Bayesian Decision Theory and Psychophysics

Alan L. Yuille
Division of Applied Sciences,
Harvard University, Cambridge, MA 02138, USA
&
Heinrich H. Bülthoff
Max-Planck Institut für biologische Kybernetik,
Spemannstr. 38
72076 Tübingen, Germany

August 12, 1994

Abstract

We argue that Bayesian decision theory provides a good theoretical framework for visual perception. Such a theory involves a likelihood function specifying how the scene generates the image(s), a prior assumption about the scene, and a decision rule to determine the scene interpretation. This is illustrated by describing Bayesian theories for individual visual cues and showing that perceptual biases found in psychophysical experiments can be interpreted as biases towards prior assumptions made by the visual system. We then describe the implications of this framework for the integration of different cues. We argue that the dependence of cues on prior assumptions means that care must be taken to model these dependencies during integration. This suggests that a number of proposed schemes for cue integration, which only allow weak interaction between cues, are not adequate and instead stronger coupling is often required. These theories require the choice of decision rules and we argue that this choice is important since these rules help capture the task dependent nature of vision. This is illustrated by analysing the generic viewpoint assumption. Finally, we suggest that the visual system uses a set of competing prior assumptions, rather than the single generic priors, or natural constraints, commonly used in computational theories of vision.

1 Introduction

1.1 The Bayesian Decision Theory Approach to Vision

We define vision as perceptual inference, the estimation of scene properties from an image or a sequence of images. 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. 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.[1] In this Chapter we argue that Bayesian decision theory provides a natural frame-

[1] The issue of precisely which scene properties need be estimated is still an open one. We will briefly discuss this in Section 6.

work for modeling perceptual inference. We will discuss the theoretical problems that arise, in particular when combining different visual cues, and propose solutions.

How are these assumptions about the world imposed in vision systems? The Bayesian formulation, see also the introductory Chapter to this book, gives us an elegant way to impose constraints in terms of prior probabilistic assumptions about the world. This approach is based on Bayes formula [1]:

$$P(S|I) = \frac{P(I|S)P(S)}{P(I)}. \tag{1}$$

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 it specifies the probability of obtaining 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 which specifies the relative probability of different scenes occurring in the world, and formally expresses 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 it 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 (1) states: 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 [2] [21] this estimate is derived by choosing a loss function which specifies the penalty paid by the system for producing an incorrect estimate.[2] 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. The loss function emphasizes that the interpretation of the image cannot be divorced from the purpose of the visual system.[3] In Section 4 we will illustrate the idea of loss functions by analyzing the generic viewpoint assumption [4], [25].

The Bayesian framework is sufficiently general to encompass many aspects of visual perception including depth estimation, object recognition and scene understanding. However, 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 [36] which only weakly interact with each other. From the Bayesian perspective, this modularization is often inappropriate, due to the interdependence between visual cues. Hence we argue in Section 3 that the visual cues should be more strongly coupled.

The choice of prior assumptions in the Bayesian framework is very important. Each visual cue, as standardly defined, contains built-in prior assumptions. If these assumptions are being used by the visual system they will inevitably bias perception, particularly for the impoverished stimuli favoured by psychophysicists. Indeed the perceptual biases detected in psychophysical

[2] For other applications of decision theory to vision see [53].

[3] Decision theory can also be used to couple vision directly to action [20].

experiments offer clues about the nature of the prior assumptions being used by the visual system.[4] However the prior assumptions used by theorists to model one visual cue may conflict with those used to model another, and consistency should be imposed when cues are combined.[5]

Moreover, the prior assumptions may be context dependent and correspond to the categorical structure of the world. Each visual module, or coupled groups of modules, will have to determine automatically which priors should be used. This can lead to a system of competitive prior assumptions, see Section 5. Bayesian Decision Theory [2] standardly deals with both competing models of this type and also complex systems of elementary priors indexed by hyper-parameters.

In this Chapter we first describe in Section 2 Bayesian theories for individual cues and argue that several psychophysical experiments can be interpreted in terms of biases towards prior assumptions. Next, in Section 3, we describe ways of combining different depth cues and argue that strong coupling between different modules is often desirable. In Section 4 we introduce the concept of loss function by analyzing the generic view assumption and argue that this concept is crucial for specifying the purpose of the visual system. Then in Section 5 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 6.

2 Bayesian Theories of Individual Visual Cues

We now briefly describe some Bayesian theories of individual visual cues and argue that psychophysical experiments can be interpreted as perceptual biases towards prior assumptions. From (1) we see that the influence of the prior is determined by the specificity of the likelihood function $P(I|S)$. In principle, as described in Section 1, the likelihood function should make no prior assumptions about the scene (though, as we will see, this is often not the case in practice).

In the following we will specifically discuss theories of stereo, 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. A number of theories described here were originally formulated in terms of energy functions [31] or regularization theory [47]. Yet the Bayesian approach incorporates, by use of the Gibbs distribution [43], these previous approaches (see also the Appendix).

Let us now look at one specific example. Shape from shading models typically assume that the scene consists of a single object with known reflectance function. It is usually assumed that there is a single light source direction $\vec{s}$ which can be estimated and that the reflectance function is Lambertian with constant albedo. This leads to an imaging 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 variance of the noise and Z is a normalization factor. The prior model for the surface geometry $P(S)$ typically assumes that the surface is piecewise smooth and biases towards a thin plate or membrane.[6]

Observe that this likelihood function contains the prior assumption that the reflectance function is Lambertian with constant albedo. Moreover, it ignores effects such as mutual illumination

[4] 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 which we need to perceive and interpret correctly.

[5] Although it is conceivable that the human visual system uses conflicting prior assumptions for different cues.

[6] These theories also assume that the occluding boundaries of the object is known. This is helpful for giving boundary conditions.

and self-shadowing. The model is therefore only applicable for a certain limited class of scenes and only works within a certain *context* [7](see Figure 1). A visual system using this module would require a method for automatically checking whether the context was correct. In this section we will assume that the context is fixed and leave the discussion of context selection to our later Section on competitive priors.[8]

Figure 1 about here

What predictions would models of this type make for psychophysical experiments? Clearly, they would predict that the perception of geometry for shape from shading would be biased by the prior assumption of piecewise smoothness (see Figure 2). If we use the models of piecewise smoothness typically used in computer vision then we would find a bias towards frontoparallel surfaces. Such a bias is found in the psychophysical shape from shading experiments by Bülthoff and Mallot[12].

Figure 2 about here

Existing shape from texture models also make similar assumptions about the scenes they are viewing. They typically assume that the scene consists of texture elements scattered on piecewise smooth surfaces. The distribution of these elements on the surface is typically assumed to be statistically homogeneous. A specific example is given in the Chapter by Blake, Bülthoff and Sheinberg. 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 Figure 1. To become well-posed, shape from texture must make strong assumptions about the world which are only valid for a limited class of scenes. If standard piecewise smoothness priors are used then texture models will also predict biases towards the frontoparallel plane, as observed experimentally [12]. Stronger predictions can be made by testing the predictions of a specific model, see Chapter by Blake, Bülthoff and Sheinberg.

Finally, we consider a simplified model of stereopsis.[9] This model again assumes that the world consists of piecewise smooth Lambertian surfaces. The imaging model is defined by saying that a surface with disparity $d(x)$ and intensity $I(x)$ will be mapped to the left and right images I_L and I_R so that $I_L(x + d(x)/2) = I(x) + N_L(x)$ and $I_R(x - d(x)/2) = I(x) + N_R(x)$, where N_L and N_R are additive Gaussian noise [17]. This defines a distribution $P(I_L, I_R | I, d)$ and by introducing a prior $P(I, d)$ and applying Bayes theorem we get

[7] Indeed 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 and 5.

[8] We also point out that ideal observer theories, see the Chapters by Kersten and Knill and by Blake, Bülthoff and Sheinberg, by necessity also operate within a specific context. The experimenter chooses a specific visual task and set of stimuli. He then models the performance of an ideal observer, who knows everything about the task and the stimuli, and compares it to that of a human observer. For the human's performance to be anywhere close to that of the ideal observer would require that humans have visual abilities tuned to this context and are able to automatically adapt to them.

[9] See the Chapter by Belhumeur for a more sophisticated model of this type.

$$P(I, d|I_L, I_R) = \frac{P(I_L, I_R|I, d)P(I, d)}{P(I_R, I_L)}. \tag{2}$$

If we assume that the prior $P(I, d)$ is uniform in I then we can integrate out[10] the surface intensity I to compute the marginal distribution

$$\begin{aligned} P(d|I_L, I_R) &= \int P(I, d|I_L, I_R)[dI] \\ &= (1/Z)e^{-\beta \int (I_L(x+d(x)/2) - I_R(x-d(x)/2))^2 dx} P(d), \end{aligned} \tag{3}$$

where $P(d)$ is the prior for the disparity, Z is a normalization constant, and β is proportional to the inverse of the variance of the noise models.

Such a model, using standard piecewise smoothness priors for $P(d)$, will once again predict the observed biases towards the frontoparallel plane, see [12]. Moreover, the strength of these biases will depend on the ambiguity of the matching between the images, see Figure 3. If the images have little variation then the likelihood function gives little constraint on $d(x)$ (many functions $d(x)$ will have non-zero probability) and the perception is strongly biased towards the prior assumptions on the geometry. Conversely, if the images have a lot of variation then there will be little ambiguity in the matching and so the likelihood function $P(I_L, I_R|d)$ will put strong constraints on the form of $d(x)$ (only one function $d(x)$ will have non-zero probability). If the image variations are periodic or semi-periodic then the likelihood function will have several peaks and there will be matching ambiguity which can result in the well-known wallpaper illusion.

Figure 3 about here

This suggests that the less the matching ambiguity then the weaker the bias towards prior assumptions. Experimental support for this comes from [16], see Figure 4, who tested the perceived depth gradient between a pair of feature points as a function of the dissimilarity between the features. The greater the dissimilarity between features then the less the perceived bias towards the frontoparallel plane. These experiments were consistent with a Bayesian theory [58] which formulated stereo as a surface reconstruction problem and interpreted the experiments as a bias towards prior assumptions which weakens as the likelihood function puts stronger constraints on the disparities.

Figure 4 about here

It seems difficult for other types of stereo theories to explain these experiments. Most theories based on feature matching (e.g. [46]) obtain depth by trigonometry after matching. They will either match the features correctly, getting one percept, or incorrectly, getting another. There seems to be no mechanism by which they can get the observed differential bias depending on the form of the features.

We stress that Bayesian theories described in this Section are intended as illustrations and only give qualitative explanations for these experiments. To give a full quantitive explanation

[10] This is possible because our assumptions have made $P(I, d|I_L, I_R)$ a Gaussian in I – which is straightforward to integrate analytically.

would require precise specifications of all the adjustable parameters in the Bayesian theory. Attempts of this type are underway, see work by [57, 29, 54] on motion perception and by Blake, Bülthoff and Sheinberg[6] on texture. This is an important research direction but it is not the main focus of this Chapter. Instead our goal is to give an overview of Bayesian theories for visual perception which contrasts them with alternative formulations and focusses on qualitative agreement with experiments.

The main focus of this Section is to give examples af visual modules, to show that it is possible to interpret some psychophysical experiments as biases towards "reasonable" prior assumptions and to stress that the less constraint the likelihood function places on the scene then the stronger the bias. Finally, we emphasize that 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.

3 Integration of Visual Cues

It has become standard practice for computational theorists and psychophysicists to assume that different visual cues are computed in separate modules [36] and thereafter only weakly interact with each other. Marr's theory [36] 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.[11] A more quantitive theory, which has experimental support [9], [23], [35], involves taking weighted averages of cues which are mutually consistent and using a vetoing mechanism for inconsistent cues. A further approach by Poggio and collaborators [48] based on Markov Random fields has been implemented on real data.

The Bayesian approach suggests an alternative viewpoint for the fusion of visual information [18]. This approach stresses the necessity of taking into account the prior assumptions used by the individual modules. These assumptions may conflict or be redundant. In either case it seems that better results can often be achieved by strongly coupling the modules in contrast to the weak methods proposed by Marr or the weighted averages theories [9], [23], [35]. See Figure 5 for an overview of weak and strong coupling.

Figure 5 about here

To see the distinction between weak and strong coupling suppose we have two sources of depth information f and g. Marr's theory would involve specifying two posterior distributions, $P_1(S|f)$ and $P_2(S|g)$, for the individual modules. Two MAP estimates of the scene S_1^* and S_2^*[12] would be determined by each module and the results would be combined in some unspecified fashion.

