). Let m_a denote the MAP estimate and $\mathcal{I}_a(m_a \mid x)$ the Fisher information for the image size cue, and m_β the MAP estimate and $\mathcal{I}_\beta(m_\beta \mid x)$ the Fisher Information for the shadow cue. Then the two cues are combined by a linear combination of the individual estimates, weighted by their inverse variances:

$$m_{best} = \frac{m_a\mathcal{I}_a(m_a \mid x) + m_\beta\mathcal{I}_\beta(m_\beta \mid x)}{\mathcal{I}_a(m_a \mid x) + \mathcal{I}_\beta(m_\beta \mid x)}. \quad (31)$$

which is a specific prediction of a confidence-driven decision (recall that confidence-driven cue integration refers to strategies that weight a cue's contribution to the inference by its ability to provide reliable estimates). In addition to being the mimimal variance estimate, m_{best} is also approximately the MAP estimate for the combined cues.

The lower bound on the variance of m_{best} is given by:

$$\frac{1}{\mathcal{I}_a(m_a \mid x) + \mathcal{I}_\beta(m_\beta \mid x)} \quad (32)$$

Because r_s and $z_r = 1 - r_s/r_b$ are related by a linear transformation, we know that probability distributions on z_r should transform gracefully to distributions on r_s. This is in fact shown by the maximum likelihood estimates for r_s and z_r. For example, combining $z_r^* = \tan(\hat{\beta})$ and $r_s^* = \mu_{r_b}(1 - \tan(\beta))$, we find $z_r^* = 1 - r_s^*/\mu_{r_b}$, which expresses the relationship between r_s and z_r, given r_b is replaced by its mean μ_{r_b}. However, note that our MAP estimate for z, $z^* = \frac{\mu_s \tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))}$ is not what we would expect from weak fusion, which would show up as a weighted linear combination of $z = \mu_{r_b}\tan(\hat{\beta})$ & $z = \mu_{r_b}(1 - \frac{\mu_{s_r}}{\hat{a}})$, nor can it be produced by converting either the z_r^* or the r_s^* to z. Thus, in this case strong (non-modular) fusion (Clark and Yuille, 1990) has resulted from marginalization.

Inspecting the Fisher information functions, we can determine how the informativeness of the cues vary as a function image size and shadow position. For all three representations, the informativeness of the shadow cue decreases with increasing distance of the shadow from the square, while the informativeness of the image size cue increases with image size. Thus shadow information is useful when an object is close to the object it casts its shadow on, while image size information is useful when an object is close to the observer.

4.2. Human Performance

We have shown how ideal performance varies given different parameterizations of the ideal's decision variable and that different choices of variable predict different patterns of cue integration. We performed a shadow and image size cue integration experiment to investigate whether or not human observers make Bayesian-like use of both cues to estimate the depth of the square (Lawson et al., 1998).

Computer graphics animations of a 2 cm by 2 cm target square moving in depth were created by a

displacement of the shadow from an initial position and by a size change of the square. Participants viewed two animations presented sequentially (the reference and test images in randomized order) and were asked to judge which of the two squares moved further in depth from the background.[5] Responses were recorded via a mouse button click. In the reference image, size change was maximal (28%) and shadow displacement was minimal (0.5 cm). In the test image, size change ranged from 16% to 28% (16%, 19%, 22%, 25%, 28%) and shadow displacement from 0.5 cm to 2.5 cm (0.5 cm, 1.0 cm, 1.5 cm, 2.0 cm, 2.5 cm). The viewing distance was 20 cm, and the simulated light source had an average α of 22.5 deg.

