The Role of Task Specification in Optimal Cue Integration [*]

Paul R. Schrater and Daniel Kersten

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

Oct 28, 1999

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, seems computationally intractable because of the high dimensionality of image measurements and scene properties. As a consequence, both artificial visual systems and models of human vision have relied on a premature commitment to computationally simple modular architectures in which a particular scene property (e.g. object depth) is estimated from a restricted set of image measurements (e.g. image size), without reference to other scene properties. Such "Depth-from-X" approaches create problems for inference that include invalid and contradictory assumptions and difficulties in fusing the potentially inconsistent estimates. The computational problem posed by optimal inference is not hopeless, however, and can be surmounted by restricting inference to particular tasks 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 $p(S, I)$. However, restricting optimal inference to particular tasks has the effect of making the choice of representation for the scene variables non-trivial. In particular, some representations lead to simpler inference algorithms than others. 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.

Keywords: Task, Data Fusion, Cue Integration, Optimal Estimation, Bayesian inference, Depth Estimation, Depth from Shadows, Depth from Image Size

[*] Based on: Schrater, P.R. & Kersten, D. (1999) Statistical Structure and Task Dependence in Visual Cue Integration. *Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling.* June 22, 1999 – Fort Collins, Colorado. http://vision.psych.umn.edu/www/kersten-lab/papers/SchraterKerstenCVPRWork99.pdf.

1. Introduction

Cue integration is the process by which we combine different kinds of image measurements (e.g. line detectors, 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)). A key problem is to determine how many different scene variables and image measurements must be considered at once. Our understanding and modeling of how these cues interact in everyday vision presents a challenge: How can we test and quantitatively model the interactions of multiple cues given the complexities of natural images? We can begin by assuming no interaction, and adding interaction corrections as needed. However, it is difficult to assess the kind or number of corrections that will be required to achieve a given level of performance, and the number of empirical tests needed to verify these corrections quickly becomes prohibitive.

Alternatively, we can study optimal cue combination strategies. This strategy, at least in theory, makes the scientific problem tractable. However, while optimal strategies best combine the information, there is no guarantee that the resulting strategies are simple enough to practically implement. In particular, optimal estimation for depth runs into the problem of dimensionality. Thus, we may have changed 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 made tractable by exploiting task dependency and the statistical structure to simplify the depth estimation problem. We illustrate these ideas with an analysis of depth estimation from image size and cast shadow position cues.

2. Probabilistic Approaches to Scene Estimation

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

2.1. Modeling $p(S, I)$

Probabilistic approaches to scene estimation require the specification of $P(S, I)$, the joint probability distribution on the set 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 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(Atick and Redlich, 1992; Ruderman and Bialek, 1994; Dong and Atick, 1995; Olshausen and Field, 1996; Zhu et al., 1997; Simoncelli and Portilla, 1998; Simoncelli, 1997; Simoncelli and Buccigrossi, 1997).

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. 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 restricting 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:

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

with respect to the posterior probability, $P(S|I)$. 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.

2.3. Why do Bayesian Cue Integration?

In a full implementation of Bayesian inference, all the variables would be considered at once. The advantage of such a procedure is that it is optimal and uses all the information. The disadvantage is combinatorial explosion and a full Bayesian approach can quickly become intractable as the size of the problem grows.

The most common solution to this problem is to try to make separate estimates of scene variables. 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 figure 2.3). There is a long tradition of treating modularity as a fundamental computational principle (as opposed to a derived principle). Marr was "...moved to elevate (modularity) to a principle"(Marr, 1982). Most 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 permit independent estimates of the quantity. The plethora of "Shape-from-X" algorithms illustrate the point. The attempt to produce independent estimates of shape from various cues frequently involves prior assumptions that are incompatible between modules. For instance, shape-from-texture typically assumes constant shading, while shape-from-shading assumes constant texture(Yuille and Bulthoff, 1996). Thus typical shape-from-shading and shape-from-texture modules will create inconsistent estimates of the shape.

The potential problems for cue combination created by an ad hoc committment 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). 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 inference insures consistent inferences and optimal combination 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.

One approach to a compromise between the benefits of optimal inference and the computational tractability is to perform optimal inference for a small set of tasks. Basically, the idea is that it is easier to build a computational efficient specialist than a generalist. We will show below how restricting inference to a particular task can simplify the computational complexity. However, building systems that are collections of specialists for particular tasks presents the danger that the system will not be able to perform tasks outside its expertise and that the specialists will not be able to co-exist in a common computational architecture. We will show that the choice of representation for scene attributes has a powerful impact on both the computational architecture required to perform optimal inference and on the ability for the same architechture to support visual inference for muliple tasks.

3. Simplifications Due to Task Dependency

The cost of modeling $p(S, I)$ for the entire ensemble of scenes and images an observer could encounter is prohibitive. However, if we tailor our modeling to what is needed for the task, then we can substantially reduce the computational burden. Fortunately, the simplifications fall directly out of the cost function and the statistical structure of $p(S, I)$.

In a Bayesian decision theory framework, tasks specify two things: a cost function and knowledge about scene attributes in the task context which is incorporated into the prior distribution. We will consider the simplifications that result from each in turn.

3.1. Reductions due to the cost function

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

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 $p(S|I)$ over the set of 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 S_{ir} 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 variables are independent of S_r, then the joint distribution factors, and hence the relevant posterior factors out of the marginalization integral i.e. $p(S \mid I) = p(S_r, S_g \mid I)p(S_{ind} \mid I)$ so that $\int_{S_{ir}} P(S_r, S_{ir} \mid I)dS_{ir} = \int_{S_g} P(S_r, S_g \mid I)dS_g \int_{S_{ind}} P(S_{ind} \mid I)dS_{ind}$ 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 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 the set of 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.

3.2. Reductions due to prior contextual knowledge

In most cases, the nature of a task fixes or restricts the values of 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 $p(S_r, S_g \mid I)$ on S_f yielding $p(S_r, S_g \mid I, S_f)$. Because conditioning increases (or leaves unchanged) statistical independence, conditioning produces independence relations which can be exploited for more efficient algorithms. In particular, we expect the conditional distribution to further decompose into relevant and irrelevant scene variables.

As an example, 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 5). Assume the

Figure 2. about here.