The weighted averages theories are not specified in a Bayesian framework. But one way to obtain them would be to multiply the models together to obtain $P(S|f,g) = P_1(S|f)P_2(S|g)$. If the MAP estimates, S_1^* and S_2^*, from the two theories are similar then it is possible to do perturbation theory and find, to first order, that the resulting combined MAP estimate $S_{1,2}^*$ is a weighted average of S_1^* and S_2^* (See Appendix).

[11] "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" ([36], page 102.).

[12] We assume for the moment that all estimates are MAP, $S^* = \arg\max_S P(S|f)$, but alternative estimates will be discussed in Section 6.

Both Marr's and the weighted averages approach would be characterized as weak [18] because they assume that the information conveyed by the a posteriori distributions of the two modules is independent. But, as we have argued, the forms of the prior assumptions may cause the information to be dependent or even contradictory.

The Markov Random Field approach by Poggio and collaborators is slightly difficult to classify in our scheme. A specific implementation [48] says that "individual modules are therefore only integrated with each other indirectly, through the brightness constraint", which would mean weak coupling. Yet the system may be improved to include feedback between the modules, which might correspond to strong coupling. This Markov Random Field approach is certainly close in spirit to the one we are advocating.

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

$$P(S|f,g) = \frac{P(f,g|S)P(S)}{P(f,g)}, \tag{4}$$

and in general will not reduce to $P_1(S|f)P_2(S|g)$. A model like (4) (which cannot be factorized) is considered a form of strong coupling [18]. An important intermediate case between weak and strong coupling occurs when the likelihood function can be factored as $P(f,g|S) = P(f|S)P(g|S)$, see Figure 5c. If the two individual cues have identical priors and the combined system is given the same prior, i.e. $P(S) = P_1(S) = P_2(S)$, then the coupling is considered weak – though it still differs from Marr's theory or the weighted averages approach. But if the combined prior differs from either of the two individual priors then the coupling is strong. It should be emphasized that it is not unusual for two modules, as formulated by Marr, to have different priors. For example, stereo uses piecewise smoothness and structure from motion uses rigidity. Moreover, because more information is available, the combined prior for two visual modules may not need to be as strong as the priors for the individual modules.[13]

The need for formulating cue combination by (4) may seem obvious to statisticians. Indeed some might argue that the need for strong coupling is only an artifact of incorrect modularization of early vision. We have sympathy for such a viewpoint.

Observe also that there is no need for a veto mechanism between cues in our framework. Such a mechanism is only needed when two cues appear to conflict. But this conflict is merely due to using mutually inconsistent priors when modeling the two cues. If we combine the cues using (4) then this conflict vanishes.

In the next two subsections we will consider some examples of cue integration. We will demonstrate that for shading and texture the likelihood function usually cannot be factored and so strong coupling is required. Next we will describe a system for coupling stereo and monocular cues so that the resulting system has no need for a prior.

3.1 Examples of Strong Coupling

We now give two examples where we argue that strong coupling is advantageous. The first example is for a case where the likelihood function of two cues are not independent. The second

[13] A strict Bayesian would argue that you should never weaken your prior just because more information is available and that the additional information should decrease the dependence on the prior automatically. However, this argument is correct only if the prior is highly accurate. Any visual prior that we can currently imagine is likely to be, at best, a poor approximation and it is sensible to try to reduce the dependence on it.

example shows that when coupling two modules the prior assumptions about the geometry can be significantly altered.

3.1.1 Shape from Shading and Texture

We now consider coupling shading with texture. Firstly, we argue that in this case the likelihood functions are not independent and that strong coupling is usually required. Secondly, we describe an experiment from [12], which shows how the integration of shading and texture information gives a significantly more accurate depth perception than that attained by shading and texture independently.

As we discussed in the previous Section, standard theories of shape from shading and texture, in particular their likelihood functions, are only valid in certain contexts. Moreover, these contexts are mutually exclusive. Shape from shading assumes that the image intensity is due purely to shading effects (no albedo variations) while shape from texture assumes that it is due only to the presence of texture.

To couple shading with texture we must consider a context where the image intensity is generated both by shading and textural processes. Such a context may be modeled by a simple reflectance model

$$I(x) = a(x)R(\vec{n}(x)) \tag{5}$$

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

Typically texture modules assume that $R(\vec{n}(x)) = 1, \forall x$, while shading modules set $a(x) = 1$, $\forall x$. For the coupled system these assumptions are invalid, see Figure 6. The shading module has to filter out the albedo $a(x)$, or texture, while the texture information must ignore the shading information $R(\vec{n}(x))$. For some images it may be possible to do this filtering independently (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.[14, 15]

Figure 6 about here

In addition we argue that, because more information is available in the likelihood term of

[14] A similar point is made by Adelson and Pentland's parable of the painter, the carpenter and the gaffer (lights technician) – see Chapter by Adelson and Pentland.

[15] This is also closely related to the concept of cooperative processes [34] where the perception of shape from shading depends very strongly on contour cues or on stereo curvature cues [10], see Chapter by Kersten and Knill.

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 towards the correct perception from the shading and texture cues.

In the experiment reported below, see Figure 7, 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 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 towards piecewise smooth surfaces.[16]

Figure 7 about here

3.1.2 Coupling Stereo with Controlled Motion

Our second example describes theoretical results where the coupling of two cues can significantly reduce the dependence on prior assumptions about the geometry of the scene.

We restrict ourselves to a world consisting of isolated point features in space. The two depth cues are binocular stereo and monocular depth cues obtained by motion parallax from small head or eye movements. This Section is based on work described in [26]

Consider the two cues independently. For binocular stereo there is the well known correspondence problem, which is illustrated in Figure 8. All the assumptions used to make stereo well-posed – the ordering constraint, piecewise smooth surfaces, the disparity gradient limit – will tend to bias the system towards a single depth plane. Although it is true that the disparity gradient limit theories have some ability to perceive transparent surfaces they will still be fooled by the double nail illusion, see Figure 8.

Figure 8 about here

On the other hand, the monocular depth cues caused by motion parallax will not have a correspondence problem since they will be able to track the feature points. The estimation of depth can then be performed by trigonometry. This estimation, however, is likely to be very inaccurate because the eye/head movements are small, so the baseline for the triangulation is small, and there may be additional uncertainty in the amount of eye/head movement. Nevertheless it is possible to define a probabilistic model for this system to give both the estimated depth values and an estimate of their variability.

Suppose we attempt to weakly couple the stereo and monocular cues for the transparent stimuli shown in Figure 8. The monocular cues would give roughly the correct depth estimates but with large variances. By contrast, the prior used by the stereo system would tend to force the data into a single surface, typically as frontoparallel as possible. Thus the monocular estimates would be more accurate than the stereo estimates but they would have larger variance. So if weak coupling is used we would expect the stereo module to override the monocular cues and the system would yield an incorrect answer.

[16] 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 combining 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 dominate it.

By contrast if we strongly couple the two cues, by multiplying together the likelihood functions for both modules, then the information from the monocular cues will be available to help solve the correspondence problem of stereo hence giving a highly nonlinear interaction between the monocular and the stereo cues. The monocular cues do not need to localize the depths of the features precisely, they only need to be accurate enough to disambiguate the stereo correspondence problem, see Figure 8.

This example illustrates several key features of the strong coupling approach: (i) the interaction between modules can become highly nonlinear, (ii) cues that contain little, or inaccurate, information may nevertheless significantly strengthen the performance of another module provided the inaccuracy can be quantified, and (iii) the dependence on priors can be reduced if more cues are available.

This example is atypical of strong coupling because the resulting combined system does not need a prior assumption. We stress that this is only because we are working in a limited context, of isolated feature points, and will not be true in general.

3.1.3 Mathematics of Monocular and Binocular Strong Coupling

This Section gives mathematical details of the theory for strongly coupling binocular and monocular cues. It can be skipped by readers who are not interested in these details.

Consider a system which has both monocular and binocular depth cues where the scenes consist of isolated feature points. Let there be N feature points, $x^l_i : i = 1, ..., N$, visible in the left image and M points, $x^r_a : a = 1, ..., M$, visible in the right image. Suppose we have a set of monocular depth values $\{x^l_i, d^l_i, \sigma^l_i : i = 1, ..., N\}$ and $\{x^r_a, d^r_a, \sigma^r_a : a = 1, ..., M\}$ where the x's are the positions in the two eyes, the d's are the corresponding monocular depth estimates, and the σ's are the standard deviations of these estimates. For details about how these estimates can be derived see [26]. So the monocular depth estimates $f(x)$ are given by Gaussian distributions:

$$P_l(\{f(x^l_i)\}|\{x^l_i\}) = \frac{1}{Z_l}\prod_{i=1}^{N} e^{-\left(f(x^l_i)-d^l_i\right)^2/2(\sigma^l_i)^2},$$

$$P_r(\{f(x^r_a)\}|\{x^r_a\}) = \frac{1}{Z_r}\prod_{a=1}^{M} e^{-(f(x^r_a)-d^r_a)^2/2(\sigma^r_a)^2}, \tag{6}$$

where Z_l and Z_r are normalization constants (i.e. $Z_l = \prod_{i=1}^{N}\{\sqrt(2\pi)\}\sigma^l_i$). For these monocular cues no priors are needed and so the distributions $P_l(\{f(x^l_i)\}|\{x^l_i\})$ and $P_r(\{f(x^r_a)\}|\{x^r_a\})$ correspond to the likelihood functions of the monocular cues. Priors are not needed because we are assuming as context that the scene consists of isolated feature points. It is straightforward to track these features and estimate their depth by motion parallax induced by eye/head movements. This is, however, a big uncertainty in the depth estimates of these points owing to the difficulty in estimating the eye/head movements (see [26]).

The binocular stereo system computes depth estimates and standard deviations $\{d_s(x^l_i, x^r_a), \sigma_{ia}\}$ assuming that a point labeled i the left image corresponds to a point labeled a in the right image. Let $\{V_{ia} : i = 1, ..., N \; a = 1, ..., M\}$ be binary matching elements which can specify the correspondences between the points in the two eyes. In other words we set $V_{ia} = 1$ if we decide that point i matches point a and set $V_{ia} = 0$ otherwise.

For binocular stereo the likelihood function $P_S(\{x^l_i, x^r_a\}|V, f)$ is given by $(1/Z)e^{-\beta E_S(V,f)}$, where Z is a normalization constant and

$$E_S(V_{ai}, f) = \sum_{a,i} \frac{1}{(\sigma_{ia})^2} V_{ai}(d(x^l_i, x^r_a) - f(x^l_i))^2$$
$$+\lambda \sum_i (1 - \sum_a V_{ia}) + \lambda \sum_a (1 - \sum_i V_{ia}), \quad (7)$$

where we require that points in each image have either one or no matches. Hence the terms $\sum_a V_{ia}$ and $\sum_i V_{ia}$ take values of either 0 or 1. The constant λ is therefore a penalty for unmatched points.

For binocular stereo the matching is ambiguous and so prior assumptions on f or V are needed. Thus the Bayesian theory is of form

$$P_S(V, f | \{x^l_i, x^r_a\}) = P_S(\{x^l_i, x^r_a\} | V, f) P_p(V, f) \quad (8)$$

where $P_p(V, f)$ is a prior assumption on V and f. As discussed previously, most standard choices for $P_p(V, f)$ will attempt to reconstruct a piecewise smooth surface (often biased towards the frontoparallel plane).

When strongly coupling the monocular and stereo cues the prior $P_p(V, f)$ becomes unnecessary and can be discarded. The reason is that, in this context of isolated feature points, there will usually be enough information in the likelihood functions to determine the correct matches. Thus the prior required by the stereo system becomes redundant and can be dropped. Observe that this differs from standard Bayesian statistics where the prior is always kept in and its influence merely degrades gracefully as the likelihood function becomes more specific.

When combining the cues we need to express all the cues in one coordinate system. We choose a coordinate system based on the left eye only and use the V variables to perform this transformation. This gives a strongly coupled theory $P_{SC}(V, f | \{x^l_i, x^r_a\}) = (1/Z) e^{-\beta E(V, f)}$ where

$$E(V, f) = \sum_i \frac{1}{(\sigma^l_i)^2} (f(x^l_i) - d^l_i)^2$$
$$+ \sum_{a,i} \frac{1}{(\sigma^r_a)^2} V_{ai} (f(x^l_i) - d^r_a)^2$$
$$+ \sum_{a,i} \frac{1}{(\sigma_{ia})^2} V_{ai} (d(x^l_i, x^r_a)$$
$$- f(x^l_i))^2, \quad (9)$$

and V is a normalization constant.

