3

A Bayesian Framework for the Integration of Visual Modules

Heinrich H. Bülthoff and Alan L. Yuille

ABSTRACT

The Bayesian approach to vision provides a fruitful theoretical framework both for modeling individual cues, such as stereo, shading, texture, and occlusion, and for integrating their information. In this formalism we represent the viewed scene by one, or more, surfaces using prior assumptions about the surface shapes and material properties. On theoretical grounds the less information available to the cues (and the less accurate it is), the more important these assumptions become. This suggests that visual illusions, and biased perceptions, will arise for scenes for which the prior assumptions are not appropriate. We describe psychophysical experiments which are consistent with these ideas. Our Bayesian approach also has two important implications for coupling different visual cues. First, different cues cannot in general be treated independently and then simply combined together at the end. There are dependencies between them that have to be incorporated into the models. Second, a single generic prior assumption is not sufficient even if it does incorporate cue interactions because there are many different types of visual scenes and different models are appropriate for each. This leads to the concept of competitive priors where the visual system must choose the correct model depending on the stimulus.

3.1 INTRODUCTION

We define vision as perceptual inference, the estimation of scene properties from an image or a sequence of images.[1] Vision is ill posed in the sense that the retinal image is potentially an arbitrarily complicated function of the visual scene and so there is insufficient information in the image to uniquely determine the scene. The brain, or any artificial vision system, must make assumptions about the real world in order to overcome this problem. These assumptions must be sufficiently powerful to ensure that vision is well posed for those properties in the scene that the visual system needs to estimate. In this chapter we argue that Bayes (1783) provides a natural framework for modeling perceptual inference. We emphasize that we are describing a framework and *not* a theory. The usefulness of such a framework is that it is powerful enough to *compactly* describe most, ideally all, visual phenomena and that it leads to specific theories (by choosing priors and likelihoods) that can be tested experimentally.

How are these assumptions imposed in vision systems? The Bayesian formulation gives us an elegant way to impose constraints in terms of prior probabilistic assumptions about the world, based on Bayes formula (Bayes 1783):

$$P(S|I) = \frac{P(I|S)P(S)}{P(I)}. \tag{3.1}$$

Here S represents the visual scene, the shape and location of the viewed objects, and I represents the retinal image. $P(I|S)$ is the *likelihood function* for the scene and specifies the probability of obtaining an image I from a given scene S; it incorporates a model of image formation and of noise and hence is the subject of computer graphics. $P(S)$ is the *prior* distribution and specifies the relative probability of different scenes occurring in the world. The probabilistic model, specified by $P(I|S)$ and $P(S)$, contains the prior assumptions about the scene structure, including the geometry, the lighting, and the material properties. $P(I)$ can be thought of as a normalization constant and can be derived from $P(I|S)$ and $P(S)$ by elementary probability theory, $P(I) = \int P(I|S)P(S)[dS]$. Finally, the *posterior distribution* $P(S|I)$ is a function giving the probability of the scene being S if the observed image is I.

In words (3.1) states that the probability of the scene S, given the image I, is the product of the probability of the image, given the scene $P(I|S)$, times the a priori probability $P(S)$ of the scene, divided by a normalization constant $P(I)$.

To specify a unique interpretation of the image I, we must make a decision based on our probability distribution, $P(S|I)$, and determine an estimate, $S^*(I)$, of the scene. In Bayesian decision theory (Berger 1985) this estimate is derived by choosing a loss function that specifies the penalty paid by the system for producing an incorrect estimate. Standard estimators like the *maximum a posteriori* (MAP) estimator, $S^* = \arg\max_S P(S|I)$ (i.e., S^* is the most probable value of S given the posterior distribution $P(S|I)$) correspond to specific choices of loss function. In this chapter we will, for simplicity and reasons of space, assume that the MAP estimator is used though other estimators are often preferable (see Yuille and Bülthoff 1996).

Although the Bayesian framework is sufficiently general to encompass many aspects of visual perception including depth estimation, object recognition, and scene understanding, to specify a complete Bayesian theory of visual perception is, at present, completely impractical. Instead, we will restrict ourselves to model individual visual cues for estimating the depth and material properties of objects and the ways these cues can be combined. It has become standard practice for computational theories of vision to separate such cues into modules (Marr 1982) that only weakly interact with each other. From the Bayesian perspective, weak coupling between modules is often inappropriate, due to the interdependence between visual cues. Hence we argue in section 3.3 that the visual cues should often be more strongly coupled. In some cases weak coupling between modules is appropriate (see Landy, et al. 1995).

In the Bayesian framework the choice of prior assumptions used to model each visual cue is very important. Each visual cue is subject to built-in prior assumptions that will inevitably bias the visual system, particularly for the impoverished stimuli favored by psychophysicists. The human visual system is very good at performing the visual tasks necessary for us to interact effectively with the world. Thus the prior assumptions used must be fairly accurate, at least for those scenes that we need to perceive and interpret correctly. The prior assumptions used to interpret one visual cue may conflict with those used to interpret another, and consistency must be imposed when cues are combined. Moreover, the prior assumptions may be context-dependent and correspond to the categorical structure of the world. A single "generic" prior is not sufficient because there are many types of visual scenes and different models are appropriate for each. Each visual module, or coupled groups of modules, will have to determine automatically which prior assumption, or model, should be used; this can lead to a system of competitive prior assumptions (see section 3.4).

In section 3.2 we first describe Bayesian theories for individual cues and argue that several psychophysical experiments can be interpreted in terms of biases toward prior assumptions. Next, in section 3.3, we describe ways of combining different depth cues and argue that strong coupling between different modules is often desirable. Then in section 3.4 we argue that it is preferable to use competing, often context-dependent, priors rather than the single generic priors commonly used. Implications of this approach are described in section 3.5.

3.2 BAYESIAN THEORIES OF INDIVIDUAL VISUAL MODULES

We now briefly describe some Bayesian theories of individual visual cues and argue that psychophysical experiments can be interpreted as perceptual biases toward prior assumptions. From (3.1) we see that the influence of the prior is determined by the specificity of the likelihood function $P(I|S)$. In principle, according to standard Bayesian statistics, the likelihood function should make no prior assumptions about the scene, yet the likelihood functions used in most visual theories often make strong context-dependent assumptions. This fact will be briefly illustrated in this section and we will describe its implications in sections 3.3 and 3.4.

We will specifically discuss theories of shape from shading and shape from texture. All these modules require prior assumptions about the scene geometry, the material properties of the objects being viewed, and, in some cases, the light source direction(s). We will concentrate on the assumptions used by the theories rather than the specific algorithms. Although a number of theories described here were originally formulated in terms of energy functions (Horn 1986) or regularization theory (Poggio, Torre, and Koch 1985), the Bayesian approach incorporates, by use of the Gibbs distribution (Parisi 1988), these previous approaches (see Yuille and Bülthoff 1996).

