Bayesian Models for Seeing Shapes and Depth

We review computational models of shape and depth perception and relate them to visual psychophysics. The Bayesian approach to vision provides a fruitful theoretical framework both for modeling individual modules, such as stereo, shading, texture, and occlusion, and for integrating their informatiom. In this formalism we represent depth by one, or more, surfaces with prior probabilities for surface shape, corresponding to natural constraints, in order to avoid the ill-posedness of vision. On theoretical grounds, the less information available to the module (and the less accurate it is), then the more important the priors become. This suggests that visual illusions, and biased perceptions, will arise for scenes for which the priors are not appropriate. We describe psychophysical experiments that are consistent with these ideas. For integration of different modules we advocate strong coupling, so that the modules can interact during computation and the priors can be modified. This framework is rich enough to accommodate straightforwardly both consonant and contradictory cue integration and different psychophysical experiments can be understood within the Bayesian approach.

INTRODUCTION

When modeling the brain it is currently impossible to study the billions of neurons and describe their individual activities. It seems better to model at a more abstract level and see what computations need to be done guided by psychophysical experiments and experience in designing computer vision systems [1].

Abstract mathematical theories, independent of their possible implementations in neural hardware, specify precisely what the problems are, what information is available, and what assumptions need to be made. These models can give predictions that can

Comments Theoretical Biology
1991, Vol. 2, No. 4, pp. 283–314

stimulate further psychophysical and neurophysiological experiments.

Visual depth perception is an interesting area in which to study the brain. There have been many psychophysical experiments and much experience building robot vision systems. It is current practice to break down the problem into manageable subunits, or *modules* [1], which are believed to be semi-independent. Examples are binocular stereo, shape from shading, and shape from texture.

Because of the ill-posedness of vision, these modules need to add constraints, or make prior assumptions, about the scene. If these assumptions are incorrect for a specific scene, then visual illusions and/or biased perceptions will result. We will describe experiments [2, 3] showing biased perceptions for shape from shading and shape from stereo. These experiments support the plausible idea that the strength of the bias decreases with the amount of accurate information in the scene.

Modules are often treated as being independent both in human psychophysics and in computer vision systems. There is, however, some psychophysical evidence of interaction and this interaction is certainly computationally desirable.

Several psychophysicists have categorized these interactions into two broad classes: one in which the cues are *consonant* and the other in which they are *contradictory*. For example, consider viewing a golf ball with both eyes. There will be consistent, or consonant, depth information from stereo, shading, and texture cues. Viewing an image of the same golf ball in a photograph, however, puts the stereo cues (which give constant depth for the entire photograph) into *conflict* with shading and texture; hence the cues are now contradictory.

People have attempted to deal with the first case by taking weighted linear combinations with some success [4, 5]. Some experiments [2], however, do not seem consistent with such a model. A qualitative demonstration of this is given in Figure 1, in which the integration of shading and texture gives a much more vivid and accurate depth perception than the individual module alone.

The case of conflicting cues seems to require significant nonlinearity and is usually assumed to require a different, and independent, mechanism. For example, this case is explicitly excluded

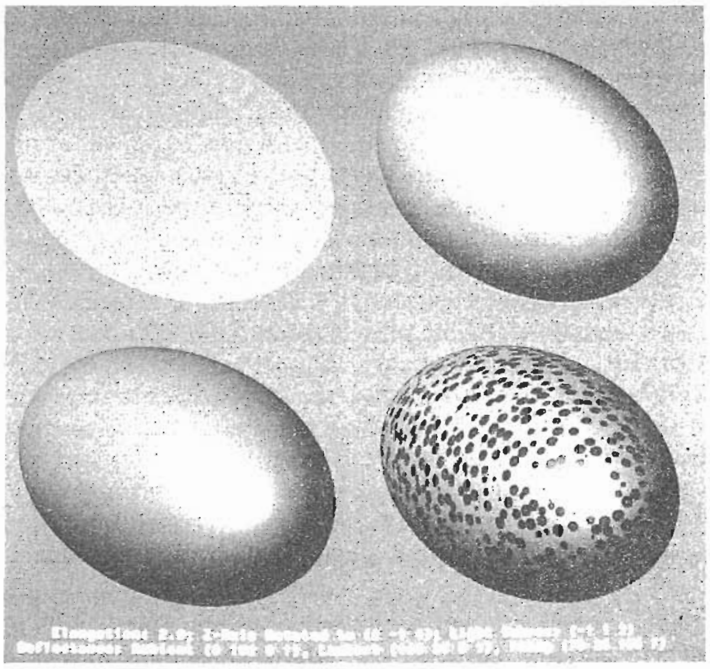

FIGURE 1 Integration of cues: The shape of the four objects looks quite different because the visual system derives different shape information from different shape cues. All four images were generated for the same 3D shape but with different simulated surface properties and under different lighting conditions.

in the statistical framework for fusion of depth information proposed by Maloney and Landy [6].

Workers in computer vision have tended to use an alternative viewpoint. A recent book on sensor fusion [7] proposed a distinction between *weak* methods in which modules compute depth independently and combine their results (with linear combination as a special case) and *strong* methods in which two modules interact during computation, usually in a very nonlinear way. They stress that, because of the ill-posedness of vision, individual modules must make assumptions about the world and may not be valid. Weak coupling may fail because individual modules may be using inconsistent assumptions. Strong coupling is usually preferable because it involves modifying the assumptions of the modules being fused to make them consistent. These theories also allow for adaptive weighting of different cues depending on the relative reliability of the cues and seem rich enough to encompass both categories defined by psychophysicists.

From this viewpoint some cues that appear conflicting might instead become consonant, provided strong coupling is used. Other cases of conflicting cues, transparency for example, can be dealt with in this formalism using binary decision units [8].

These methods are expressed in a Bayesian framework that can be used both for describing the individual modules and for their integration. Although there are many other methods for dealing with individual modules, the Bayesian approach subsumes a number of these methods by isolating the key assumptions used by these theories.

BAYESIAN MODELS FOR INDIVIDUAL MODULES

In this section we concentrate on individual, noninteracting, modules. We describe the need for natural constraints and psychophysical demonstrations for biased perception. We introduce the Bayesian approach and show that it gives a natural explanation of these biases.

The Need for Constraints: Why We Do Not See Depth Correctly

Vision is ill-posed. There is insufficient information in the retinal image to uniquely determine the visual scene. The brain must make certain assumptions about the real world to resolve this ambiguity [1], and visual illusions can result when these assumptions are invalid.

Psychophysicists design situations where the human visual system has a limited number of depth cues, usually from a single module. It is easy to produce situations of this type for which humans will get wrong shape or depth perception. An example would be photographs of the moon for which the perception of crater or hill depends on the assumption we make about the light source direction. It is plausible that assumptions such as "light comes from above" are innately built into our perceptual system [9].

Bülthoff and Mallot [2] used a shape probe to measure quantitatively the shape and depth perception for singular or multiple depth cues. For example, a subject viewed a shaded object (ellipsoid of rotation viewed end on with one eye) and manipulated the shading of the object in real time in order to match the 3D appearance of a reference shape (without shading) viewed with both eyes (stereoscopically). The results show a systematic tendency to underestimate shape from shading, compared to shape from stereo; in other words, the perception is biased toward the frontoparallel plane. The same holds true for other depth cues, e.g., shape from texture (see Figure 2).

This result can be interpreted as saying that the visual system has prior assumptions about the surfaces. The weaknesses of the shading information, discussed in the next section, mean that the perception is significantly biased by these prior assumptions.

An Example of Ill-Posedness: Shape from Shading

Shape from shading assumes a relationship, described by the *reflectance model* (see Reference [10]), between the geometry of the surface being viewed and the image intensity.

A standard model, Lambertian Reflectance, assumes that:

$$I(\vec{\mathbf{x}}) = \vec{\mathbf{s}} \cdot \vec{\mathbf{n}}(\vec{\mathbf{x}}), \tag{1}$$

where $I(\vec{\mathbf{x}})$ is the intensity at a point $\vec{\mathbf{x}}$ in the image corresponding to a point on the surface with surface normal $\vec{\mathbf{n}}(\vec{\mathbf{x}})$, and $\vec{\mathbf{s}}$ is the direction to the light source.