Figures 8 and 9 show data for two naive subjects. The probability the observer chose the test as moving further in depth is plotted against the shadow displacement β. Each of the five curves corresponds to a different test image size, shown in the legend box in the upper right panel. Discounting the shadow information would result in constant curves as a function of β with

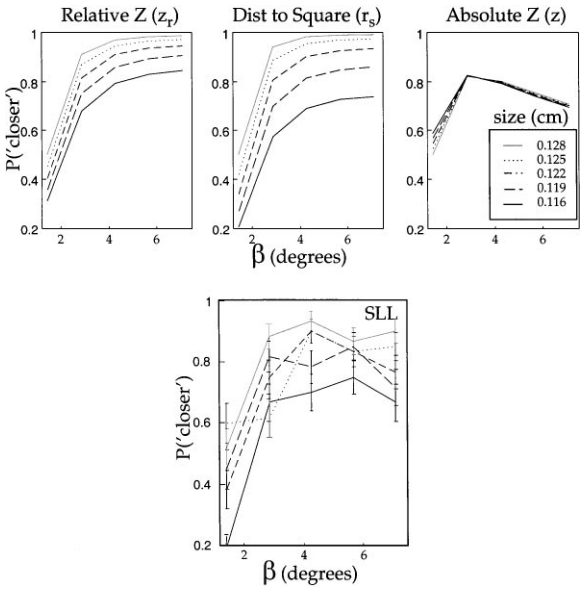

Figure 8. Data for one observer is shown in the bottom panel. The probability that the observer chose the test as moving further in depth is plotted against the shadow displacement β. Each of the five curves corresponds to a different test image size. Each probability is an estimate from 60 trials, and the error bars represent the standard errors of the estimate. The reference stimulus is the same as the test stimulus with the maximal image size and the minimal shadow displacement. The upper three panels show the probabilities predicted by the approximate cue integration models for the three representations. The model free parameters were set by maximum likelihood fits to the data.

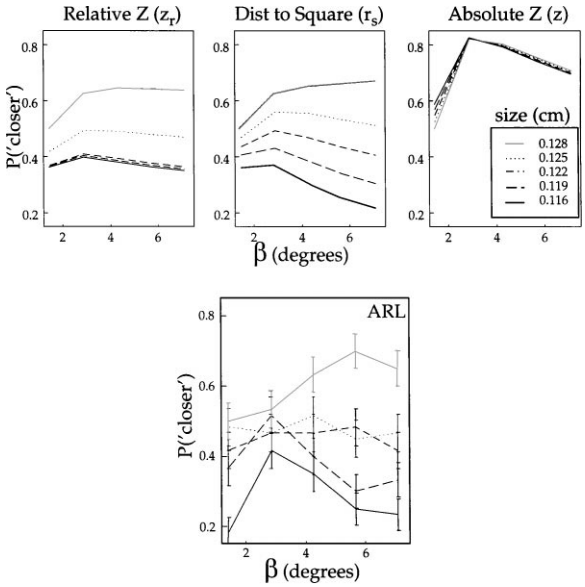

Figure 9. Data for a second observer is shown in the bottom panel. See Fig. 8 for details.

all the probabilities less than 0.5 (because the test image sizes are all less than the reference image size), while discounting image size information would result overlapping curves. For both subjects the curves are neither overlapping nor flat, demonstrating that observers do use both kinds of information. To assess whether observers were weighting the cues based on their reliability, we compared the human data to approximate performance of the three cue integration models.

The performance of the three different estimators on the task was approximated using the MAP estimate and Fisher Information equations. The optimal decision rule for the task is to choose the interval with the larger (smaller) MAP estimate of the distance from the background (from the observer). If we approximate the MAP estimates μ as being Gaussian, then we can use the fact that the inverse of the Fisher information is a lower bound on the variance of an unbiased estimator to write an approximate upper bound on performance. The decision variable is then normally distributed with mean given by the difference in map estimates, and the variance given by the sum of the reciprocals of the Fisher Information. This performance approximation is quite coarse. However, simulations showed that the networks had similar qualitative behavior. The performance of the three estimators is illustrated in the upper panels with the model free parameters set by maximum likelihood fits of the models to the data. The relative distance observer (Task 1) has two parameters, the sum

of the image size variance and the variance of the prior on square size, $\sigma_a^2 + \sigma_s^2$, and the mean of the prior on square size μ_s. The distance to square observer (Task 2) has both these free parameters and a third for the mean of the prior on r_b. The absolute distance observer (Task 3) has two free parameters μ_s and μ_{r_b}. Note that the behavior of the relative distance and the depth-from-observer models are qualitatively similar to both subjects' data, with the depth-from-observer model being the better predictor for the data sets of both subjects.