Shape from Shading

Let us now look at one specific example. Standard models of image formation assume that the observed intensity depends on the tendency of the viewed surface to reflect light, its albedo, and a geometric reflectance factor that depends on the orientation of the surface, the viewing direction, and the light source(s) direction. Shape from shading models (Horn 1986) typically assume that the scene consists of a single surface with constant albedo, a single light source direction $\vec{s}$ that can be estimated, and a *Lambertian* reflectance function. This leads to an image formation model $I = \vec{s} \cdot \vec{n} + N$, where $\vec{n}$ denotes the surface normals and N is additive Gaussian noise. In this case the likelihood function can be written as $P(I|S) = (1/Z)e^{-(1/2\sigma^2)(I-\vec{s}\cdot\vec{n})^2}$, where σ^2 is the variance of the noise and Z is a normalization factor.[2] The prior model for the surface geometry $P(S)$ typically assumes that the surface is piecewise smooth and biases toward a thin plate or membrane. These theories also assume that the occluding boundaries of the object are known, which is helpful for giving boundary conditions.

This likelihood function contains the prior assumption that the reflectance function is Lambertian with constant albedo. Moreover, it ignores effects such as mutual illumination and self-shadowing. The model is therefore only applicable for a certain limited class of scenes and only works within a certain *context* (see fig. 3.1). A visual system using this module would require a

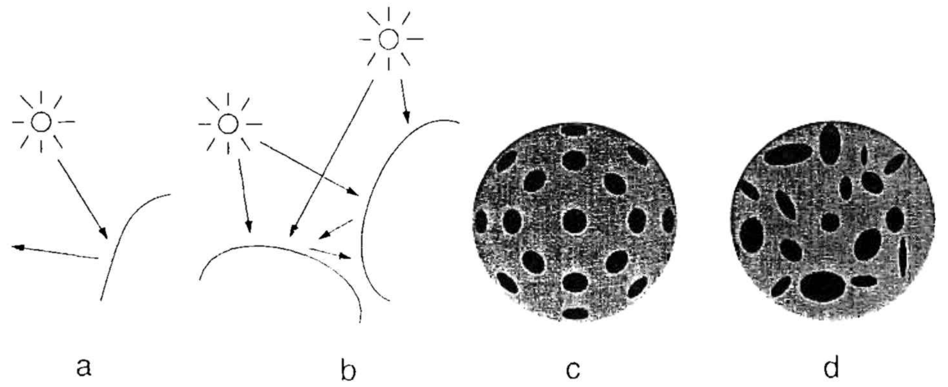

Figure 3.1 Cues are valid only in certain contexts. In (a) we sketch a Lambertian object illuminated by single light source and no mutual illumination; thus standard shape from shading algorithms will work. However, in (b) mutual illumination will prevent shape from shading from working. Similarly, shape from texture is possible for (c) but not for (d), where homogeneity assumption for texture elements is violated. Thus both shading and texture shape cues are only valid in certain contexts.

method for automatically checking whether the context was correc return to this issue in section 3.4 on competitive priors.

What predictions would models of this type make for psycl experiments? Clearly, they would predict that the perception of gec shape from shading would be biased by the prior assumption of smoothness (see fig. 3.2). If we use the models of piecewise s typically used in computer vision, then we would find a bias towa parallel surfaces. Such a bias is found for example in the psychophy: from shading experiments by Bülthoff and Mallot (1988), Mam: Kersten (1994), and Koenderink, van Doorn, and Kappers (1992). (not all smoothness priors cause such a bias (see, for exampl Mayhew, and Frisby 1985); nevertheless, the bias appears to be th mentally. The more impoverished the stimuli, the greater the bia: might expect this effect to be larger for psychophysical experim for realistic stimuli (realistic stimuli, i.e., natural images, are typ impoverished).

Shape from Texture

Existing shape from texture models also make similar, though inc assumptions about the scenes they are viewing. They assume th variations can be modeled by spatial changes in the albedo and geometric reflectance factor can be neglected, or filtered out in s

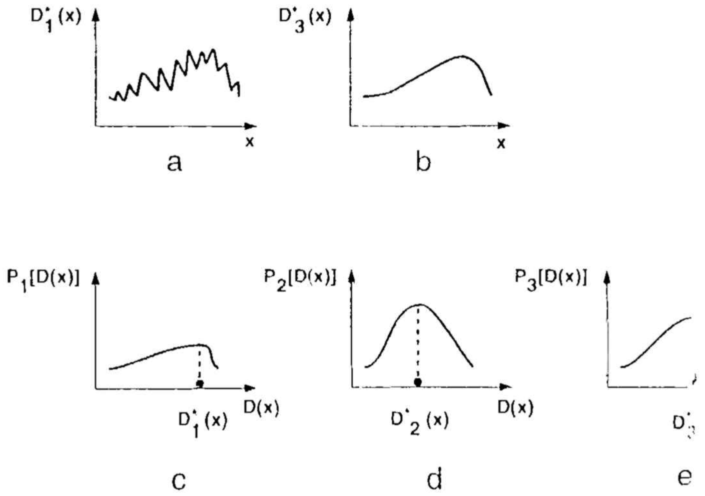

Figure 3.2 Prior assumption bias perception. Graph (a) shows true depth $D_1^*(x)$ ar shows biased depth percept $D_3^*(x)$ after smoothing. In (c) we assume that likelih $P_1[D(x)]$ is weakly peaked at true depth $D_1^*(x)$. Prior in (d), however, is peake Resulting posterior distribution $P_3[D(x)]$ is shown in (e) and yields biased percept l

Texture elements are assumed to be "painted" onto piecewise smooth surfaces in a spatially statistically homogeneous manner. A specific example is given by Blake, Bülthoff, and Sheinberg (1993). Therefore the imaging model, or likelihood function, will assume that these texture elements are generated from a homogeneous distribution on the surface and then projected onto the image plane. Assumptions about the geometry, such as piecewise smoothness, are then placed in the prior.

Once again, the nature of the likelihood term means that the models will only be appropriate in certain contexts (see fig. 3.1). To become well posed, shape from texture must make strong assumptions about the world that are only valid for a limited class of scenes. If standard piecewise smoothness priors are used, then texture models will also predict biases toward the frontoparallel plane, as observed experimentally (Bülthoff and Mallot 1988). Stronger predictions can be made by testing the predictions of a specific model (see Blake, Bülthoff, and Sheinberg 1993).