The viewer receives the input $I(\vec{\mathbf{x}})$ and needs to determine the surface normal $\vec{\mathbf{n}}(\vec{\mathbf{x}})$. There are several reasons why this problem is ill-posed: (1) the viewer will not know what reflectance function to assume unless the viewer already has information about the object (although Lambertian might be a good default assumption); (2) the viewer does not know the direction of the source $\vec{\mathbf{s}}$; there may often be several sources and multiple reflections from other

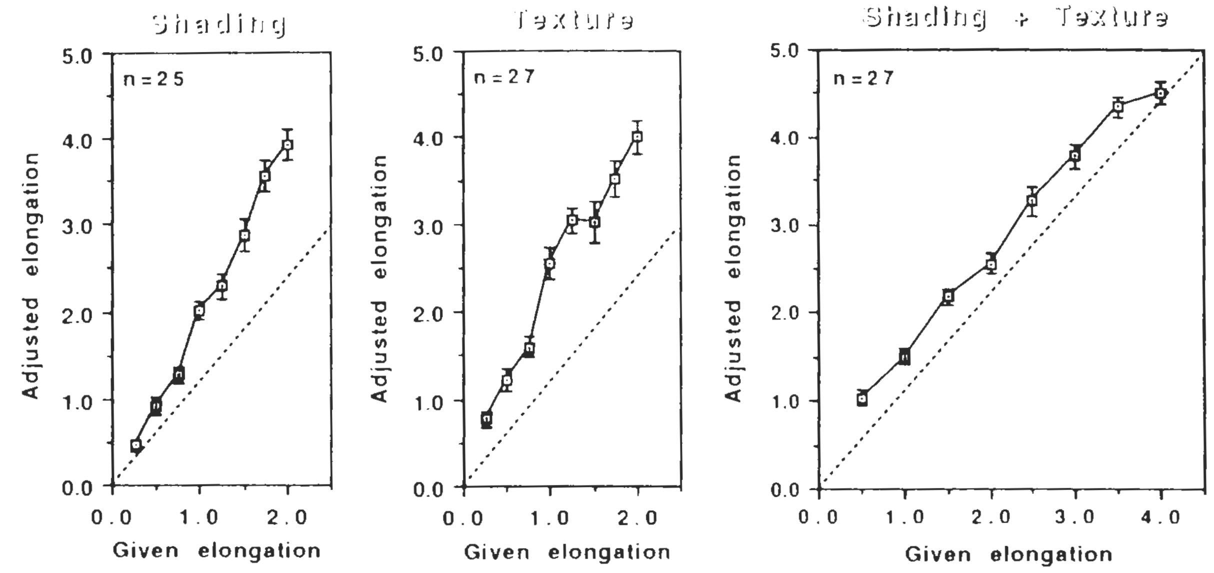

FIGURE 2 Integration of shading and texture: In an adjustment task subjects interactively adjusted the shading or texture of a simulated ellipsoid of rotation (seen by one eye) in order to match the form of a given ellipsoid seen with both eyes (in stereo). The ellipsoids were seen end on so that the outline was the same for both surfaces. Shape from shading and shape from texture 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). For further details about this experiment, see Bülthoff and Mallot [16].

objects; and (3) even assuming that (1) and (2) are solved, Eq. (2) is only one equation for the two components of $\vec{n}(\vec{x})$. Although because of technical properties of surface normals, there is sufficient information to solve Eq. (2), in some cases [10], it is not clear exactly how robust the solution is.

To make shape from shading well posed, it is usually assumed that the reflectance function and the light source direction are known and that the surface normal $\vec{n}(\vec{x})$ varies smoothly with $\vec{x}$; in other words, that the surface is smooth. Moreover, the surface normal is typically assumed to be known at the boundaries of the object. This last assumption is fairly weak because unless the surface normal is discontinuous at the boundary it can be deduced from the silhouette.

There is a common paradigm for enforcing smoothness for shape from shading, using energy functions [10]. The basic idea is to minimize a function $E[\vec{n}|I]$ with respect to $\vec{n}(\vec{x})$. Using the Gibbs distribution [11], we can define a corresponding probabilistic model (see Appendix I) so that minimizing the energy function with respect to $\vec{n}$ is equivalent to maximizing the probability distribution. In the next section we interprete this in a Bayesian framework.

A typical energy function for shape from shading is of the form:

$$E[\vec{n}|I] = \int \{I(\vec{x}) - \vec{s} \cdot \vec{n}(\vec{x})\}^2 \, dx + \lambda \int \{S\vec{n}\} \cdot \{S\vec{n}\} \, dx, \quad (2)$$

where S is a differential operator that causes the second term to penalize nonsmooth surfaces and λ controls the amount of smoothing. Minimizing $E[\vec{n}|I]$ will, therefore, give a compromise between fitting the data (the first term) and giving a smooth surface (second term). It is usually not desirable to fit the data perfectly, because of noise effects. The measure of smoothness is typically viewpoint dependent, although it is often chosen to be an approximation to a viewpoint-independent measure, such as the curvature of the surface. This viewpoint dependency results in biases toward the frontoparallel plane for many commonly used smoothness measures. In the probabilistic interpretation the smoothness term corresponds to the prior (see Appendix I).

A consequence of this paradigm, and in particular the smoothness constraint, is typically to bias the surface toward the frontoparallel plane. The amount of this bias is determined by the parameter λ and depends on the amount of information in the data term and its degree of accuracy. The less information available and the less accurate it is, the more the need for the prior assumptions and the greater the bias in the perception.

This bias agrees qualitatively with that found by Bülthoff and Mallot [2] (see Figure 2). These experiments showed that this bias is severe. Perhaps (because of the ill-posedness discussed above) the visual system puts relatively little weight on the data term and puts more faith in smoothness, an *a priori* assumption about the surface. It is interesting that if the reflectance function of the viewed object is modified to include a specular component, then the observer overestimates the depth (or curvature of the ellipsoid) [12]. A possible explanation for this overestimation is that the observer assumes a Lambertian reflectance model for the surface as a default. A specular highlight would, therefore, cause curvature overestimation. This suggests further experiments to determine the prior assumptions for surface reflectance functions.

A Way to Impose Constraints: The Bayesian Approach

How are natural constraints imposed in vision systems? For example, Marr [1] proposes that the visual system uses smoothness of surfaces as a natural constraint. But there are many possible definitions of smoothness and ways to incorporate it into the theory. The Bayesian formulation gives an elegant way to impose constraints in terms of prior probabilistic assumptions about the surfaces and, moreover, requires us to specify precisely what smoothness is. In general the choice of priors can be guided by psychophysical experiments and experience building computer vision systems.

The Bayesian approach is based on Bayes formula [13] and for surfaces can be written as:

$$P(\vec{\mathbf{n}}(\vec{\mathbf{x}}) \,|\, I(\vec{\mathbf{x}})) = \frac{P(I(\vec{\mathbf{x}}) \,|\, \vec{\mathbf{n}}(\vec{\mathbf{x}}))P(\vec{\mathbf{n}}(\vec{\mathbf{x}}))}{P(I(\vec{\mathbf{x}}))}. \tag{3}$$

In words: the probability of the surface $\vec{\mathbf{n}}(\vec{\mathbf{x}})$ given the data $I(\vec{\mathbf{x}})$ is the product of the probability of $I(\vec{\mathbf{x}})$ given $\vec{\mathbf{n}}(\vec{\mathbf{x}})$, $P(I(\vec{\mathbf{x}}) \,|\, \vec{\mathbf{n}}(\vec{\mathbf{x}}))$, times the *a priori* probability of $\vec{\mathbf{n}}(\vec{\mathbf{x}})$, $P(\vec{\mathbf{n}}(\vec{\mathbf{x}}))$, divided by a normalization constant.