Note that the data from the two subjects are qualitatively different.[6] Subject ARL shows an initial increase in $p(\text{`}closer\text{'})$ with increasing shadow displacement β, followed by a decrease, especially for the smaller image sizes. The depth-from-observer model shows qualitatively similar behavior, when the prior expectation on the distance to the background μ_{r_b} is reduced by about 20% and the estimate of distance from image size has less uncertainty. Moving the assumed background forward has the effect of increasing the reliability of the shadow cue close to the background, which accounts for the initial increase in $p(\text{`}closer\text{'})$, while the reduction in assumed size-change noise causes the flattening of the curves.

The absolute distance model provides a poor fit to the data. The main cause is that the absolute distance model estimates are rather insensitive to changes in the values of the cues. This insensitivity causes a strong flattening and lack of variability between the performance curves. This inflexibility is not a result of the approximations, in that simulations show the same behavior. This suggests that the visual system is not optimized to compute the metric distances between objects.

5. Discussion

Given the computational cost of doing Bayes inference over traditional estimation (e.g. need to compute whole posterior, not just estimate), why might the expense be worth it? One reason could be that ensuring consistency is practical. Doing optimal cue integration with consistent cues allows very good estimation of scene variables from data, even when the number of data samples are less than the number of unknown scene variables and with very little prior knowledge. As an example, Fig. 10 shows the marginal distributions for all of the scene variables in the depth-from-observer network given only two image size and shadow position measurements, and flat priors on all the variables.

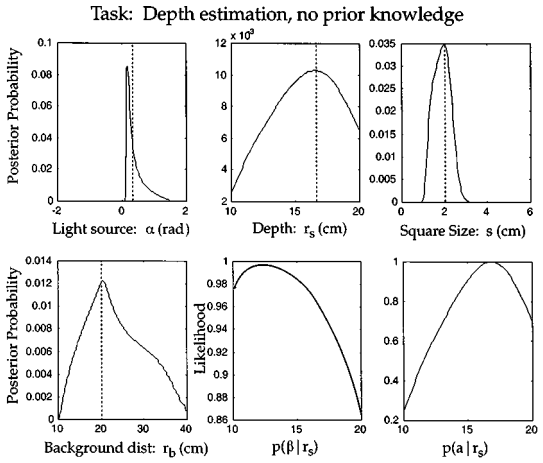

Figure 10. Simulation of the depth-from-observer network for just two data samples of a and β and uniform priors on all the variables. Curves in the first four panels represent the posterior distributions across each of scene variables. The dashed lines show the true value of each of the variables. The last two panels show the likelihood functions for r_s from the image size and shadow position data.

Dashed lines mark the true values of the scene variables, which were chosen from among the values used in the experiments. Notice that the MAP estimates are nearly correct for all four variables despite the broad posteriors. This result is in fact typical, and the plots shown were randomly generated (i.e. no intentional selection bias).

While the data do not unequivocally establish the human depth representation, the fact that the depth-from-observer model is more similar to the subjects' data is somewhat surprising. After all, we make perceptual decisions about the relative distances between objects all the time. Further, although the perception of depth from shadows and size is phenomenally quite strong (Kersten et al., 1997), observers can readily see the animations as simulations on a flat screen and hence unreachable. On the other hand, the visual system is highly adapted for reach. If the visual system can only optimize for one depth variable, then distance from the observer is a sensible one.

While the idea that the visual system is only optimized for certain variables may seem counter to the visual system being well described as an optimal estimator, this possibility would be a natural consequence of trying to perform optimal inference with a fixed computational architecture.

5.1. *Multiple Tasks*

Although we have shown how the computational burden of optimal inference can be reduced, our solution

required specialization by restricting inference to particular tasks. However, it is possible for an optimal inference system to remain computationally efficient and perform multiple tasks.