In this case weak coupling will simply correspond to multiplying the distribution P_{SC} by the prior P_p. This gives

$$P_W(V, f | \{x^l_i, x^r_a\}) = \frac{P_{SC}(V, f | \{x^l_i, x^r_a\}) P_p(V, f)}{Z_w}, \quad (10)$$

where Z_w is the normalization factor.

Thus the weakly coupled system will show a bias towards the prior assumptions in $P_p(f, V)$ but the strongly coupled system will show no bias.

4 Decision Theory

In this Section we develop the concept of a loss function which we briefly mentioned in the Introduction. This is a key ingredient of Bayesian Decision theory and, by specifying a penalty for making an incorrect perceptual inference, emphasizes the task dependent nature of vision. We illustrate the importance of the choice of loss function by reformulating Freeman's original Bayesian treatment of the generic viewpoint assumption [25]. We argue that decision theory gives the correct framework for treating this assumption. Freeman and Brainard [8] have independently reached a similar conclusion, making similar choices of Gaussian loss functions, see Freeman's Chapter in this book.

Given a Bayesian distribution $P(S|I)$ we must make a decision about the viewed scene. Let the set of allowable decisions be $D = \{d_\mu : \mu \in \Lambda\}$ (i.e. μ labels a decision d_μ and these labels lie in a set Λ.). These decisions will correspond to the set $\{S\}$ of possible scenes. We introduce a loss function $L(S, d)$, which is the penalty for making a decision d when the true scene is S. The loss function can be used to specify which scenes the visual system considers important or the type of errors that it considers acceptable.

If we have enough visual information to determine the scene S uniquely then the optimal decision corresponds to the d^* which minimizes $L(S, d)$. Typically, however, we will only have a probability distribution $P(S|I)$ for the scene. In this case the Bayes' decision minimizes the expected loss, or *risk*, defined by:

$$R(d) = \int L(S, d) P(S|I)[dS]. \tag{11}$$

Conventional statistical estimators can be obtained by an appropriate choice of loss function. If we decide to penalize equally every time we make the incorrect decision and set $L(d, S) = -\delta(S - d)$ (where δ is the Dirac delta function), then we find that $R(d) = -P(d|I)$ and the Bayes decision is the scene d^* that maximizes $P(d|I)$, the maximum a posteriori (MAP) estimator.

The MAP estimator, i.e. the mode of the posterior distribution, is often used in vision but, because it only rewards the system if it attains precisely the right solution, it is suspect to statisticians.[17] If the task requires us only to get precisely the right solution then it should be used, otherwise alternatives are better. One alternative is the minimal variance (MV) estimator whose loss function is $L(S, d) = (S - d)^2$. Thus decisions which are close, but not identical, to the right solution get rewarded. In this case the risk function becomes $R(d) = \int (S - d)^2 P(S|I)[dS]$ and so, by differentiating with respect to d, we see that the optimal decision $d^* = \int S P(S|I)[dS]$ is simply the mean of the distribution. Some typical loss functions are shown in Figure 9.

Figure 9 about here

It should be emphasized that the choice of loss function depends on the visual task that the system is designed to accomplish. To illustrate this we consider Freeman's original formulation of generic viewpoint assumption [25]. This assumption states that the interpretation of the image should not be sensitive to some of the variables, the *generic variables*, which are estimated. Freeman gives an example of shape from shading with unknown light source direction. Thus the image I is considered to be a function of the surface geometry G and the light source direction S.

[17] "... This means that the mode is hard to find and need not be a good summary of the posterior distribution. It is a Bayes rule, but under a rather perculiar loss function..." page 95, [50].

He defines a Bayesian theory

$$P(G,S|I) = \frac{P(I|G,S)P(G,S)}{P(I)}. \tag{12}$$

We must now decide on what we want the system to do. Do we want it to estimate the geometry only and ignore the light source direction? Or do we want to estimate both geometry and source direction simultaneously? If so, how accurate do we want to estimate these variables? Do we want to estimate the light source direction precisely, or do we only need to know them to within a few degrees? Is there sufficient information in $P(G,S|I)$ to provide reliable answers to these questions?

Clearly there are many possible tasks we could ask the system to do and we must choose a loss function suitable for the task. We should also only consider tasks for which we believe that $P(G,S|I)$ contains enough information to accomplish it.

One possibility is that we should attempt to find the geometry exactly but only estimate the light source direction approximately. Thus we could pick a loss function $L(d_G, d_S, G, S) = -G_{ass}(d_S - S)\delta(d_G - G)$ where G_{ass} is a Gaussian function. This strongly penalizes errors in the geometry, d_G, but is tolerant to errors in the light source direction, d_S.

Such a loss function is consistent with the generic viewpoint assumption. It will effectively prefer fat peaks in the probability distribution of S to thin spikes – see Figure 10. Thin peaks clearly do not obey the generic viewpoint assumption since small changes in the estimators lead to very improbable images.[18]

Figure 10 about here

Thus from a decision theoretical standpoint the generic views assumption is equivalent to saying that some parameters need to be measured very accurately and others need only be estimated roughly. This can be achieved by picking the appropriate loss function.[19]

For another example remember the double nail illusion in the previous Section. Consider a Bayesian theory which tries to estimate the orientation of a line joining the two dots in space. Suppose that the variable we are interested in is the precise orientation of the line. There are two possibilities, frontoparallel and frontoperpendicular, depending on the correspondence between features – and each is equally likely if we use a MAP estimator. Now suppose we are only interested in estimating the orientation of the line to within a few degrees, and use a loss function that only penalizes errors greater than this. Then the frontoparallel interpretation (plus or minus a few degrees) becomes far more likely since it is far more stable with respect to orientation changes, see Figure 11.

Figure 11 about here

Finally, we should add that picking the correct loss function is necessary for any Bayesian

[18] Observe that Freeman's original interpretation [25] is different but, would lead to a similar interpretation for this example. He proposes integrating out the S variable by doing a saddle point approximation. This yields a *generic viewpoint factor* which would also favour fat peaks to thin ones.

[19] We are grateful for discussions with P. Belhumeur, S. Geman, D. Mumford and B. Ripley which helped clarify these points.

theory and is far more general than the generic viewpoint assumption. It critically depends on what task the visual system wants to achieve and how badly the system will be penalized if the task is not completed successfully.

5 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 [36] proposed that vision should proceed in a feedforward way. Low level vision should be performed by vision modules which each used a single general purpose prior[20] such as rigidity for structure from motion and surface smoothness for stereo. Low level vision culminated in the 2-1/2 D sketch, a representation of the world in terms of surfaces. Finally, object specific knowledge was used to act on the 2-1/2 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 towards achieving the task. Hence it can be argued [18] [59] that a set of different systems geared towards 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 first sketch how the idea applies to competing models for prior geometries, then develop the theory more rigorously and give an example of competing priors for material properties.

To make this more precise consider the specific example of shape from shading. Methods based on energy function,[21] such as Horn and Brooks [30], assume a specific form of smoothness for the surface. The algorithm is therefore biased towards the class of surfaces defined by the exact form of the smoothness constraint. This prevents it from correctly finding the shape of surfaces such as spheres, cylinders and cones.

On the other hand there already exist algorithms that are guaranteed to work for specific types of surfaces. Pentland [45] designed a local shape from shading algorithm which, by the nature of its prior assumptions, is ensured to work for spherical surfaces. Similarly Woodham [56] has designed a set of algorithms that are guaranteed to work on developable surfaces, a class of surfaces which 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, Pentland and Woodham's, in parallel. A goodness of fitness criterion is required for each theory to determine how well it fits the data. These fitness criteria can then be used to determine which theory should be applied.

[20] Such priors were called natural constraints by Marr [36].

[21] Which can therefore be directly interpreted as Bayesian by using the Gibbs distribution, see the Appendix.

5.1 Theory of Competitive Priors

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

$$P_1(f|I) = \frac{P_1(I|f)P_1(f)}{P_1(I)},$$
$$\vdots$$
$$P_N(f|I) = \frac{P_N(I|f)P_N(f)}{P_N(I)}. \tag{13}$$

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

$$R(d, i) = \sum_a \int L(d, i : f, a) P_a(f|I) P_p(a)[df], \tag{14}$$

where, for example, we might set $L(d, i : f, a) = -\delta(f - d)\delta_{ia}$ (i.e. we are penalized by δ_{ia} for not finding the right model and by $-\delta(f - d)$ for not finding the right surface). Here $\delta(f - d)$ denotes the Dirac delta function and δ_{ia} is the Kronecker delta, where $\delta_{ia} = 1$ if $i = a$ and is 0 otherwise.

The Bayes decision corresponds to picking the model i and the scene d that minimizes the risk, see Figure 12.

Figure 12 about here

It is straightforward to adapt this model if a input sequence of images are available as will usually be the case. We simply replace I in the formulae above by the set $\{I_1, I_2, ..., I_M\}$ of images. For some scenes a single image may not yield enough information to decide between competing models and yet an image sequence may give the correct result (see [19] for preliminary results showing this). In some situations the system may initially make an incorrect decision which it later corrects as more information becomes available, see Section 5.3.

5.2 Determining the Fitness of Prior Models for Material Properties

We now give a specific example for determining the shading model for a surface [18]. In this example the two competing image formation models are Lambertian reflectance and specular reflectance.

We label the competing models by a and the surface shape by f. Let $P(I|f, a)$ be the probability of generating the image by model a when the surface shape is f. Let $P_p(a)$ be the prior probability that model a is correct.

For simplicity, we initially assume that the surface shape is known (this assumption will be relaxed later in this Section). The problem of deciding which model is most appropriate is now considered as one of deciding, in the presence of noise, whether we have one signal or another (the binary decision problem of statistical communication theory). This involves specifying a *decision rule* $\Delta(i|I)$ which tells us which model i to pick as a function of the input image I.

The optimal Bayesian decision rule, $\Delta(i|I)$ for this problem is that which minimizes the expected risk [37]:

$$R(\Delta) = \sum_a P_p(a) \int [dI] P(I|f,a) \sum_i L(a,i)\Delta(i|I) \tag{15}$$

This differs from our previous formulation because: (i) we are finding a decision rule $\Delta(i|I)$ for a class of images instead of making a single decision for a single image (these are equivalent – [21]) and (ii) we are not interested in determining the scene so we have fixed the f variable.

We label the possibilities $a = 1, 2$ for whether the surface is Lambertian or specular. $P(I|f,a)$ is the image formation model – hence $P(I|f,1) = (1/Z)e^{-\int[dx](I-\vec{n}\cdot\vec{s})^2}$ and $P(I|f,2) = (1/Z)e^{-\int[dx](I-(\vec{h}\cdot\vec{k})^m)^2}$, where $m, \vec{k}, \vec{h}, \vec{s}, \vec{n}$ take their standard meanings for the Phong shading model. The surface normal $\vec{n}$ can be calculated directly from the surface shape f. Let $P_p(1) = p$ and $P_p(2) = q$.

It is straightforward algebra to derive the optimal Bayes rule for this problem. It corresponds to deciding that the image is specular if $\Lambda(I) < K$, and Lambertian otherwise. $\Lambda(I)$ is the *likelihood ratio*, and is given by

$$\Lambda(I) = \left(\frac{p}{q}\right)\left(\frac{p(I|f,1)}{p(I|f,2)}\right) \tag{16}$$

and K is a decision threshold given by

$$K = \frac{L(2,1) - L(2,2)}{L(1,2) - L(1,1)}. \tag{17}$$

Suppose we set $L(1,1) = L(2,2) = 0$ (i.e. no cost for correct decision) and $L(1,2) = L(2,1)$ (i.e. both possible errors have equal cost), then $K = 1$. The decision rule can be rephrased as: decide specular when $\log \Lambda(I) < 0$ and Lambertian otherwise, where

$$\begin{aligned}\log\Lambda(I) = {} & [\int [dx](I - \hat{n}\cdot\hat{s})^2 - \log p] \\ & -[\int [dx](I - (\hat{h}\cdot\hat{k})^m)^2 - \log q].\end{aligned} \tag{18}$$

$\log\Lambda(I)$ is a very intuitive quantity because it depends on the difference in energies of the two possible reflectance models. Essentially, we choose the Lambertian model if its energy is lower than that of the specular model, with a correction factor to adjust for the priors p and q.