For shape from shading $P(I(\vec{\mathbf{x}}) \,|\, \vec{\mathbf{n}}(\vec{\mathbf{x}}))$ is specified by the reflectance model, which can be thought of as synthesizing an image knowing the shape. $P(\vec{\mathbf{n}}(\vec{\mathbf{x}}))$ specifies the prior assumptions and natural constraints. The most probable interpretation is obtained by maximizing $P(\vec{\mathbf{n}}(\vec{\mathbf{x}})|I(\vec{\mathbf{x}}))$. We can get specific forms for these distributions using the energy function defined in the previous section and the Gibbs distribution (see Appendix I).

The prior probability distribution is needed to ensure that $P(\vec{\mathbf{n}}(\vec{\mathbf{x}})|I(\vec{\mathbf{x}}))$ has a unique and robust global minima, rather than a degenerate global minimum due to the ill-posedness of vision.

The Bayesian formulation, as is well known, can incorporate and extend previous theories, such as regularization [14] and the energy function minimization paradigm [10]. It is more general than regularization theory, because, for example, it can include decision units (so the answer does not have to depend smoothly on the input data—as regularization theory requires). Moreover, the Bayesian approach contains a far larger class of constraints than those that can be imposed by regularization techniques. For these reasons it seems rich enough to deal with the integration of both consonant and conflicting cues.

THE BAYESIAN APPROACH TO CUE INTEGRATION

This section describes the need for cue integration and how it can be accomplished within the Bayesian framework.

The Need for Integration: Why One Module Is Not Enough

By restricting the set of depth cues, psychophysicists can easily construct situations where depth is incorrectly perceived, as seen in the previous section. In natural scenes, however, more cues are available and humans tend to do better. Figure 2, from Bülthoff and Mallot [16], shows how the integration of shading and texture information gives a significantly more accurate depth perception.

It is generally believed that, in most situations, the integration of different depth cues will improve the accuracy of the perception (more information is available so there will be less need for priors). The accuracy may also be improved by recognition processes (recognizing an object as a baseball strongly suggests that it is spherical). For some natural scenes, however, integration and recognition may not be sufficient to obtain the correct interpretation. For example, they will not get us out of the Ames room [15]. It is not known how veridical depth perception has to be in order to solve the principal visual tasks of recognition and navigation. True accuracy may only be needed for a few tasks, such as threading a needle. Nevertheless, most visual tasks require some depth estimation and integration should improve task performance.

Integration raise a number of important issues: What information does one get from a depth module? How accurate is it? How robust is it to errors in measurement? How much does it depend on its assumptions, which will not always be correct?

Clearly a theoretical basis, which can deal with these questions, is needed for the integration of modules. In recent years the Bayesian approach has been proposed as a promising formulation for both individual modules and their integration. Clark and Yuille [2] in their book on sensor fusion describe such a formulation and review previous theories.

A Bayesian Formulation for Integration

The Bayesian approach gives an elegant way for combining information from different modules. Suppose we have two sources of depth information $f(\vec{x})$ and $g(\vec{x})$. Then we can write:

$$P\vec{n}(\vec{x})|f(\vec{x}), g(\vec{x})) = \frac{P(f(\vec{x}), g(\vec{x})|\vec{n}(\vec{x}))P(\vec{n}(\vec{x}))}{P(I(\vec{x}))}, \qquad (4)$$

with $P(I(\vec{x}))$ as a normalization constant. If these sources are independent then we can write:

$$P(f(\vec{x}), g(\vec{x})|\vec{n}(\vec{x}))P(\vec{n}(\vec{x})) = P(f(\vec{x})|\vec{n}(\vec{x}))P(g(\vec{x})|\vec{n}(\vec{x}))P(\vec{n}(\vec{x})). \qquad (5)$$

This corresponds to weak coupling (as defined in Clark and Yuille, [7]). This contrasts with strong coupling that occurs when the sources are dependent and Eq. (4) is not satisfied.

The more independent depth cues there are, the less the need for prior constraints. In some situations $P(f(\bar{x}))|\bar{n}(\bar{x}))P(g(\bar{x}))|\bar{n}(\bar{x}))$ may have a well-defined global minima and the $P(\bar{n}(\bar{x}))$ will be unnecessary.

The Bayesian formulation using the Gibbs distribution (see Appendix I) and weak coupling by the addition of quadratic energy functions leads naturally to a linear role for combining depth information (because minimizing a quadratic energy function leads to a linear equation). There are a number of psychophysical experiments on cue integration that are consistent with this linear rule [4, 5]. Maloney and Landy [6] for example, propose such a theory for integrating consonant cues (because they propose an adaptive way of estimating the relative importance of cues their approach is a special case of an adaptive Bayesian method.)

However, even for weak coupling, cues need not be combined linearly either because quadratic energy functions may not be used or because the prior for the combined system may differ from those of the individual modules. Because more information is available in the combined situation we might expect the prior to become less important, in certain situations, and the perception to be more accurate.

This is supported by the following experiment [12, 16], which shows coupling between shading and texture. The underestimation of depth by about 50% for individual cues (shading or texture) and the almost veridical perception for both cues combined (Figure 2) is inconsistent with the linear rule.

This experiment might still be explained by weak coupling, though with modified priors. The combined shading and texture system will need weaker priors than the individual systems, because more information is available. Hence there is both less bias toward the frontoparallel plane from the priors and more bias toward the correct perception from the shading and texture cues. The two modules would combine more than if a linear rule were used. Strong coupling, however, will allow even stronger interactions.

A theoretical example of strong coupling occurs in the work of Geiger and Yuille [17] combining stereo information with monocular depth cues, obtained from small head or eye movements. These monocular depth cues can be used to help solve the correspondence problem of stereo. This gives a highly nonlinear interaction between the monocular and the stereo cues. In contrast, a weak coupling method would solve the stereo correspondence problem without the use of the monocular cues. Note that the monocular cues need not localize the depths of the features precisely; they need only be accurate enough to disambiguate the stereo correspondence problem.

To illustrate the difference consider the following "Gedanken Experiment," which demonstrates how contradictory cues can become consonant if strong coupling is used. Suppose we have a random dot stereogram with monocular depth cues. The stereogram is set up so that it would appear to be flat, if the monocular cues are suppressed. It is straightforward to design such a stereogram, because there is a severe correspondence problem for the random dots that is resolved by using the smooth surface assumption for stereo [1]. We now arrange the monocular cues so that the surface is very jagged. If we try to couple the stereo weakly with the monocular cues, then the cues are contradictory. The problem lies in the smooth surface assumption used by the stereo algorithm. For strong coupling, as in Geiger and Yuille [17] the smoothness assumption is not needed, and the stereo and monocular information complement each other. The cues are now consonant.

This example illustrates several key features of the strong coupling approach: (1) the interaction between modules can become highly nonlinear, (2) cues that contain little, or inaccurate, information may nevertheless significantly strengthen the performance of another module provided the inaccuracy can be quantified, (3) the dependence on priors is reduced when more cues are available, and (4) strong coupling is particularly important for situations that require decisions, such as stereo correspondence or transparency.

Another example of strong coupling can be found in the work of Blake and Bülthoff [18] on specular stereo. They show that the human visual system can use the information about the 3D position of highlights in order to resolve the convex/concave ambiguity of shape from shading.

A BAYESIAN FORMULATION OF STEREO

In this section we introduce a theoretical formulation for stereo in terms of the Bayesian approach to vision. We briefly describe techniques from statistical physics which allow us to relate our theory to different existing theories and to develop fast heuristic algorithms.