What we have seen in this section are examples of individual visual modules. Although it is possible to interpret some psychophysical experiments as biases toward "reasonable" prior assumptions, we have stressed that the less constraint the likelihood function places on the scene, the stronger the bias. All these theories make strong contextual assumptions, and the visual system must be able to automatically verify whether the context is correct before believing the output of the model. In the next section we will look at how several different visual modules can be integrated to achieve a more robust interpretation of the visual world.

3.3 BAYESIAN THEORIES OF MULTIPLE VISUAL MODULES

It has become standard practice for computational theorists and psychophysicists to assume that different visual cues are computed in separate modules (Marr 1982) and thereafter only weakly interact with each other. Marr's theory did not fully specify this weak interaction but seemed to suggest that each module separately estimated scene properties, such as depth and surface orientation, and then combined the results in some way.[3] A more quantitative theory, which has experimental support (Bruno and Cutting 1988; Dosher, Sperling, and Wurst 1986; Maloney and Landy 1989), involves taking weighted averages for mutually consistent cues and using a vetoing mechanism for inconsistent cues. A further approach by Poggio and collaborators (Poggio, Gamble, and Little 1988) based on Markov random fields has been implemented on real data.

Coupling of Modules

The Bayesian approach suggests an alternative viewpoint for the fusion of visual information (Clark and Yuille 1990). This approach stresses the necessity of taking into account the prior assumptions used by the individual cues.

These assumptions may conflict or be redundant. In either case, i
better results can often be achieved by strongly coupling the
contrast to the weak methods proposed by Marr or the weight
theories, though weak coupling may indeed be appropriate in son
(Landy et al. 1995).

To see the distinction between weak and strong coupling,
have two sources of depth information, f and g. Marr's theory wo
specifying two posterior distributions, $P_1(S|f)$ and $P_2(S|g)$, for th
modules. Two MAP estimates of the scene, S_1^* and S_2^*, would be
by each module and the results combined in some unspecified
figure 3.3 for an overview of weak and strong coupling.

Weak Coupling

Although the weighted averages theories are not specified in
framework, one way to obtain them would be to multiply the
gether to obtain $P(S|f,g) = P_1(S|f)P_2(S|g)$. If the MAP estimates
from the two theories are similar, then it is possible to use p
theory and find, to first order, that the resulting combined MAP e
is a weighted average of S_1^* and S_2^* (see Yuille and Bülthoff 1996).

Both Marr's and the weighted averages approach would be cl
as weak (Clark and Yuille 1990) because they assume that the i
conveyed by the a posteriori distributions of the two modules is in
But, as we have argued, the forms of the prior assumptions ma
information to be dependent or even contradictory.

By contrast, the Bayesian approach would require us to specify
likelihood function $P(f,g|S)$ for the two cues and a single prior
$P(S)$ for the combined system. This will give rise to a distributi
given by

$$P(S|f,g) = \frac{P(f,g|S)P(S)}{P(f,g)}$$

and in general will not reduce to $P_1(S|f)P_2(S|g)$. A model like
cannot be factorized is considered a form of strong coupling (Clark
1990).

We now discuss an important intermediate case between weak
coupling. Consider two modules $P_1(f|S)$, $P_1(S)$ and $P_2(g|S)$, $P_2(S)$.
pose that the likelihood function for the combined cues can be
$P(f,g|S) = P_1(f|S)P_2(g|S)$. Then we get correct Bayesian integrati
by using model (fig. 3.3c) provided the priors for the two m
identical (i.e., $P_1(S) = P_2(S)$) and the prior for the coupled mod
changed (i.e., $P(S) = P_1(S) = P_2(S)$). For historical reasons, we refe
"weak" coupling. It is not unusual, however, for existing vision n
use different priors (for example, binocular stereo modules typically
wise surface smoothness assumptions, while structure-from-motion

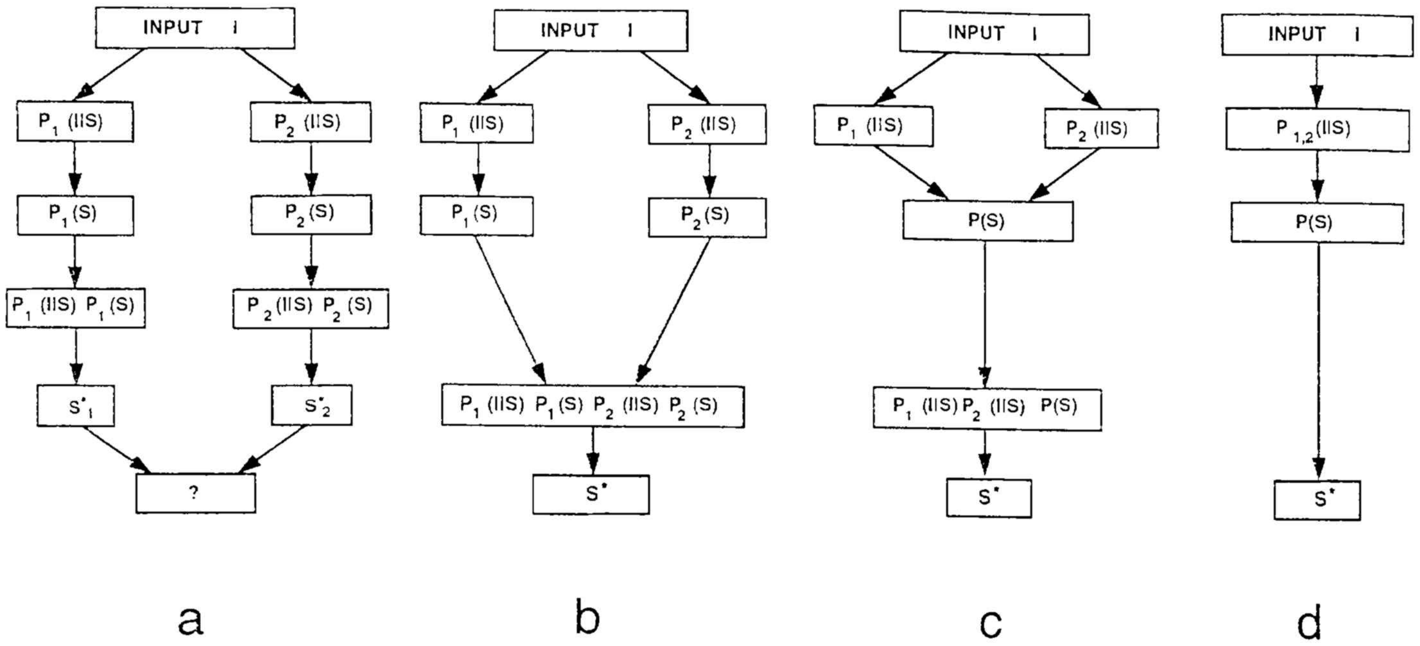

Figure 3.3 Different types of coupling between modules. Diagram (a) shows form of weak coupling where two modules act independently, with their own likelihood functions $P(I|S)$ and priors $P(S)$, producing MAP estimators, $S^* = \arg\max_S P(I|S)P(S)$, as outputs, which are then combined in unspecified manner. Observe that this method provides no information at all about uncertainty of each estimate. Diagram (b) shows weak coupling where likelihood functions and priors of two modules are multiplied together and then MAP estimator is calculated. Such coupling would yield weighted combination of cues in some circumstances (see Yuille and Bülthoff 1995). In (c) likelihood functions of modules are combined with single prior for combined modules and then MAP estimator is found. This case is on borderline between weak and strong coupling. It is weak if prior $P(S)$ is same as that used for individual modules, and it is strong otherwise. Diagram (d) shows strong coupling where it is impossible to factor the likelihood function of the combined modules into the likelihood functions for the individual modules.

often assume rigidity). In this case, the prior for the coupled m different from the priors for the individual modules and we say the are "strongly" coupled.