This discussion has assumed that the surface shape, represented by f, is already known. We now relax this assumption and show how to estimate f and a simultaneously. First we define prior distributions $P(f|a)$ for the surface shape as a function of the model a. The posterior distribution for the model *and* the surface shape is now:

$$P(f, a|I) = \frac{P(I|f,a)P(f|a)P_p(a)}{P(I)}. \tag{19}$$

Our risk function becomes:

$$R(d,i) = \sum_{a=1}^{2} \int L(d,i:f,a)P(f,a|I)[df]. \tag{20}$$

Let us set $L(d,i:f,a) = -\delta(f-d)L(a,i)$, where $L(a,i)$ is the loss function defined above (i.e. L(1,1) = L(2,2) = 0 and $L(0,1) = L(1,0) = 1$). Then the risk simplifies to

$$\begin{aligned} R(d,1) &= -P(d,1|I), \\ R(d,2) &= -P(d,2|I). \end{aligned} \tag{21}$$

To find the optimal decision we calculate $d_1^* = \arg\max_d P(d,1|I)$ and $d_2^* = \arg\max_d P(d,2|I)$. Then, if $L(d_1^*,1) < L(d_2^*,2)$ we choose model 1 and surface shape d_1^*, otherwise we pick model 2 and shape d_2^*. In other words, we find the best estimate for the surface shape for each of the models and compare the probabilities of these estimates to determine which model is correct.

5.3 Psychophysics of Competitive Priors

It seems that a number of psychophysical experiments, some of which are described in other Chapters, seem to require explanations in terms of competitive priors. In all cases the perception of the stimuli can be made to change greatly by small changes in the stimuli. Some of these experiments would also seem to require strong coupling.

Kersten, et al [33] describe a transparency experiment in which the scene can be interpreted as a pair of rectangles rotating rigidly around a common axis or as two independent rigid rectangles rotating around their own axis (Fig. 13). The competitive priors correspond to assuming that the rectangles are coupled together to form a rigid object or that the rectangles are uncoupled and move independently. By adjusting the transparency cues either perception can be achieved. Interestingly, the perception of the uncoupled motion is only temporary and seems to be replaced by the perception of the coupled motion. We conjecture that this is due to the build up of support for the coupled hypothesis over time, as described in Section 5.1. The uncoupled interpretation is initially supported because it agrees with the transparency cue. Over a long period of time, however, the uncoupled motion is judged less likely than coupled motion. This hypothesis does require a relative ordering of competing explanations (see also the Chapter by Richards, Jepson, Feldman), 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.

Figure 13 about here

Blake and Bülthoff's [5] work on specular stereo, see Figure 14, 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. If the specularity is at the center it is seen as a light bulb and the sphere appears transparent. If the specularity lies in the physical correct position within the sphere (halfway between the center and the surface) then the sphere is perceived as being a glossy, metallic object.[22] If the specularity lies in front of the sphere then it is seen as a cloud floating in front of a matte sphere. We can interpret this as saying that there are three competing assumptions for the material of the sphere: (i) transparent, (ii) glossy, (iii) matte. The choice of model depends on the data. In addition if the sphere is arranged so that its Lambertian part has no disparity then the stereo cue for the specularity resolves the concave/convex ambiguity from the shading cues, see [5] for details.

Figure 14 about here

Nakayama and Shimojo [41, 42] describe an impressive set of stereo experiments which seem to imply that the visual system attempts to interpret the world in terms of surfaces that can partially occluded each other (see also Chapter by Nakayama and Shimojo). The visual system often performs significant interpolation in regions that are partially hidden. For example, one can obtain a strong perception of a Japanese flag, see Figure 15, even when the stimulus contains very little information, provided that the missing parts of the flag are occluded by another surface. Nakayama and Shimojo themselves [41] argue that their experiments can be described by having a set of competing hypotheses $i = 1, ..., N$ about the possible scene and corresponding image formation models $P_i(I|S_i)$. 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, see Chapter by Freeman – if a regularity appears in an image then the regularity is due to a regularity in the scene rather than being an accidental result of the viewpoint. See also Barlow's notion of suspicious coincidences in his Chapter.

Figure 15 about here

6 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, competing priors, that best fit the data. We envision a far larger and richer set of competing priors than the natural constraints proposed in [36] or the regularizers occurring in regularization theory [47]. These priors arise from the categorical structure of the world, see discussion in Chapter 1 of this book.

How sophisticated must these contextural priors be? In this Chapter we have only considered priors for low level tasks such as surface estimation. But we see no reason why they should not reach up to object recognition and scene interpretation. At an intermediate stage we should mention the interesting results of Kersten et al [33] which showed that humans make use of shadow information for depth perception. In these experiments the perceived motion of a ball in a box was strongly affected by the motion of its shadow. But for this shadow information to be meaningful the visual system must have decided that the geometry of the scene was a box. In other words, that the shadow was projected from a ball onto the planar surface at the bottom of the box.

[22] It is interesting that, before doing the experiment, most people think that the specularity should lie on the convex surface and not behind. You can convince yourself otherwise by looking, for example, at the reflection of a candle appearing inside a wineglass at a candle light dinner.

It is clear that the most effective computer vision systems are those which strongly exploit contextural knowledge and are geared to achieving specific tasks. To what extent should the competing priors be geared towards specific tasks? Ideally one would like to have priors which accurately model all aspects of the visual scene, but this may be unrealistic. Instead it would be simpler to have priors which accurately model the aspects of the world that the visual system needs to know about. Though this will mean that the decision rules must be sophisticated enough to prevent the system from constantly hallucinating the things that it desires to see.[23] Building up priors in this task dependent way seems a sensible strategy for designing a visual system, but is there any evidence that biological systems are designed like this? It may be hard to test for humans, since our visual system appears very general purpose, but it is possible that experiments might be designed for animals with simpler visual systems. This emphasize on task dependence is at the heart of recent work on active vision [7]. By making very specific prior assumptions about certain structures in the scene, and ignoring everything else, it has proven possible to design automatic vehicles capable of driving at high speeds on the Autobahn [22]. In this case the outputs of the visual system are used directly to control the vehicle, thereby giving another link to decision theory.[24] Decision theory can also be used at a higher level for planning tasks [20]. We argue that it is also useful for vision itself because, by means of loss functions, it builds in the preferences of the system and hence can incorporate task dependent vision.

Clearly the range of visual tasks that we can achieve is determined by the information, $P(S|I)$, we have about the scene. A cleverly chosen loss function can, at best, allow us to make the most use of the information available. 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 by 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 in Mumford's Chapter. 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 which might be used to make the decisions. In this we are following Marr's levels of explanation [36] where a distinction is made between the high level information processing description of a visual system and the detailed algorithms for computing it. 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 [3] 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 [52], even though precise models for how these computational tasks are achieved is often 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

[23] It is tempting to consider the hallucinations induced by sensory deprivation as an example of the prior imposing nonexistent structure on the data.

[24] Control theory and decision theory are equivalent when applied to such problems as how much to turn the steering wheel of a car.

algorithm might require first extracting textural features which are then used to determined surface shape. Thus Bayesian theories certainly do not imply "direct perception" [27] 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, in this Chapter we have been using a broad brush and have 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 real images. Bayesian decision theory gives a framework but there are many details that need to be filled in. For example, the Bayesian approach emphasizes the importance of priors but does not give any prescription for finding them. Although workers in computational vision have developed a number of promising priors for modeling the world, it is an open research task to try to refine and extend these models in order to build systems of the type outlined here. Fortunately the Bayesian framework is able to incorporate learning.[25], see [32], and the success of (Bayesian) Hidden Markov Models for speech recognition [44] suggests that it may be practical to learn Bayesian theories[26]

7 Conclusion

In this Chapter we have argued for a framework for Vision based on Bayesian Decision theory. From this perspective, vision consists of specifying priors, likelihood functions and decision rules. Such theories will inevitably causes biases towards the prior assumptions of the theory, particularly for the impoverished stimuli used by psychophysicists.

This approach suggests that when coupling visual cues care must be taken with the dependence between the cues and, in particular, on the prior assumptions which they use. In many cases this will lead to strong coupling between visual cues rather than the weak coupling proposed by other theorists.

We also argue that the prior assumptions used by the visual system must be considerably more complex than the natural constraints and generic priors commonly used. Instead there seems to be evidence for a competing sets of prior assumptions or contexts. This also seems to be a sensible pragmatic way to design a visual system to perform visual tasks. It may be better to design visual systems in terms of modules that are geared towards specific visual tasks in restricted contexts rather than modules based on the traditional concepts of visual cues. This can be incorporated into the Bayesian framework used hyperpriors (or priors with hyperparameters) and decision rules to determine which prior is suitable.

Picking the correct decision rule is also important and is directly tied to the task that the visual system is trying to solve. Certain properties of the visual scene need only be known approximately and undesirable, non-generic, interpretations may result if the decision rule is badly chosen.

Acknowledgements

We would like to thank Jim Clark, Dan Kersten, David Knill, Whitman Richards, David Mumford, Peter Belhumeur, Ken Nakayama, Andrew Blake and Stuart Geman for helpful discussions. We would also like to thank the four reviewers of this Chapter for their extremely useful com-

[25] Neural network learning is also relevant here.

[26] It is particularly interesting to ask whether priors can be learnt for new task.

ments. One of us, ALY, would like to thank the hospitality of the Isaac Newton Institute for Mathematical Sciences and support from the Brown/Harvard/MIT Center for Intelligent Control Systems with U.S. Army Research Office grant number DAAL03-86-K-0171.

Appendix

Bayesian Theory subsumming Regularization Theory

The Bayesian approach subsumes work based on regularization theory and minimizing energy functions [47] [31]. In such theories a problem can be made well-posed by adding a regularizing term. Once again we need to estimate a scene S given an input I. The problem is "solved" by minimizing, with respect to S, an energy function

$$E(S;I) = E_{data}(S;I) + E_{regularizer}(S), \tag{22}$$

where $E_{data}(S;I)$ measures the consistency of a scene S with the data I and $E_{regularizer}(S)$ biases the solution to a particular set of scenes.

Minimizing (22) is equivalent to maximizing a probability function $P(S|I)$ defined by

$$P(S|I) = \frac{1}{Z} e^{-E(S;I)}, \tag{23}$$

where Z is a normalization constant.

Observe that by substituting (22) into (23) we can interpret the the data term and the regularizer as corresponding to the likelihood function and the prior of a Bayesian theory respectively. More precisely, $E_{data}(S:I) = -\log P(I|S)$ and $E_{regularizer}(S) = -\log P(S)$. Finding the MAP estimator of $P(S|I)$ corresponds to minimizing $E(S;I)$.

We can also reverse this argument to re-express Bayesian theories in terms of energy minimization. Take the logarithm of both sides of Bayes theorem

$$P(S|I) = \frac{P(I|S)P(S)}{P(I)}, \tag{24}$$

to obtain

$$-\log P(S|I) = -\log P(I|S) - \log P(S) + \log P(I). \tag{25}$$

By comparing to (22) we can interpret this as an energy function theory where $-\log P(I|S)$ is the data term and $-\log P(S)$ is the prior. The term $\log P(I)$ is independent of S and so can be ignored. Thus doing MAP on Bayes can be interpreted as minimizing an energy which is the sum of a data term and a regularizer.

Thus regularization theory, in its energy function formulation, is simply a special case of Bayes. But the Bayesian framework is far richer and gives greater insight by making clear the statistical assumptions underlying regularization theory. For example, many regularization theories in vision use quadratic energy functions. From the Bayesian perspective this is equivalent to assuming Gaussian distributions and is only justifiable if this assumption is correct. Similarly regularization theories usually combine sources of evidence by adding together energy terms.

This is equivalent to multiplying probability distributions together and is only appropriate if the sources are independent.

Weighted Averages from Weak Coupling

We now show that some forms of weak coupling give a weighted combination of cues to first order approximation provided that the MAP estimates S_1^* and S_2^* of the two cues are similar.

We start with the formula for weak coupling, $P(S|f,g) = P_1(S|f)P_2(S|g)$, and take the logarithm of both sides to obtain

$$\log P(S|f,g) = \log P_1(S|f) + \log P_2(S|g). \tag{26}$$

Performing Taylor series expansions of $\log P_1(S|f)$ and $\log P_2(S|g)$ about their MAP estimators S_1^* and S_2^* gives

$$\begin{aligned}\log P(S|f,g) = \log P_1(S_1^*|f) - (1/2)(S - S_1^*)^2 w_1 \\ + \log P_2(S_2^*|f) - (1/2)(S - S_2^*)^2 w_2 \\ + O\{(S - S_1^*)^3, (S - S_2^*)^3\},\end{aligned} \tag{27}$$

where $w_1 = -(d^2 \log P_1(S|f)/dS^2)(S_1^*)$ and $w_2 = -(d^2 \log P_2(S|f)/dS^2)(S_2^*)$. The first order terms in the Taylor expansion vanish because S_1^* and S_2^* are extrema. Moreover w_1 and w_2 are positive since the extrema are maxima. Extremizing $\log P(S|f,g)$, ignoring terms higher than second order, gives

$$S^* = \frac{w_1 S_1^* + w_2 S_2^*}{w_1 + w_2}. \tag{28}$$

If the distributions are Gaussians then the higher order terms in (27) vanish and (28) is exact. In this case the weights are proportional to the inverse of the variances of the distributions. Thus the sharper the distribution then the more it is weighted.

