Proceedings of the IEEE Workshop on Statistical and Computational Theories of Vision, Fort Collins, CO, June, 1999. (Published on Web).

Statistical Structure and Task Dependence in Visual Cue Integration

Paul R. Schrater[1] and Daniel Kersten[1]

Abstract

A full Bayesian approach to vision requires consideration of potential interactions between all the variables in both the scene and image. A complete model of the interactions, however, would seem computationally intractable because of the large dimensionality of image measurements and scene properties. As a consequence, both experimental studies and theoretical models of human vision have relied on an assumption of modularity in which a particular scene property, such as object depth, is estimated from a restricted set of image measurements, such as image size. The computational problem is not hopeless, however, and can be surmounted by restricting the task and taking advantage of the statistical structure of the problem. In a Bayesian context, modularity falls out of the conditional independencies in the joint distribution of scenes and images $p(S, I)$. By conditioning the joint distribution with respect to particular inference tasks, further modularity is possible while preserving optimal cue combination. We illustrate the problem of modularity and cue combination for the perception of depth from two highly disparate cues, cast shadow position and image size. While strong modularity would suggest ad hoc or no cue combination, we find that the performance of human subjects is better predicted by near-optimal cue combination.

1. Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 E. River Dr., Minneapolis, MN 55455, USA.

1 Introduction

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.[8]). Our understanding of depth perception has largely developed through the experimental study of single cues in isolation. Although, there have been a few experimental studies of cue combination pairs[6][19][25], testing all possible cue combinations is impossible. Thus our understanding and modeling of how these cues interact in everyday vision presents an empirical challenge: How can we test and quantitatively model the interactions of multiple cues given the complexities of natural images? Our proposed solution is to develop ideal observer models of optimal cue combination which provide the bases for specific testable hypotheses of human perception. We, of course, expect departures from optimality for any real system; but an ideal observer provides the baseline default model from which new models are created. This strategy, at least in theory, makes the scientific problem tractable. But, one could argue, that all we have done is to change the impossible empirical problem into a theoretically intractable one. Our primary goal is to argue that optimal Bayesian theories of depth cue integration can be developed by exploiting task dependency and the statistical structure of the depth estimation problem. We illustrate these ideas with an analysis of depth estimation from image size and cast shadow position cues.

1.1 Why do Bayesian Cue Integration?

There is a long tradition of treating modularity as fundamental. Marr was "...moved to elevate (modularity) to a principle"[22]. Most ad hoc modularity schemes begin with several different image measurements which are related to the scene variable to be estimated, and then assume that if the image measurements are functionally separable, they should produce independent estimates. However, the statistical independence of image measurements with respect a scene variable depends on the joint distribution, $p(S, I)$, of scene and image variables. We will show that modularity is determined by the statistical independence structure of the joint distribution.

The use of ad hoc modularity creates problems for cue combination[7, 19]. 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 several scene variables which depend on the same image measurements. In this case, optimal estimation must treat 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 it is in principle impossible to derive separate optimal estimators of surface geometry from different image measurements[18].

In contrast, Bayesian inference insures consistent inferences and combination of cues based on the confidence in the estimates.

2 Probabilistic Approaches to Scene Estimation

2.1 Modeling $p(S, I)$

Probabilistic approaches to scene estimation require the specification of $P(S, I)$, the joint probability distribution of scene, S and image variables I. This joint distribution contains all the required information for making optimal inferences and doing optimal encoding of the image information. For example, marginalizing the joint distribution over S yields $p(I) = \int_S P(S, I)dS$, which specifies the distribution of images. A great deal of recent work on image coding has involved seeking compact representations of $p(I)$, typically using redundancy reduction principles[2, 27, 9, 23, 33, 30, 28, 29].

For the problem of inferring scene descriptions from image measurements, we use Bayes rule to write the posterior probability as:

$$P(S|I) = \frac{P(S, I)}{P(I)} = \frac{P(I|S)P(S)}{\int_S P(I|S)P(S)dS} \tag{1}$$

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

2.2 Task Dependency

Although modeling $p(S, I)$ is theoretically possible, the cost of doing so for the entire ensemble of scenes and images an observer could encounter is prohibitive. However, for most inference tasks we are only interested in a small subset of the variables contained in the set S. Thus, for particular tasks, S can be replaced by a subset of related variables (e.g. object motion, light source direction, etc.), and I by a small set of required image measurements. Even considering the union of the set of tasks the visual system performs, the number of variables required by this union will be orders of magnitude less (arguably a different cardinality) than the variables required to describe $p(S, I)$ completely. While it is clear that restricting the domain of expertise of the visual system to a limited number of tasks appreciably relaxes the computational burden, the complexity can be further reduced if we take advantage of the statistical structure of $p(S, I)$ restricted to the task.

We consider a task as specifying four things, the required set of scene variables S_r, the nuisance (e.g. generic [11]) scene variables S_g, the scene variables which are presumed known S_f, and the decision to be made. Each of the four components of a task plays a role in determining the structure of the optimal inference computation. We show that S_r and S_f can be used to simplify the joint distribution through independence relations, while S_g and the decision rule can make one choice of S_r simpler than another.

2.2.1 Factoring Distributions and Conditional Independence

When the joint distribution factors due to statistical independence:

$$p(S, I) = p(I|S_1)p(I|S_2)p(S_1)p(S_2),$$

then we can ignore the variables in S_2 when making inferences on variables in the set S_1. Thus, the first simplification is to factor $p(S, I)$ into two parts, one of which contains all the variables which are statistically independent of S_r and the other which contains all of the dependent variables, $p(S, I) = p(I_{ind}|S_{ind})p(I_{dep}|S_{dep})p(S_{ind})p(S_{dep})$.

In most cases, the nature of a task fixes some of the scene variables S_f. 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 sizes 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.

Since the variables in S_f are presumed known, we can condition $p(I_{dep}, S_{dep})$ on S_f, $p(I_{dep}, S_{dep}|S_f)$, which increases the statistical independence of the variables. In general, conditioning produces independence relations which can be exploited for cue combination and cooperative computation. We expect the conditional distribution to further decompose into relevant and irrelevant scene variables.