The need for formulating cue combination by (3.2) may seem ol statisticians. Indeed, some might argue that the need for strong co only an artifact of incorrect modularization of early vision. We hav thy for such a viewpoint.

Observe also that there is no need for a veto mechanism betwee our framework. Such a mechanism is only needed when two cues a conflict. But this conflict is merely due to using mutually inconsisten for the two cues; if we combine the cues using (3.2), this conflict va

In the next subsection we will give one example of cue integra will demonstrate that for shading and texture the likelihood functio cannot be factored, and thus strong coupling is required.

Strong Coupling of Shading and Texture

We now consider coupling shading with texture. First, we argue th case the likelihood functions are not independent and that strong co usually required. Second, we describe an experiment from Bülthoff an (1990), which shows how the integration of shading and texture inf gives a significantly more accurate depth perception than that att shading and texture independently.

As we discussed in the previous section, standard theories of sh shading and texture, in particular their likelihood functions, are only certain contexts; moreover, these contexts are mutually exclusive. Sh shading assumes that the image intensity is due purely to shading c albedo variations), while shape from texture assumes that it is due on presence of texture.

To couple shading with texture, we must consider a context w image intensity is generated both by shading and textural processe context may be modeled by a simple reflectance model

$$I(x) = a(x)R(\vec{n}(x)),$$

where the texture information is conveyed by the albedo term $a(x$ shading information is captured by $R(\vec{n}(x))$. It is typically assumed reflectance function is Lambertian $\vec{s}\cdot\vec{n}$ and that there are a class of el texture elements painted onto the surface in a statistically uniform tion. This will induce a distribution on the albedo, $a(x)$, that depend geometry of the surface in space.

Typically, texture modules assume that $R(\vec{n}(x)) = 1$, $\forall x$, while modules set $a(x) = 1$, $\forall x$. For the coupled system, these assumpt invalid (see fig. 3.4). The shading module has to filter out the albed texture, while the texture information must ignore the shading inf $R(\vec{n}(x))$. For some images, it may be possible to do this filtering indep

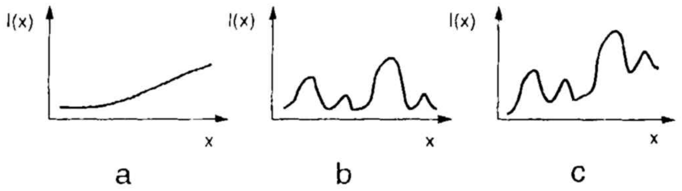

Figure 3.4 Difficulty of decoupling shading and texture cues. Graph (a) shows typical intensity profile for Lambertian surface with constant albedo, context in which shape from shading can be computed. Graph (b) shows intensity profile for surface with strong albedo variation, context for shape from texture. Graph (c) shows intensity profile when both cues are present. Separating this profile into its shading in (a), and textural components in (b), is hard in general. In Bayesian terms this is because likelihood function for combined shading and texture cannot, in general, be factored into likelihood functions for two individual cues.

(i.e., the texture model can filter out $R(\vec{n}(x))$ without any input from the shading module, and vice versa). In general, however, distinguishing between $R(\vec{n}(x))$ and $a(x)$ is not at all straightforward. Consider an object made up of many surface patches with Lambertian reflectance functions and differing albedos. For such a stimulus, it seems impossible to separate the intensity into albedo and shading components *before* computing the surface geometry. Thus we argue that the likelihood functions for the combined shading and texture module usually cannot be factored as the product of the likelihood functions for the individual modules, and hence strong coupling is required (a similar point is made by Adelson and Pentland 1991).

Other examples of "unfactorizable cues" include the phenomena of *cooperative processes*, where the perception of shape from shading depends very strongly on contour cues (Knill and Kersten 1991) or on stereo curvature cues (Buckley, Frisby, and Freeman 1993).

In addition we argue that, because more information is available in the likelihood term of the combined module, the prior assumption on the surface geometry can be weakened. Hence there is both less bias towards the fronto-parallel plane from the priors and more bias toward the correct perception from the shading and texture cues.

In the experiment reported below (fig. 3.5), shape from shading and shape from texture alone gave strong underestimations of orientation, yet the combined cues gave almost perfect orientation. Such a result seems inconsistent with Marr's (1982) theory or with coupling by weighted averages. Instead, it seems plausible that this is an example of strong coupling between texture and shading, with a weak prior toward piecewise smooth surfaces. The only way that these results might be consistent with weak coupling would be if simple filters could decompose the image into texture and shading parts, hence factorizing the likelihood function, and then combine the cues using the same prior used by both modules. This prior would have to be so weak that the likelihood functions of the two modules would dominate it.