A consequence of (28) is that the combined estimate S^* is a convex combination of S_1^* and S_2^*. Thus if S^* represents a single number, such as depth, it must be bigger than $\min(S_1^*, S_2^*)$ and smaller than $\max(S_1^*, S_2^*)$.

We note that this analysis becomes invalid unless $S_1^* \approx S_2^*$. Also the weighting constants w_1 and w_2 correspond to the Fisher information and are a measure of the reliability of the different cues.

Other forms of weak coupling such as setting $P(S|f,g) \propto P_1(f|S)P_2(g|S)P(S)$ (with $P(S) = P_1(S) = P_2(S)$) might also lead to a weighted combination of cues.[27] We rewrite this as $P(S|f,g) \propto P_1(S|f)P_2(S|g)/P(S)$ and perform a Taylor series expansion of $P_1(S|f)$, $P_2(S|g)$, and $P(S)$. This yields, to first order,

$$S^* = \frac{w_1 S_1^* + w_2 S_2^* - w_3 S_3^*}{w_1 + w_2 - w_3}, \tag{29}$$

where S_3^* is the MAP of $P(S)$ and $w_3 = -(d^2 \log P(S)/dS^2)(S_3^*)$. This approximation, however, is less valid than that used to derive (28). It requires that not only must S_1^* and S_2^* be similar but

[27] We are grateful to A. Blake for this argument.

also that both of these are close to the estimate given from the prior S_3^*, which is independent of the input data! Moreover, we might expect that the distributions $P(S|f)$ and $P(S|g)$ convey more information than $P(S)$ and hence are sharper. This would imply that w_3 is much less than w_1 and w_2. This casts doubts on our ignoring the higher order terms in the Taylor series expansion since the third order terms in the expansions of $\log P_1(S|f)$ and $\log P_2(S|g)$ may be larger than the second order terms of $\log P(S)$.

References

[1] T. Bayes. "An essay towards solving a problem in the doctrine of chances." *Phil. Trans. Roy. Soc.* **53**, pp 370-418, 1783.

[2] J.O. Berger. **Statistical Decision Theory and Bayesian Analysis**. (2nd Edition), Springer-Verlag, New York, 1985.

[3] W. Bialek. "Physical Limits to Sensation and Perception." *Ann. Rev. Biophys. Biophys, Chem.*, **16**, pp 455-78, 1987.

[4] T.O. Binford. "Inferring surfaces from images." *Artificial Intelligence*, **17**, pp 205-244, 1981.

[5] A. Blake and H. Bülthoff. "Shape from specularities: computation and psychophysics." *Philosophical Transactions of the Royal Society of London B* , **331**, pp 237-252, 1991.

[6] A. Blake, H.H. Bülthoff and D. Sheinberg, "An ideal observer model for inference of shape from texture." *Vision Research*, **33**, pp 1723-1737, 1993.

[7] A. Blake and A.L. Yuille. (Editors). **Active Vision**. MIT Press, Cambridge, MA, 1992.

[8] D. H. Brainard and W. T. Freeman. "Bayesian Method for Recovering Surface and Illuminant Properties from Photosensor Responses", *Proceedings of SPIE, Human Vision, Visual Processing and Digital Display V.* Eds. B. Rogowitz and J. Allebach, San Jose, CA, 1994.

[9] Bruno, N. and Cutting, J.E. "Minimodularity and the perception of layout." *J. Exp. Psychology: General*, **117**, pp 161-170, 1988.

[10] D. Buckley, J.P. Firsby and J. Freeman. "Lightness Perception can be Affected by Surface Curvature From Stereopsis." Artificial Intelligence Vision Research Unit preprint. Dept. Psychology, University of Sheffield, 1993.

[11] Bülthoff, H.H. and Yuille, A.L. "Bayesian Models for Seeing Shapes and Depth." *Comments on Theoretical Biology*, **2**, pp 283-314, 1991.

[12] H. Bülthoff and H.A. Mallot. "Interaction of different modules in depth perception." *J. Opt. Soc. Am.*, **5**, pp 1749-1758, 1988.

[13] H. Bülthoff and H.A. Mallot. "Integration of stereo, shading and texture." In *AI and the Eye.* A. Blake and T. Troscianko eds., Wiley & Sons, Chicester, 1990.

[14] Bülthoff, H. and Kersten, D. "Interactions between transparency and depth." *Perception*, **18**, A22b, 1989.

[15] H. Bülthoff, J. Little and T. Poggio. "A parallel algorithm for real-time computation of optical flow." *Nature*, **337**, pp 549-553, 1989.

[16] H. Bülthoff, M. Fahle and M. Wegmann. "Disparity gradients and depth scaling." *Perception*, **20**, pp 145-153, 1991.

[17] B. Cernushi-Frias, D.B. Cooper, Y-P hung and P.N. Belhumeur. "Towards a model-based Bayesian theory for estimating and recognizing parameterized 3D objects using two or more images taken from different positions." *IEEE Trans. Pattern Anal. Machine Intell.*, **11**, pp 1028-1052, 1989.

[18] J.J. Clark and A.L. Yuille. **Data Fusion for Sensory Information Processing Systems**. Kluwer Academic Press. Boston/ Dordrecht/ London, 1990.

[19] J.J. Clark, M.J. Weisman and A.L. Yuille. "Using Viewpoint Consistency in Active Stereo Vision" *Proceedings SPIE.* Boston, November, 1992.

[20] T.L. Dean and M.P. Wellman. **Planning and Control**, Morgan Kaufmann, 1991.

[21] M.H. DeGroot. **Optimal Statistical Decisions**. McGraw-Hill, New York, 1970.

[22] E.D. Dickmanns, B. Mysliwetz and T. Christians. "An integrated spatio-temporal approach to automated visual guidance of autonomous vehicles." *IEEE Trans. on Systems, Man and Cybernetics*, **20**(6), pp 1273-1284, 1990.

[23] B.A. Dosher, G. Sperling and S. Wurst. "Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure." *Vision Research*, **26**, pp 973–990, 1986.

[24] J. Earman. **Bayes or Bust**. MIT Press, Cambridge, MA, 1992.

[25] W. Freeman. "Exploiting the generic view assumption to estimate scene parameters." In *Proc. 4th Intl. Conf. Computer Vision*, pp 347-356, Berlin, Germany, 1993.

[26] D. Geiger and A.L. Yuille. "Stereo and Eye Movement." *Biological Cybernetics*, **62**, pp 117-128, 1989.

[27] **The Ecological Approach to Visual Perception**. Houghton Mifflin, 1979.

[28] R.L. Gregory. **Eye and Brain**. 3d ed., McGraw-Hill, New York, 1978.

[29] N.M. Grzywacz, J.A. Smith, and A.L. Yuille. "A computational theory for the perception of inertial motion". *Proceedings IEEE Workshop on Visual Motion*, Irvine, 1989.

[30] B.K.P. Horn and M.J. Brooks. "The Variational Approach to Shape from Shading." *CVGIP*, (2), pp 174-208, 1986.

[31] B.K.P. Horn. **Robot Vision**. MIT Press, Cambridge, MA, 1986.

[32] D. Kersten, A.J. O'Toole, M.E. Sereno, D.C. Knill and J.A. Anderson. "Associative learning of scene parameters from images." *Optical Society of America*, **26** (23), pp 4999-5006, 1987.

[33] D. Kersten, H.H. Bülthoff, B. Schwartz, and K. Kurtz. "Interaction between transparency and structure from motion." *Neural Computation*, **4**, pp 573-589, 1991.

[34] D. Knill and D. Kersten. "Apparent surface curvature affects lightness perception." *Nature*, **351**, pp 228-230, 1991.

[35] L.T. Maloney and M.S. Landy. "A statistical framework for robust fusion of depth information."*Proceedings of the SPIE: Visual Communications and Image Processing* Part2, pp 1154–1163, 1989.

[36] D. Marr. **Vision**. W.H. Freeman and Company. San Francisco, 1982.

[37] D. Middleton. **An Introduction to Statistical Communication Theory**. Peninsula Publishing, Los Altos, CA, 1987.

[38] D. Mumford. "Pattern Theory: a unifying perspective." Dept. Mathematics Preprint. Harvard University, 1992.

[39] K. Nakayama and G.H. Silverman. "The aperture problem – I. Perception of nonrigidity and motion direction in translating sinusoidal lines." *Vision Research*, **28**, pp 739-746, 1988.

[40] K. Nakayama and G.H. Silverman. "The aperture problem – II. Spatial integration of velocity information along contours." *Vision Research* **28**, pp 747-753, 1988.

[41] K. Nakayama and S. Shimojo. "Experiencing and Perceiving Visual Surfaces." *Science*, **257**, pp 1357-1363, 1992.

[42] K. Nakayama and S. Shimojo. "Towards a Neural Understanding of Visual Surface Representation." In *Cold Spring Harbour Symposia on Quantitative Biology*, Volume LV, 1990.

[43] G. Parisi. **Statistical Field Theory**. Addison-Wesley, Reading, MA, 1988.

[44] D.B. Paul. "Speech Recognition Using Hidden Markov Models." *The Lincoln Laboratory Journal*, Vol. 3, No. 1, 1990.

[45] A. Pentland. "Local Shading Analysis." In **Shape from Shading**. Eds. B.K.P. Horn and M.J. Brooks, MIT Press, 1989.

[46] S.B. Pollard, J.E.W. Mayhew and J.P. Frisby. "A stereo correspondence algorithm using a disparity gradient limit." *Perception*, **14**, pp 449-470, 1985.

[47] T. Poggio, V. Torre and C. Koch. "Computational vision and regularization theory." *Nature*, **317**, pp 314-319, 1985.

[48] T. Poggio, E.B. Gamble and J.J. Little. "Parallel Integration of Vision Modules." *Science*, **242**, pp 436-440, 1988.

[49] V.S. Ramachandran and S. Anstis. "Displacement thresholds for coherent apparent motion random dot-patterns." *Vision Research*, **24**, pp 1719-1724, 1983.

[50] B.D. Ripley. “Classification and Clustering in Spatial and Image Data.” In **Analyzing and Modeling Data and Knowledge**. Ed. M. Schader, Springer-Verlag, 1992.

[51] J. Risannen. “A universal prior for integers and estimation by minimum description length.” *Annals of Statistics*, **11** (2), pp 416–431, 1983.

[52] B. Sakitt. “Counting every Quantum.” *J. Physiol.*, **284**, 261, 1972.

[53] G. Sperling and B.A. Dosher. “Strategy and optimization in human information processing.” In K. Boff, L. Kaufman and J. Thomas (Eds), *Handbook of Perception and Performance. Vol. 1.* NY: Wiley. Chapter 2, pp 1-65, 1986.

[54] S.N.J. Watamaniuk, N.M. Grzywacz, and A.L. Yuille. “Dependence of Speed and Direction Perception on Cinematogram Dot Density.” *Spatial Vision*, in press, 1993.

[55] D.W. Williams and R. Sekuler. “Coherent global motion percepts from local stochastic motion”, *Nature*, **324**, pp 253-255, 1986.

[56] R.J. Woodham. “Analysing Images of Curved Surfaces.” *A.I. Journal*, Vol. 17, No.s 1-3, pp 117-140, 1981.

[57] A.L. Yuille and N.M. Grzywacz. “A Computational Theory for the Perception of Coherent Visual Motion”. *Nature*, **333**, pp 71-74, 1988.

[58] A.L. Yuille, D. Geiger and H. Bülthoff. “Stereo integration, mean field theory and psychophysics.” *Network*, **2**, pp 423-442, 1991.

[59] A.L. Yuille and J.J. Clark. “Bayesian Models, Deformable Templates and Competitive Priors.” To appear in **Spatial Vision in Humans and Robots**. Eds. L. Harris and M. Jenkin, Cambridge University Press, 1993.