Thus given the task, we can first factor $p(S, I|S_f) = \prod_{i=1}^{N} p(S_i, I|S_f)$. To do inference we need only consider the factors in which the S_i contain the variables in S_r. Let S_j denote the minimal set of statistically dependent variables containing S_r. The variables in S_j excluding S_r are just the nuisance variables S_g. Then, $p(S_g, S_r, I|S_f)$ contains all the information we need to perform the inference task, and has automatically specified the task relevant vs. irrelevant variables. Thus the independence structure determines which variables should be involved in an inference computation. However, conditional independence structure also determines which variables interact, which has consequences for data fusion and cue combination.

The probabilistic structure of the joint probability distribution $p(S, I)$ can be represented by a Bayes Net[24, 12], which is simply a graphical model which expresses the conditional independences between the variables. Using labels to represent variables and arrows to represent conditioning (with $a \rightarrow b$ indicating b is conditioned on a[1]), independence can be represented by the absence of connections between variables. Using these graphical models we can determine the interactions between variables by inspection. 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 data cues to produce independent estimates of variable x, what determines modularity in a Bayesian inference is whether or not the data are conditionally independent given x. When this is true, we can produce separate likelihood functions for x, which can be combined by multiplication, a property we call *Bayesian modularity*. Graphically, this requirement is that the data are singly connected to the variable of interest. Figure 1 shows examples of a singly connected net and a non-singly connected net. The non-singly connected net corresponds to the case in which more than one scene variable depends on the data cues, which is exactly the case that calls for cooperative computation.

[1] In graph theory, a is called the *parent* of b

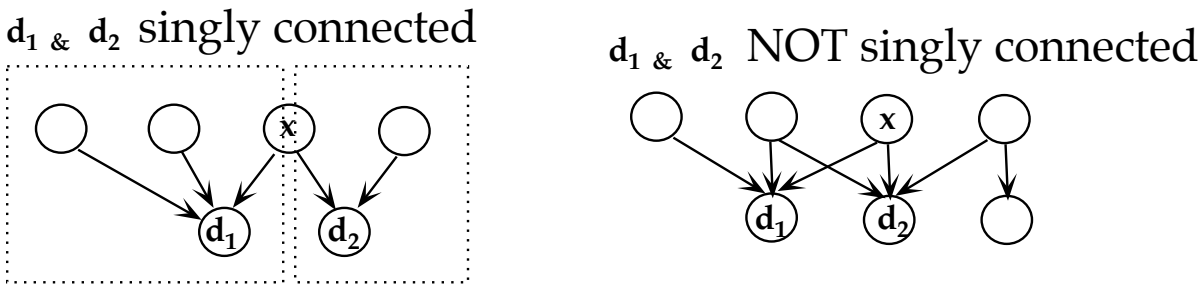

Figure 1: Whether independent data measures are singly connected to the estimated variable x determines whether or not estimation modules can be created for x.

2.2.2 Marginalization and Decision Rule

Bayesian decision theory provides a precise language to model the costs of errors determined by the choice of visual task[32][5]. The *risk* $R(\Sigma; I)$ of guessing Σ when the image measurement is I is defined as the expected loss:

$$R(\Sigma; I) = \int_S L(\Sigma, S) P(S \mid I) dS,$$

with respect to the posterior probability, $P(S|I)$. The best interpretation of the image can then be made by finding the Σ which minimizes the risk function. One possible loss function is a delta function $\delta(\Sigma - S)$. In this case the risk becomes $R(\Sigma; I) = -P(\Sigma \mid I)$, 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. For simplicity, we will assume the former or latter of these loss functions depending on whether the variable is needed or not. Thus, we estimate the most probable relevant scene value (MAP estimation), while marginalizing with respect to the irrelevant generic variables. While the statistical structure of the joint distribution determines which variables interact, the choice of decision rule and marginalization variables determine the details of how they interact.

3 Implications for Psychophysics

Does the visual system do Bayesian inference? If we assume the visual system is optimized for a limited number of tasks, there are two kinds of predictions: characteristic successes and characteristic failures. Characteristic successes denote cases when the visual system behaves optimally. One of the key predictions is *confidence-driven* cue combination, in which observers use information based on its reliability. Evidence for confidence driven use of texture information in judgements of surface orientation has been shown in several studies[4, 16, 17, 31] by several authors. Another key prediction is that consistent interpretations of related scene properties like surface geometry and shading are preferred over inconsistent ones. Several lightness illusions rely on exactly this property [15, 1]. We should also be able to predict which variables interact directly from the conditional independence relations.

Both Bayesian and non-Bayesian visual systems will show sub-optimal performance for tasks which they are not designed for. However, a Bayesian system will show characteristic failures for a set of related tasks which require optimal inference on different parametrizations of a set of scene variables. For instance, a visual system which is optimized to compute the relative depths of objects will show characteristic failures when asked to compute absolute depth.

The set of scene variable we do our inference on matter because Bayesian inference is not invariant to reparametrizations. Thus if we perform optimal inference on one variable, we cannot just transform the result to get optimal inference on another 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 [10]. While this fact has been used to critique Bayesian inference [10], 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.

In the next section we perform a detailed analysis of Bayesian inference on a simple scene, to compare several of these predictions to psychophysical data. In particular, we investigate whether we can predict which variables interact, whether cue combination is confidence-driven, and how ideal performance varies given different parametrizations of the observer's decision variable.

4 Estimating Depths from Image Size and Shadows

We illustrate the dependence of Bayesian cue combination on task demands and conditional independences with a simple scene due to Kersten et al.[13]. 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 of the stimuli is shown in figure 2.

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 coordinate frame. Cast shadow position is determined by the *allocentric* distance of the square from the background and the position of the light source. Thus to combine the shadow and image size data, we must convert one of the variables into a common coordinate frame.