For a system that must perform multiple tasks using a common architecture, the scene variables that are shared between tasks should be constrained to have a common representation. This is an issue of great importance in trying to understand the trade-offs made by the human visual brain between flexibility and specialization (Goodale et al., 1994). For example, consider an object recognition task that only requires estimating the relative depths of points on the object. If another task involves a ballistic reach that requires metric depth information, then a metric depth representation may be more desirable for both tasks.

Because the estimation complexity of optimal inference varies as a function of the nuisance variables and conditional independence, it is computationally advantageous to choose the representation that results in the simplest architecture for the set of tasks.

Although we have discussed the simplifications afforded by complete statistical independence, we have laid the groundwork for the use of principled approximations. Basically we can trade off performance against estimation complexity. If variables are nearly independent, then there will be little cost to performance in treating them as independent. In practice we can break the links between variables by evaluating the weakly dependent nuisance variables S_x at their most probable values S_x^* to produce a Bayes modular system: $p(I_a, I_b \mid S_x, S_y) \rightarrow p(I_a, I_b \mid S_y, S_x^*) = p(I_a \mid S_y, S_x^*)\, p(I_b \mid S_y, S_x^*)$. Thus the order in which links should be broken can be determined by a measure of independence like the mutual information, and the cost of assuming independence can be assessed by comparing the approximate performance to the optimal performance.

6. Summary

This paper starts from the premise that a fundamental goal of a visual system is to make optimal statistical estimates of scene variables given some image data. We showed how a specific task and representation chosen for the scene variables affects the modularity and performance of an inference computation. We argued that these ideas lay the foundation for introducing approximations that may yield more efficient algorithms for optimal cue integration. We analyzed in detail Bayesian inference for a simple depth estimation task involving two disparate cues, image size and cast shadow position, for three different depth representations. From the analysis we generate predictions for human performance on a simple depth discrimination task from the optimal estimator using each representation. We found that human observers' decisions are near-optimal for certain depth representations, in that they weight the information from the two cues in accord with their informativeness.

Appendix A: Laplace Approximations to an Integral

A.1. Case 1: Likelihood Much Narrower than the Prior

We used Laplace's method (Tanner, 1996; Freeman, 1994) to construct analytic approximations to the required integrals, which are of the form:

$$\int_a^b p(x)\, p(I \mid x, S)\, dx \tag{33}$$

where $p(I \mid x, S)$ is the likelihood function of the data I given a particular nuisance scene variable x and the remaining (both required and nuisance) scene variables S. Let $p(I \mid x, S) = \exp(nh(I \mid x, S))$, and suppose $h(I \mid x, S)$, is smooth and unimodal as a function of x,[7] with a maximum at $\hat{x}(I, S)$, and x is a scalar. The parameter n represents the number of independent samples or the reciprocal of the measurement noise, that can go to infinity (i.e. the high sample/low noise limit).[8] Performing a Taylor series expansion of h around the point $\hat{x}(I, S)$ to second order, we have:

$$h(I \mid x, S) \approx h(I \mid \hat{x}, S) + (x - \hat{x}) \frac{\partial h(I \mid \hat{x}, S)}{\partial x} + \frac{(x - \hat{x})^2}{2!} \frac{\partial^2 h(I \mid \hat{x}, S)}{\partial x^2} \tag{34}$$

Because $\hat{x}(I, S)$ is the location that maximizes $h(I \mid \hat{x}, S)$, the linear term is zero. If $\hat{x}(I, S)$ is in (a, b) and $p(\hat{x}(I, S)) \neq 0$,[9] then the exponential term dominates the integral and we have the approximation:[10]

$$\int_a^b p(x)\, \exp(nh(I \mid x, S))\, dx \approx p(\hat{x})\, \exp(nh(I \mid \hat{x}, S))$$

$$\times \int_a^b \exp\left(n\frac{(x-\hat{x})^2}{2}\frac{\partial^2 h(I\mid\hat{x},S)}{\partial x^2}\right)dx$$
$$\approx p(\hat{x})\,\exp(nh(I\mid\hat{x},S))\sqrt{\frac{2\pi}{-n\frac{\partial^2 h(I\mid\hat{x},S)}{\partial x^2}}}$$

A.2. Case 2: Prior Width on Order of the Likelihood

In the approximation above, the assumption that the prior was broad with respect to the likelihood allowed us to neglect the prior except for the contribution of a scale factor at the maximum likelihood. We can handle the case where the width of the prior is on the order of the likelihood with a slight modification of the previous procedure. Let $p(I\mid x,S)=\exp(nk(I\mid x,S))$ and $p(x)=\exp(nm(x))$. Then let $h(I\mid x,S)=k(I\mid x,S)+m(x)$, and proceed as before, performing a Taylor expansion around the maximum of h.