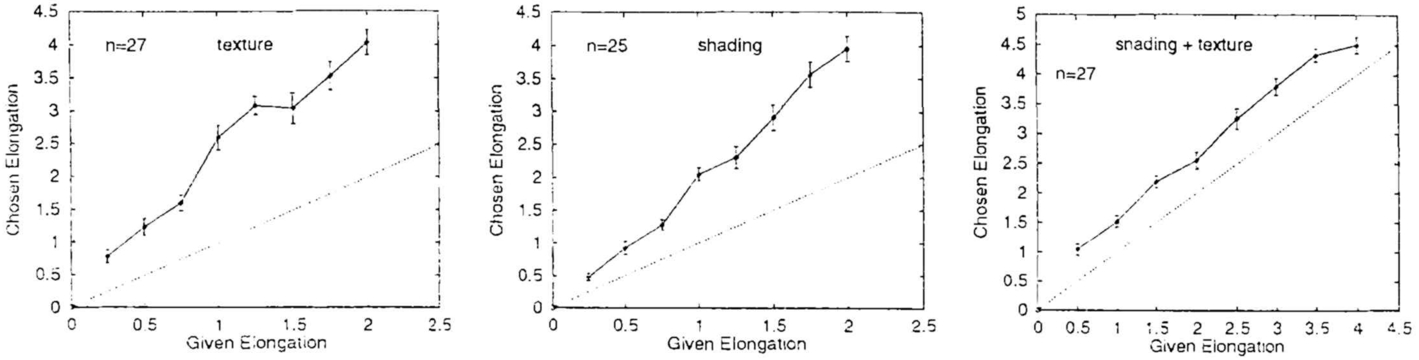

Figure 3.5 Psychophysical experiments on integration of shading and texture. In adjustment task subjects interactively adjusted shading or texture of simulated ellipsoid of rotation (seen by one eye) in order to match form of given ellipsoid seen with both eyes (in stereo). Ellipsoids were seen end-on so that outline was same for both surfaces. Shape from shading and shape from texture individually lead to strong underestimation of shape, that is, shading or texture of ellipsoid with much larger elongation had to be simulated in order to match given ellipsoid (slope $\gg 1$). If shading and texture are presented simultaneously, shape is adjusted almost correctly (slope $= 1$). Redrawn from Bülthoff and Mallot (1990).

3.4 CONTEXTS AND COMPETITIVE PRIORS

As we have seen, the current models for visual cues make prior assumptions about the scene. In particular, the likelihood function often assumes a particular context—for example, Lambertian surfaces. The choices of priors and contexts is very important; they correspond to the "knowledge" about the world used by the visual system. In particular, the visual system will only function well if the priors and the contexts are correct.

What types of priors or contexts should be used? The influential work of Marr (1982) proposed that vision should proceed in a feedforward way. Low-level vision was performed by vision modules that each used a single general purpose prior such as rigidity for structure from motion or surface smoothness for stereo. Such priors were called "natural constraints" by Marr (1982). Low-level vision culminated in the $2\frac{1}{2}$-D sketch, a representation of the world in terms of surfaces. Finally, object specific knowledge was used to act on the $2\frac{1}{2}$-D sketch to perform object recognition and scene interpretation. Because the types of priors suggested for low-level vision are general-purpose we will refer to them as generic priors.

The question naturally arises whether models of early vision should have one generic prior. It is clear that when designing a visual system for performing a specific visual task, the prior assumptions should be geared toward achieving the task. Hence it can be argued (Clark and Yuille 1990; Yuille and Clark 1993) that a set of different systems geared toward different tasks and competing with each other is preferable to a single generic prior.

These competitive priors should apply both to the material properties of the objects and their surface geometries. We will sketch how the idea applies to competing models for prior geometries and then give a general mathematical formulation. An example of competing priors for material properties is described in Yuille and Bülthoff (1996).

To make this more precise, consider the specific example of shape from shading. Methods based on an energy function, such as Horn and Brooks (1986), assume a specific form of smoothness for the surface. The algorithm is therefore biased toward the class of surfaces defined by the exact form of the smoothness constraint, which prevents it from correctly finding the shape of surfaces such as spheres, cylinders, or cones.

On the other hand, there already exist algorithms that are guaranteed to work for specific types of surfaces. Pentland (1989) designed a local shape from shading algorithm that, by the nature of its prior assumptions, is guaranteed to work for spherical surfaces. Similarly, Woodham (1981) has designed a set of algorithms guaranteed to work on developable surfaces, a class of surfaces that includes cones and cylinders.

Thus, instead of a single generic prior, it would seem more sensible to use different theories; in this case, Horn and Brooks's, Pentland's, and Woodham's, in parallel. A fitness criterion is required for each theory to determine how well it fits the data; these criteria can then be used to determine should be applied.

More formally, let $P_1(f), P_2(f), \ldots, P_N(f)$ be the prior assump of competing models with corresponding imaging models $P_1(I|f$ We assume prior probabilities $P_p(a)$ that the ath model is the c so $\sum_{a=1}^{N} P_p(a) = 1$. This leads to a set of different modules, each the solution that maximizes their associated conditional probabil

$$P_1(f|I) = \frac{P_1(I|f)P_1(f)}{P_1(I)},$$

$$P_N(f|I) = \frac{P_N(I|f)P_N(f)}{P_N(I)}.$$

Let our space of decisions be $D = \{d, i\}$, where d specifies th labels the model we choose to describe it. We must specify a $L(d, i:f, a)$, the loss for using model i to obtain scene d when th should be a and the scene is f, and define a risk

$$R(d, i) = \sum_{a} \int L(d, i:f, a)P_a(f|I)P_p(a)[df],$$

where, for example, we might set $L(d, i:f, a) = -\delta(f - d)\delta_{ia}$ penalized by δ_{ia} for not finding the right model and by $-\delta(f$ finding the right surface). Here $\delta(f - d)$ denotes the Dirac delta δ_{ia} is the Kronecker delta, where $\delta_{ia} = 1$ if $i = a$ and 0 otherwis decision corresponds to picking the model i and the scene d th the risk.

A number of psychophysical experiments seem to require in terms of competitive priors. In all cases the perception of th be made to change greatly by small changes in the stimuli; so experiments would also seem to require strong coupling.

Transparency

Kersten et al. (1991) describe a transparency experiment in whi can be interpreted either as a pair of rectangles rotating rigidl common axis or as two independent rigid rectangles rotating a own axis (fig. 3.6). The competitive priors correspond to assumi rectangles are coupled together to form a rigid object or that th are uncoupled and move independently; by adjusting the transp either perception can be achieved. Interestingly, the perception o pled motion is only temporary and seems to be replaced by the p the coupled motion. We conjecture that this is due to the buildup for the coupled hypothesis over time. The uncoupled interpretatio supported because it agrees with the transparency cue. Over a lon time, however, the uncoupled motion is judged less likely th

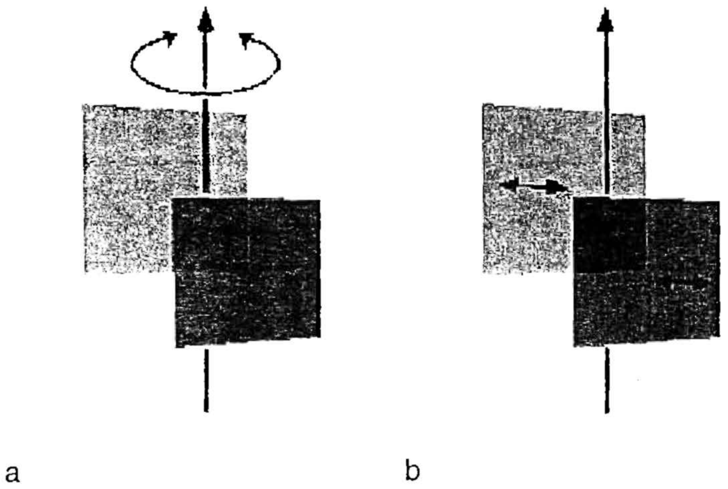

Figure 3.6 In order to study how structure from motion interacts with transparency cues, two rigidly coupled planar surfaces rocking back and forth about common vertical axis midway between them were simulated. Intensity relationships of various regions of overlapping faces bias apparent transparency faces and therefore depth ordering of faces. a. If simulated depth ordering was consistent with depth ordering suggested by transparency cue rigid motion around common axis was perceived. b. If motion parallax and transparency cue were contradictory, nonrigid motion of two faces slipping and sliding over one another was perceived. Redrawn from Kersten et al. (1991).

motion, although this hypothesis does require a relative ordering of competing explanations, which could be implemented by prior probabilities. It is not hard to persuade oneself that coupled motion is more natural, and hence should have higher prior probability, than uncoupled motion.