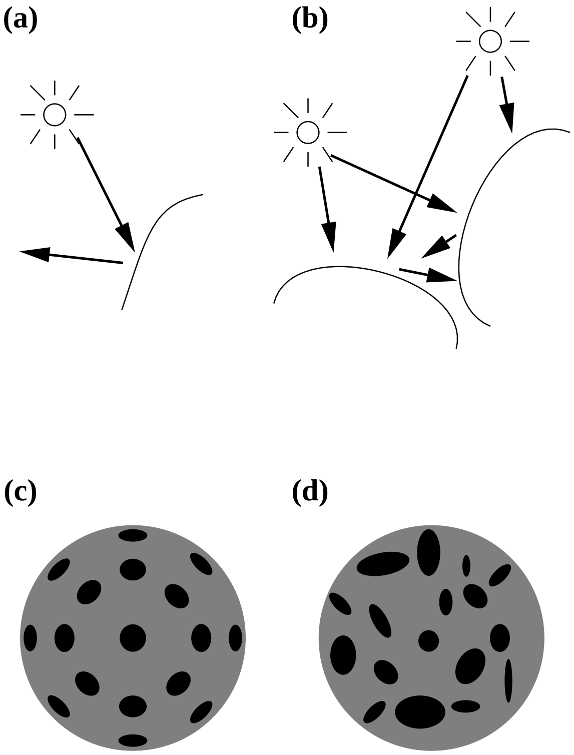

Figure 1: Cues are valid only in certain contexts. In (**a**) we sketch a Lambertian object illuminated by a single light source and no mutual illumination, so standard shape from shading algorithms will work. However, in (**b**) the mutual illumination will prevent shape from shading from working. Similarly, shape from texture is possible for (**c**) but not for (**d**) where the homogeneity assumption for the texture elements is violated. Thus both shading and texture depth cues are only valid in certain contexts.

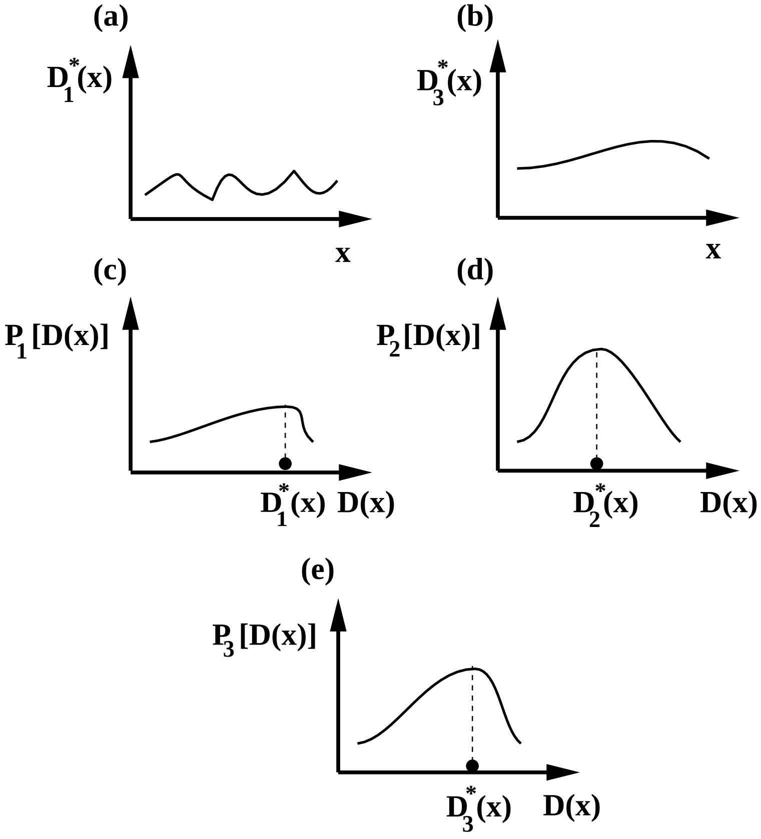

Figure 2: Prior assumption bias perception. (**a**) shows the true depth $D_1^*(x)$ and (**b**) shows the biased depth percept $D_3^*(x)$ after smoothing. In (**c**) we assume that the likelihood function $P_1[D(x)]$ is weakly peaked at the true depth $D_1^*(x)$. The prior in (**d**), however, is peaked at $D_2^*(x)$. The resulting posterior distribution $P_3[D(x)]$ is shown in (**e**) and yields a biased percept $D_3^*(x)$.

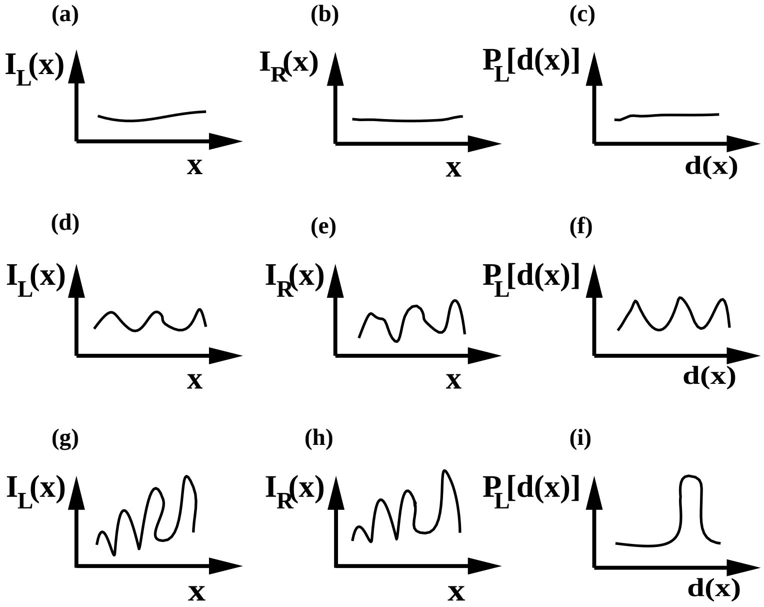

Figure 3: The ambiguity of the likelihood function for binocular stereo. For the intensity profiles of left (**a**) and right (**b**) eye there is considerable matching ambiguity and so the likelihood function $P_L[d(x)]$ is almost flat in (**c**). For the inputs in (**d**) and (**e**) there is less ambiguity, because the bumps in the two images must match, yet there are several possible correspondences and hence the likelihood function has several peaks in (**f**). However, the images in (**g**) and (**h**) are sufficiently structured so that only one match is likely and therefore the likelihood function has only a single peak as shown in (**i**).

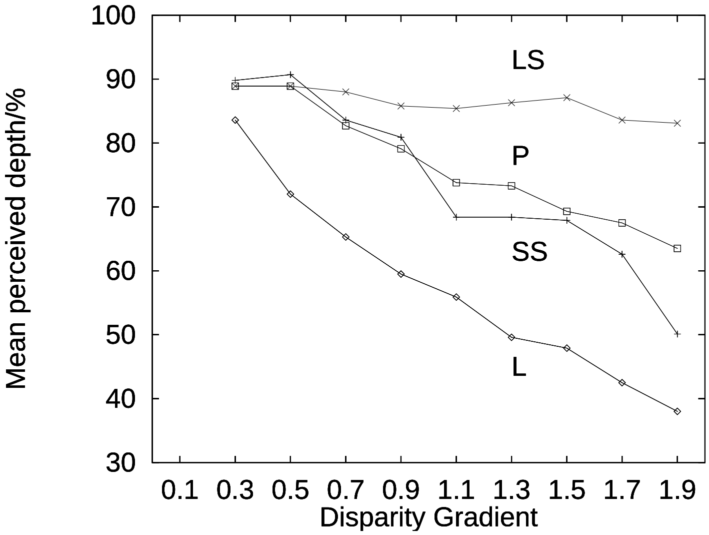

Figure 4: Perceptual bias and matching ambiguity. Perceived depth in percent of displayed depth as a function of depth gradient for points (P), lines (L), small symbols (SS) and large symbols (LS). Each data item represents the mean of nine different disparities (3 – 27 arc min) tested with 10 subjects. The standard errors of the means are in the order of the symbol size. Redrawn from [16].

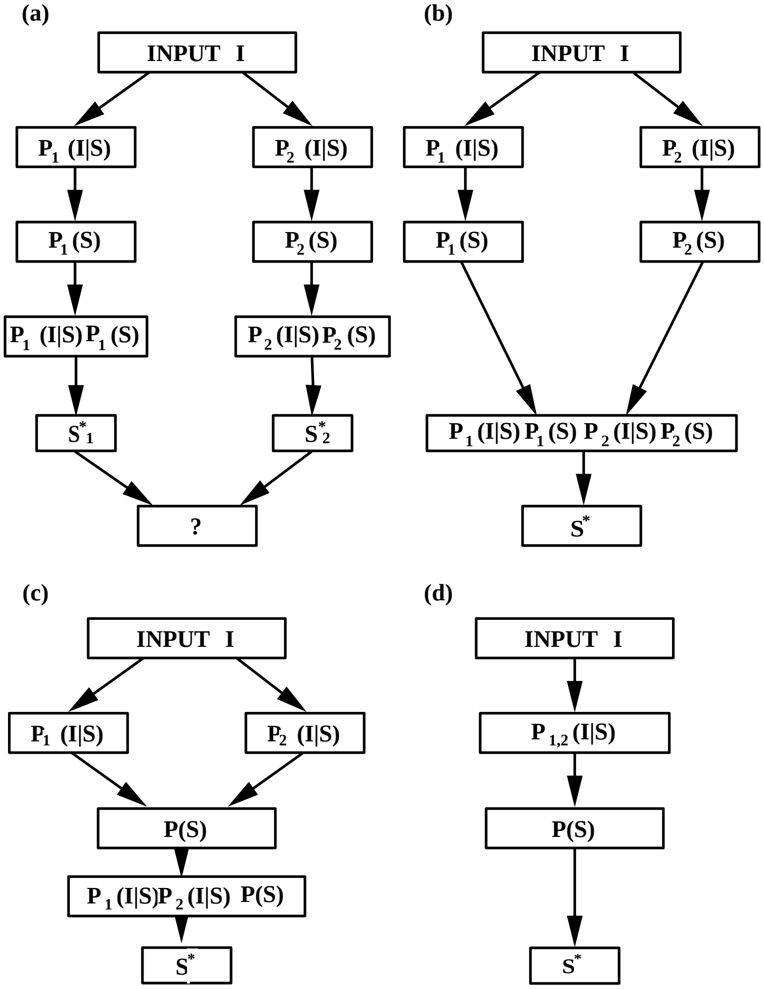

Figure 5: Different types of coupling between modules. (**a**) shows a form of weak coupling where the two modules act independently, with their own likelihood functions $P(I|S)$ and priors $P(S)$, producing MAP estimators, $S^* = \arg\max_S P(I|S)P(S)$, as outputs which are then combined in an unspecified manner. (**b**) shows weak coupling where the likelihood functions and priors of the two modules are multiplied together and then the MAP estimator is calculated. Such coupling would yield a weighted combination of cues in some circumstances, see Appendix 2. In (**c**) the likelihood functions of the modules are combined with a single prior for the combined modules and then the MAP estimator is found. This case is on the borderline between weak and strong coupling. It is weak if the prior $P(S)$ is the same as that used for the individual modules and it is strong otherwise. (**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.

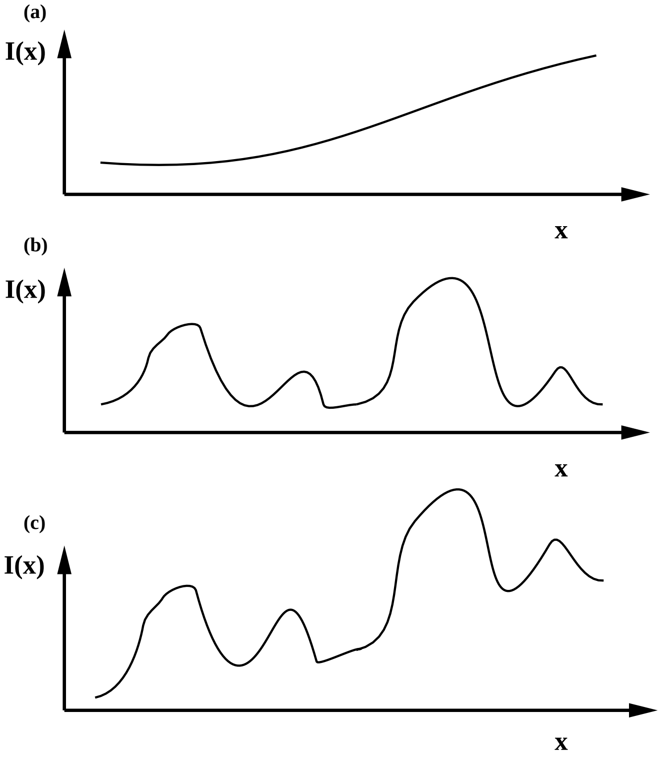

Figure 6: The difficulty of decoupling shading and texture cues. (**a**) shows a typical intensity profile for a Lambertian surface with constant albedo, the context in which shape from shading can be computed. (**b**) shows the intensity profile for a surface with strong albedo variation, the context for shape from texture. (**c**) shows the 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 the likelihood function for combined shading and texture cannot, in general, be factored into the likelihood functions for the two individual cues.