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 figure 2). Further assume the observer has two cues available: the cast shadow position, 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 shadow position can be produced by a family of object depths and light source positions.

To summarize our discussion of task-dependence we have shown that mapping the notion of a task into Bayesian inference produces four components:

- a decision rule (loss function)
- a required set of scene variables S_r
- the generic (nuisance) scene variables S_g.
- a set of scene variables that are fixed (specified) by the nature of the task S_f

Each of the four components of a task plays a role in determining the structure of the optimal inference computation.

4. Cue Combination

In this section, we discuss four ways in which task specification affects the computational architecture for an optimal cue integration problem. We show, given the task: 1) the natural modularity of the inference (and hence the computational complexity) falls out of conditional independence between image cues; 2) that the shape of the posterior distribution can be dominated by the effects of marginalization with respect to the generic variables; 3) that a particular task, such as discriminating depth, does not unambiguously specify the representation of the scene variable, and the choice of variable representation results in differences in the properties of the optimal estimators; and, 4) the choice of variable representation should take into account the possibility of being used for more than one task.

The third point in particular will bring us to our central idea, illustrated with an example in Section 5, that the choice of scene variable representation for a decision determines the modularity and the pattern of performance for optimal cue integration.

4.1. Conditional independence

Having discussed how to simplify the optimal inference problem by restricting to a task, we can now show how the statistical structure of the simplified problem determines how variables interact in the inference, which has consequences for data fusion. The key element in the optimal inference is the posterior, which we can rewrite in terms of the joint distribution and then in terms of likelihoods and priors:

$$p(S_r, S_g \mid I) = p(I, S_r, S_g)/p(I) = p(I \mid S_r, S_g)p(S_r, S_g)/p(I) \tag{3}$$

The probabilistic structure of the joint probability distribution $p(I, S_r, S_g)$ can be represented by a Bayes Net(Pearl, 1988; Jensen, 1996), 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 image cues to produce independent estimates of 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 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-

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

Figure 3. about here.

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.

Although we have discussed the simplifications afforded complete statistical independence, we have layed the groundwork for the use of principled approximations. Basically we can trade off performance against computational 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 variables S_x but are strongly dependent given S_y, we can produce a Bayes modular system by evaluating S_x at its most probable value S_x^*: $p(I_a, I_b \mid S_x, S_y) \rightarrow p(I_a, I_b \mid S_y, S_x^*) = p(I_a \mid S_y, S_x^*)p(I_b \mid S_y, S_x^*)$. Thus the order in which links should be broken can be determined by a measure of independence like the mutal information, and the cost of assuming independence can be assessed by comparing the approximate performance to the optimal performance.

4.2. Marginalizing with respect to the generic variables, S_g

In order to assess the effect of marginalization of the generic variables on the inference, we used Laplace's method (Bender and Orszag, 1978; Freeman, 1994; MacKay, 1992) to construct analytic approximations to the required integrals. In Laplace's method integrals of the form:

$$F(\sigma^2, \vec{y}) = \int_a^b f(x)\exp(h(x, \vec{y})/v^2)dx \tag{4}$$

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

$$F(v^2, \vec{y}) \sim \frac{\sqrt{2\pi v^2}}{\sqrt{|h''(c(\vec{y}), \vec{y})|}} f(c(\vec{y}))\exp(h(c(\vec{y}), \vec{y})/v^2) \tag{5}$$

For our application, $f(x) = p(x)$, and $h(x, \vec{y}) = \log(p(I|x, S))v^2$, i.e. it is the log likelihood function of the data I given a particular generic scene variable x and the remaining (both required and generic) scene variables S.

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

If the measurement noise distribution is unbiased and the log likelihood is amenable to a quadratic approximation about the data, then the effects of marginalization dominate the distribution. We can show this dominance by expanding the log likelihood in a Taylor series to second order in I about $I_{max} = \mu(x, S)$, yielding $p(I|x, S) \simeq \exp(-\frac{1}{2}(I - \mu(x, S))^2/\sigma(x, S)^2)$, where $\sigma(x, S)^2 = \frac{\partial^2 \log(p(I|x,S))}{\partial I^2}|_{I_{max}}$. Thus the h function in the integral expansion is approximately $-\frac{1}{2}(I - \mu(x, S))^2/\sigma(x, S)^2$. Note that the exponential factor in equation 5 equals one when evaluated at the maximum, so the integral only depends on the second derivative of h. Computing the second derivative with respect to x and evaluating at the maximum $x_{max} = c(I, S)$, we find $\frac{\partial^2}{\partial x^2}(-\frac{1}{2}(I - \mu(x, S))^2/\sigma(x, S)^2) = 2(\partial\mu(x, S)/\partial x)^2/\sigma(x, S)^2$.

Using these simplifications, equation 5 reduces to:

$$p(I \mid S) \sim \frac{\sqrt{2\pi\sigma(c, S)^2}}{|\partial\mu(c, S)/\partial x|} p(c(I, S)) \tag{6}$$

Assuming that $\sigma(c, S)^2 \sim \sigma^2$ (i.e. is nearly a constant), this last expression shows that the likelihood after marginalization is dominated by values of S where where $|\partial\mu(c, S)/\partial x|$ is close to zero. For an unbiased measurement noise, the maximum function $\mu(c, S)$ is determined by the imaging equations, e.g. $I = \mu(x, S)$, hence the maximum function is determined by the physics and geometry of the problem rather than the noise. What we have shown is that under reasonable conditions the shape of the likelihood function is dominated by the effects of marginalization and furthermore, given small image measurement noise and enough uncertainty on the generic variable (i.e. the prior $p(x)$ approachs a constant), the likelihood is dominated by the physical and geometrical constraints determining the image cues.

In general, the variables that are marginalized 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 generic variables can have a greater impact on the estimator than the particular properties of the measurement noise distribution. Thus, what we have shown is that the properties of optimal estimators are largely determined by the conditional dependence structure, and the generic variables that are marginalized.

4.3. Dependence of cue combination 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.

First, 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.

Second, the particular scene variables we estimate 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 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.