Appendix B. Effect of Marginalizing Nuisance Variables on the Posterior

Problems frequently occur in which the image data given the scene variables can be expressed as

$$I = f(x,S)+\nu(x,S)$$

where the function f expresses the deterministic imaging equations, x represents the nuisance variable, S represents the remaining set of scene variables and $\nu(x,S)$ is a term representing the imaging noise, which is frequently a constant or a slowly-varying function of x and S.

If, in addition to being slow-varying, the measurement noise distribution has zero mean, and the likelihood admits a quadratic approximation,[11] then the effects of marginalization can be shown to dominate the resulting distribution.

We write the likelihood function $p(I\mid x,S)=\exp(-nk(I\mid x,S))$, where n is a parameter like the sample size that can be made large. To show the dominance of marginalization, we expand $k(I\mid x,S)$ in a Taylor series to second order in I about $I_{\max}=\arg\max_I k(I\mid x,S)=f(x,S)$:

$$k(I\mid x,S)\approx k(I\mid x,S)\mid_{f(x,S)} + \frac{1}{2}\frac{\partial^2 k(I\mid x,S)}{\partial I^2}\bigg|_{f(x,S)}(I-f(x,S))^2 \tag{35}$$

where we have used the fact the linear term goes to zero at $f(x,S)$.

Thus the likelihood can be written

$$p(I\mid x,S) \simeq g(x,S)\exp\left(-\frac{n}{2}(I-f(x,S))^2/\sigma(x,S)^2\right)$$

where $\frac{1}{\sigma(x,S)^2}=\frac{\partial^2 k(I\mid x,S)}{\partial I^2}\mid_{f(x,S)}$, and $g(x,S)=\exp(-\frac{n}{2}k(f(x,S)\mid x,S))$.

We wish to marginalize the distribution across x, and hence approximate the integral:

$$p(I\mid S)\approx \int p'(x,S)\times\exp\left(-\frac{n}{2}(I-f(x,S))^2/\sigma(x,S)^2\right)dx \tag{36}$$

where $p'(x,S)=g(x,S)p(x,S)$, which absorbs the g factor into the prior.

We now use Laplace's method to approximate this integral. In the previous section, we show that integrals of the form of Eq. (36) can be approximated by:

$$p(I\mid S)\approx p'(\hat{x},S)\,\exp(nh(I\mid\hat{x},S))\sqrt{\frac{2\pi}{-n\frac{\partial^2 h(I\mid\hat{x},S)}{\partial x^2}}} \tag{37}$$

where $\hat{x}(I,S)=\arg\max_x h(I\mid x,S)$, and $h(I\mid x,S)=-\frac{1}{2}(I-f(x,S))^2/\sigma(x,S)^2$.

However, because $I-f(\hat{x},S)=0$ for all I and S, $\exp(nh(I\mid\hat{x},S))$ is equal to one. In addition, it is easy to show that the second derivative of h with respect to x evaluated at $\hat{x}$ simplifies to:

$$\frac{\partial^2 h(I\mid\hat{x},S)}{\partial x^2}=\frac{(\partial f(\hat{x},S)/\partial x)^2}{\sigma(x,S)^2}$$

Using these simplifications, Eq. (37) reduces to:

$$p(I\mid S)\approx\frac{\sqrt{2\pi\sigma(\hat{x}(I,S),S)^2}}{|\partial f(\hat{x}(I,S),S)/\partial x|}p'(\hat{x}(I,S),S) \tag{38}$$

Acknowledgments

The authors thank Alan Yuille, James Coughlan, James Elder, and two anonymous reviewers for helpful comments and criticisms of previous drafts of this paper. This research was supported by NIH RO1 EY11507-001.