From the standpoint of traditional estimation, a strong case can be made not to combine 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

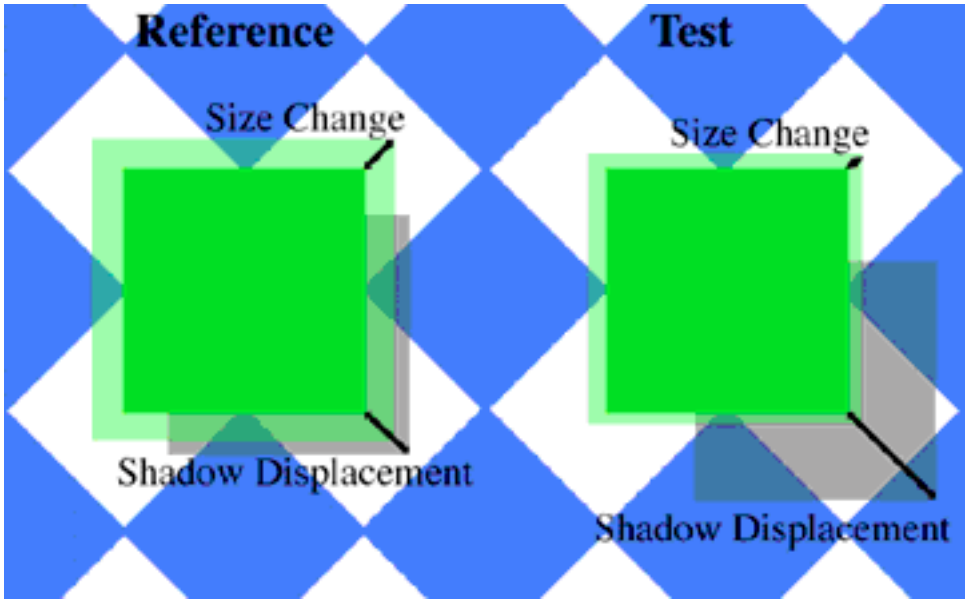

Figure 2: An illustration of the stimuli used in the experiment. Two movies depicting a square moving in depth are sequentially shown to the observer. The image size of the square becomes larger and the shadow moves away from the square with decreasing depth from the checkerboard background. The image 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 appeared closer at the end of the movie in a two-alternative forced-choice method.

larger image closer. Similarly, treating the shadow information and 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 combine the cues,and instead make separate judgements of depth from the cues.

In contrast, Bayesian inference requires choosing a common coordinate frame to combine the cues. However, to combine the cues the size of the square and the light source direction can no longer be neglected. We considered three possible common coordinate frames to do the inference. Each of these leads to a different Bayes net and different optimal inference structure. For each of the three tasks, however, the best way to judge the depths of the two squares is to compute decision variables consisting of MAP depth estimates for both intervals and choose the smaller (closer to the observer) value.

4.1 Task 1: Estimating Relative Distance from Background (z_r)

The geometric diagram in figure 3 defines all of the relevant variables for the task. One way of judging the depths of the two squares is to compute the distance from the background. This leaves 4 unknowns, $\alpha, s, z, \& r_b$ with only two data variables. If the observer scales the distance from the background z, and the object size s by the distance from the observer

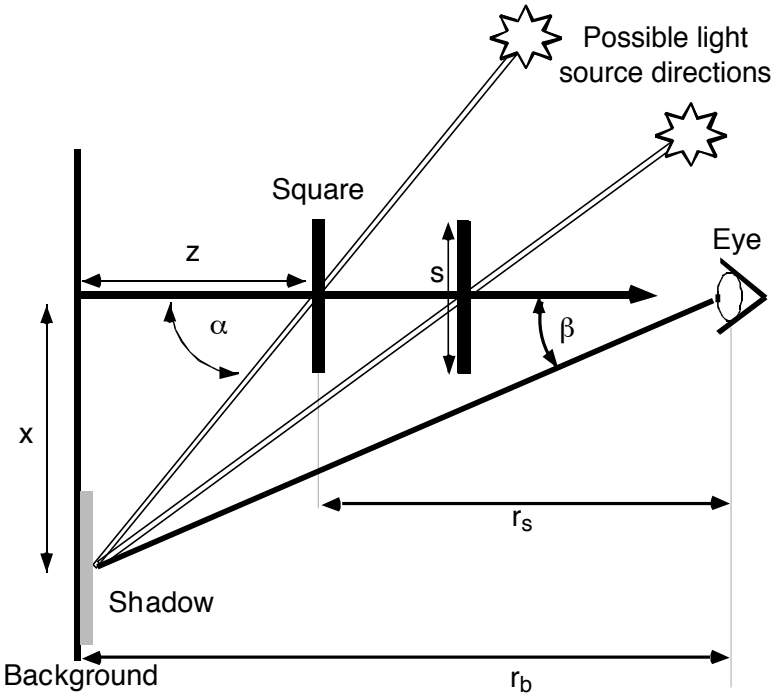

Figure 3: Diagram illustrating the problem of inferring depth from image size and cast shadow position in 1-D for the central square in front of a checkerboard background (see figure 2). There are three depth variables, distance to the background r_b, distance to the square r_s, and the distance of the 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 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.

to the background r_b, then estimates of the relative distance from the background can be made without having to deal with 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. 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 task 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), is a function of z_r and light source position α:

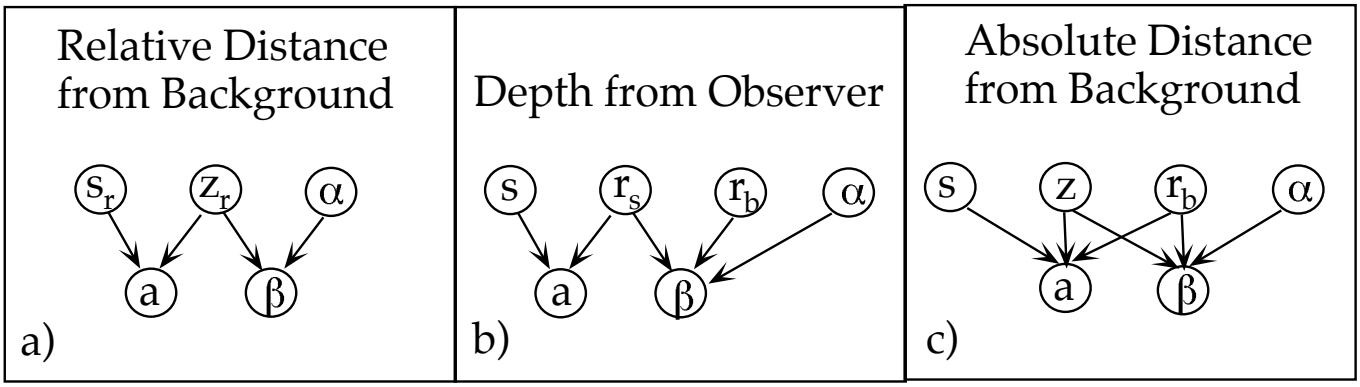

Figure 4: Bayes nets for the three tasks. **a)** Bayes net for relative distance to the background. This task involves estimating object relations (world centered), and requires the least prior knowledge. **b)** Bayes net for distance to observer. Notice that the use of the shadow information requires 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.