Specular Stereo

Blake and Bülthoff's (1990, 1991) work on specular stereo shows how small changes in the stimuli can dramatically change the perception. In these experiments a sphere is given a Lambertian reflectance function and is viewed binocularly. A specular component is simulated and is adjusted so that it can lie in front of the sphere, between the center and the surface of the sphere or at the center of the sphere (fig. 3.7). If the specularity is at the center, it is seen as a lightbulb and the sphere appears transparent. If the specularity lies in the physically correct position within the sphere (halfway between the center and the surface), the sphere is perceived as being a glossy, metallic object. It is interesting that, before doing the experiment, most people think that the specularity should lie on the convex surface and not behind. If the specularity lies in front of the sphere, it is seen as a cloud floating in front of a matte sphere. We can say that there are three competing assumptions for the

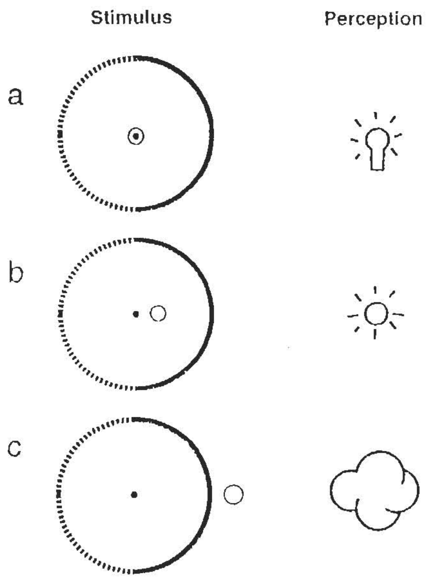

Figure 3.7 Specular stereo where hemisphere is viewed binocularly. In (a) spe white ellipsoid) is adjusted to lie behind center of sphere; it is perceived as ligh behind transparent sphere. In (b) specularity lies in approximately the correct p hemisphere is perceived to be metallic, with specularity appearing as image of lig specularity lies in front of hemisphere, as in (c), it is perceived as cloud floating hemisphere.

material of the sphere: (1) transparent, (2) glossy, and (3) matte; the model depends on the data. In addition, if the sphere is arranged s Lambertian part has no disparity, the stereo cue for the specularity the concave/convex ambiguity from the shading cues (see Blake an 1990, 1991) for details.

Amodal Completion

Nakayama and Shimojo (1992) describe an impressive set of sterc ments that suggest the visual system can interpret the world in surfaces that may partially occlude each other. The visual system forms significant interpolation in regions that are partially hidden. I ple, one can obtain a strong perception of a Japanese flag (see fig. when the stimulus contains very little information, provided that th parts of the flag are occluded by another surface.

Nakayama and Shimojo (1992) themselves argue that their ex can be described by having a set of competing hypotheses, $i =$ about the possible scene and corresponding image formation model

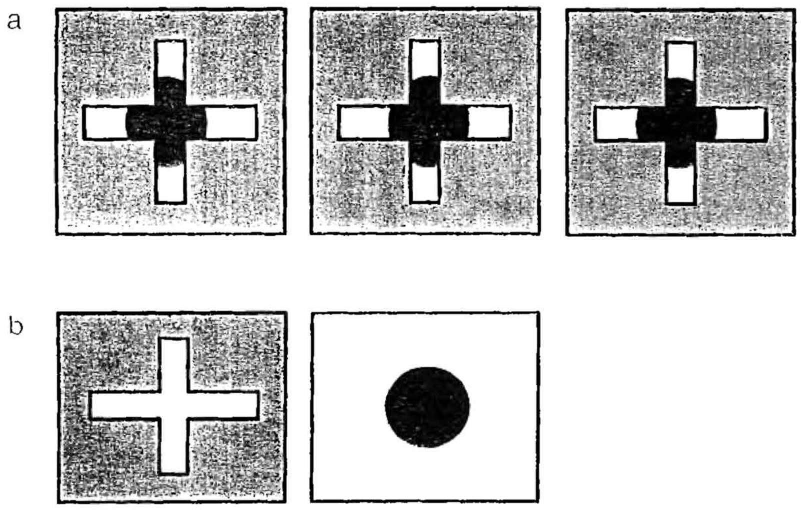

Figure 3.8 Binocular stereo cues for surfaces occluding each other. Observers of stereo pair (left pair for uncrossed fusion and right pair for crossed fusion) usually see planar surface, with cross-shaped hole in its central region, floating above surface with circle at its center (see decomposition below).

They suggest picking the interpretation j that maximizes $P_j(S_j|I) = P_j(I|S_j)/\{\sum_k P(I|S_k)\}$—which can be seen as a special case of our competitive prior formulation. They also argue that this is related to the generic viewpoint hypothesis (Freeman 1993)—if a regularity appears in an image, then the regularity is due to a regularity in the scene rather than an accidental result of the viewpoint. Recently, Bülthoff, Kersten, and Bülthoff (1994) showed that the presumption of a generic viewpoint can be extended also to the domain of illumination and the resulting shadow and lighting effects. Given an accidental view or a sequence of views of an object, the human visual system can make use of global information from the illumination (shadows) to determine the object's shape and properties. For example, shadow information strongly biases the perception of the horizontal bar in Nakayama and Shimojo's stereogram of a cross to appear nonplanar (fig. 3.9).

3.5 DISCUSSION

The competitive prior approach assumes that there is a large set of possible hypotheses about scenes in the world and that these scenes must be interpreted by the set of hypotheses, or competing priors, that best fit the data. We envision a far larger and richer set of competing priors than the natural constraints proposed in Marr (1982) or the regularizers occurring in regularization theory (Poggio, Torre, and Koch 1985).[4] These priors arise from the categorical structure of the world (see also Knill, Kersten, and Yuille, 1996).