The geometry for stereo is illustrated in Figure 3 assuming a pinhole camera model. If the correspondence problem is solved,

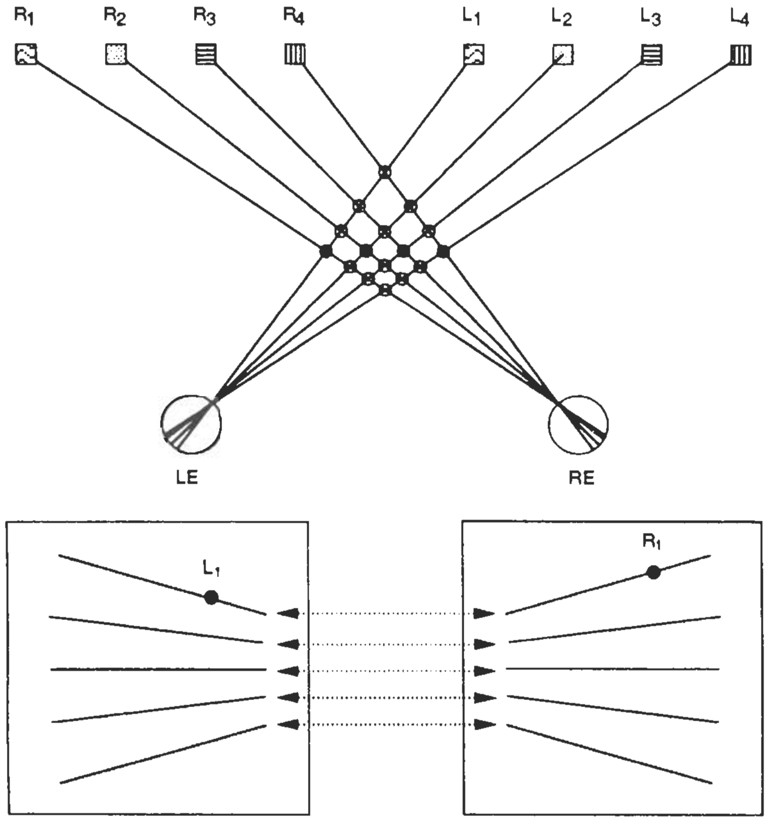

FIGURE 3 Stereo geometry: Matching corresponding points on the two retina is difficult because, in principle, any point in the left eye (LE) could match any other point in the right eye (RE). Only 4 of the 16 matching possibilities are correct (filled circles) and correspond to projections of 4 points in space. Without further constraints (priors on surfaces) the inverse optics problem of stereopsis cannot be resolved. One constraint that reduces the matching possibilities is the epipolar line constraint, illustrated in the lower part. A point on the left retina (L_1) can only be matched to a point on the right retina (R_1) on the corresponding epipolar line which is given by the geometry of the optics.

allowing us to match features between the two eyes, then depth can be determined by triangulation, assuming the eye directions are known. Although if there is any uncertainty in the positions of the image features, there will be uncertainty in the depth perception.

The epipolar line constraint ensures that a given feature in either eye can only match features lying on a specific line in the other eye; this line can be determined if the orientation of the cameras is known. It can be shown, to a first-order approximation, that the depth of a feature relative to the fixation depth is proportional to the relative distance between the images of the feature in the two eyes, the *disparity*. It is usually more convenient to describe stereo theories in terms of disparity rather than depth.

The fundamental issues of stereo are: (1) what primitives are matched between the two images; (2) what *a priori* assumptions are made about the scene to determine the matching and thereby compute the depth; and (3) how is the geometry and calibration of the stereo system determined. For this section we assume that (3) is solved, and so the corresponding epipolar lines between the two images are known. Thus we use the epipolar line constraint for matching; some support for this is discussed in the next section.

The Bayesian approach suggests using prior assumptions about the surface to solve the correspondence problem and to compute the disparity. Intuitively the correspondence problem is solved to give the disparity field that best satisfies the *a priori* constraints about the surface. This differs from some previous theories of stereo that first solved the correspondence problem and then constructed a surface by interpolation [19].

For this section we will restrict ourselves to priors given by quadratic smoothness measures, sometimes with line processes [20] used to break the smoothness constraint. This class of priors seems adequate for the psychophysics described later. The formalism, however, is not restricted to this class of priors [21].

Our framework combines cues from different matching primitives to obtain an overall perception of depth. These primitives can be weighted according to their robustness. For example, depth estimates obtained by matching intensity are sometimes unreliable, because small fluctuations in intensity (due to illumination or detector noise) might lead to large fluctuations in depth; hence they are less reliable than estimates from matching edges. The formalism can also be extended to incorporate information from other depth modules.

We will introduce the model by the simplest version, referred to as the Level 1 theory. The theory uses the epipolar line constraint to reduce the problem to one dimension. Ways to extend this to a two-dimensional theory are discussed in Yuille *et al.* [21].

The Energy Function for the Level 1 Theory

The basic idea is that there are a number of possible primitives that could be used for matching and that these all contribute to a disparity field $d(x)$. This disparity field exists even where there is no source of data. The primitives we will consider here are features, such as edges in image brightness. Edges typically correspond to object boundaries, and other significant events in the image. Other primitives, such as peaks in the image brightness or texture features, can also be added. We will describe the theory for the one-dimensional case.

We assume that the edges and other features have already been extracted from the image in a preprocessing stage. The matching elements in the left eye consist of the features x_{i_L}, for $i_L = 1, \ldots, N_l$. The right eye contains features x_{a_R}, for $a_R = 1, \ldots, N_r$. We define a set of binary matching elements $V_{i_L a_R}$, the matching field, such that $V_{i_L a_R} = 1$ if point i_L in the left eye corresponds to point a_R in the right eye, and $V_{i_L a_R} = 0$ otherwise. A *compatibility field* $A_{i_L a_R}$ is defined over the range $[0, 1]$. For example, it is small if i_L and a_R are compatible (i.e., features of the same type), and large if they are incompatible (an edge cannot match a peak).

We now define a cost function $E(d(x), V_{i_L a_R})$ of the disparity field and the matching elements. This can define a probabilistic model by the Gibbs distribution which we can interpret in terms of Bayes' theorem. The probabilistic formalism has two important advantages over the energy formalism: (1) we can use it to show relationships between different theories, and (2) it suggests good methods to compute desired quantities.

$$E(d(x), V_{i_L a_R}) = \sum_{i_L, a_R} A_{i_L a_R} V_{i_L a_R} (d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + \gamma \int_M (Sd)^2 dx. \quad (6)$$

The first term gives a contribution to the disparity obtained from matching i_L to a_R. The second term imposes a smoothness constraint on the disparity field imposed by a smoothness operator S.

Minimizing the energy function with respect to $d(\vec{x})$ and $V_{i_L a_R}$ will cause the matching that results in the smoothest disparity field. The $V_{i_L a_R}$ must satisfy global constraints to ensure that each feature usually has only one match. We discuss ways of doing this minimization subject to these global constraints later.

The coefficient γ determines the amount of *a priori* knowledge required. If all the features in the left eye have only one compatible feature in the right eye then little *a priori* knowledge is needed and γ may be small. If all the features are compatible, then there is matching ambiguity which the *a priori* knowledge is needed to resolve, requiring a larger value of γ and hence more smoothing. This is consistent with the psychophysics described in the next section.

Finally, and perhaps most importantly, we must choose a form for the smoothness operator S. As discussed previously Marr [1] proposes that the human visual system assumes that the world consists of smooth surfaces. This suggests that we should choose a smoothness operator that encourages the disparity to vary smoothly spatially. We will, therefore, use $S = \partial | \partial x$ as a default choice for our theory.

Compatibility is enforced multiplicatively in the first term of Eq. (1). An alternative would be to choose

$$\Sigma_{i_L, a_R} V_{i_L a_R} \{(d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2 + A_{i_L a_R}\}.$$

In most situations we will want to allow the smoothness constraint to break, for example, at the boundaries of objects. This is done in the next section.

In some cases stereo data cannot be fit to smooth surfaces even with discontinuities. Some examples are: (1) really jagged surfaces such as Bryce Canyon, (2) discontinuous surfaces such as the bristles on a hair brush or the railings in a picket fence, maybe with an object behind the fence, or (3) overlapping transparent surfaces such as a face behind a window.

The first example is a situation in which compatibility constraints are far more important than prior surface assumptions. Fortunately, the more jagged the surface the more structure there will be in the image. The last two cases may be thought of as examples of conflicting cues (as defined in the introduction), some of which correspond to one surface and some to the other. It is even more complicated, however, because the correspondence problem must be solved first before a conflict becomes apparent. These ideas are discussed further in Bülthoff and Yuille [22].