$$\beta = \tan^{-1}(z_r \tan(\alpha)) + n_\beta \tag{2}$$

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:

$$p(\beta|z_r, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp(-\frac{(\beta - \tan^{-1}(z_r \tan(\alpha))^2}{2\sigma_\beta^2}) \tag{3}$$

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. Since both s_r and $1 - z_r$ are physically constrained to be positive, we modeled the size measure noise as log normal. Then the likelihood for a is given by:

$$p(a|z_r, s_r) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp(-\frac{\log(\frac{s_r}{1-z_r})^2}{2\sigma_a^2}) \tag{4}$$

To estimate z_r we compute $p(z_r|\beta, a)$. Assuming that the measurements of the image size a and the shadow position β are independent, $p(z_r|\beta, a)$ can be written:

$$
\begin{aligned}
p(z_r|\{\beta\},\{a\}) &= \frac{p(\{\beta\}|z_r)p(\{a\}|z_r)p(z_r)}{p(\{\beta\},\{a\})} \\
p(z_r|\{\beta\},\{a\}) &\propto p(\{\beta\}|z_r)p(\{a\}|z_r)p(z_r) \\
&= \left(\int_\alpha \prod_{i=1}^{N} p(\beta_i|z_r,\alpha)p(\alpha)d\alpha\right)\left(\int_{s_r}\prod_{i=1}^{N} p(a_i|z_r,s_r)p(s_r)ds_r\right)p(z_r),
\end{aligned}
$$

where N is the number of measurements. The Bayes net which corresponds to this inference is shown in figure 4a. Note that this network is Bayes modular, which shows up in the factoring of the likelihoods above.

4.2 Task 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 from the observer to the squares r_s. The Bayes net for this inference is shown in figure 4b. In this case the shadow position must be converted to an observer coordinate frame. Using $r_b = z + r_s$, we can write the shadow position measurement as:

$$
\beta = \tan^{-1}(\left(1 - \frac{r_s}{r_b}\right)\tan(\alpha)) + n_\beta \tag{5}
$$

The likelihood function is given by:

$$
p(\beta|r_s,r_b,\alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta}\exp(-\frac{(\beta - \tan^{-1}(\left(1-\frac{r_s}{r_b}\right)\tan(\alpha))^2}{2\sigma_\beta^2}) \tag{6}
$$

The image size a is given by:

$$
a = \frac{s}{r_s} + n_a
$$

Hence the likelihood for a is given by:

$$
p(a|r_s,s) = \frac{1}{\sqrt{2\pi}\sigma_a a}\exp(-\frac{\log(\frac{s}{r_s})^2}{2\sigma_a^2}) \tag{7}
$$

To base the decision on r_s, we compute $p(r_s|\beta,a)$:

$$
\begin{aligned}
p(r_s|\{\beta\},\{a\}) &\propto p(\{\beta\}|r_s)p(\{a\}|r_s)p(r_s) \\
&= \left(\int_{r_b}\int_\alpha \prod_{i=1}^{N} p(\beta_i|r_s,r_b,\alpha)p(\alpha)p(r_b)d\alpha\, dr_b\right)\left(\int_s \prod_{i=1}^{N} p(a_i|r_s,s)p(s)ds\right)p(r_s)
\end{aligned} \tag{8}
$$

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 task, the uncertainty in our shadow depth estimates increases as compared with the relative distance task (Task 1).

4.3 Task 3: Estimating Absolute Distance to Background (z)

Finally, the observer could compute z, the absolute distance from the square to the background. This requires conversion of the image size information into object coordinates. The computation also involves a second unknown for both cues, the distance to the background r_b. The Bayes net which corresponds to this inference is shown in figure 4c. The measurements can be written in terms of z as:

$$\beta = \tan^{-1}(z\tan(\alpha)/r_b) + n_\beta \tag{9}$$

$$a = \frac{s}{r_b - z} + n_a.$$

The likelihood functions are:

$$p(\beta|z, r_b, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta}\exp(-\frac{(\beta - \tan^{-1}(z\tan(\alpha)/r_b))^2}{2\sigma_\beta^2}) \tag{10}$$

$$p(a|z, r_b, s) = \frac{1}{\sqrt{2\pi}\sigma_a a}\exp(-\frac{\log(\frac{s}{r_b - z})^2}{2\sigma_a^2}) \tag{11}$$

To estimate z we compute $p(z|\beta, a)$:

$$\begin{aligned} p(z|\{\beta\},\{a\}) &\propto p(\{\beta\},\{a\}|z)p(z) \\ &= \left(\int_{r_b}\left(\int_\alpha \prod_{i=1}^n p(\beta_i|z, r_b, \alpha)p(\alpha)d\alpha\right)\left(\int_s \prod_{i=1}^n p(a_i|z, r_b, s)p(s)ds\right)p(r_b)\,dr_b\right)p(z) \end{aligned} \tag{12}$$

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 combination that we explore below.