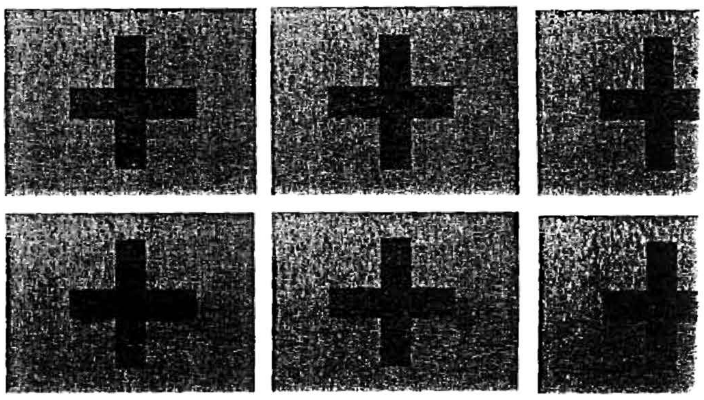

Figure 3.9 Upper stereogram (left pair for uncrossed fusion and right pair for cros shows "stereocross" by Nakayma and Shimojo (1992), which can be seen either as frontoparallel horizontal bar or as cross with wings (horizontal bar bent toward the Lower stereogram shows same three-dimensional layout but with added shadow information can disambiguate many interpretations of upper stereogram and most will see horizontal bar bent toward the observer (Bülthoff, Kersten, and Bülthoff 19

How sophisticated must these contextural priors be? In this ch have only considered priors for low-level tasks such as surface estima we see no reason why they should not reach up to object recogni scene interpretation. At an intermediate stage we should mention t esting results of Kersten, Mamassian, and Knill (1994), which sho humans make use of shadow information for depth perception. In t periments the perceived motion of a ball in a box was strongly aff the motion of its shadow. But for this shadow information to be m the visual system must have decided that the geometry of the sce box—in other words, that the shadow was projected from a ball planar surface at the bottom of the box.

It is clear that the most effective computer vision systems strong contextural knowledge and are geared to achieving specific tasks. extent should the competing priors be geared toward specific tasks one would like to have priors that accurately model all aspects of t scene, but this may be unrealistic. Instead, it would be simpler to ha that accurately model the aspects of the world the visual system know about, although the decision rules must be sophisticated e prevent the system from constantly hallucinating the things it desir It is tempting to consider the hallucinations induced by sensory de as an example of the prior imposing nonexistent structure on the da ing up priors in this task-dependent way seems a sensible strategy fo ing a visual system, but is there any evidence that biological sys designed like this? It may be hard to test for humans because o system appears very general-purpose, but we believe that many lower animals behavior can be interpreted this way already. This emphasis on task dependence is at the heart of recent work on active vision (Blake and Yuille 1992). By making very specific prior assumptions about certain structures in the scene, and ignoring everything else, it has proven possible to design autonomous vehicles capable of driving at high speeds on the autobahn (Dickmanns, Mysliwetz, and Christians 1990).

Clearly the range of visual tasks that we can achieve is determined by the information, $P(S|I)$, we have about the scene. Thus the issue of what visual tasks we can achieve, or what scene parameters we can estimate, is determined by the form of $P(S|I)$, assuming we have exploited all our prior knowledge. It may well be that $P(S|I)$ contains enough information for us to make a reliable decision about whether one object is in front of another, but not enough to decide on the absolute depth values of the objects themselves.

In its current formulation the competitive prior approach leaves many questions unanswered. In particular, how many priors should there be, and how can one search efficiently through them? We believe that the answer to the first question is largely empirical and that, by building increasingly sophisticated artificial vision systems and performing more psychophysical experiments, it will be possible to determine the priors required. To search efficiently between competing priors seems to require a sophisticated mixed bottom-up and top-down strategy of the type described by Mumford (1992). In such an approach, low-level vision is constantly generating possible interpretations while, simultaneously, high-level vision is hypothesizing them and attempting to verify them.

In this Bayesian framework we have said nothing about the algorithms that might be used to make the decisions. In this we are following Marr's (1982) levels of explanation, where a distinction is made between the high-level information processing description of a visual system and the detailed algorithms for computing it.[5] Thus we may hypothesize that a specific visual ability can be modeled by a Bayesian theory without having to specify the algorithm. In a similar style, Bialek (1987) describes various experiments showing that the human visual system approaches optimal performance for certain tasks, such as estimating the number of photons arriving at the retina (Sakitt 1972), even though precise models for how these computational tasks are achieved are currently lacking. Certainly the algorithms used to compute a decision may be complex and require intermediate levels of representation. For example, a shape from texture algorithm might require first extracting textural features, which are then used to determined surface shape. Thus Bayesian theories certainly do not imply "direct perception" (Gibson 1979) in any meaningful sense. The issues of when to introduce intermediate levels of representations and of finding algorithms to implement Bayesian theories are important unsolved problems.

Finally, we have used a broad brush and not given specific details of many theories. Though much progress has been made, existing vision theories are

still not as successful as one would like when implemented on r
Bayesian decision theory gives a framework, but there are many
need to be filled in. For example, the Bayesian approach emph
importance of priors but does not give any prescription for find
Although workers in computational vision have developed a
promising priors for modeling the world, it is an open research tas
refine and extend these models in order to build systems of the typ
here. Fortunately, the Bayesian framework is able to incorporate lea
Kersten et al. 1987), and the success of (Bayesian) hidden Markov
speech recognition (Paul 1990) suggests that it may be practica
Bayesian theories. It is particularly interesting to ask whether pri
learned for new task.

3.6 CONCLUSION

In this chapter we have argued for a framework for vision
Bayesian theory. Such a theory will inevitably cause biases toward
assumptions of the theory, particularly for the impoverished stimul
psychophysicists.

This approach suggests that, when coupling visual cues, one mu
mind the interdependence between the cues and, in particular, the
sumptions they might be subject to. In many cases, this will lead
coupling between visual cues rather than the weak coupling pro
other theorists.

We also argue that the prior assumptions used by the visual sys
be considerably more complex than the natural constraints and gen
commonly used. Instead, there seems to be evidence for a competi
prior assumptions or contexts, which also seems to be a pragmati
design a visual system to perform visual tasks. It may be better t
visual systems in terms of modules that are geared toward speci
tasks in restricted contexts than modules based on the traditional co
visual cues.

ACKNOWLEDGMENTS

We would like to thank Andrew Blake, Stuart Geman, Dan Kerste
Knill, Larry Maloney, David Mumford, Ken Nakayama, Whitman
and Pawan Sinha for helpful discussions. Some of the material in this
appears also in the book *Perception as Bayesian Inference*, ed. D. Knill
Richards (Cambridge University Press, 1995). We would like to th
reviewers for their very helpful comments.