Notes

1. The change in dependence between the remaining variables depends on the causal and statistical influence relations between the variables.
2. E.g. the measurement noise distribution is dominated by its second moment or there are multiple samples, so that a quadratic approximation is justified on the basis of the central limit theorem.
3. In graph theory, a is called the *parent* of b.
4. The decision to measure the angle rather than some other related quantity like the projected shadow distance $l = \tan(\beta)$ matters little because the posterior is dominated by the marginalizations. It can be shown that the effect of measuring l amounts to replacing every instance of β in the formula with $\tan^{-1}(l)$.
5. For another set of trials whose data does not appear here, observers were asked to judge which square *appeared closer*. Although the two questions constituted different tasks, there was no measurable effect of question type on performance.
6. Kersten et al. (1997) report size change and shadow displacement results in a different experiment which also showed statistically significant differences between subjects in cue integration strategies.
7. Unimodal can be relaxed to having a dominant maximum, i.e. a global maximum orders of magnitude larger than the local maxima.
8. Although this technical condition is required for the validity of the asymptotic expansion, Laplace's method frequently works well even when $n = 1$.
9. For maxima at end points or vanishing $p(\hat{x}(I, S))$, the method yields slightly different approximations.
10. We have neglected a term involving the integration limits because for all the integrals we consider the term evaluates to 1.
11. E.g. the measurement noise distribution is dominated by its second moment or there are multiple samples, so that a quadratic approximation is justified on the basis of the central limit theorem.

References

Blake, A., Bulthoff, H.H., and Sheinberg, D. 1996. Shape from texture: Ideal observers and human psychophysics. In *Perception as Bayesian Inference*, D.C. Knill and W. Richards (Eds.). Cambridge University Press: New York, pp. 287–321.

Brainard, D.H. and Freeman, W.T. 1997. Bayesian color constancy. *J. Opt. Soc. Am. A*, 14(7):1393–1411.

Clark, J.J. and Yuille, A.L. 1990. *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Publishers: Boston.

Cutting, J.E. and Vishton, P.M. 1996. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of Space and Motion*, W. Epstein and S. Rogers (Eds.). Academic Press: San Diego, pp. 69–117.

Edwards, A.W.F. 1992. *Likelihood*. Johns Hopkins University Press: Baltimore.

Freeman, W.T. 1994. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368:542–545.

Goodale, M.A., Meenan, J.P., Bulthoff, H.H., Nicolle, D.A., Murphy, K.J., and Racicot, C.I. 1994. Separate neural pathways for the visual analysis of object shape in perception and prehension. *Current Biology*, 4(7):604–610.

Jensen, F.V. 1996. *An Introduction to Bayesian Networks*. Springer: New York.

Kersten, D., Knill, D., Mamassian, P., and Buelthoff, I. 1996. Illusory motion from shadows. *Nature*, 379(6560):31.

Kersten, D., Mamassian, P., and Knill, D.C. 1997. Moving cast shadows induce apparent motion in depth. *Perception*, 26: 171–192.

Knill, D.C. and Kersten, D. 1991. Apparent surface curvature affects lightness perception. *Nature*, 351:228–230.

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

Lawson, S., Madison, C., and Kersten, D. 1998. Depth from cast shadows and size-change: Predictions from statistical decision theory. *Investigative Opthamology and Visual Sciences (ARVO)*, 39(5):827.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann: San Mateo, CA.

Rao, C. 1973. *Linear Statistical Inference and Its Applications*. John Wiley and Sons: New York.

Stevens, S.S. 1957. On the psychophysical law. *The Psychological Review*, 64(3):153–181.

Tanner, M.A. 1996. *Tools for Statistical Inference*. Springer: New York.

Yuille, A.L. and Bulthoff, H.H. 1996. Bayesian decision theory and psychophysics. In *Perception as Bayesian Inference*, D.C. Knill and W. Richards (Eds.). Cambridge University Press: New York, pp. 123–161.