4.4 MAP Estimates

To derive formula for the MAP estimates of square depth for the three models, we found analytic approximations to the required marginalization integrals using Laplace's method [3, 11, 21]. In Laplace's method integrals of the form:

$$F(\sigma^2) = \int_a^b f(x)\exp(h(x)/\sigma^2)dx \tag{13}$$

can be well approximated in the low noise limit $\sigma^2 \to 0$. If the maximum c of $h(x)$ is in (a, b) and $f(c) \neq 0$,[2] then by expanding $h(x)$ in a second order Taylor series about c, the integral is asymptotically:

$$F(\sigma^2) \sim \frac{\sqrt{2\pi\sigma^2}}{\sqrt{|h''c|}}f(c)\exp(h(c)/\sigma^2) \tag{14}$$

[2]For maxima at end points or vanishing $f(c)$, the method yields slightly different approximations.

4.4.1 Task 1: MAP Estimate for z_r (Relative z)

When the prior on α is uniform, the marginalization step can be approximately evaluated:

$$\int_\alpha \prod_{i=1}^{n} p(\beta_i|z_r,\alpha)p(\alpha)d\alpha \simeq \frac{\sqrt{2}z_r}{\pi(z_r^2\cos(\hat{\beta})^2+\sin(\hat{\beta})^2)} \tag{15}$$

where $\hat{\beta}$ is the mean of the N sample βs.
The maximum z_r occurs at:

$$max_{z_r}\left(p(\beta|z_r)\right) = \tan(\hat{\beta}). \tag{16}$$

For the size change cue, some knowledge of the relative size is crucial to compute the relative distance. In the absence of a peaked prior, it is easy to show that the optimal estimate of z_r is always zero. Thus we marginalized with respect to a log normal prior on s_r yielding:

$$p(\{a\}|z_r) = \frac{1}{\sqrt{\pi(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)}\hat{a}} exp(-\frac{\log(\hat{a}(1-z_r)/\mu_{s_r})^2}{2(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)}) \tag{17}$$

where $\hat{a}$ is the geometric mean of the N samples, and $\sigma_{\hat{a}}^2 = \sigma_a^2/N$. The maximum z_r with respect to image size occurs at

$$max_{z_r}\left(p(\{a\}|z_r)\right) = 1-\mu_{s_r}/\hat{a} \tag{18}$$

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

4.4.2 Task 2: MAP Estimate for r_s

We find the optimal estimator for the shadow cue as we did previously, with the exception that we need to marginalize over the additional unknown r_b, the distance to the background. We assumed a log normal prior on r_b. This yields two asymptotic approximations, one for small uncertainty on the prior $\sigma_{r_b}^2$ and one for large $\sigma_{r_b}^2$. The small $\sigma_{r_b}^2$ approximation was used in our data analysis and is shown below:

$$\int_{r_b} p(\{\beta\}|r_s,r_b)p(r_b)dr_b \simeq \frac{2(1-r_s/\mu_{r_b})}{\pi((1-r_s/\mu_{r_b})^2\cos(\hat{\beta})^2+\sin(\hat{\beta})^2)} \tag{19}$$

The maximum r_s occurs at:

$$max_{r_s}\left(p(\{\beta\}|r_s)\right) = \mu_{r_b}(1-\tan(\hat{\beta})). \tag{20}$$

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

$$p(\hat{a}|r_s) = \frac{1}{\sqrt{\pi(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)}\hat{a}} exp(-\frac{\log(\hat{a}r_s/\mu_{s_r})^2}{2(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)}). \tag{21}$$

The maximum r_s with respect to image size occurs at

$$max_{z_r}\left(p(\{a\}|z_r)\right) = r_s = \frac{\mu_s}{\hat{a}}. \tag{22}$$

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	$z = \frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))}$		$\frac{2\hat{a}^2(1-\tan(\hat{\beta}))^4}{(\mu_s^2\tan(\hat{\beta})^2}$	

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

4.4.3 Task 3: MAP Estimate for z

In optimal estimation of z we cannot consider the shadow cue and image size cues separately. Instead the joint distribution must be marginalized over r_b. The asymptotic approximation to the posterior is:

$$p(z|\{\beta\},\{a\}) \propto \frac{\mu_s z\csc(\hat{\beta})\sec(\hat{\beta})}{\sqrt{2}\hat{a}\sqrt{\hat{a}^2z^2+\mu_s^2(\sigma_{\hat{a}}^2+\sigma_s^2)\tan(\hat{\beta})^2}}\exp\left(-\frac{(z-\frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))})^2}{2\frac{\hat{a}^2z^2+\mu_s^2(\sigma_{\hat{a}}^2+\sigma_s^2)\tan(\hat{\beta})^2}{\hat{a}^2(1-\tan(\hat{\beta}))^2}}\right)p(z) \tag{23}$$

The exact MAP estimator for this equation is complicated, but can be approximated by:

$$max_z\left(p(z|\{\beta\},\{a\})\right) \simeq \frac{\mu_s\tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))} \tag{24}$$

for the range of $\hat{a}$ and $\hat{\beta}$ used in the experiments.

4.5 Fisher Information

A lower bound on the variance of unbiased estimators is given by the reciprocal of the Fisher Information [26]. The Fisher Information is given by:

$$\mathcal{I}(x) = -N\int_{data} p(data|x)(\partial^2\log p(data|x)/\partial x^2)d(data) \tag{25}$$

Recognizing the second derivative of the log of $p(data|x)$ as an estimate of the inverse variance of the Gaussian approximation to the likelihood function on x, we can interpret the Fisher Information as the expected approximate variance of the likelihood function.

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 [4, 26]. Let m_a denote the MAP estimate and $\mathcal{I}_a(x|m_a)$ the Fisher information for the image size cue, and m_β the MAP estimate and $\mathcal{I}_\beta(x|m_\beta)$ 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:

$$m_{best} = \frac{m_a\mathcal{I}_a(x|m_a)+m_\beta\mathcal{I}_\beta(x|m_\beta)}{\mathcal{I}_a(x|m_a)+\mathcal{I}_\beta(x|m_\beta)}. \tag{26}$$

which is a specific prediction of a confidence-driven decision.

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

$$\frac{1}{\mathcal{I}_a(x|m_a) + \mathcal{I}_\beta(x|m_\beta)} \tag{27}$$

These estimates are also the expected MAP estimates for cues which are consistent (i.e. the likelihood functions have similar maxima). We computed Fisher information for each of the independent depth likelihood functions. The MAP estimates and Fisher information values are summarized in table 1.