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

5. Estimating Depths from Image Size and Shadow Displacement

Using the idea of proof by prototype, we can ask the question: Does the human visual system do task-dependent Bayesian inference? 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(Blake et al., 1996; Knill, 1998a; Knill, 1998b; Weiss and Adelson, 1998) 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 (Knill, 1990; Adelson, 1993). In this section we perform a detailed analysis of Bayesian inference on a simple scene, to compare several of these predictions to human 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. One of the challenges of testing models of human vision is that we do not have direct access to the variable representations on which a decision is being made. We show that different choices of variable predict different patterns of cue integration.

5.1. Theory

We illustrate the dependence of Bayesian cue combination on task demands and conditional independences with a simple scene due to Kersten et al.(Kersten et al., 1996). 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

Figure 4. about here.

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 figure 4.

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 combine the shadow and image size data, we must convert one of the variables into a common depth representation.

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 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 depth representation 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 depth representations to do the inference. Each of these leads to a different Bayes net and different optimal inference structure. For each of the three representations, 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.

Figure 5. about here.

Figure 6. about here.

The geometric diagram in figure 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 figure 6.

5.1.1. *Representation 1: Estimating relative distance from background (z_r)*

One way of judging the depths of the two squares is to compute the distance from the background. This leaves 4 unknowns, $\alpha, s, z, \ and, r_b$ 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 $s_r = s/r_b$, 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)[3], is a function of z_r and light source position α:

[3] 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 replaced every instance of β in the formula with $\tan^{-1}(l)$.

Figure 7. about here.

$$\beta = \tan^{-1}(z_r \tan(\alpha)) + n_\beta \tag{7}$$

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:

$$p(\beta|z_r, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp(-\frac{(\beta - \tan^{-1}(z_r \tan(\alpha))^2}{2\sigma_\beta^2}) \tag{8}$$

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:

$$p(a|z_r, s_r) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp(-\frac{(\log(a) - \log(\frac{s_r}{1-z_r}))^2}{2\sigma_a^2}) \tag{9}$$

We assume that the observer can potentially have 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 $p(z_r|\{\beta\}, \{a\})$. Assuming that the repeated 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 7a. Note that this network is Bayes modular, which shows up in the factoring of the likelihoods above.

5.1.2. *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 figure 7b. In this case, we must convert the shadow position cue's dependence on the allocentric distance z to the egocentric depth r_s. Using $r_b = z + r_s$, we can write the shadow position measurement as:

$$\beta = \tan^{-1}(\left(1 - \frac{r_s}{r_b}\right) \tan(\alpha)) + n_\beta \tag{10}$$

The likelihood function is given by:

$$p(\beta|r_s, r_b, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp(-\frac{(\beta - \tan^{-1}(\left(1 - \frac{r_s}{r_b}\right) \tan(\alpha))^2}{2\sigma_\beta^2}) \tag{11}$$

The image size a is given by:

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

Hence the likelihood for a is given by:

$$p(a|r_s, s) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp(-\frac{(\log(a) - \log(\frac{s}{r_s}))^2}{2\sigma_a^2}) \tag{12}$$

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

$$\begin{aligned} p(r_s|\{\beta\}, \{a\}) &\propto p(\{\beta\}|r_s)p(\{a\}|r_s)p(r_s) \\ &= \left(\int_{r_b} \int_\alpha \prod_{i=1}^{N} p(\beta_i|r_s, r_b, \alpha)p(\alpha)p(r_b)d\alpha\, dr_b\right) \left(\int_s \prod_{i=1}^{N} p(a_i|r_s, s)p(s)ds\right) p(r_s) \end{aligned} \tag{13}$$

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 shadow depth estimates increases as compared with the relative distance representation (Representation 1).

5.1.3. *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 of the image size cue's dependence on the egocentric depth r_s to the allocentric distance z. 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 7c. The measurements can be written in terms of z as:

$$\beta = \tan^{-1}(z \tan(\alpha)/r_b) + n_\beta \tag{14}$$

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

The likelihood functions are:

$$p(\beta|z, r_b, \alpha) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp(-\frac{(\beta - \tan^{-1}(z \tan(\alpha)/r_b))^2}{2\sigma_\beta^2}) \tag{15}$$

$$p(a|z, r_b, s) = \frac{1}{\sqrt{2\pi}\sigma_a a} \exp(-\frac{(\log(a) - \log(\frac{s}{r_b - z}))^2}{2\sigma_a^2}) \tag{16}$$

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

$$\begin{aligned} p(z|\{\beta\}, \{a\}) &\propto p(\{\beta\}, \{a\}|z)p(z) \\ &= \left(\int_{r_b} \left(\int_\alpha \prod_{i=1}^{n} p(\beta_i|z, r_b, \alpha)p(\alpha)d\alpha \right) \left(\int_s \prod_{i=1}^{n} p(a_i|z, r_b, s)p(s)ds \right) p(r_b)\, dr_b \right) p(z) \end{aligned} \tag{17}$$

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.

5.1.4. *MAP estimates*

Maximum a posteriori estimates of the depth decision variable were computed from the posterior distribution after marginalization. Marginalizations were approximated using Laplace's method, described above. Because the three depth representations of the depth estimation task differ primarily in the marginalizations, we expect that they will yield estimators with different properties. We show that these expectations are correct below.

5.1.5. *Representation 1: MAP Estimate for z_r (Relative z)*

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

$$\int_\alpha \prod_{i=1}^{n} p(\beta_i|z_r,\alpha)p(\alpha)d\alpha \simeq \frac{\sqrt{2}z_r}{\pi(z_r^2\cos(\hat{\beta})^2+\sin(\hat{\beta})^2)} \tag{18}$$

where $\hat{\beta}$ is the mean of the N sample βs.