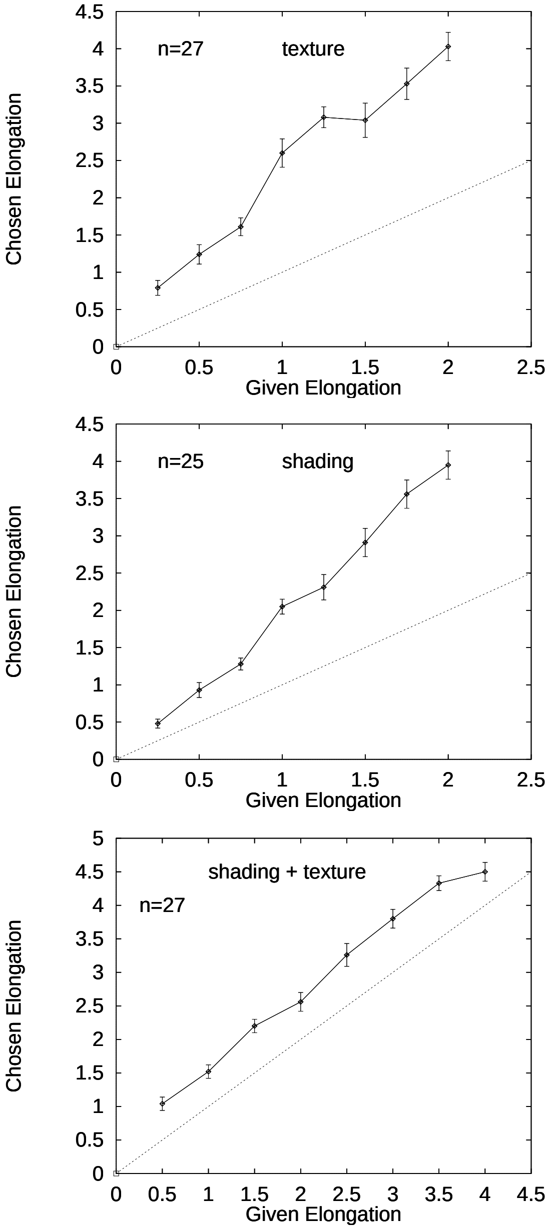

Figure 7: Psychophysical experiments on the integration of shading and texture. In an adjustment task subjects interactively adjusted the shading or texture of a simulated ellipsoid of rotation (seen by one eye) in order to match the form of a given ellipsoid seen with both eyes (in stereo). The ellipsoids were seen end-on so that the outline was the same for both surfaces. Shape from shading and shape from texture individually lead to a strong underestimation of shape, i.e., shading or texture of an ellipsoid with much larger elongation had to be simulated in order to match a given ellipsoid (slope >> 1). If shading and texture are presented simultaneously the shape is adjusted almost correctly (slope = 1). Redrawn from [13].

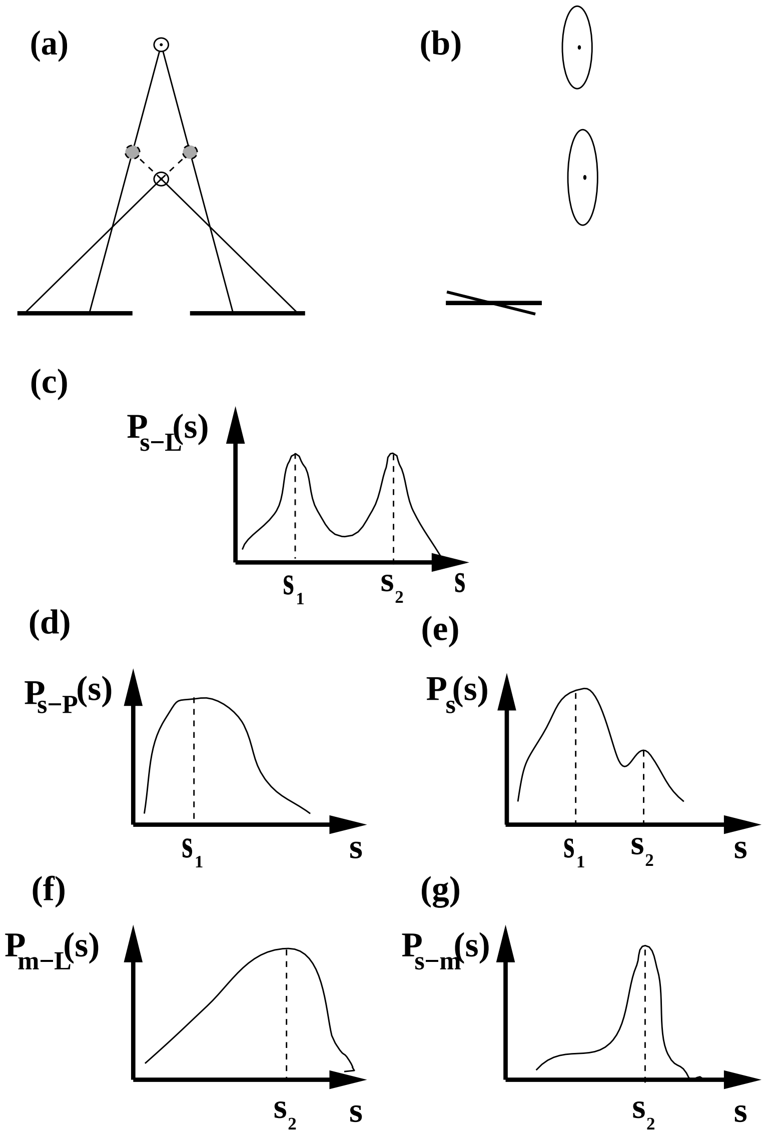

Figure 8: Monocular and stereo cues can combine to solve the double nail illusion. The s variable represents the positions of the two dots in space with s_1 and s_2 denoting the horizontal (frontoparallel) and vertical configurations respectively. The dots are in the vertical configuration s_2. Binocular stereo has a correspondence problem and (**a**) shows the two possible solutions s_1 and s_2, illustrated by the grey and white ellipses respectively. The monocular cues (**b**) have no correspondence problem but only yield approximate depth estimates, the sizes of the ellipses show the magnitude of the uncertainty. (**c**) shows that the stereo likelihood function $P_{s-L}(s)$ has two maxima corresponding to the two possible solutions. A typical prior $P_{s-P}(s)$ for stereo (**d**) will favour the frontoparallel interpretation s_1. So the stereo module with posterior distribution $P_s(s) \propto P_{s-L}(s)P_{s-P}(s)$ will be biased towards the incorrect solution s_1 shown in (**e**). The monocular likelihood function $P_{m-L}(s)$ is peaked at the correct interpretation s_2 but the distribution is so broad that there is considerable uncertainty in the estimated position, see (**f**). However, combining the likelihood functions for the stereo and monocular cues, $P_{s-m}(s)$, yields a sharp peak at the correct solution s_2, see (**g**).

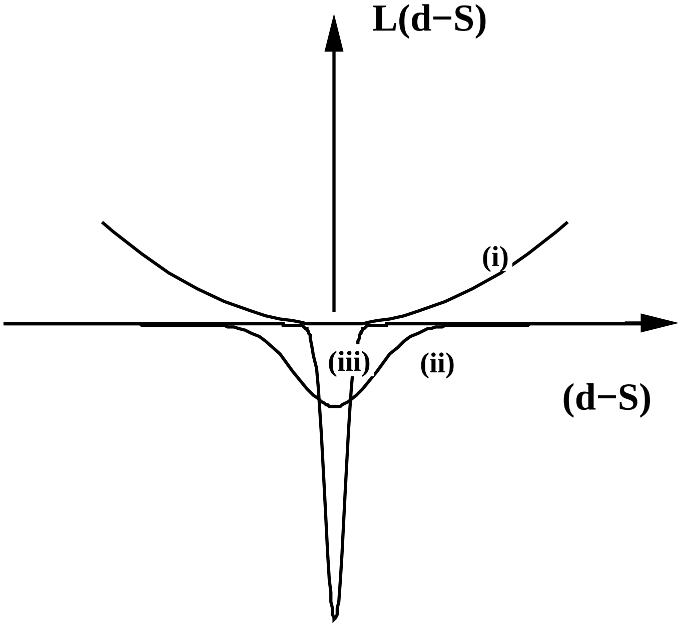

Figure 9: Several standard loss functions. They depend only on the difference between the decision d and the scene S so we write them as functions of $(d - S)$. The quadratic loss function, labeled (i), has $L(d - S) = (S - d)^2$, and its estimator is the mean of the distribution. Curves (ii) and (iii) are the negatives of a Gaussian and a delta function respectively. Observe that the delta function, which corresponds to MAP estimation, only rewards interpretations which are absolutely correct, with $d = S$, while the other two loss functions are more tolerant. Unless the probability distribution $P(S)$ is very sharply peaked it is unrealistic to attempt to estimate S to absolute precision, so MAP estimation is often inappropriate.

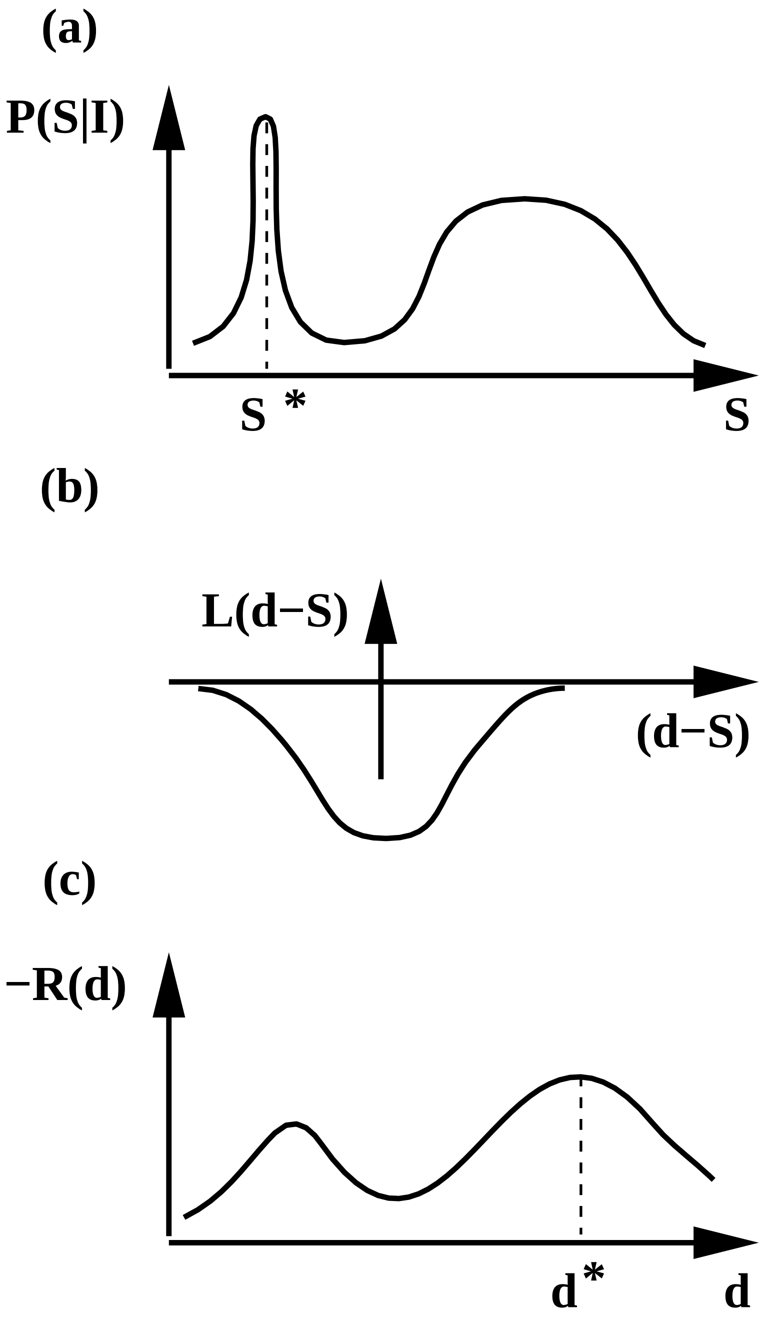

Figure 10: Loss functions can enforce generic viewpoint constraints. The posterior probability $P(S|I)$ in (**a**) has a high narrow peak, at S^*, and a lower broad peak. If we only want to estimate S to within a certain broad tolerance, as in Freeman's original formulation of generic viewpoints, then we should prefer the broad peak to the thin one. Using a negative Gaussian loss function $-G(d-S)$ in (**b**) will introduce the necessary tolerance because the risk, obtained by multiplying $P(S|I)$ by $-G(d-S)$ and integrating, is now minimized near the broad peak $-G(d-S)$. This is demonstrated by plotting the negative of the risk in (**c**). Note that, because the loss function is a function of $(d-S)$, the risk is obtained by convolving the posterior with the loss function.

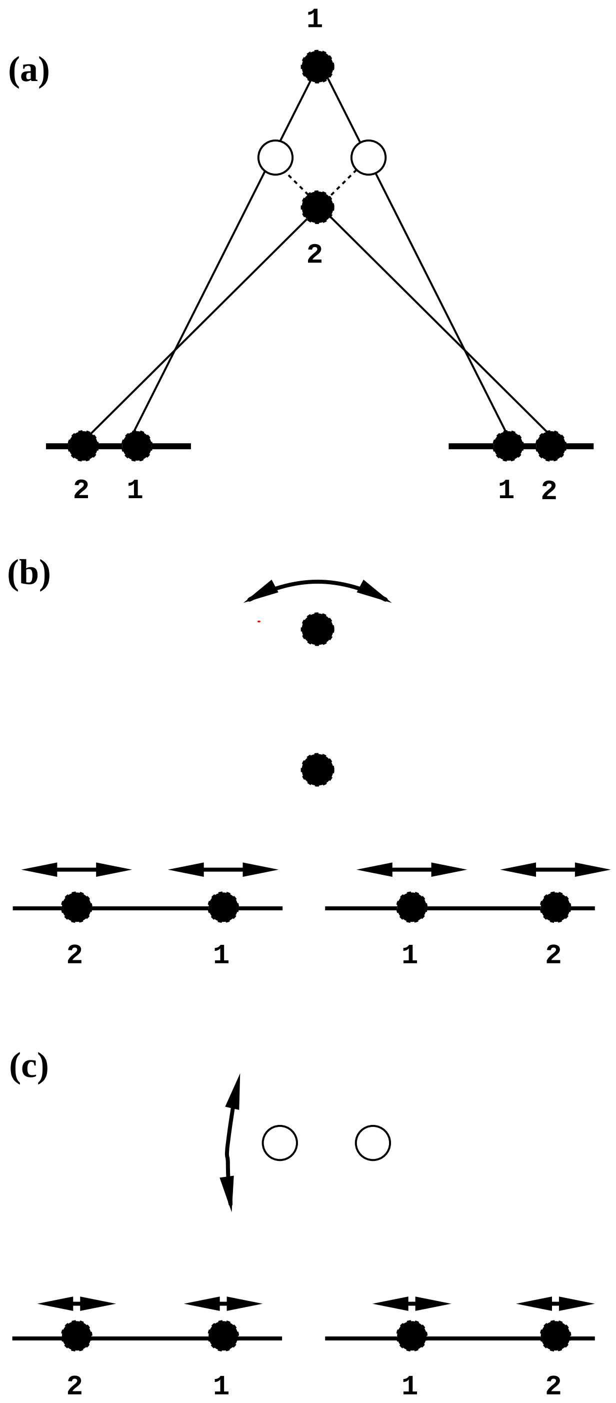

Figure 11: Using the generic viewpoint assumption to "solve" the double nail illusion shown in (**a**). The "black solution" in (**b**), is less stable than the white solution in (**c**). This is because small rotations of the black solution will induce larger changes in the positions of the image points than will corresponding rotations of the white solution. In Bayesian terms the posterior distribution is narrowly peaked about the black solution but broadly peaked about the white solution.

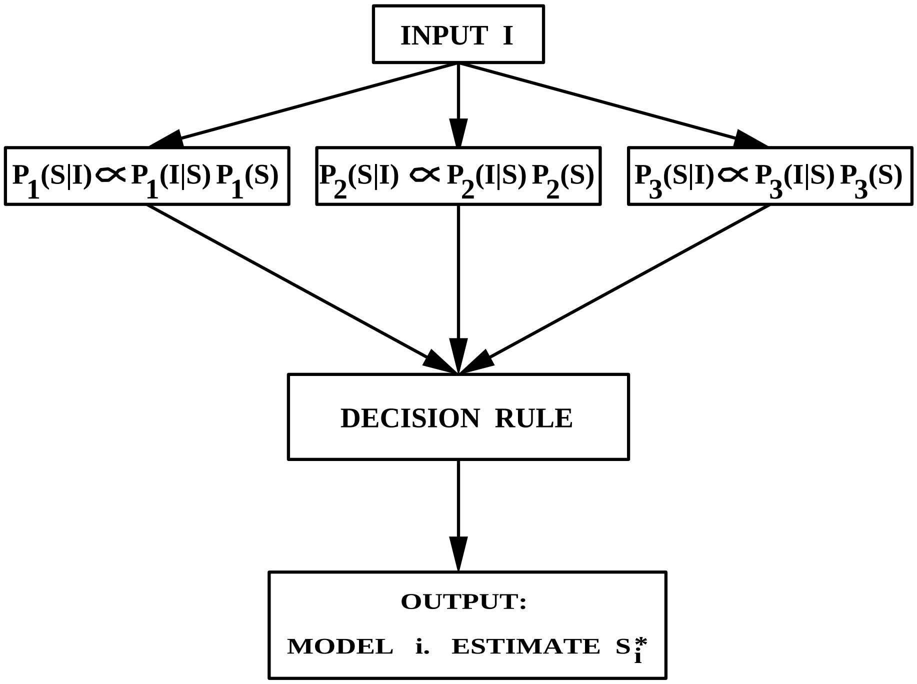
INPUT I
$P_1(S|I) \propto P_1(I|S)\, P_1(S)$
$P_2(S|I) \propto P_2(I|S)\, P_2(S)$
$P_3(S|I) \propto P_3(I|S)\, P_3(S)$
DECISION RULE
OUTPUT:
MODEL i. ESTIMATE S_i^*

Figure 12: Competitive priors. Three models with different priors compete to explain the input I. The winner is decided by a decision rule. The output is the choice of model i and the estimate given by the winner S_i^*.

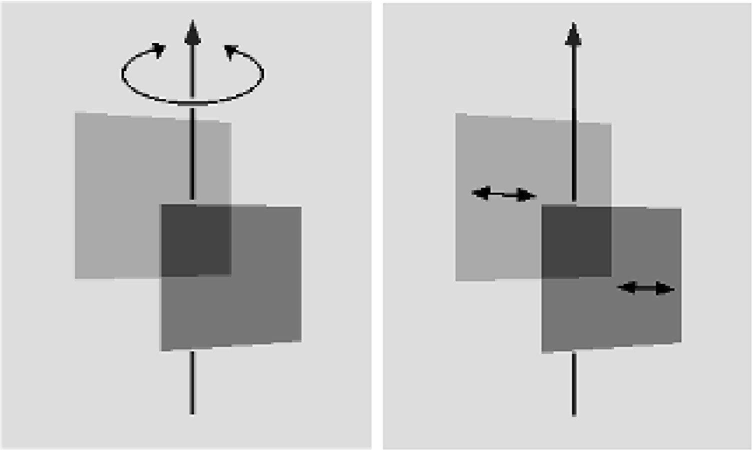

Figure 13: Different types of motion are perceived depending on transparency cues. In (**a**) the two planes are perceived to rotate rigidly together. However in (**b**) they are seen to slide across each other in a periodic motion.

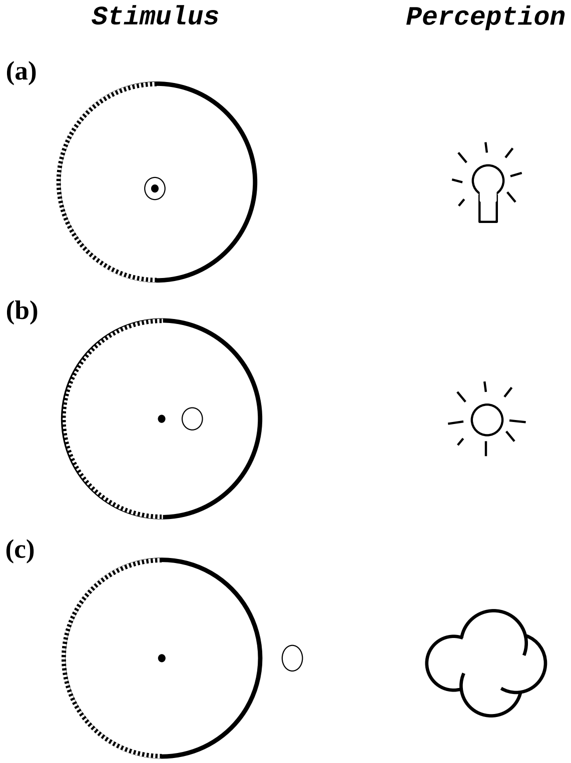

Figure 14: Specular stereo where a hemisphere is viewed binocularly. In (**a**) the specularity, the white ellipsoid, is adjusted to lie behind the center of the sphere. It is perceived as a light bulb lying behind a transparent sphere. In (**b**) the specularity lies in approximately the correct position and the hemisphere is perceived to be metallic with the specularity appearing as the image of the light source. If the specularity lies in front of the hemisphere (**c**), then it is perceived as a cloud floating in front of the hemisphere.

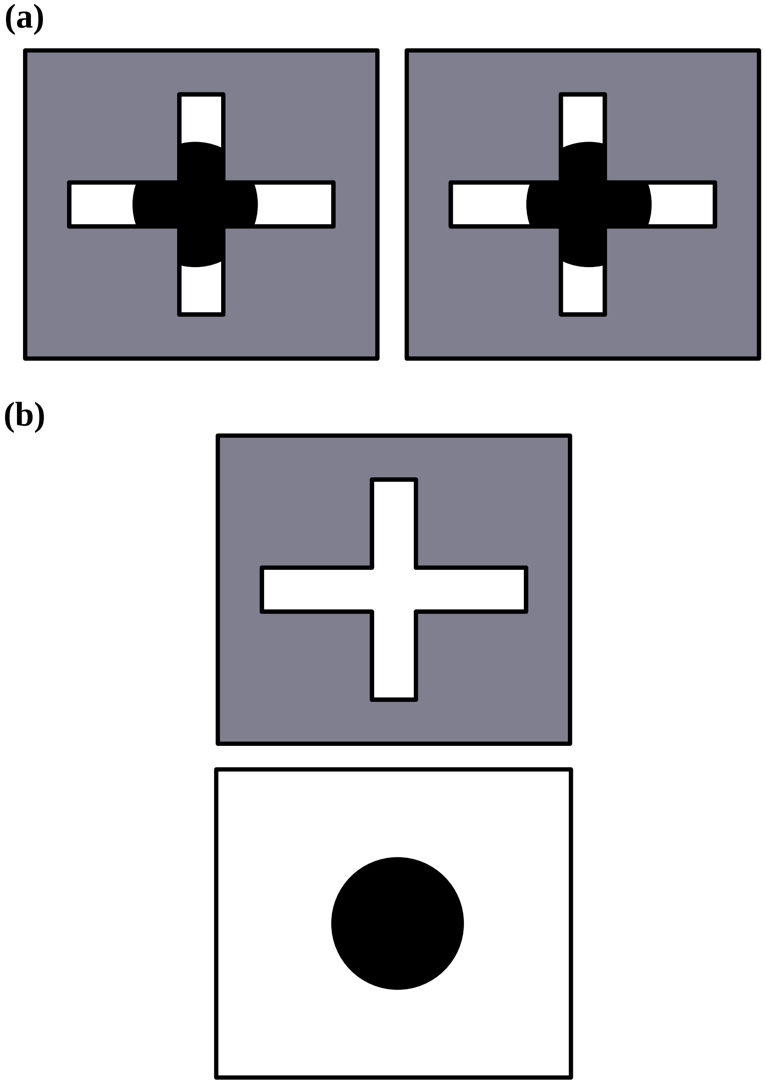

Figure 15: Binocular stereo cues for surfaces occluding each other. The stereo pair (**a**), is perceived as a planar surface, with a cross-shaped hole in its central region, floating above a surface with a circle at its center, see (**b**).