The Level 2 Theory: Adding Discontinuity Fields

The Level 1 theory is easy to analyze but makes the *a priori* assumption that the disparity field is smooth everywhere, which is false at object boundaries. The Level 2 theory introduces a discontinuity process to break the smoothness constraint [20, 23, 24]. Our energy function becomes:

$$E(d(x), V_{iLaR}, C) = \sum_{iL,aR} A_{iLaR} V_{iLaR} (d(x_{iL}) - (x_{aR} - x_{iL}))^2 + \gamma \int_{M - C} (Sd)^2 \, dx + M(C). \quad (7)$$

We use the formulation of Mumford and Shah [24] to enforce smoothness only within the domains $M - C$ but not across the boundaries C of the domains. The third term gives a cost $M(C)$ for the creation of the boundaries.

The Level 3 Theory: Adding Intensity Fields

The Level 3 theory incorporates information from the intensity fields $L(x)$ and $R(x)$ in the left and right eyes. It adds a term that attempts to match points in the two eyes with similar intensity values.

$$E(d(x), V_{i_L a_R}, C) = \sum_{i_L, a_R} A_{i_L a_R} V_{i_L a_R}(d(x_{i_L}) - (x_{a_R} - x_{i_L}))^2$$
$$+ \mu \int \{L(x) - R(x + d(x))\}^2 \, dx$$
$$+ \gamma \int_{M - C} (Sd)^2 \, dx + M(C). \quad (8)$$

If certain terms are set to zero in Eq. (8), it reduces to previous theories of stereo. If the second and third terms are kept, without allowing discontinuities, it is similar to work by Gennert [25] and Barnard [26]. If we add the fourth term and allow discontinuities, we get connections to some work described in Yuille [27] (done in collaboration with T. Poggio). Keeping the first, third, and fourth terms gives a theory somewhat similar to the disparity gradient limit theories [28, 29].

Stereo with Monocular Cues

Suppose we have a set of monocular measurements $(x_i^l, d_i^l, \sigma_i^l)$ and $(x_a^r, d_a^r, \sigma_a^r)$ and a depth from stereo function $d_s(x_s^l, x_a^r)$ with standard deviation $\sigma_s(i, a)$ (which is a measure of the uncertainty of the value of the depth obtained with the stereo module). Here x_i^l is a point in the left image with a depth estimate d_i^l, given by eye movements or focus changes or some other monocular cue, and σ_i^l is the standard deviation of this estimate. Similarly, x_a^r is a point in the right image (a possible match to x_i^l) that also has a depth estimate, d_a^r, associated with it obtained from some monocular (single view) method, having a standard deviation given by σ_a^r.

If we know that points i correspond to points a_i, the best mean squared estimate of the depth $d(x_i)$, assuming gaussian errors, is obtained by minimizing:

$$E(f) = \sum_i \frac{1}{(\sigma_i^l)^2} (f(x_i^l) - d_i^l)^2 + \sum_i \frac{1}{(\sigma_{a_i}^r)^2} (f(x_i^l) - d_{a_i}^r)^2$$
$$+ \sum_i \frac{1}{(\sigma_{s(i,a_i)})^2} (f(x_i^l) - d_s(i, a_i))^2. \quad (9)$$

Because the correspondence is usually unknown, we can define binary matching elements V_{ai} as before and an energy function:

$$E(V_{ai}, f) = \sum_{i} \frac{1}{(\sigma_i^l)^2} (f(x_i^l) - d_i^l)^2 + \sum_{a,i} \frac{1}{(\sigma_a^r)^2} V_{ai}(f(x_i^l) - d_a^r)^2 + \sum_{a,i} \frac{1}{(\sigma_s)^2} V_{ai}(d(x_i^l) - f_s(i, a))^2. \quad (10)$$

We set $V_{ai} = 1$ if point a in one eye is matched to point i in the other eye; otherwise $V_{ai} = 0$. If the correct or optimal matching is known or given *a priori* this cost function reduces to the previous cost function. This energy function should be minimized over the set of all V_{ai} satisfying the global constraints. Note that Eq. (9) is asymmetric between the left and right eyes; it assumes we are matching from right to left.

Once again we can also include *a priori* terms in the energy function, typically smoothness terms allowing discontinuities. These terms are not needed to give the energy function a unique minimum, but they are required to give dense depth values. Because they are not needed to solve the correspondence problem they will give little perceptual bias.

Statistical Mechanics, Mean Field Theory, and Marginal Distributions

An additional advantage of using the probabilistic formalism is that we can use techniques from statistical physics and probability theory to: (1) design algorithms to calculate the estimators, and (2) relate existing theories.

There are a variety of stochastic and deterministic algorithms that can be used to calculate the statistical estimators. Perhaps the most famous is simulated annealing [20, 30], but to guarantee convergence it requires considerable computer time.

More recently a number of other techniques from statistical physics have been adapted [31–35], which lead to fast heuristic algorithms. They can be thought of as deterministic annealing and lead to good empirical results.

These techniques allow two alternative methods for imposing the global constraints on the matching fields V. One can impose the constraints *softly* by including additional cost terms in the energy function such as $\lambda\{\Sigma_{iL}(\Sigma_{aR} V_{iLaR} - 1)^2 + \Sigma_{aR}(\Sigma_{iL} V_{iLaR} - 1)^2\}$. An alternative is to sum only over configurations that satisfy the global constraints when computing the marginal probability distributions. This is referred to as *hard constraints* [33, 34, 35] and leads to far more efficient deterministic algorithms than soft constraints.

These techniques can also be used to show relations between different theories. For example, by computing the marginal probability distribution $P_M(V) = \Sigma_d P(V, d)$ of the theory described earlier, we can show a relation between a version of our theory, using soft constraints, and the cooperative stereo algorithms of Dev [36] and Marr and Poggio [37]. Applying a similar approach to Eq. (10) gives a relation to the theory of House [38] for the fusion of stereo and accommodation depth cues.

STEREO PSYCHOPHYSICS AND THE BAYESIAN FRAMEWORK

We now describe two different experiments on depth from stereo that seem consistent with the Bayesian approach.

Perceived Depth Scales with Disparity Gradient

In the experiments of Bülthoff and Fahle [39] subjects were asked to estimate the disparity of stereo stimuli relative to a set of reference lines. The stimuli were either lines at various angles or pairs of dots or other features. The perceived depth was plotted as a function of the true disparity and the true disparity gradient. These were calculated, assuming features at $x_1, x_2 (x_1 > x_2)$ and $y_1, y_2 (y_1 > y_2)$ in the left and right eyes, using the formulae [40] for: (1) disparities $d_1 = (x_1 - y_1)$, $d_2 = (x_2 - y_2)$, (2) the binocular disparity $d_b = d_1 + d_2$, and (3) the disparity gradient $d_{\text{grad}} = d_b/\{(x_1 + y_1) - (x_2 + y_2)\}$.

The experiments showed that the perceived disparity decreased as a function of the disparity gradient. This effect was: (1) strongest for horizontal lines, (2) strong for pairs of dots or similar features,

(3) weak for dissimilar features, and (4) nonhorizontal lines (Figure 4).

Our explanation assumes these effects are due to the matching strategy and is based on the Level 2 theory, with energy function given by Eq. (7). The idea is that the smoothness term (the third term) is required to give unique matching but that its importance, measured by γ, increases as the features become more similar. If the features are sufficiently different (perhaps preattentively discriminable) then there is no matching ambiguity, so the correct disparities are obtained. If the features are similar, then smoothness (or some other *a priori* assumption) must be used to obtain a unique match, leading to biases toward the frontoparallel plane. The greater the similarity between features, the more the need for smoothness and hence the stronger the bias toward the frontoparallel plane. Smoothness is only imposed between the two points.