The maximum z_r occurs at:

$$max_{z_r}\,(p(\beta|z_r)) = \tan(\hat{\beta}). \tag{19}$$

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:

$$p(\{a\}|z_r) = \frac{1}{\sqrt{\pi(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)\hat{a}}}exp(-\frac{\log(\hat{a}(1-z_r)/\mu_{s_r})^2}{2(\sigma_{\hat{a}}^2+\sigma_{s_r}^2)}) \tag{20}$$

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

$$max_{z_r}\,(p(\{a\}|z_r)) = 1-\mu_{s_r}/\hat{a} \tag{21}$$

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

5.1.6. *Representation 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:

$$\int_{r_b} p(\{\beta\}|r_s,r_b)p(r_b)dr_b \simeq \frac{2(1-r_s/\mu_{r_b})}{\pi((1-r_s/\mu_{r_b})^2\cos(\hat{\beta})^2+\sin(\hat{\beta})^2)} \tag{22}$$

The maximum r_s occurs at:

$$max_{r_s}\,(p(\{\beta\}|r_s)) = \mu_{r_b}(1-\tan(\hat{\beta})). \tag{23}$$

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

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

$$p(\hat{a}|r_s) = \frac{1}{\sqrt{\pi(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2)\hat{a}}} exp(-\frac{\log(\hat{a}r_s/\mu_{s_r})^2}{2(\sigma_{\hat{a}}^2 + \sigma_{s_r}^2)}). \tag{24}$$

The maximum r_s with respect to image size occurs at

$$max_{z_r}\left(p(\{a\}|z_r)\right) = r_s = \frac{\mu_s}{\hat{a}}. \tag{25}$$

5.1.7. *Representation 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:

$$p(z|\{\beta\},\{a\}) \propto \frac{\mu_s z \csc(\hat{\beta})\sec(\hat{\beta})}{\sqrt{2}\hat{a}\sqrt{\hat{a}^2 z^2 + \mu_s^2(\sigma_{\hat{a}}^2 + \sigma_s^2)\tan(\hat{\beta})^2}} \exp\left(-\frac{(z - \frac{\mu_s \tan(\hat{\beta})}{\hat{a}(1-\tan(\hat{\beta}))})^2}{2\frac{\hat{a}^2 z^2 + \mu_s^2(\sigma_{\hat{a}}^2+\sigma_s^2)\tan(\hat{\beta})^2}{\hat{a}^2(1-\tan(\hat{\beta}))^2}}\right) p(z) \tag{26}$$

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

$$max_z\left(p(z|\{\beta\},\{a\})\right) \simeq \frac{\mu_s \tan(\hat{\beta})}{\hat{a}(1 - \tan(\hat{\beta}))} \tag{27}$$

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

5.1.8. *Fisher Information*

A lower bound on the variance of unbiased estimators is given by the reciprocal of the Fisher Information (Rao, 1973). The Fisher Information is given by:

$$\mathcal{I}(x) = -N \int_{data} p(data|x)(\partial^2 \log p(data|x)/\partial x^2) d(data) \tag{28}$$

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 (Blake et al., 1996; Rao, 1973). 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:

$$m_{best} = \frac{m_a \mathcal{I}_a(x|m_a) + m_\beta \mathcal{I}_\beta(x|m_\beta)}{\mathcal{I}_a(x|m_a) + \mathcal{I}_\beta(x|m_\beta)}. \tag{29}$$

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

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

$$\frac{1}{\mathcal{I}_a(x|m_a) + \mathcal{I}_\beta(x|m_\beta)} \tag{30}$$

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 I.

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 (non-modular) fusion has resulted from marginalization. In addition, because the shadow likelihood and the size likelihood are both marginalized or averaged over the same variable, we expect less variation in the optimal estimate of z with changes in the size and shadow cues.

In a Bayesian context, linear combination is only appropriate for Bayes nets with certain properties. For Bayes nets which are modular, the data are consistent, and the noise distributions are Gaussian, 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 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. Note that the information mirrors our expectation about the natural depth representations for the two cues.

5.2. 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 (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. 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.

Figure 8. about here.

Figures 8 & 9 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 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 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.

Figure 9. about here.

Note that the data from the two subjects are qualitatively different[4]. 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 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 (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.

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 10 shows the marginal distributions for all of the scene variables in the depth-from-observer network given only two image size

[4] 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 combination strategies.

Figure 10. about here.

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

This paper starts from the premise that a fundamental goal of a 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 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. These ideas lay the foundation for introducing approximations that may yield more efficient algorithms for optimal cue integration. 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 depth representations. From the analysis we predict performance on a simple depth discrimination task from the optimal cue combination in each representation. We find that human observers' decisions are confidence-driven, in that they weight the information from the two cues in accord with their informativeness.

Acknowledgements

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

References

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

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

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

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

Brainard, D. H. and W. T. Freeman: 1997, ‘Bayesian color constancy’. *J Opt Soc Am A* **14**(7), 1393–411.

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

Cutting, J. E. and P. M. Vishton: 1996, ‘Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Information about Depth.’. In: W. Epstein and S. Rogers (eds.): *Perception of Space and Motion.* San Diego: Academic Press, Inc., pp. 69–117.

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

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

Freeman, W. T.: 1994, ‘The generic viewpoint assumption in a framework for visual perception’. *Nature* **368**, 542–545.

Goodale, M. A., J. P. Meenan, H. H. Bulthoff, D. A. Nicolle, K. J. Murphy, and C. I. Racicot: 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.* New York: Springer.

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

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

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

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

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

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

Landy, M. S., L. T. Maloney, E. B. Johnson, and M. Young: 1995, 'Measurement and modeling of depth cue combination: In defense of weak fusion'. *Vision Research* **35**, 389–412.

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

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

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

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

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

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

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

Simoncelli, E. P.: 1997, 'Statistical Models for Images: Compression, Restoration and Synthesis'. *31st Asilomar Conference on Signals, Systems and Computers*.

Simoncelli, E. P. and R. W. Buccigrossi: 1997, 'Embedded Wavelet Image Compression Based on a Joint Probability Model'. *4th IEEE International Conference on Image Processing*.

Simoncelli, E. P. and J. Portilla: 1998, 'Texture Characterization via Joint Statistics of Wavelet Coefficient Magnitudes'. *Proceedings 5th IEEE Int'l Conf on Image Processing*.

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

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

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

List of Footnotes:

Affiliation of author: Department of Psychology, University of Minnesota

1. Based on: Schrater, P.R. & Kersten, D. (1999) Statistical Structure and Task Dependence in Visual Cue Integration. *Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling.* June 22, 1999 – Fort Collins, Colorado. http://vision.psych.umn.edu/www/kersten-lab/papers/SchraterKerstenCVPRWork99

2. In graph theory, a is called the *parent* of b

3. For maxima at end points or vanishing $f((c(\vec{y}))$, the method yields slightly different approximations.

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 replaced every instance of β in the formula with $\tan^{-1}(l)$.

5. 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 combination strategies.

List of Figure Captions:

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" (Clark and Yuille, 1990).

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

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.

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 appeared closer 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 figure 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.