NOTES

1. This is somewhat similar to the idea of "unconscious inference" developed by von
(1910) and Gregory (1970).

2. That is, we assume the observed intensity at each point in the image is modeled by a Gaussian distribution with a mean given by the $\vec{n} \cdot \vec{s}$, where $\vec{n}$ is the normal of the corresponding point in space, and variance σ^2.

3. "The principle of modular design does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular" (Marr 1982, 102.).

4. The Bayesian approach to vision, and to statistics in general, emphasizes the importance of specifying precisely which prior assumptions are being used. Thus it intrinsically leads to the search and identification of priors/constraints.

5. Note that we treat the choice of modules and their coupling as being high-level descriptions rather than algorithmic ones.

REFERENCES

Adelson, E., and Pentland, A. (1991). The perception of shading and reflectance. In Blum, B. (Ed.), *Channels in the visual nervous system*. London: Freud.

Bayes, T. (1783). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, London, 53*, 370–418.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. 2d ed. New York: Springer.

Bialek, W. (1987). Physical limits to sensation and perception. *Annual Review of Biophysics and Biophysical Chemistry, 16*, 455–478.

Blake, A., and Bülthoff, H. (1990). Does the brain know the physics of specular reflection? *Nature, 343*, 165–168.

Blake, A., and Bülthoff, H. (1991). Shape from specularities: Computation and psychophysics. *Philosophical Transactions of the Royal Society, London, B331*, 237–252.

Blake, A., Bülthoff, H., and Sheinberg, D. (1993). An ideal observer model for inference of shape from texture. *Vision Research, 33*, 1723–1737.

Blake, A., and Yuille, A. (Eds.). (1992). *Active vision*. Cambridge, MA: MIT Press.

Bruno, N., and Cutting, J. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General, 117*, 161–170.

Buckley, D., Frisby, J., and Freeman, J. (1993). Lightness perception can be affected by surface curvature from stereopsis. Artificial intelligence vision research unit preprint, Department of Psychology, University of Sheffield.

Bülthoff, I., Kersten, D., and Bülthoff, H. (1994). General lighting can overcome accidental viewing. *Investigative Ophthalmology and Visual Science, 35*(4), 1741.

Bülthoff, H., and Mallot, H. (1988). Interaction of different modules in depth perception. *Journal of the Optical Society of America, A5*, 1749–1758.

Bülthoff, H., and Mallot, H. (1990). Integration of stereo, shading and texture. In A. Blake and T. Troscianko (Eds.), *AI and the eye*, 119–146. Chichester: Wiley.

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

Dickmanns, E., Mysliwetz, B., and Christians, T. (1990). An integrated spatiotemporal approach to automated visual guidance of autonomous vehicles. *IEEE Transactions on Systems, Man, and Cybernetics, 20*, 1273–1284.

Dosher, B., Sperling, G., and Wurst, S. (1986). Trade-offs between stereopsis and
luminance covariance as determinants of perceived 3-D structure. *Vision Research, 2*6.

Freeman, W. (1993). Exploiting the generic view assumption to estimate scene par
Proceedings of the Fourth International Conference on Computer Vision, 347–356. IEEE
Society Press. Los Alamitos, CA: Berlin.

Gibson, J. (1979). *The ecological approach to visual perception*. Boston: Houghton Miffl

Gregory, R. (1970). *The intelligent eye*. New York: McGraw-Hill.

Horn, B. (1986). *Robot vision*. Cambridge, MA: MIT Press.

Horn, B., and Brooks, M. (1986). The variational approach to shape from shadin
Vision, Graphics, and Image Processing, 2, 174–208.

Kersten, D., Bülthoff, H., Schwartz, B., and Kurtz, K. (1991). Interaction between tr
and structure from motion. *Neural Computation, 4*, 573–589.

Kersten, D., Mamassian, P., and Knill, D. (1994). Moving cast shadows and the pe
relative depth. Technical report no. 6. Max Planck Institute for Biological Cyber
ingen, Germany.

Kersten, D., O'Toole, A., Sereno, M., Knill, D., and Anderson, J. (1987). Associative
scene parameters from images. *Journal of the Optical Society of America, A26*(23), 499

Knill, D., and Kersten, D. (1991). Apparent surface curvature affects lightness
Nature, 351, 228–230.

Knill, K., Kersten, D., and Yuille, A. (1996). A Bayesian formulation of visual
In Knill, D. and Richards, W., (Eds.), *Perception as Bayesian inference*. Cambridge:
University Press.

Koenderink, J., van Doorn, A., and Kappers, A. (1992). Surface perception in picture
and Psychophysics, 52, 487–496.

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

Maloney, L., and Landy, M. (1989). A statistical framework for robust fusion of dep
tion. *Proceedings of the SPIE: Visual Communications and Image Processing*, Part 2,
Boston.

Mamassian, P., and Kersten, D. (1994). Perception of local orientation on shad
surfaces. *Vision Research* Submitted.

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

Mumford, D. (1992). On the computational architecture of the neocortex. II. T
cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.

Nakayama, K., and Shimojo, S. (1992). Experiencing and perceiving visual surfaces.
1357–1363.

Parisi, G. (1988). *Statistical field theory*. Reading, MA: Addison Wesley.

Paul, D. (1990). Speech recognition using hidden Markov models. *Lincoln Laborator*
(1).

Pentland, A. (1989). Local shading analysis. In B. Horn, B. and M. Brooks (Eds.).
shading, 443–487. Cambridge, MA: MIT Press.

Poggio, T., Gamble, E., and Little, J. (1988). Parallel integration of vision modules.
436–440.

Poggio, T., Torre, V., and Koch, C. (1985). Computational vision and regularization theory. *Nature, 317,* 314–319.

Pollard, S., Mayhew, J., and Frisby, J. (1985). A stereo correspondence algorithm using a disparity gradient limit. *Perception, 14,* 449–470.

Sakitt, B. (1972). Counting every quantum. *Journal of Physiology, 284,* 261.

von Helmholtz, H. (1910). *Treatise on physiological optics.* Trans. from 3d German ed. Vol. 3, ed. J. P. C. Southall. Reprint, New York: Dover, 1962.

Woodham, R. (1981). Analyzing images of curved surfaces. *AI Journal, 17,* 117–140.

Yuille, A., and Bülthoff, H. (1996). Bayesian decision theory and psychophysics. In D. Knill and W. Richards (Eds.), *Perception as Bayesian inference,* Cambridge: Cambridge University Press.

Yuille, A., and Clark, J. (1993). Bayesian models, deformable templates, and competitive priors. In L. Harris and M. Jenkin (Eds.), *Spatial vision in humans and robots.* Cambridge: Cambridge University Press.