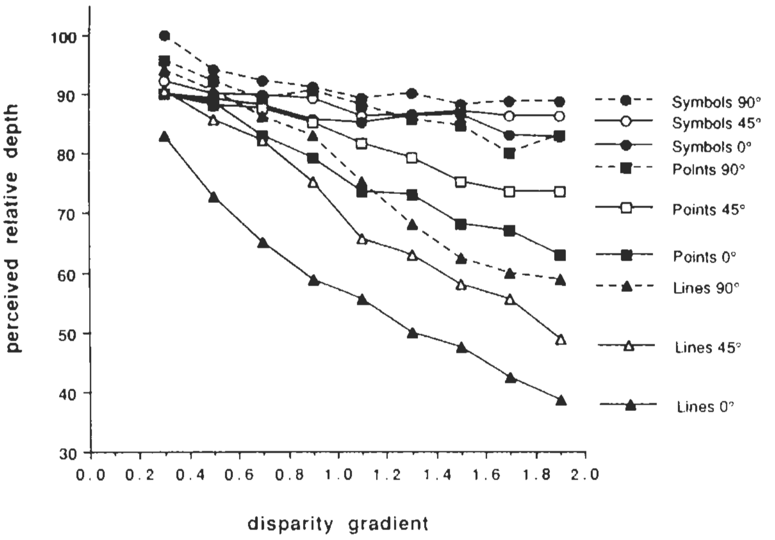

FIGURE 4 Disparity gradient and matching ambiguity: Perceived depth in percent of displayed depth as a function of depth gradient for points, lines, and large symbols in horizontal (0 deg), oblique (45 deg), and vertical orientation (90 deg). Each data item represents the mean of nine different disparities (3 – 27 arc min) tested with 10 subjects. The standard errors of the means are in the order of the symbol size. Redrawn from Bülthoff and Fahle [39].

Thus the two points are considered the boundaries of an object and only the object itself is smoothed.

An analysis [21] of the Level 2 theory shows that it predicts the falloff of perceived disparity with disparity gradient, provided the smoothness operator is chosen to be the first derivative of the disparity. The change of rate of falloff for different types of features is due to varying γ as described above.

The results are not consistent with several possible choices of the smoothness operator, such as the second derivative of the disparity $\partial^2 d/\partial x^2$. It is straightforward to calculate that this choice does not bias toward the frontoparallel plane. It is likely that the smoothness operator S must contain a $\partial/\partial x$ term to ensure the observed frontoparallel bias.

Edge Versus Intensity-Based Stereo

The experiments of Bülthoff and Mallot [2, 41] compared the relative effectiveness of image intensity and edges as matching primitives. The stimuli were chosen to give a three-dimensional perception of an ellipsoid. The observer used a stereo depth probe to make a pointwise estimate of the perceived shape.

The experiments showed that depth could be derived from images with disparate shading even in the absence of disparate edges. The perceived depth, however, was weaker for shading disparities (70% of the true depth). Figure 5 shows the perception of depth as a function of the stimuli.

Putting in edges or features helped improve the accuracy of the depth perception. But in some cases these additional features appeared to decouple from the intensity and were perceived to lie above the depth surface generated from the intensity disparities.

These results are again in general agreement with our model using the Level 3 theory. The edges give good estimates of disparity and so little *a priori* smoothness is required and an accurate perception results. The disparity estimates from the intensity, however, are far less reliable (small fluctuations of intensity might yield large fluctuations in the disparity). Therefore, more *a priori* smoothness is required to obtain a stable result. This gives rise to a weaker perception of depth.

The use of the peak as a matching feature is vital (at least for

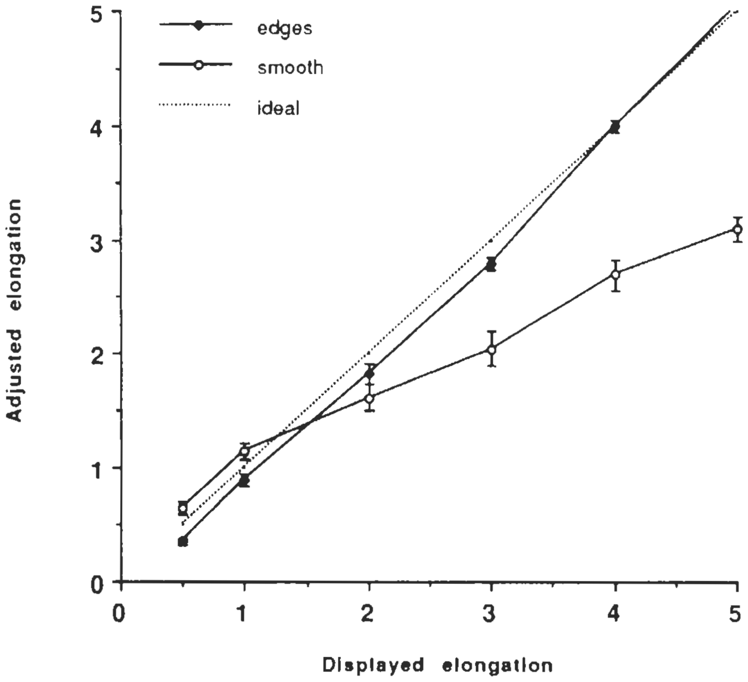

FIGURE 5 Edge versus intensity-based stereo: Ellipsoidal surfaces with or without edge information (facet or smooth shading) were stereoscopically displayed on a CRT monitor. Perceived elongation of the ellipsoids was measured with a local stereo depth probe as a function of displayed elongation. Less reliable information (smooth surfaces without edges) puts more weight on surface priors and leads to a frontoparallel bias of surface perception. Redrawn from Bülthoff and Mallot [2].

the edgeless case) because it ensures that the image intensity is accurately matched (some stereo theories based purely on intensity give an incorrect match for these stimuli (M. Gennert, personal communication). For these images, however, the peak is difficult to localize and depth estimates based on it are not very reliable. Thus the peak is not able to pull the rest of the surface to the true depth.

Bülthoff and Mallot [2] found that pulling up did occur for the edgeless case if a dot were added at the peaks of the images. This is consistent with our theory, because, unlike the peaks, the dots are easily localized and matching them would give a good depth estimate. Our present theory, however, is not consistent with a perception that sometimes occurred for this stimulus. In some cases the dots were perceived as lying above the surface rather than being part of it. This may be explained by the extension of our theory to transparent surfaces [8].

Ongoing Issues in Transparency and Strong Coupling

Strong coupling will be particularly important for situations that require decisions, such as stereo correspondence and transparency. These ideas are developed further in Bülthoff and Yuille [22]. It can be shown that transparency can be modeled by having two or more surfaces, and we describe psychophysical experiments [48] on transparency, which seem consistent with strong coupling. Preliminary experiments are reported suggesting that strong coupling might also be able to speed up stereo correspondence. This is related to the importance of compatibility for stereo matching. Other interesting experiments on transparency that seem consistent with strong coupling are described in Nakayama and Shimoyo [42].

DISCUSSION: WHY BAYESIAN INTEGRATION?

The Bayesian approach offers a framework in which many problems in low level and high level vision can be addressed. To make it into a theory, specific priors, sensor models, and methods for computing the statistical estimates must be specified. Two classic examples of successfuly implemented theories derived in this way include the work of Geman and Geman [20] on image segmentation (low level) and the work of Grenander [43] and his collaborators on object recognition. Much work on Markov Random Field models [44, 45] falls within this framework.

Many people have found this framework stimulating and a good basis for constructing theories. However, as in statistics, others have doubted its validity and/or usefulness. In this section we will discuss some arguments for and against this approach.

Computational vision [1] suggests that when studying a specific vision problem, one should abstract out the assumptions necessary to solve the problem, independent of implementation issues (and algorithms). Bayesian models extend this idea by providing a framework within which these assumptions need to be specified more precisely, as prior probabilities. For example, Marr proposed the smoothness constraint to solve stereo correspondence, but there are many different ways of imposing this. A precise formulation of smoothness as a prior probability on the surface should make it clear exactly what type of smoothness was being imposed and hence to determine for which situations the constraint would be appropriate. In fact the specific implementation of "smoothness" in Marr and Poggio's cooperative stereo algorithm [37] biases the system more to frontoparallel surfaces (surfaces of constant depth) than to smooth curved surfaces.

Thus as required by computational theories, Bayesian models force one to be specific about models and prevent assumptions from being smuggled in. This specification tells us under which situations the specific model should work, without bothering with the implementation issues. It is often harder to determine this from theories specified purely as algorithms or differential equations.