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.

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.

Figure 8. Data for one observer is shown in the bottom panel. The probability that 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 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 figure 8 for details.

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.

List of Keywords:

1. Task
2. Data Fusion
3. Cue Integration
4. Optimal Estimation
5. Bayesian inference
6. Depth Estimation,
7. Depth from Shadows
8. Depth from Image Size

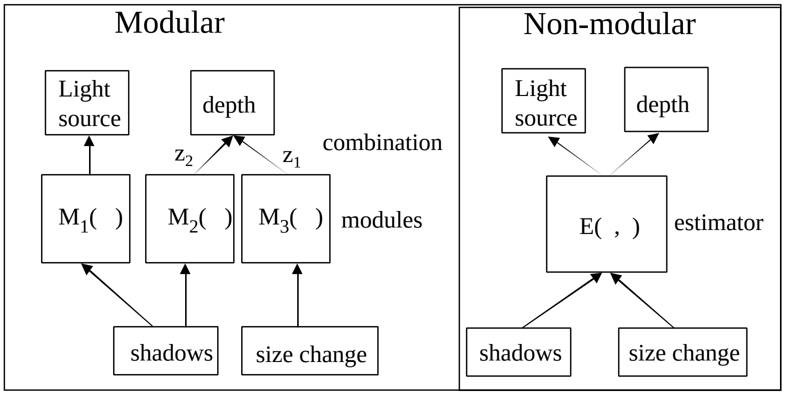
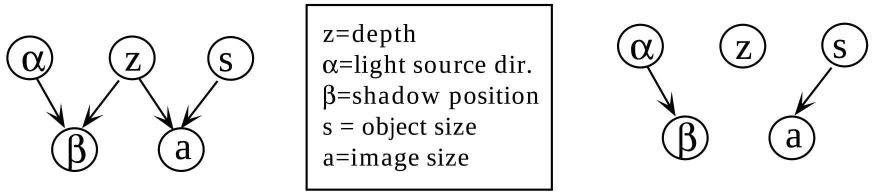

Modular
Light source
depth
z2
z1
combination
M1( )
M2( )
M3( )
modules
shadows
size change
Non-modular
Light source
depth
E(,)
estimator
shadows
size change

Figure 1:



$$p(\beta,a,\alpha,z,s) = p(\beta \mid z,\alpha)p(a \mid z,s)p(\alpha)p(z)p(s)$$

$$p(\beta,a,\alpha \mid z^{*}) = \Big(p_{z^{*}}(\beta \mid \alpha)p(\alpha)\Big)\Big(p_{z^{*}}(a \mid s)p(s)\Big)$$

Figure 2:

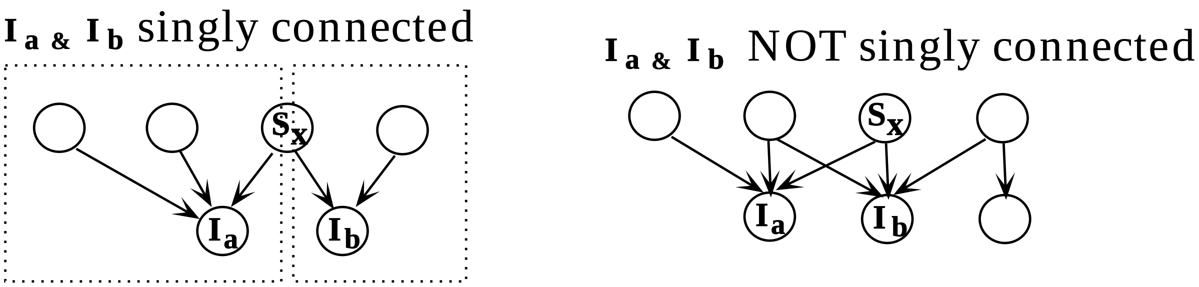

I a & I b singly connected
S x
I a
I b
I a & I b NOT singly connected
S x
I a
I b

Figure 3:

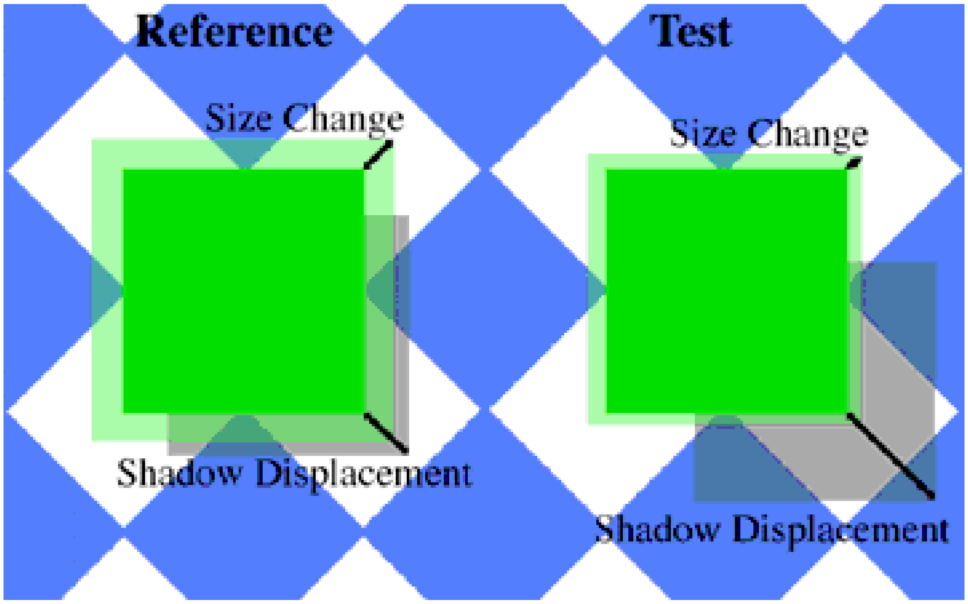

Reference
Test
Size Change
Size Change
Shadow Displacement
Shadow Displacement

Figure 4:

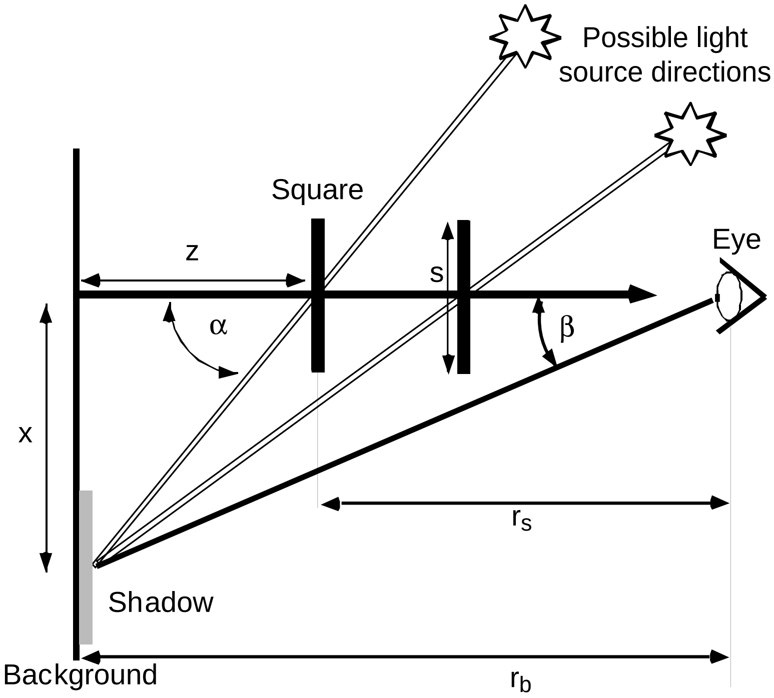

Possible light source directions
Square
z
s
Eye
α
β
x
r_s
Shadow
Background
r_b

Figure 5:

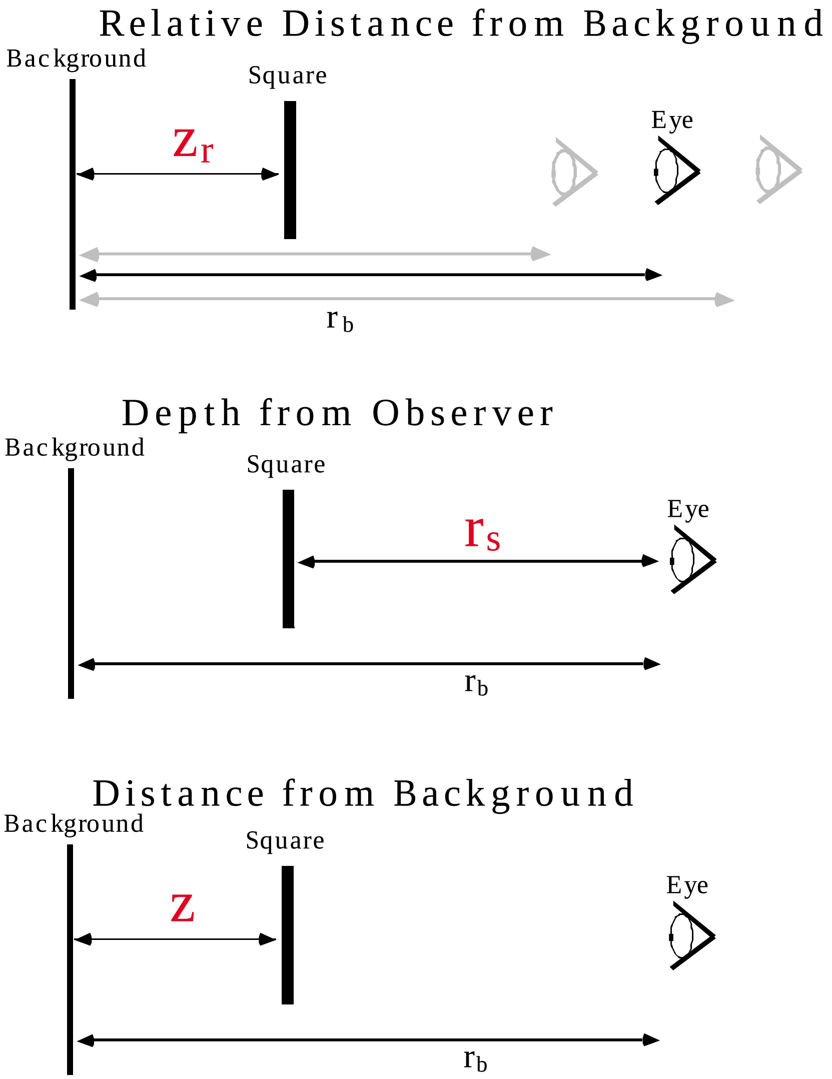

Relative Distance from Background
Background
Square
Z_r
Eye
r_b
Depth from Observer
Background
Square
r_s
Eye
r_b
Distance from Background
Background
Square
Z
Eye
r_b

Figure 6:

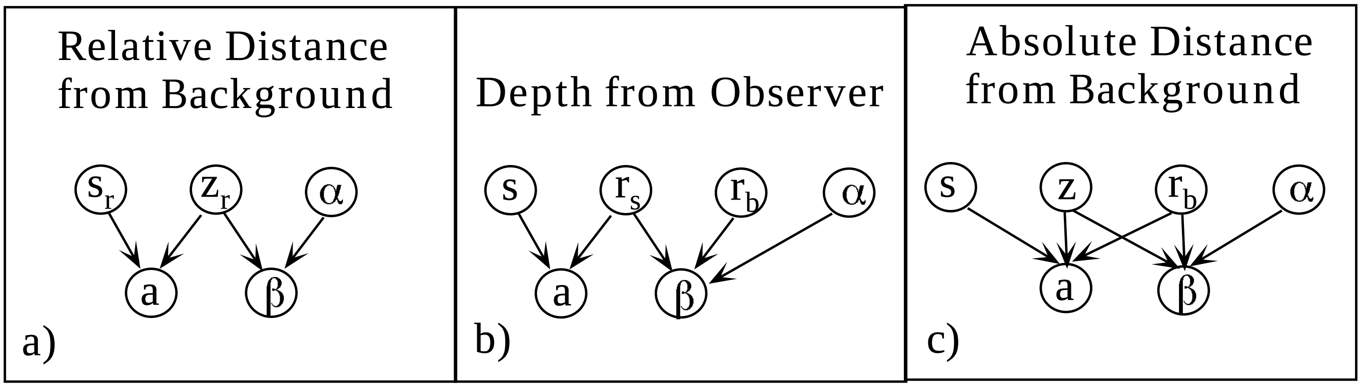

Relative Distance from Background
s_r
z_r
α
a
β
a)
Depth from Observer
s
r_s
r_b
α
a
β
b)
Absolute Distance from Background
s
z
r_b
α
a
β
c)