Because r_s and z_r are related by a linear transformation we know the probability distributions should transform gracefully. However, note that our MAP estimate for z is not what we would expect from weak fusion, nor can it be produced by converting either the $z_{r_{best}}$ or the $r_{s_{best}}$ to z. Thus, in this case strong fusion has resulted from marginalization.

In a Bayesian context, linear combination is only appropriate for Bayes nets with certain properties. For Bayes nets which are modular and the data are consistent, a linear combination rule, inversely weighted by the variances of the estimates is optimal. When the Bayes net is modular, we can compute the estimates for linear changes of variables directly from the linear transform of the estimates, given precise knowledge of any unknowns involved in the transform. Although the z_r and r_s estimates are compatible in this way, it is important to point out that depth decisions based on these estimates can substantially differ.

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 estimation tasks, 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. Note that the information mirrors our expectation about the natural coordinate frames for the two cues.

5 Human Performance

We performed a shadow and image size cue combination experiment to investigate whether or not human observers make Bayesian-like use of both cues to estimate the depth of the square [20].

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. Responses were recorded via a mouse button click. In the reference image, size change was maximal (128%) and shadow displacement was minimal (0.5 cm). In the test image, size change ranged from 116% to 128% (116%, 119%, 122%, 125%, 128%) 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 5 & 6 show data for two naive subjects. The probability the observer chose the test as appearing closer 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 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 combination models.

The performance of the three different estimators on the task was approximated using the estimator 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 informations. 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[3]. Subject ARL shows an initial increase in $p('closer')$ followed by a decrease 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. The decreased uncertainty in the image size cue coupled with the decreasing effectiveness of the shadow cue with β cause the flattening of the curves and the downward trend. The downward trend can be briefly offset, however, by decreasing the expected background distance, which increases the informativeness of the β cue.

Although the absolute distance approximation is poorer than the other two, the qualitative behavior of the model and the simulations least resembles the subjects' data. This suggests that the visual system may not be optimized to compute the metric distances

[3]Kersten et al. [14] report size change and shadow displacement results in a different experiment which also showed statistically significant differences between subjects in cue combination strategies.

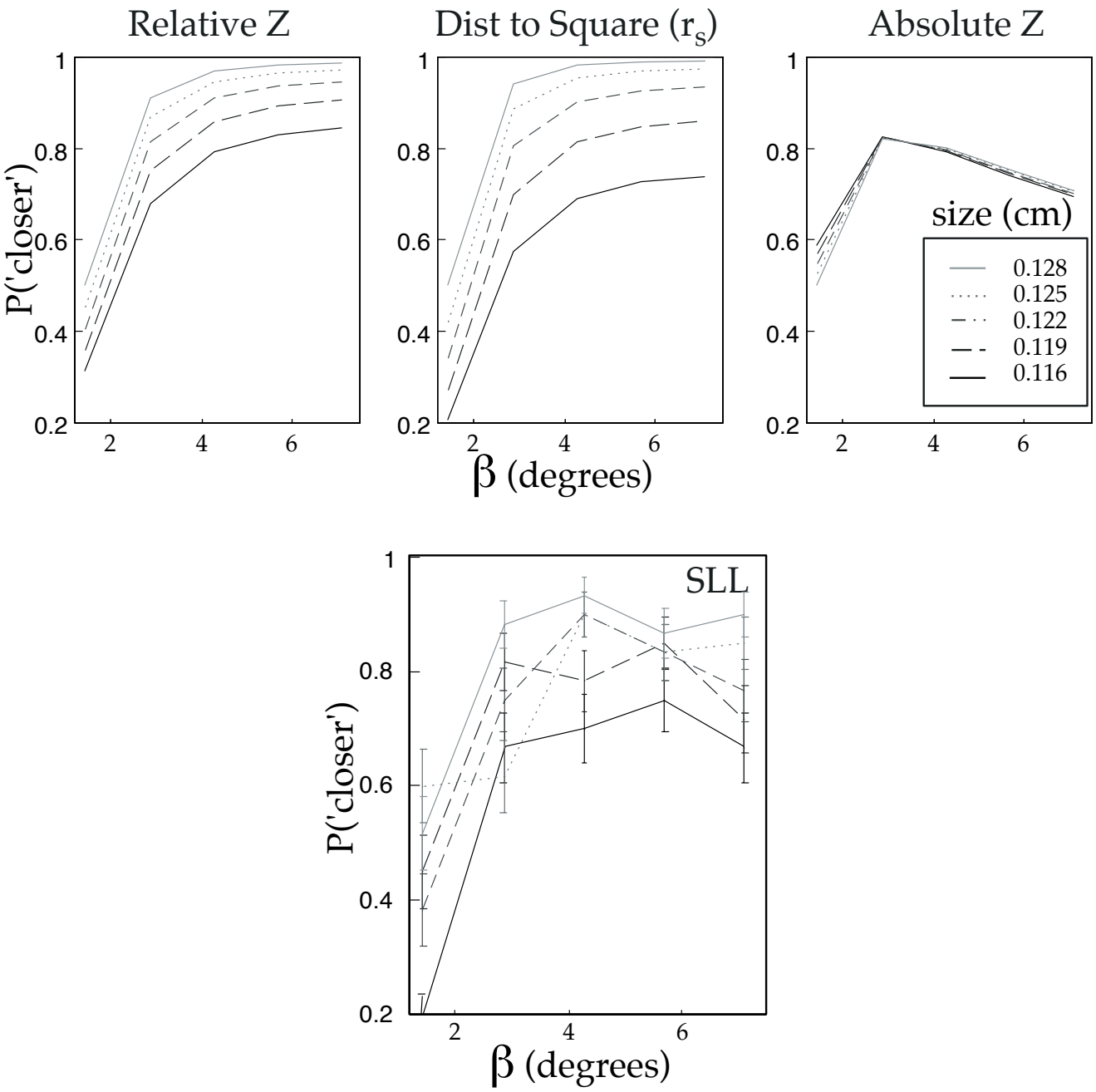

Figure 5: Data for one observer is shown in the bottom panel. The probability the observer chose the test as appearing closer is plotted against the shadow displacement β. Each of the five curves corresponds to a different test image size. Each probability is an estimate from 60 trials, and the error 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 combination models for the three tasks. The model free parameters were set by maximum likelihood fits to the data.

between objects.

While the data are preliminary, 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 [14], 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.

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 combination 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, figure 7 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. Dashed lines mark the true values of the scene variables. Notice that the MAP estimates are nearly correct for all four variables.

6 Summary

We have argued that a fundamental goal of the visual system is to model the joint distribution $p(I, S)$ subject to task constraints. While modeling $p(I, S)$ completely is intractable, a visual system which which is only required to be optimal on a limited number of tasks can considerably simplify the problem by exploiting conditional independence to reduce the number of required variables and the complexity of the relations between variables. We contrasted Bayes inference and more traditional estimation schemes which are driven by an early, and sometimes premature, commitment to modularity. We analyze in detail Bayesian inference for a simple depth estimation task involving two disparate cues, image size and cast shadow position, for three different coordinate frames. From the analysis we predict performance on a simple depth discrimination task from the optimal cue combination in each coordinate frame. We find that observers' decisions are confidence-driven, in that they weight the information from the two cues in accord with their informativeness.

References

[1] E. H. Adelson. Perceptual organization and the judgement of brightness. *Science*, 262:2042–2044, 1993.

[2] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992.

[3] C. M. Bender and S. A. Orszag. *Advanced mathematical methods for scientists and engineers*. McGraw-Hill, Inc., New York, 1978.

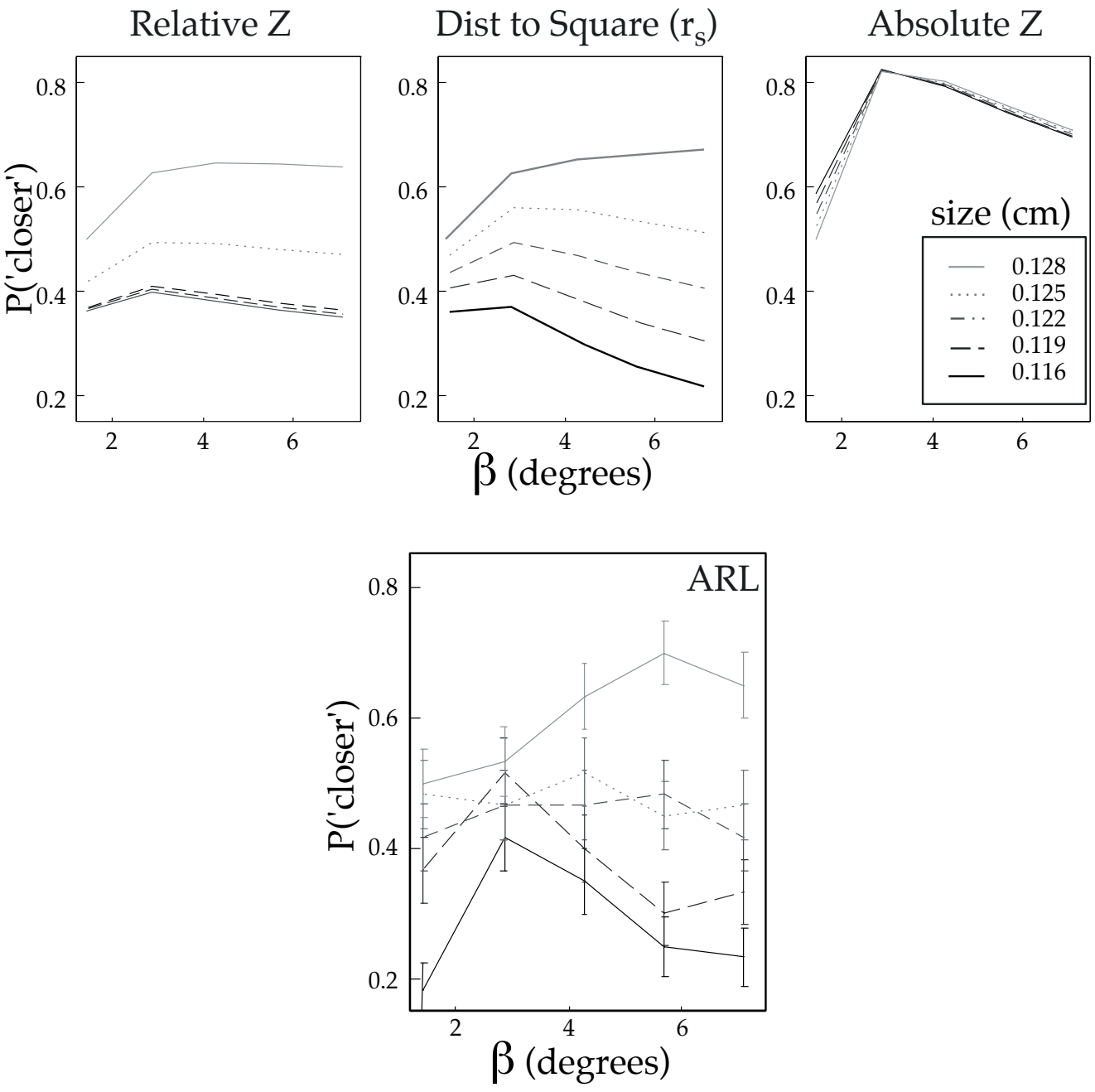

Figure 6: Data for a second observer is shown in the bottom panel. See figure 5 for details.

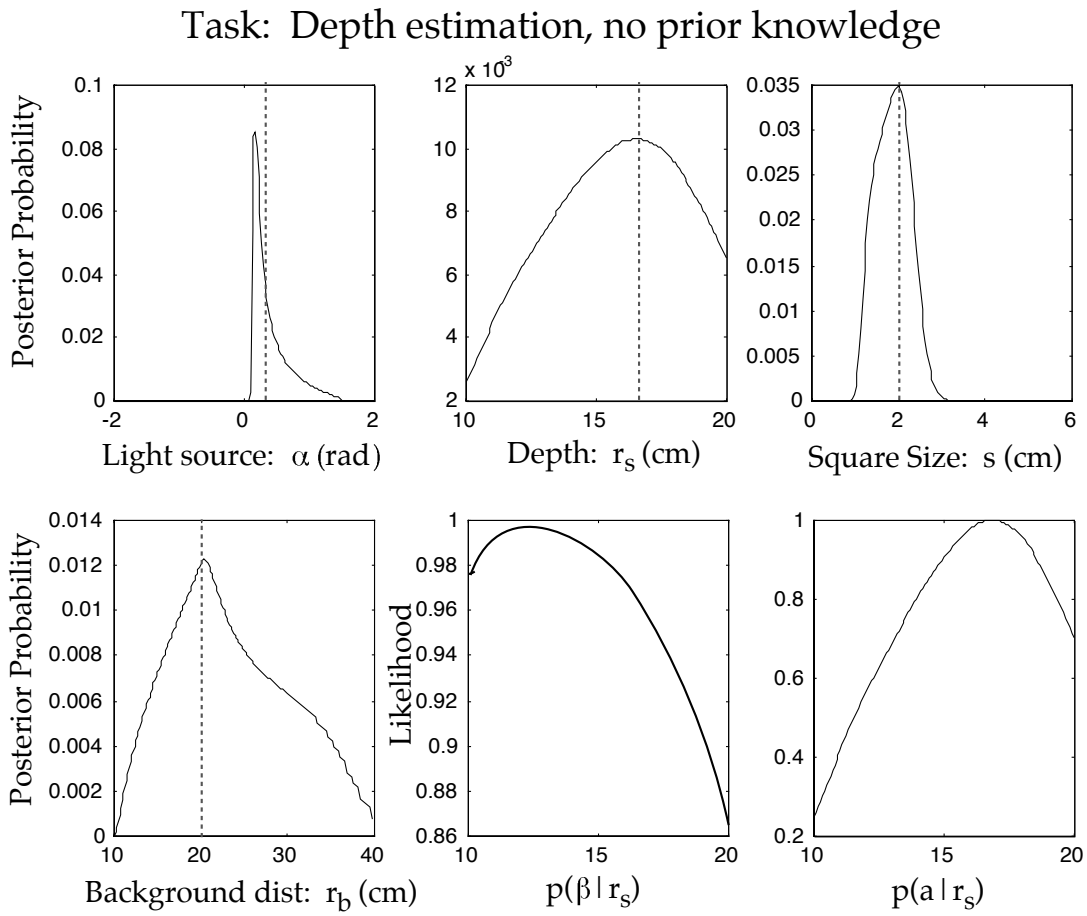

Figure 7: Simulation of the depth-from-observer network for just two data samples of a and β and uniform priors on all the variables. Curves in the first four panels represent the posterior distributions across each of scene variables. The dashed lines show the true value of each of the variables. The last two panels show the likelihood functions for r_s from the image size and shadow position data.

[4] A. Blake, H. H. Bulthoff, and D. Sheinberg. Shape from texture: Ideal observers and human psychophysics. In D. C. Knill and W. Richards, editors, *Perception as Bayesian inference*, pages 287–321. Cambridge University Press, New York, 1996.

[5] D. H. Brainard and W. T. Freeman. Bayesian color constancy. *J Opt Soc Am A*, 14(7):1393–411, 1997.

[6] H. H. Bülthoff and H. A. Mallot. Integration of depth modules: stereo and shading. *Journal of the Optical Society of America, A*, 5(10):1749–1758, 1988.

[7] J. J. Clark and A. L. Yuille. *Data fusion for sensory information processing systems*. Kluwer Academic Publishers, Boston, 1990.

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

[9] D. W. Dong and J. J. Atick. Statistics of natural time varying images. *Network Computation in Neural Systems*, 6:345–358, 1995.

[10] A. W. F. Edwards. *Likelihood*. Johns Hopkins University Press, Baltimore, 1992.

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

[12] F. V. Jensen. *An introduction to Bayesian networks*. Springer, New York, 1996.

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

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

[15] D. C. Knill. *The role of cooperative processing in the perception of surface and reflectance*. Ph.D. Thesis, Brown University, 1990.

[16] D. C. Knill. Discriminating planar surface slant from texture: Human and ideal observers compared. *Vision Research*, 38:1683–1711, 1998.

[17] D. C. Knill. Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision Res*, 38(11):1683–711, 1998.

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

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

[20] S. Lawson, C. Madison, and D. Kersten. Depth from cast shadows and size-change: Predictions from statistical decision theory. 1998.

[21] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[22] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, New York, 1982.

[23] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network Computation in Neural Systems*, 7:333–339, 1996.

[24] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[25] J. Porrill, J. P. Frisby, W. J. Adams, and D. Buckley. Robust and optimal use of information in stereovision. *Nature*, 97:63–66, 1999.

[26] C. Rao. *Linear statistical inference and its applications*. John Wiley and Sons, New York, 1973.

[27] D. Ruderman and W. Bialek. Statistics of natural images - scaling in the woods. *Physical Review Letters*, 73:814–817, 1994.

[28] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. *31st Asilomar Conference on Signals, Systems and Computers*, 1997.

[29] E. P. Simoncelli and R. W. Buccigrossi. Embedded wavelet image compression based on a joint probability model. *4th IEEE International Conference on Image Processing*, 1997.

[30] E. P. Simoncelli and J. Portilla. Texture characterization via joint statistics of wavelet coefficient magnitudes. *Proceedings 5th IEEE Int'l Conf on Image Processing*, 1998.

[31] Y. Weiss and E. H. Adelson. Slow and smooth: a bayesian theory for the combination of local motion signals in human vision. (A.I. Memo No. 1624), February, 1998 1998.

[32] A. L. Yuille and H. H. Bulthoff. Bayesian decision theory and psychophysics. In D. C. Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 123–161. Cambridge University Press, New York, 1996.

[33] S. C. Zhu, Y. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9:1627–1660, 1997.