An advantage for those interested in constructing a theory of low level vision is that the Bayesian approach can be applied to most problems in early vision. There is hence no need for independent hacks for separate problems. Although the probability models would have to be specified for each problem, these models may well be applied to different problems, and the Bayesian approach may allow us to distill out the assumptions used by vision algorithms developed by trial and error, or evolution. For example, the Geman and Geman [20] idea of using line processes to cut smoothness constraints for image segmentation can be related to [32] the graduated nonconvexity algorithm of Blake and Zisserman [45] for visual reconstruction. Moreover a number of non-Bayesian theories for vision reconstruction and stereo can be approximated by Bayesian theories [21, 32, 46] allowing their relative probability assumptions and their effectivenesses to be compared. The Bayesian approach can also be related to ideal observer theory [47].

As discussed earlier the Bayesian approach offers a nice method for fusing different sources of information. Higher level information, if available, about the objects being viewed can also be incorporated [43]. Moreover, learning these models can be attractively framed as estimating the probabilities both for low level vision [48] and for high level vision [49].

Finally, the Bayesian is philosophically attractive, because it is nice to think of vision as providing the most probable estimate of a scene given the data and prior assumptions.

On the other hand, the Bayesian approach can be criticized on a number of grounds. A criticism of the use of these models in statistics depends on how the priors should be specified unless there is an objective way of determining them. This is a particularly serious problem in vision, because it is by no means clear what the priors should be, e.g. which precise smoothness assumption should be used for stereo (see above). One can further question whether the human visual system uses surfaces as a basic representation and whether it is appropriate to put priors on them. To some extent these are experimental issues. There is, for example, some evidence [50] that surfaces are, indeed, represented. Moreover, although its may be possible to determine the general form of natural constraints, for example, surface smoothness or object rigidity, by "Gedanken Experimente," the precise form of the constraints may require learning, trial and error experimentation, or quantitative psychophysics. If the domain of application of the vision system can be modeled precisely, then the probabilities can be determined from this model [43]. Thus, the increased sophistication of computer graphics models of real world scenes may also help.

An alternative approach [51] suggests that there is no real theory of vision, but only a bag of tricks perhaps developed over the course of evolution [52]. However, a number of the tricks suggested by Ramachandran and Anstis [51] can be derived as special cases of the constraints used in computational vision theories [53, 54]. For a bag of tricks to be a useful theory, it must specify which tricks are used and how the vision system decides to use them. This, however, is precisely what computational theories try to achieve. It does raise the interesting point that maybe the visual system has a set of possible prior constraints and adaptively uses different constraints in different circumstances [7]. Theories of this type are difficult to implement as Markov models, but can still fall within the Bayesian framework.

The Bayesian approach uncritically used may lead to rather mindless theories of sensor fusion by merely writing an enormous energy function, without specifying how the statistical estimates can be calculated or how the relative weighting of terms can be decided (in the same way, as Feynman joked, that one can produce a unified theory of physics in one equation by summing the squares of all the individual laws [53]). It is important to analyze the situation carefully and estimate the dependence and robustness of different cues. If energy functions are used, then the relative weighting of terms should depend on their reliability [7, 56].

One can also question the precise form of the probabilities used in these theories, particularly if the correct model is only approximately known. Robust statistics [57] proposes a number of techniques to make reliable estimates of quantities when the underlying model is only partially known and where there may be outliers in the data that do not fit the model. The usefulness of these types of techniques has been suggested for computational vision [58, 59] and the use of some of these methods might significantly strengthen the Bayesian approach. It is interesting that some of these techniques can be straightforwardly adapted into the Markov Random Field formalism [8]. Other relevant statistical ideas include minimal length encoding [60] but, although controversial, this can also be interpreted as Bayesian if the lengths of the encodings are considered as prior assumptions.

CONCLUSION

The theoretical ideas outlined in this paper emphasize that when fusing information from different vision modules one must pay careful attention to the assumptions used by the modules, the reliability, and robustness of their information and the degree of independence of the information they supply. Strong coupling between modules is usually desirable. This leads to a Bayesian framework, which seems rich enough to deal with both consonant and contradictory depth cues.

The psychophysical experiments reported here seem consistent with this framework. In particular, they suggest that the weaker and less accurate the information, the more the *a priori* assumptions are used. The experiments on stereo integration and on transparency seem to require explanations based on strong coupling.

Following Penrose, in his recent best seller [61], we can classify scientific theories into three categories: superb, useful, and tentative.

We believe that the Bayesian approach falls within the second category. It is general enough to be somewhat flexible, but it also suggests a way of thinking about vision problems that captures their crucial features. The arguments for and against it are discussed earlier. It seems to offer a promising framework for computer vision and is good for stimulating psychophysical experiments. Whether it turns out to be useful for modeling biological vision systems is an experimental question.

We do not, however, agree completely with Popper's falsification principle for Science. The visual system is very complex, and it is inevitable that our initial attempts to model it are going to be simplistic and unable to deal with all its aspects. Thus, if the theory succeeds in capturing the important aspects it should not be discarded for failing to deal with others; at least not unless a better theory is found.

Acknowledgments

This work was supported by Grant N00014-88-K-0164 from the Office of Naval Research, Cognitive and Neural Sciences Division, and by the Brown Harvard MIT Center for Intelligent Control Systems with U.S. Army Research Office Grant DAAL03-86-K-0171. We would like to thank Jim Clark, Davi Geiger, Dave Sheinberg, and Satoru Suzuki for helpful discussions.

HEINRICH H. BÜLTHOFF
Department of Cognitive Science and Linguistic Sciences,
Brown University, Providence, Rhode Island 02912

and

ALAN L. YUILLE
Division of Applied Science,
Harvard University, Cambridge, Massachusetts 02138

(Received November 15, 1990; in final form February 6, 1991)

APPENDIX I. STATISTICAL MODELS AND THE BAYESIAN FORMULATION

We can define a statistical theory from any energy function model (see, for example, [20]). Given the energy $E(\vec{\mathbf{n}}:I)$ specified in Eq. (2), then (using the Gibbs distribution [11]) we can define an associated probability distribution by:

$$P(\vec{\mathbf{n}}|I) = \frac{e^{-\beta E(\vec{\mathbf{n}}:I)}}{Z} \qquad (11)$$

where β is a parameter (which can be interpreted, by analogy to physics, as the inverse of the temperature) and Z is a normalization constant (or partition function in physics terminology).

This implies that every state of the system has a finite probability of occurring. The more likely ones are those with low energy. This statistical approach is attractive because the β parameter gives us a measure of the uncertainty of the model temperature parameter $T = 1/\beta$. At zero temperature ($\beta \rightarrow \infty$) there is no uncertainty. In this case the only state of the system that has nonzero probability, hence probability 1, is the state that globally minimizes $E(\vec{\mathbf{n}}:I)$.

We can now interpret minimizing the energy $E(\vec{\mathbf{n}}|I)$ as maximizing the probability. Thereby giving the most probable (assuming our model) interpretation $\vec{\mathbf{n}}$ of the data I, this is referred to as the MAP (maximum *a posteriori*) estimator.

There are alternative estimators that are often preferable to the MAP. One possibility is the mean field solution:

$$\bar{\vec{\mathbf{n}}} = \sum_{\vec{\mathbf{n}}} \vec{\mathbf{n}} P(\vec{\mathbf{n}}:I), \qquad (12)$$

which is more general and reduces to the MAP as $T \rightarrow 0$. It corresponds to defining the solution to be the mean fields; the averages of the $\vec{\mathbf{n}}$ field over the probability distribution. This enables us to obtain different solutions depending on the uncertainty.

In this paper we concentrate on the mean quantities of the field. A justification to use the mean field as a measure of the fields resides in the fact that it represents the minimum variance Bayes estimator [62].

Observe that we can write $E(\vec{\mathbf{n}}:I) = E_{\text{data}}(\vec{\mathbf{n}}:I) + E_{\text{smooth}}(\vec{\mathbf{n}})$ where E_{data} and E_{smooth} correspond to the first and second terms on the right hand side of Eq. (2). This addition of terms in the energy function space corresponds to multiplication of probabilities. Thus we can write:

$$P(\vec{\mathbf{n}}|I) = P_{\text{data}}(\vec{\mathbf{n}}:I)P_{\text{smooth}}(\vec{\mathbf{n}}), \tag{13}$$

where $P_{\text{data}}(\vec{\mathbf{n}}:I) = e^{-\beta E_{\text{data}}(\vec{\mathbf{n}}:I)}/Z_1$ and $P_{\text{smooth}}(\vec{\mathbf{n}}) = e^{-\beta E_{\text{smooth}}(\vec{\mathbf{n}})}/Z_2$, where Z_1 and Z_2 are normalization constants.

This equation is reminiscent of Bayes theorem Eq. (3). Relating the two equations we identify $P_{\text{data}}(\vec{\mathbf{n}}:I)$ with $P(I|\vec{\mathbf{n}})$ and $P_{\text{smooth}}(\vec{\mathbf{n}})$ with $P(\vec{\mathbf{n}})$. Thus, the data term corresponds, ideally, to a model of how the image I should appear given an object with surface normal $\vec{\mathbf{n}}$ (i.e., a reflectance function and a model of noise). Similarly, the smoothness term corresponds to the *a priori* probability of a surface before we receive any data (i.e., our assumptions about the world).

References

1. D. Marr, *Vision* (W. H. Freeman and Company, San Francisco, 1982).
2. H. Bülthoff and H. Mallot, J. Opt. Soc. Am., **5**, 1749–1758 (1988).
3. H. Bülthoff, M. Fahle and M. Wegman, Perception, in press 1991.
4. B. A. Dosher, G. Sperling and S. Wurst, Vision Res., **26**, 973–990 (1986).
5. N. Bruno and J. E. Cutting, J. Exp. Psychol. Gen., **117**, 161–170 (1988).
6. L. T. Maloney and M. S. Landy, Proc. SPIE: Visual Communications and Image Processing, Part 2, pp. 1154–1163 (1989).
7. J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing Systems* (Kluwer Academic Press, 1990).
8. A. L. Yuille, T. Yang and D. Geiger, Harvard Robot. Lab. Tech. Rep. 90-7 (1990).
9. I. A. Rock, *Perception* (Scientific American Library, Freeman and Company, New York, 1984).
10. B. K. P. Horn, *Robot Vision* (MIT Press, Cambridge Massachusetts, 1986).

11. G. Parisi, *Statistical Field Theory* (Addison-Wesley, Reading, Massachusetts, 1988).
12. H. Bülthoff, In *Computational Models of Visual Processing* (M. Landy and A. Movshon, eds., MIT Press, Cambridge, Massachusetts, 1991).
13. T. Bayes, Phil. Trans. R. Soc., **53**, 370–418 (1783).
14. M. Bertero, T. Poggio and V. Torre, *Regularization of Ill-posed Problems* (A.I. Memo. 924, MIT A.I. Lab. Cambridge, Massachusetts, 1987).
15. R. L. Gregory, *Eye and Brain, 3d ed.* (McGraw-Hill, New York, 1978).
16. H. Bülthoff and H. Mallot, In *AI and the Eye* (T. Troscianco and A. Blake, eds., John Wiley & Sons, London, UK, 1989).
17. D. Geiger and A. L. Yuille, Biol. Cybernet., **62**, 117–128 (1989).
18. A. Blake and H. Bülthoff, Nature, **343**, 165–168 (1990).
19. E. Grimson, Phil. Trans. R. Soc. (Lond.) Ser., **B 298**, 395–427 (1982).
20. S. Geman and D. Geman, IEEE Trans. PAMI, **6**, 721–741 (1984).
21. A. L. Yuille, D. Geiger and H. Bülthoff, Harvard Robot. Lab. Tech. Rep., **89-10** (1989).
22. H. Bülthoff and A. L. Yuille, Harvard Robotics Lab. Tech. Rep., **90-11** (1990).
23. A. Blake, Pattern Recog. Lett., **1**, 393–399 (1983).
24. D. Mumford and J. Shah, Proc. IEEE Conf. Comput. Vision Patt. Recog. (San Fransisco, 1985).
25. M. Gennert, M.I.T. AI Lab PhD. Thesis (1987).
26. S. Barnard, Int. J. Comput Vision, **3**, 17–33 (1989).
27. A. L. Yuille, Biol. Cybernet., **61**, 115–123 (1989).
28. K. Prazdny, Biol. Cybernet., **52**, 93–99 (1985).
29. S. B. Pollard, J. E. W. Mayhew and J. P. Frisby, Perception, **14**, 449–470 (1985).
30. S. Kirkpatrick, C. D. Gelatt, Jr. and M. P. Vecchi, Science, **220**, 671–680 (1983).
31. R. Durbin and D. Willshaw, Nature, **326** (April, 1987).
32. D. Geiger and F. Girosi, MIT Artificial Intelligence Laboratory Memo 1114 (Cambridge, Massachusetts, 1989).
33. C. Peterson and B. Söderberg, Int. J. Neural. Systems, **1**, 3–22 (1989).
34. P. Simic, Network: Comput. Neural Syst., **1**, 1–15 (1990).
35. A. L. Yuille, Neural Comput., **2**, 1–24 (1990).
36. P. Dev, Int. J. Man-Machine Stud., **7**, 511–528 (1975).
37. D. Marr and T. Poggio, Science, **194**, 283–287 (1976).
38. D. H. House, Ph.D. Thesis. University of Massachusetts at Amherst (1984).
39. H. Bülthoff and M. Fahle, MIT Artificial Intelligence Memo 1175 (Cambridge, Massachusetts, 1989).
40. P. Burt and B. Julesz, Science, **208**, 615–617 (1980).
41. H. Bülthoff and H. Mallot, In *Proceedings of the First International Conference on Computer Vision* (London, 1987).
42. K. Nakayama and S. Shimojo, Proc. Cold Spring Harbor, in press, 1990.
43. U. Grenander, Ann. Stat., **17**, **1**, 1–30 (1989).
44. J. Marroquin, S. Mitter, and T. Poggio, J. Am. Stat. Assoc., **82** (**397**), 76–89 (1987).
45. A. Blake and A. Zisserman, *Visual Reconstruction* (MIT Press, Cambridge, Massachussetts, 1987).
46. D. Geiger and A. Yuille, Int. J. Comput. Vision, in press (1990).
47. D. Kersten, In Vision: Coding and Efficiency (C. Blakemore, ed., Cambridge University Press, Cambridge England, 1990).
48. D. Kersten, H. Bülthoff, B. Schwartz, and K. Kurtz, Submitted for publication (1991).

49. A. Knoerr, Patt. Anal. Tech. Rep. No. 148 (Brown University, Providence, RI 1989).
50. M. Husein, S. Treue and R. Anderson, *Neural Computation*, **1**, 324–333 (1989).
51. V. S. Ramachandran and S. M. Anstis, Sci. Am., **254**, 102–109 (1986).
52. F. H. C. Crick, *What Mad Pursuit* (Basic Books, New York, 1988).
53. H. Bülthoff, J. Little and T. Poggio, Nature, **337**, 549–553 (1989).
54. A. L. Yuille and N. M. Grzywacz, Int. J. Comp. Vision, **3**, 155–175 (1989).
55. R. Feynman, *The Feynman Lectures on Physics* (Addison-Wesley, 1963).
56. D. Sheinberg, H. Bülthoff and A. Blake, *Perception*, **19**, A87b (1990).
57. P. J. Huber, *Robust Statistics* (John Wiley & Sons, New York, 1981).
58. T. Pavlidis, Proc. Eight Int. Conf. Patt. Recog. (Paris, 1986).
59. R. McKendall and M. Mintz, *Robust Fusion of Location Information* (Preprint. Department of Computer and Information Science, University of Pennsylvania, 1989).
60. J. Risansnen, Annals Stat., **11** **(2)**, 416–431 (1983).
61. R. Penrose, *The Emperor's New Mind* (Oxford University Press, Oxford, England, 1989).
62. A. Gelb, *Applied Optimal Estimation* (MIT Press, Cambridge, Massachusetts, 1974).