Figure 7:

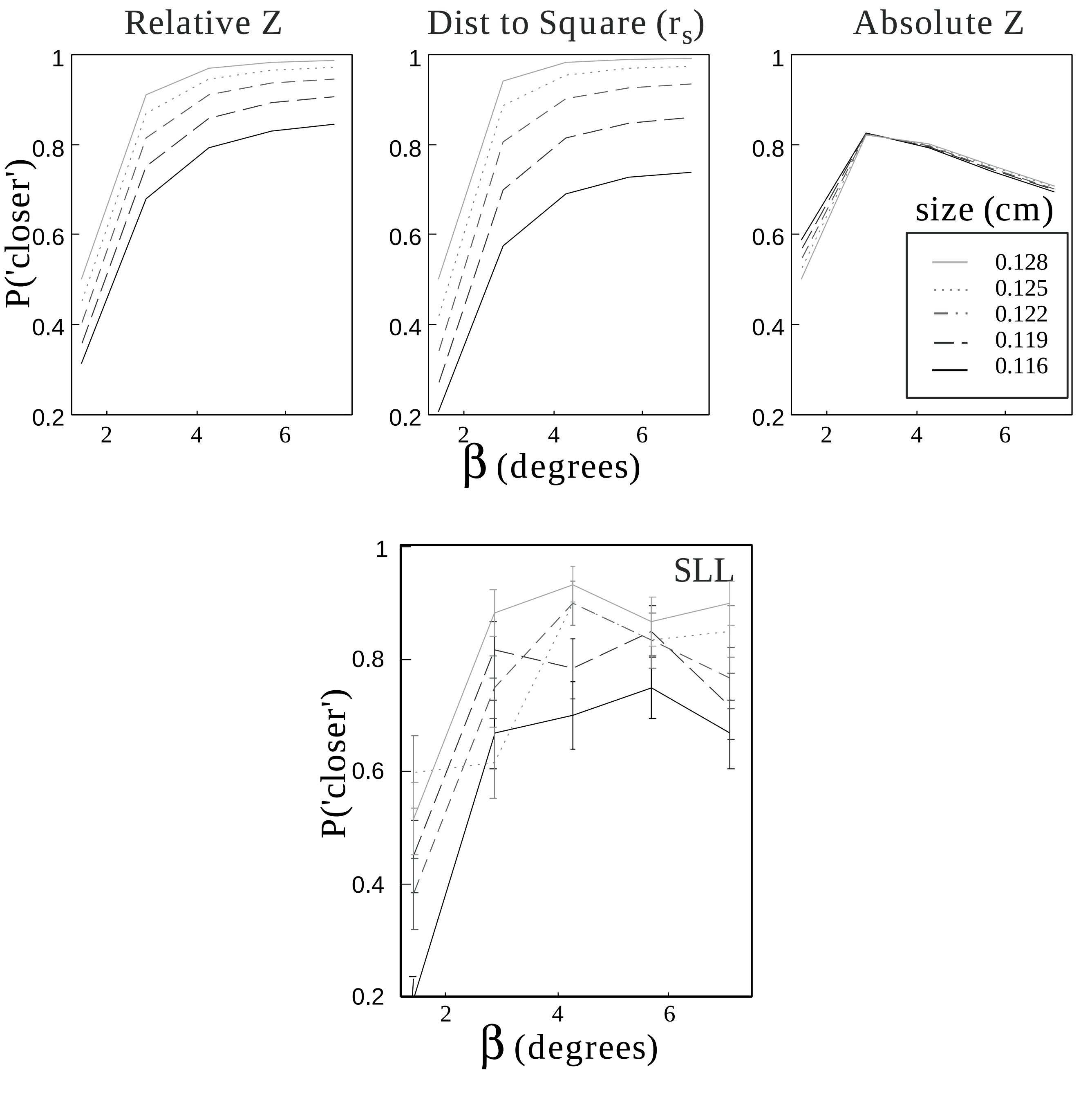

Relative Z
Dist to Square (r_s)
Absolute Z
P('closer')
1
0.8
0.6
0.4
0.2
2
4
6
β (degrees)
size (cm)
0.128
0.125
0.122
0.119
0.116
SLL
P('closer')
1
0.8
0.6
0.4
0.2
2
4
6
β (degrees)

Figure 8:

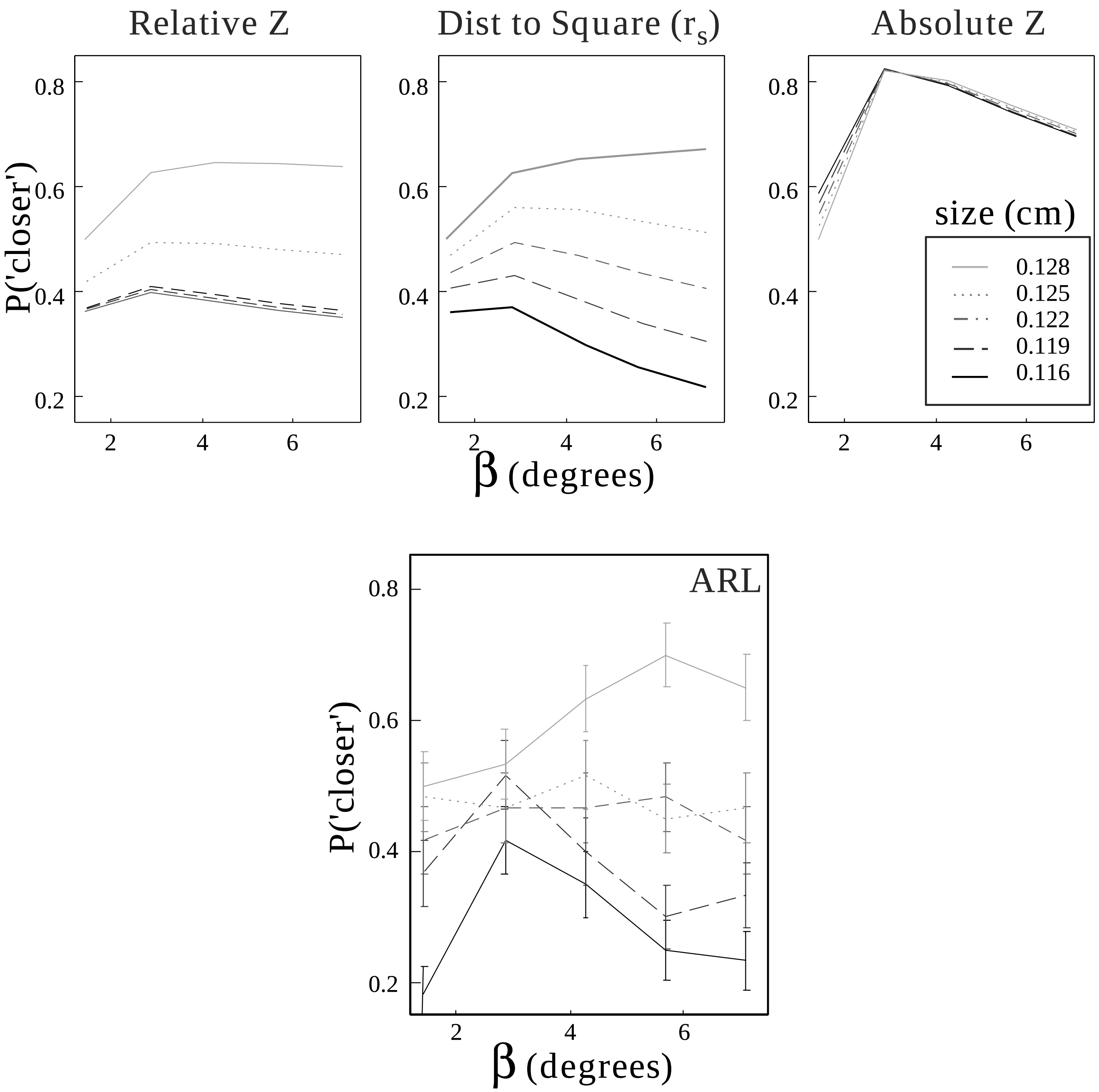

Relative Z
Dist to Square (r_s)
Absolute Z
P('closer')
0.2
0.4
0.6
0.8
2
4
6
β (degrees)
size (cm)
0.128
0.125
0.122
0.119
0.116
ARL
P('closer')
0.2
0.4
0.6
0.8
2
4
6
β (degrees)

Figure 9:

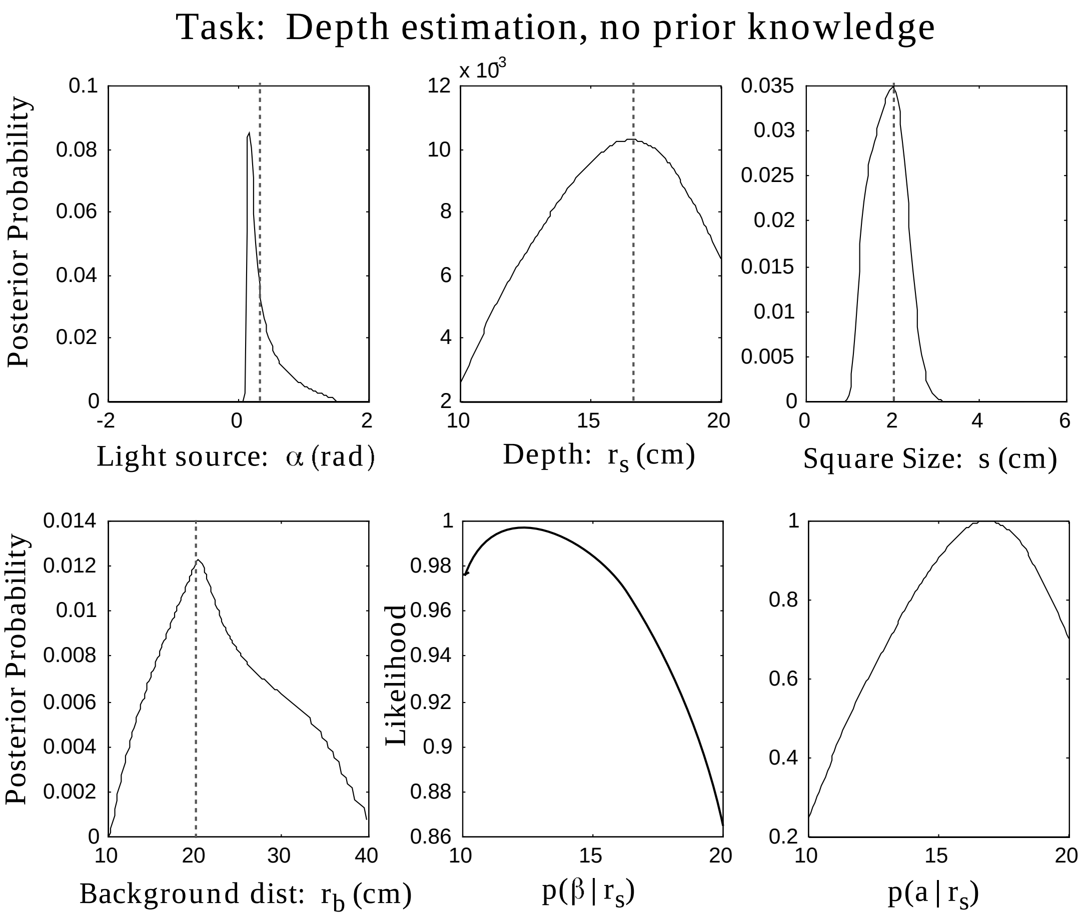

Task: Depth estimation, no prior knowledge
Posterior Probability
0.1
0.08
0.06
0.04
0.02
0
-2
0
2
Light source: α (rad)
x 10^3
12
10
8
6
4
2
10
15
20
Depth: r_s (cm)
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
2
4
6
Square Size: s (cm)
Posterior Probability
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
10
20
30
40
Background dist: r_b (cm)
Likelihood
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
10
15
20
p(β | r_s)
1
0.8
0.6
0.4
0.2
10
15
20
p(a | r_s)

Figure 10: