0

Introduction: A Bayesian formulation of visual perception

DAVID C. KNILL
Dept. of Psychology, University of Minnesota,

DANIEL KERSTEN
Dept. of Psychology, University of Minnesota

ALAN YUILLE
Division of Applied Sciences, Harvard University

0.1 Overview

Bayesian approaches have enjoyed a great deal of recent success in their application to problems in computer vision (Grenander, 1976-1981; Bolle & Cooper, 1984; Geman & Geman, 1984; Marroquin *et al.*, 1985; Szeliski, 1989; Clark & Yuille, 1990; Yuille & Clark, 1993; Madarasmi *et al.*, 1993). This success has led to an emerging interest in applying Bayesian methods to modeling human visual perception (Bennett *et al.*, 1989; Kersten, 1990; Knill & Kersten, 1991; Richards *et al.*, 1993). The chapters in this book represent to a large extent the fruits of this interest: a number of new theoretical frameworks for studying perception and some interesting new models of specific perceptual phenomena, all founded, to varying degrees, on Bayesian ideas. As an introduction to the book, we present an overview of the philosophy and fundamental concepts which form the foundation of Bayesian theory as it applies to human visual perception. The goal of the chapter is two-fold: first, it serves as a tutorial to the basics of the Bayesian approach to readers who are unfamiliar with it, and second, to characterize the type of theory of perception the approach is meant to provide. The latter topic, by its meta-theoretic nature, is necessarily subjective. This introduction represents the views of the authors in this regard, not necessarily those held by other contributors to the book.

First, we introduce the Bayesian framework as a general formalism for specifying the information in images which allows an observer to perceive the world. Such a specification, however, is only one side of the story of perception, written from a point of view outside an observer's head. It characterizes the information available to observers for perception, not how observers use this information. To characterize how observers use visual information requires a description of how the visual system makes inferences about the world based on image data, and is the point of view most commonly associated with information processing approaches

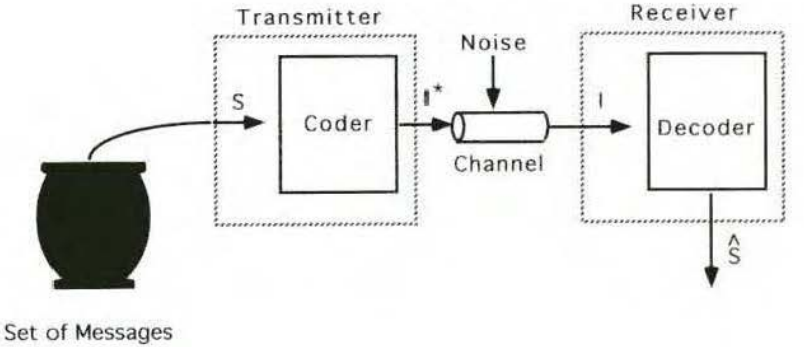
to perception. Secondly, therefore, we re-introduce the Bayesian framework in the context of modeling perceptual inference. By taking both points of view, we hope to highlight the fact that a Bayesian approach provides a useful framework for modeling both information and inference, and that the elements used to model information are equivalent to those used to model perceptual inference. In particular, we will see that explicit models of world structure (i.e. regularities in properties of the world) are needed to completely characterize both the information provided in images for perception and the actual inferences made by the visual system in the course of perception. The information problem demands of us models of the "true" structure of the world, whereas the inference problem demands models of the implicit assumptions about the world which the human visual system relies on for perception.

The introduction is organized into four parts: a qualitative formulation of the general problem of perception as communication, a brief tutorial on the Bayesian formulation of information, a reconceptualization of Bayesian formulations in terms of perceptual inference and a brief discussion of some of the issues involved in modeling visual perception within a Bayesian framework.

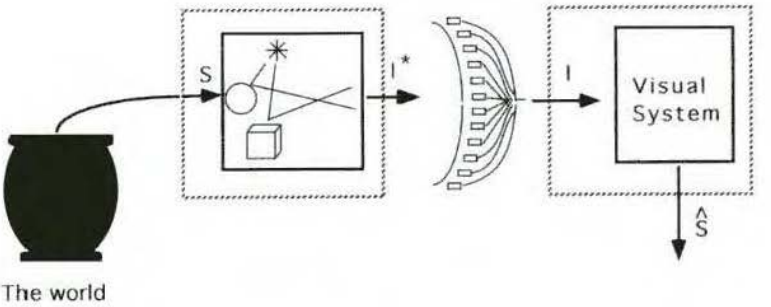
0.2 Perception as communication

Formulating visual perception as communication provides a useful metaphor for illustrating the nature of the information processing problem faced by the human visual system. A generic communication system (see figure 0.1a) consists of a *message set*, from which a *transmitter* draws messages, which it codes and sends as signals down a *channel* to a *receiver*, which decodes the signal to determine the message which was sent. Consider how this maps onto visual perception (figure 0.1b). For simplicity of discussion, we will consider the message set as the set of all possible physical configurations of scenes in our world[†]. While there is no identifiable physical transmitter, we can consider the messages (physical scenes) to be coded in the pattern of light reflected from surfaces and projected on a retinal receiving surface. The coding rules are the physical laws of light reflection, refraction and transmittance and the geometric laws of perspective projection. The receiver is the visual system, which processes the pattern of light impingent on the retina

[†] One can generalize the notion of "visual messages" to more abstract properties of a scene, such as the moods and intentions of biological organisms. For such abstract messages, we must conceive of the coder as including the processes by which these abstract properties are mapped to physical properties of a scene (e.g. facial expressions), as well as the image formation process which encodes these physical properties. In some sense, then, the set of messages is determined in part by exactly what an observer wants to "perceive". This is not a flaw in the metaphor, but does suggest caution in fixing our notions of what elements of the communication metaphor map to corresponding elements of perception.



(a)



(b)

Fig. 0.1 (a) A general communication system model: A transmitter draws a message **S** from a bin according to some probability law, codes it into a signal $\mathbf{I}^*$ and transmits the signal down a noisy channel. A receiver receives the noisy signal, **I** and attempts to decode it to determine the messages sent; that is, to estimate **S**. (b) The analogy with perception – **S** is a description of a particular scene in the world. An imaginary coder codes this description of scene properties in the form of an idealized image, $\mathbf{I}^*$. The visual system receives a noisy, bandlimited version of this image, **I**, which it must use to estimate properties of the scene **S**.

to "decode" the message; that is, it determines as best it can at least some of the properties of the scene which are projected to an image, or set of images.

Communication systems, at the level of abstraction used here, seem simple enough; however, as any communication engineer will tell you, the details of most real systems are quite complex. The code may not be complete (it may not be invertible to uniquely determine the original, coded message) and the physical channel will generally be bandlimited and noisy, so that the signal which arrives at the receiver is a degraded version of the original. The job faced by the receiver is, therefore, highly non-trivial. The same is true in vision. If we take as the received signals

patterns of photon capture (in space and time) in the retinal receptor mosaic, we see immediately that the signal is bandlimited and noisy. This is due to the purely physical properties of the imaging process, such as optical aberrations and diffraction at the pupil and the inherently probabilistic nature of photon emission and absorption. More noise is added in the transduction of light energy to electrochemical energy by receptors in the retina. Even if idealized as being uncorrupted by these influences, the received signals are not completely invertible, since the mapping from a three-dimensional scene description to a two-dimensional image description can potentially result in a loss of specificity. Moreover, the coding scheme embodied in physical image formation is inordinately complex: it includes highly nonlinear, and sometimes non-local effects; partial occlusion of one object behind others, interreflections between surfaces and shadows, just to name a few. Thus, even in cases where the decoding problem is theoretically well-defined, actually solving it is an extremely difficult computational problem.

Two related properties of engineered communication systems can help ameliorate a receiver's decoding problems; the set of messages often has a high degree of statistical structure, the knowledge of which can aid in the decoding of a signal; and the receiver often does not require a complete reconstruction of the transmitted message, but rather is concerned with estimating high-level features of the message (the existence of which result from the regularities in the message set). Consider, for example, a satellite surveillance system which tracks the movements of military ships and transmits the positions and trajectories of the ships to an intelligence station on earth. The set of possible messages is the set of possible positions and motions of military ships on the seas' surfaces. The sender is the satellite computer/radio system which codes the information and transmits it to a radio receiver on earth. The signal received by the radio operator on earth will be corrupted by noise; thus, some of the reported ships' coordinates may be in error or may be missing altogether. The set of messages has a very strong structure imposed by the constraints on positioning and movement of military ships. Besides physical constraints (for example, on speed), military ships are often clustered into groups whose motions are very strongly correlated. A well-designed decoding system, when doing error correction on an individual ship's motion, should estimate it based not only on the data transmitted for that ship, but also on the data transmitted for ships in its group. Moreover, the military planners who ultimately will use the information received may only be interested in fleet movements, thus the system could "average" the data for each ship in a group to produce an estimate of the fleet's motion which is more reliable than the estimate of any individual ships' motion.

The same situation holds for visual perception. The world has a tremendous amount of structure. A simple and obvious example is that matter coheres into objects, the shapes of whose surfaces structure the light projected to images. Moreover,

these shapes are not arbitrary, first being clustered in different classes (landscapes, plants, rocks, man-made objects, etc.), and within these classes having certain regularities (mountains being fractal, man-made objects tending to be symmetric, etc.). The same holds true for other scene properties; for example, surface material is constrained by natural laws, most objects are rigid, and when they aren't, deform in specific ways (e.g. the articulated motion of animate objects), ballistic movements follow Newton's laws of motion, etc.. This structure helps to make the information in images about scenes more reliable than it would be in a less structured world. It also plays a significant role in determining what scene properties a visual observer might be designed to estimate from images.

The perceptual problem faced by any visual system, like the decoding problem faced by the receiver in a general communication system, requires four basic components (see column 1, table 0.1) for its specification:

(1) The elements of interest in messages – for visual perception these are the properties of scenes the visual system attempts to estimate. As mentioned above, the structure of the environment plays some role in determining this, but so do the functional needs of the organism. An excellent example is the importance of surface properties to perception, which arises in part from the fact that matter coheres into objects and in part from the fact that the surface properties of objects determine in large part how they interact with each other and with observers (e.g. balls roll more easily than cubes).

(2) The structure of the message set; that is, the regularities which messages have – for visual perception this is the structure of scenes in our environment (regularities in object shape, etc.).

(3) The coding scheme used by the transmitter – in the context of visual perception, "the transmitter" encodes scenes as an image signal. While in some absolute sense, one should model the image signal as the pattern of photon capture over time in retinal receptors, many problems in perception are more conducive to high-level descriptions of the signal. This could be in terms of features such as optic flow, image contours or texture gradients, to name a few. In these cases, the coding scheme would map high-level features of a scene to high-level features of an image (e.g. edges of surfaces map to contours along luminance discontinuities in images). Whatever the case, the coding scheme is ultimately based on the physics of light reflection, refraction and transmission and the geometric laws of perspective projection.

(4) The form of signal corruption – again, this depends on what one considers to be the signal for a particular problem. A signal represented as the pattern of photon capture over time in retinal receptors would be "corrupted" by the uncertainty of photon emission and capture. For analyses in which the signal is treated as a collection of higher-level image features, the effects of physical corruption of the image are often considered to be negligible for purposes of the problem at hand or are approximated as noise added to the coding of the high-level features; for example, noise added to the orientations or curvatures of image contours.

Taken together, these four components define what properties of a scene a visual system attempts to estimate in the course of perception and how these scene properties are encoded in images. In a deeper sense, components (2), (3) and (4) specify the information content of images; that is, what images can potentially tell one about the world. Note the role of the second component, the structure of the environment, in the definition of information. It is not a second source of information which is "added" by an observer to image information, but rather it is an integral part of a specification of what information images carry about scenes.

The discussion so far can be thought of as describing a particular way to characterize perceptual problems posed to an observer. We can summarize this in the following statement:

Perceptual problems posed to an observer are characterized by (1) the properties of the world which an observer makes inferences about (e.g. shape), and (2) the information provided by images about those properties, as determined by the prior structure of the world, the coding scheme and the form of image data corruption.

A complete characterization of a communication system also requires specifying how the receiver actually decodes the signals it receives to determine what message was sent; that is, how it solves the decoding problem. Analogously, we are interested in how an observer solves perceptual problems in the act of perception:

An observer's solutions to perceptual problems are characterized by (1) the properties of the world which an observer makes inferences about, (2) the image data actually used by observers as the basis for perceptual inferences and (3) the assumptions about image coding and about the prior structure of the world used by the observer to make inferences.

The quality of an observer's solution of a perceptual problem depends on how well the observers' assumptions about the world and about image coding match the world in which it exists; that is, on the similarity between corresponding elements of the perceptual problem and perceptual solution specifications.

The communication metaphor does not completely capture the difficulty of perception. In prototypical communication systems, both man-made and biological, senders and receivers are designed, or evolve, together; that is, the coding and decoding schemes are designed hand-in-hand to match one another. The classic example of this in the biological domain is human language, for which production and comprehension systems evolved together. Moreover, the coding schemes are often designed to ameliorate the problems imposed by signal corruption in the transmission channel (for example, by adding appropriate forms of redundancy in the code). Visual perception, on the other hand, involves the evolution of an organism's visual system to match the structure of the world and the coding scheme provided by nature. Unlike usual communication systems, the coding scheme (light reflection and perspective projection) has not been designed a-priori to maximize the reliability of the information transmitted about message features of interest to an organism

(scene properties), nor to minimize the computational problems of decoding the signals. It simply exists as a property of our environment, and the visual system has to make do with what nature has provided. All of our experience attempting to build artificial vision systems tells us that the computational problems of decoding images are actually quite difficult.

0.3 The Bayesian formulation of the problem of perception

0.3.1 The Bayesian characterization of information

The basic idea behind the Bayesian approach is to characterize the information about the world contained in an image as a probability distribution which characterizes the relative likelihoods of a viewed scene being in different states, given the available image data. The exact form of the distribution, called the "posterior" conditional probability distribution, is determined in part by the image formation process, including the nature of the noise added in the image coding process, and in part by the statistical structure of the world. As we will see shortly, Bayes' rule provides the mechanism for combining these two factors into a final calculation of the posterior distribution. The Bayesian approach distinguishes itself from other statistical formulations of information by taking into account the contributions of both factors to the specification of information. In particular, the approach is notable for its reliance on explicit models of world structure. While this forms the basis for most attacks on the approach, we emphasize that modeling this aspect of visual information is a fundamental necessity, and is always implicitly done, if not explicitly.

Table 0.1 summarizes the Bayesian formalization of the decoding problem posed to the receiver in a communication system. Referring back to our original discussion of the four major components of a model of information, we have for visual perception;

(1) A formal representation of the scene properties of interest – $\mathbf{S}$. $\mathbf{S}$ might include such things as surface shape, object motion, observer motion, the projected time of collision between objects, and so on.

(2) A model of the structure of scenes which defines the *prior* probability distribution, $p(\mathbf{S})$. $p(\mathbf{S})$ embodies the large number of statistical dependencies which exist between scene properties.

(3) A model of image formation, which we write as a function applied to $\mathbf{S}$, $\pi(\mathbf{S})$. π can be thought of as an idealized model of image formation which incorporates the laws of light reflection, refraction and emission as well as the laws of perspective projection. More realistically, π could be modeled so as to take into account physical effects of imaging such as blur, optical aberrations in the eye and sampling.

(4) A model of image noise, $\mathbf{N}$, which we can think of as being added to the result of the image formation function, $\mathbf{I} = \pi(\mathbf{S}) + \mathbf{N}$. It need not, of course, be strictly additive,

<table>
<tr><th colspan="2">Communication system</th><th colspan="2">Bayesian framework for perception</th><th></th></tr>
<tr><td colspan="2">Elements of interest in messages</td><td colspan="3">Scene properties of interest
$\mathbf{S}$</td></tr>
<tr><td rowspan="3">Information</td><td>Structure of the message set</td><td colspan="2">Prior
$p(\mathbf{S})$</td><td rowspan="3">Posterior
$p(\mathbf{S} \mid \mathbf{I})$</td></tr>
<tr><td>Coding</td><td>Image Formation
$\pi(\mathbf{S})$</td><td rowspan="2">Likelihood
$p(\mathbf{I} \mid \mathbf{S})$</td></tr>
<tr><td>Noise</td><td>Image Noise
$\mathbf{N}$</td></tr>
</table>

Table 0.1 *Column 1 shows the qualitative components of a communication problem specification. Column 2 shows the corresponding formal components within the Bayesian framework (see text for details of variable and function meanings).*

and, depending on what one is modeling as the input to the visual system, it may involve complex models of noise induced at various stages in neural processing.

Sticking to our metaphor of perception as communication, we say that images, $\mathbf{I}$, are signals which provide information about transmitted messages, which are taken to be specific configurations of scene properties, $\mathbf{S}$. The posterior conditional probability distribution $p(\mathbf{S} \mid \mathbf{I})$, characterizes this information. If an image uniquely specifies the scene (e.g. their is no uncertainty induced by noise), then the posterior distribution is trivial, being zero for all scene configurations but the one actually being viewed. More commonly, images have some ambiguity, and this is reflected in the "spread" of probability over the space of possible scenes. The posterior distribution depends on the structure of the set of possible scenes ($p(\mathbf{S})$), the image formation function ($\pi(\mathbf{S})$) and the noise added to images ($\mathbf{N}$). Bayes' rule specifies a way to partially decompose the posterior into these parts. According to Bayes' rule, the posterior is given by

$$p(\mathbf{S} \mid \mathbf{I}) = \frac{p(\mathbf{I} \mid \mathbf{S}) p(\mathbf{S})}{p(\mathbf{I})}. \tag{0.1}$$

For our purposes, we can treat $p(\mathbf{I})$, the probability of occurrence of an image, as a normalizing constant, so we have

$$p(\mathbf{S} \mid \mathbf{I}) \propto p(\mathbf{I} \mid \mathbf{S}) p(\mathbf{S}). \tag{0.2}$$

$p(\mathbf{I} \mid \mathbf{S})$, for a given value of $\mathbf{S}$ (a given scene), is a probability distribution specifying

the relative probability of obtaining different images from that scene. It is a function of the image formation function and the corrupting noise (thus incorporating two of the components of information described above). $p(\mathbf{I} \mid \mathbf{S})$ is generally referred to as the likelihood function for $\mathbf{S}$. $p(\mathbf{S})$ we have described above, and is the prior distribution on scene configurations $\mathbf{S}$.

Equation (0.2) is the foundation of the Bayesian approach to visual perception. It shows how to factor out the relative effects on image information of the coding scheme and noise on the one hand, and the prior structure of the environment on the other. Consider what the likelihood function and prior distribution represent for problems of visual perception. The likelihood function reflects the noisiness of the data and the loss of specificity implicit in the projection from three dimensions to two. If the image were uncorrupted by noise and unaffected by optical distortions, then $p(\mathbf{I} \mid \mathbf{S})$ would be non-zero only for those scenes which would project, under perspective projection, to a given image $\mathbf{I} = I$; that is, it would select a set of candidate scene interpretations for a given image[†]. Noise has the effect of spreading the likelihood function over a larger range of possible scenes, making the information provided by an image about scene properties more unreliable. The distribution $p(\mathbf{S})$ is the prior probability of different collections of scene properties actually occurring in our environment. It embodies knowledge of the structure of the environment which constrains the perceptual estimate of scene properties. A good example of a prior constraint is the assumption that object motions tend to be rigid. The rigidity constraint is often hard-wired into structure-from-motion models, leading to an effective assumption that $p(\mathbf{S}) = 0$ for non-rigidly moving objects (Koenderink & van Doorn, 1975; Ullman, 1979; Bennett *et al.*, 1989). Other examples of prior constraints are the smoothness constraints often used in computer vision models (Ikeuchi & Horn, 1981; Julesz, 1971; Marr & Poggio, 1979; Yuille, 1989). Typically, when formulated in probabilistic terms they characterize particular probabilistic models of surfaces (Szeliski, 1989). (See chapter 5, by Yuille and Bülthoff and chapter 8 by Belhumeur for complete discussions of the relationship between smoothness constraints and Bayesian priors).

0.3.2 A tutorial example

In this section, we illustrate the Bayesian formulation of an information processing problem with a simple example for which we can compute the posterior function exactly, but which retains key similarities to real problems in perception. In our example communication system (see figure 0.2), the set of messages consists of four

[†] Transactionalist theory, a school of perceptual psychology popularized by Ames with "illusions" such as the Ames' room and Ames' trapezoidal window, referred to the set of scenes which could project to a given image or images as "equivalent configurations" (Ittleson, 1960)

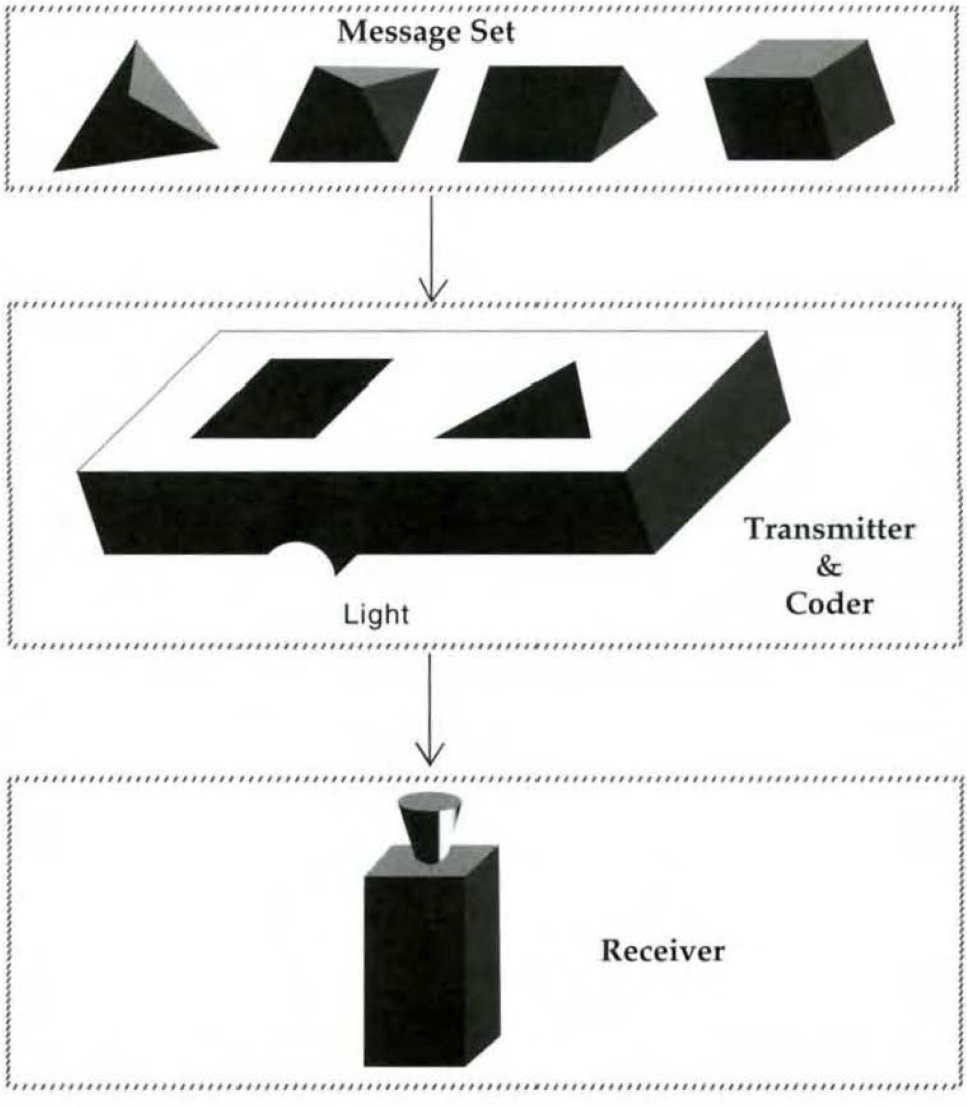


Fig. 0.2 The communication system for the shape sorter example. See text for description.

objects: a tetrahedron, a pyramid, a prism and a cube. Each object is a "message". The tetrahedron has four triangular sides, the pyramid has a square base with four triangular sides, the prism has three square sides and two triangular sides, and the cube has six square sides. Each of the square and triangular sides has the same shape for all the objects. The transmitter selects objects for coding and transmission with the probabilities given in table 0.2. These probabilities form the prior distribution characterizing the structure of the message set, or what we have referred to for vision

Prior distribution of objects p(object)	
tetrahedron	0.1
pyramid	0.3
prism	0.4
cube	0.2

Table 0.2.

as the structure of the world. The coding device used by the transmitter has two stages. The first is like the toy for toddlers in which only certain three-dimensional shapes fit through two-dimensional holes. The selected object is dropped in a box and can fall through one of two slots; a triangle slot, whose shape matches the triangular sides of all the objects, or a square slot, whose shape matches the square sides of the objects. The shape of the side of an object which faces down determines which slot an object falls through. For simplicity, we assume that each side of an object has equal probability of facing down. The laws governing this device are crudely analogous to the process of geometric projection in vision; thus, we refer to the "output" of this stage of the coding device (which slot an object falls through) as an object's silhouette. The second stage of the coder sends a color signal to the receiver based on the object's silhouette: red if the silhouette is a triangle, and blue if it is square. The final component of the system is a receiver, which we will take to be a photodetector which is sensitive to the wavelength of light it absorbs. The photodetector signals whether a red or blue light is received.

As a first step in our analysis of the information provided by the signals in this system, let us ignore the color coding and treat the silhouettes as the received signal. The problem for a receiver detecting these silhouettes is that they do not uniquely determine the shapes of the objects selected by the transmitter, since, unlike the child's toy, two of the objects (the pyramid and the prism) can fall through either of the two slots. For a given silhouette, therefore, there is more than one possible message which could have given rise to the silhouette, and the information provided by the silhouette is ambiguous and probabilistic. The information is characterized by the posterior function, $p(\mathbf{S} \mid \mathbf{I}) = p(\textbf{object} \mid \textbf{silhouette})$, where **object** is a random variable specifying the object chosen by the transmitter, and **silhouette** is a random variable specifying the silhouette received as a signal. For now, we are assuming a noiseless signal, so the posterior function is determined by the coding scheme and the prior distribution of objects. Having specified the prior, we turn to a probabilistic specification of the coding scheme.

	Likelihood function $p(\textbf{silhouette} \mid \textbf{object})$			
	Tetrahedron	Pyramid	Prism	Cube
triangle	1.0	0.8	0.4	0.0
square	0.0	0.2	0.6	1.0

Table 0.3 *For a given silhouette, the sum of the likelihood function taken over the objects (within a row) is not 1, reflecting the fact that fixing the signal does not make the likelihood function a probability distribution on set of possible messages. Fixing the message, however, does make it a probability distribution on the set of possible signals, as seen by summing within a column.*

We use the likelihood function, $p(\mathbf{I} \mid \mathbf{S})$, to model the probabilistic properties of the coding scheme. Since we have assumed that each side of an object, when dropped in our imaginary coding box, has an equal probability of facing down, the probability that an object will be coded as a given silhouette is simply the proportion of sides of the object which have that silhouette's shape. A simple calculation gives the probabilities shown in table 0.3.

To obtain the posterior function, we combine the likelihood function and the prior distribution using Bayes' rule, giving

$$p(\textbf{object} \mid \textbf{silhouette}) \propto p(\textbf{silhouette} \mid \textbf{object})\, p(\textbf{object}). \tag{0.3}$$

Table 0.4 summarizes the results of calculating the posterior function for all possible signals and messages in our example. While both silhouettes allow three possible interpretations of the object selected by the transmitter, a receiver which had to choose one and be correct as often as possible would choose the object with the highest probability conditional on the silhouette received: for a triangle, it would be the pyramid, and for the square, it would be the prism.

We now turn to a consideration of the effects of noise on the posterior and consider the full example system, including the color coder and the photodetector receiver. In vision, we do not directly receive information about the geometrical shape of objects, rather, the signal received by the retina is a more indirectly coded form of the shape information than is given by silhouettes. In a similar way, the transmitter in our full example codes objects in the form of the color of light it transmits. If there is no noise in the coding or in the transduction of light by the photodetector, the posterior for objects conditional on the color signal is equivalent to the one derived for a silhouette signal, with red replacing triangle, and blue replacing square. Suppose, however, that noise is added to the signal, either in the coder or in the photodetector,

Posterior distribution $p(\textbf{object} \mid \textbf{silhouette})$		
	Triangle	Square
tetrahedron	0.2	0.0
pyramid	0.48	0.12
prism	0.32	0.48
cube	0.0	0.4

Table 0.4 *For a given silhouette, the sum of the posterior function over the different objects (within a column) is 1, reflecting the fact that fixing the signal, makes the posterior a probability distribution defined over the set of possible messages.*

so that the mapping from silhouettes to received color signals is not one-to-one. We then need to compute a different likelihood function, $p(\textbf{color} \mid \textbf{object})$, and hence a different posterior, $p(\textbf{object} \mid \textbf{color})$. Assuming the color noise is independent of the process used to select which silhouette matches an object, we can write the likelihood function as

$$\begin{aligned} p(\textbf{color} \mid \textbf{object}) = {} & p(\textbf{color} \mid \textbf{silhouette} = \text{triangle}) \\ & \times p(\textbf{silhouette} = \text{triangle} \mid \textbf{object}) \\ & + p(\textbf{color} \mid \textbf{silhouette} = \text{square}) \\ & \times p(\textbf{silhouette} = \text{square} \mid \textbf{object}), \end{aligned} \tag{0.4}$$

where $p(\textbf{color} \mid \textbf{silhouette})$ is determined by the color noise. Values of $p(\textbf{color} \mid \textbf{silhouette})$ for the noise-free case and an example noisy case are tabulated in table 0.5. If we use the likelihood function obtained in the noisy example, we obtain the posterior function shown in table 0.6. Note that the noise has the effect of making the posterior distribution more similar to the prior distribution of shapes. This reflects the loss of reliability of the signal's information induced by the addition of noise. In the limit, as the noise increases, the posterior distribution approaches the prior distribution showed in table 0.2. In the example we have described, the noise has also changed the peaks of one of the distributions, so that the most likely interpretations given our example noisy color signals are different from those obtained with noise-free data (in fact the most likely interpretations given either signal are the same, suggesting that a receiver which uses the strategy of picking the most likely interpretation will do no better with the information provided by the received signal than without).

Noise-free color signal p(**color** | **silhouette**)		
	Triangle	Square
red	1.0	0.0
white	0.0	1.0

Noisy color signal p(**color** | **silhouette**)		
	Triangle	Square
red	0.6	0.4
white	0.4	0.6

Table 0.5.

Posterior function for noisy color signal p(**object** | **color**)		
	Red	White
tetrahedron	0.12	0.08
pyramid	0.336	0.264
prism	0.384	0.416
cube	0.16	0.24

Table 0.6.

The example illustrates a number of points about the problem of visual perception and the Bayesian approach to characterizing visual information. First, the form of the received signal is not simply related to the form of the messages. Image intensities are a coded form of what we "see" and are as qualitatively different from scene properties as the color signal was from the nature of the objects in our example. Of course, in our example the mapping from messages to signals was quite simple. The same is not true for the mapping from scene properties to image data. Second, both the lack of a one-to-one inverse mapping from images back to scene properties and the presence of image noise make the information provided by images about scenes inherently probabilistic. In our example, not only could different objects fit through different slots in the coder, but noise further increased the ambiguity of the received color signal. If we had included a stellate shaped object in the message set and a similarly shaped slot in the coder, that particular silhouette would provide unambiguous information about the object chosen as a message (since only the stellate-shaped object would fall through it). The addition of noise would impose some ambiguity on the final color signal. While searching for such invariants is a good research strategy, we should not be surprised to find that few exist in images. Finally, just as in the example, the prior structure of the environment plays a crucial role in determining the information provided by images about scene

properties. In the example, treating the likelihood function as a characterization of signal information would lead an observer to make irrational inferences (compare tables 0.3 and 0.6) in that the maxima of the likelihood function occur for different objects than the maxima of the posterior distributions.

0.4 Perception as unconscious inference

0.4.1 Bayesian models of inference

We have described the Bayesian framework as a language for specifying what information images provide about the world. From this perspective, the framework provides a way to objectively specify the information content of images for the estimation of scene properties or more generally for the performance of perceptual tasks. Consideration of human perceptual performance, however, generally suggests a somewhat different perspective; namely, the characterization of perception as a process of unconscious inference, as suggested by Helmholtz (1925). From this point of view, Bayesian probability provides a normative model for how prior knowledge should be combined with sensory data to make inferences about the world[†]. Specification of the functions $p(\mathbf{I} \mid \mathbf{S})$ and $p(\mathbf{S})$ form the basis of what would be an "ideal" perceptual inference device. One more element is needed, however, to completely model an inference process: a specification of a decision rule for selecting an estimate of $\mathbf{S}$ based on $p(\mathbf{S} \mid \mathbf{I})$. Common rules applied in the literature include selection of the peak of the distribution (Maximum A-Posteriori, or MAP, estimation) or selection of the mean of the distribution (Minimum Mean Squared-Error, or MMSE, estimation). More general decision rules can be incorporated using cost functions to weight the relative cost of making errors in an inference (see chapter 5, by Yuille & Bülthoff, and chapter 9, by Freeman). A complete functional model of an ideal perceptual inference device, then, consists of a model of the information in images, as characterized by $p(\mathbf{S} \mid \mathbf{I})$, and a model of the decision rule to be applied to this function to make inferences.

We make the jump from building ideal inference devices to modeling human perception by recognizing that one can treat the human visual system as making perceptual inferences on an implicitly assumed model of $p(\mathbf{S} \mid \mathbf{I})$, which we will refer to as $p_h(\mathbf{S} \mid \mathbf{I})$ (Kersten, 1990; Knill & Kersten, 1991). This model incorporates assumptions about image formation and the structure of scenes in our environment. In some sense, one could say that a model of $p_h(\mathbf{S} \mid \mathbf{I})$ (along with a model of perceptual decision rules) characterizes the world to which the human visual system

[†] Classical Bayesian inference in the sciences interprets probabilities as degrees of belief. Jaynes (1986) has shown that given some elementary and reasonable assumptions about how degrees of belief should be formulated, one arrives at the probabilistic calculus, or a class of monotonic derivatives of the calculus, as the appropriate way to combine information to modify degrees of belief.

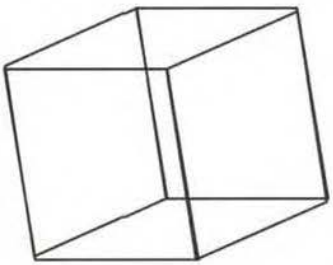


Fig. 0.3 The Necker cube. The line drawing can appear in one of two orientations, depending on which face is seen in front, however, it always appears as a cube.

is "tuned" and in which humans would be ideal perceptual inference makers (see chapter 6, by Knill *et al.*). To demonstrate this, consider the example of the previous section, and assume the presence of an observer viewing the outputs of the light. Our hypothetical observer might implicitly (and incorrectly) assume that each of the four objects was equally likely to fall into the shape sorter. this would lead the observer to "perceive" the shape of objects based on a posterior function which has the same form as the likelihood function (see table 0.3), leading to more mistakes than if the observer assumed the correct prior.

Terms like prior knowledge and inference suggest to many people the view that perception is strongly influenced by cognitive factors. We do not mean to do so here. While we readily acknowledge the possibility of cognitive effects on perception, this is not what we mean in our conception of perception as inference. Much of what we refer to as prior knowledge may be built into low-level, automatic perceptual processes which do not have access to our cognitive database of explicit knowledge about the world. For example, in some contexts, pior knowledge about the world can be implicitly built into relatively simple filters for the estimation of scene properties (Kersten *et al.*, 1987; Knill & Kersten, 1990). More generally, work in neural networks has shown that many network models can be conceived of as particular implementations of Bayesian inference (Golden, 1988; MacKay, 1991). The prior knowledge in these cases is "represented" by the connection strength between cooperative computational elements.

The Necker cube provides a simple example of a probabilistic inference made by the human visual system (see figure 0.3) which is classically "perceptual" and automatic. Though often used to illustrate the bistability of some percepts, the more impressive phenomenon is the obvious one – that we see it as a cube at all. Consideration of the ambiguity imposed by mapping a three-dimensional object onto two dimensions shows that an infinite number of possible polyhedral shapes could have given rise to an image of a Necker cube (just as multiple objects could fall through the square slot in our toy example). The visual system selects as its

estimate of the shape the most symmetric of the possibilities – a cube. As simple as it is, this is an impressive demonstration of the visual system's use of prior constraints on object shape (the nature of the prior constraints needed for such a percept to be accurate is discussed in chapter 3, by Richards *et al.*).

0.4.2 Bayesian theories and levels of explanation

The information processing approach to modeling perceptual inference typically leads to theories about perceptual *process*; that is, about the architecture and algorithms of the system which makes the inferences. This is true despite Marr's prescription to build computational theories for perceptual problems before modeling the processes which implement the theories (Marr, 1982). One reason for this state of affairs is that most researchers have not had available a formal framework for building computational theories with enough specificity to usefully constrain models of process (more so, that is, than informal statements of principles). Moreover, there is some confusion about the nature of what comprises a computational theory, the term itself being rather vague. Marr was unclear as to whether a computational theory should characterize the problems posed to a perceiving organism or some aspects of an organism's solution to these problems. At various times he seems to have meant it to characterize one or the other (or both). In considering the Bayesian approach as a framework for building what we think of as computational theories, we have found that a new conceptualization of Marr's three-fold heirarchy of levels of explanation (computational / algorithmic / implementation) has naturally emerged which resolves the ambiguity. This is summarized in figure 0.4. In essence, we split the computational level into two components: theories of information and rational theories of inference[†]. The former can be thought of as characterizing the computational problems posed to an observer, while the latter characterizes the computational aspects of an observer's solution to these problems. Below the two computational levels is the impementation level of explanation, which describe properties of the processeses of perception. For our purposes here, we treat this as one level, though it may usefully be partitioned into more, as Marr did. What is notable about the formulation is that the Bayesian framework applies both to the information level of description and to the rational level. By providing a common language for building theories of both types, the framework supports a strong interaction between theoretical analyses of information and the process of modeling human perceptual behavior. Just as importantly, it provides the formal tools with which to build theories at these levels without necessarily having to make recourse to lower levels of explanation. Ultimately, the levels must interact and constrain

[†] We have borrowed the term "rational" to characterize Bayesian theories of inference from a related proposal by Anderson (1991) in the context of explaining cognitive function

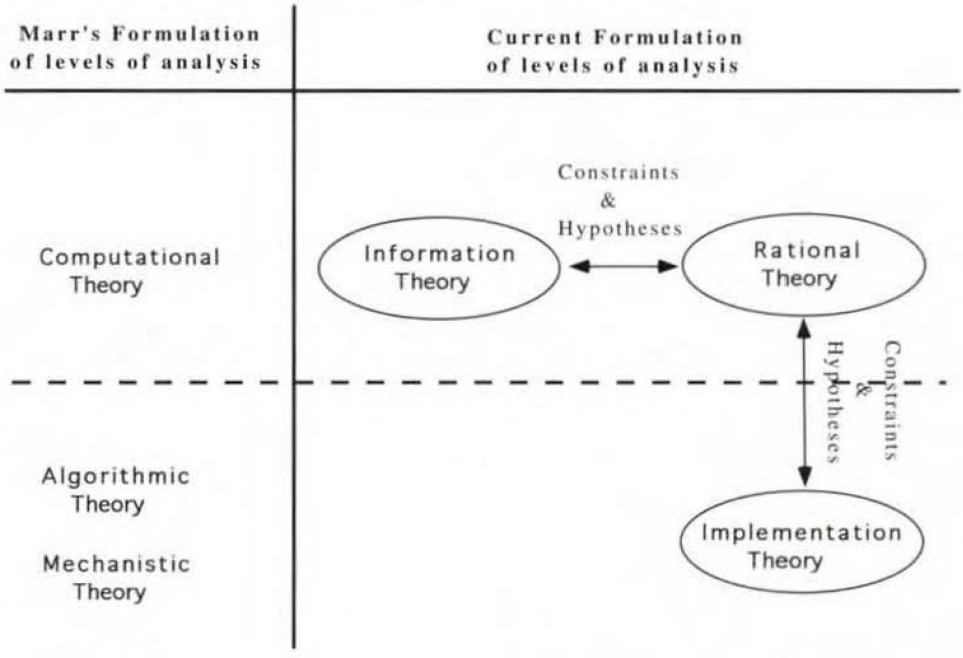


Fig. 0.4 The discussion in the text suggests that Marr's computational level of theory for visual processes really consists of two parallel modes of explanation: theories of the information available for perception and rational theories of the inferences made by the visual system in the course of perception. The Bayesian framework provides the structure for building both types of theory, with the addition of decision criteria for models of inference. Theories of information constrain theories of inference, because they limit the reliability with which inferences can be made. Both types of theories can suggest hypotheses for the other, since, on the one hand, the information and constraints in our world are always available to the visual system to use, and on the other hand, information and constraints employed by the visual system are likely to have evolved to match those available in the world. While we have focused on the two upper levels of theory-building, theories at the implementation level clearly interact strongly with those at the level of rational inference, mutually constraining and informing one another.

one another, but it is important for the development of perceptual models that we be able to build predictive theories of human performance at the rational level. This level suggests its own questions and modes of explanation which cannot be easily characterized at other levels.

0.5 Conclusions

A number of arguments support using the Bayesian framework for modeling perception. First, it provides a normative framework within which to formulate objective

theories of the information provided by images in our world for different perceptual tasks[†]. The form of the function $p(\mathbf{S}|\mathbf{I})$ which describes our environment provides a theoretical absolute limit on the reliability with which an observer can make perceptual inferences (Kersten, 1990). A function made up of strongly peaked posterior distributions, $p(\mathbf{S}|\mathbf{I} = I)$, supports inferences which are likely to be correct or to be made with small errors, while one composed of broad distributions does not. Secondly, formulation of computational models within the framework requires making explicit many assumptions which are often left implicit. It is, we feel, the natural framework in which to formulate computational descriptions of many problems. In particular, it distinguishes between functionally different aspects of the computational problems facing an observer: The nature of the uncertainty in the data for performing a perceptual task and the prior constraints on scene structure which serve to reduce this uncertainty. Thirdly, the framework provides a means for formalizing experimentally testable hypotheses about functional aspects of human perceptual processing. Building objective theories of the information available for perception and theories of human perceptual performance within the same framework supports a strong degree of cooperation between formal, mathematical analyses and psychophysical experimentation.

We have attempted to introduce the main concepts of a Bayesian framework for modeling perception and have highlighted three of its features: that it provides the tools for a full mathematical description of the problems of perception, that these same tools may be used to build functional models of perceptual performance, and that it suggests a new conceptualization of perception which provides a novel structure for asking questions about perceptual function. We have argued for the usefulness of the framework as a paradigm for investigating and modeling human perception, but have done so at a fairly abstract level, never actually discussing particular applications of the approach to real problems in perception. The success or failure of such applications will be the ultimate test of the framework's usefulness and will help define the domains to which it is best suited. We also have not elucidated many of the specific problems which arise from considering perception within a Bayesian framework. The remaining chapters of the book flesh out these gaps and should leave the reader with a greater appreciation and understanding of the approach and its application to visual perception.

[†] Jaynes (1986) has shown that some basic qualitative criteria on how measures of belief are enough to derive the probabilistic calculus (or some monotonic derivative of it) as the appropriate mechanism for combining and manipulating degrees of belief. We refer the reader to his paper for a proof and discussion and simply not that the criteria he proposes for degrees of belief are exactly those which one would want to apply to measures of information.

References

Anderson, J.R. (1991). The adaptive nature of human categorization, *Psychological Review*, **98**, 409-429.

Bennett, B.M., Hoffman, D.D. & Prakash, C. (1989). *Observer Mechanics*. London: Academic Press.

Bennett, B.M., Hoffman, D.D., Nicola, J.E. & Prakash, C. (1989). Structure from two orthographic views of rigid motion, *Journal of the Optical Society of America, A*, **6**, no. 7, 1052-1069.

Bolle, R.M. & Cooper, D.B. (1984). Bayesian recognition of local 3-D shape by approximating image intensity functions with quadric polynomials, *IEEE Trans. PAMI*, **PAMI-6**, 418-429.

Clark, J.J. & Yuille, A.L. (1990). *Data Fusion for Sensory Information Processing Systems*. Kluwer Academic Press.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. PAMI*, 721-741.

Golden, R. (1988). A unified framework for connectionist systems, *Biological Cybernetics*, **59**, 109-120.

Grenander, U. (1976-1981). *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures*. Springer-Verlag.

Helmholtz, H. (1925). *Physiological Optics, Vol. III: The Perceptions of Vision (J. P. Southall, Trans.)*. Optical Society of America, Rochester, NY. (Original publication in 1910).

Ickeuchi, K. & Horn, B.K.P. (1981). Numerical shape from shading and occluding boundaries, *Artificial Intelligence*, **17**, 141-184.

Ittleson, W.H. (1960). *Visual Space Perception*. New York: Springer Publishing Co.

Jaynes, E.T. (1986). Bayesian methods: general background. In *Maximum Entropy and Bayesian Methods in Applied Statistics*, ed. J.H. Justice. Cambridge University Press.

Jepson, A. & Richards, W. (1993). What is a percept? *Dept. of Computer Science Tech. Report RBCV-TR-93-43*.

Julesz, B. (1971). *Foundations of Cyclopean Perception*. University of Chicago Press.

Kersten, D. (1990). Statistical limits to image understanding. In *Vision: Coding and Efficiency*, ed. C. Blakemore. Cambridge Univ. Press.

Kersten, D., O'Toole, A.J., Sereno, M., Knill, D.C. & Anderson, J. (1987). Associative learning of scene parameters from images, *Applied Optics*, **26**, 4999-5006.

Knill, D.C. & Kersten, D. (1990). Learning a near-optimal estimator for surface shape from shading, *Computer Vision, Graphics and image Processing*, **50**, 75-100.

Knill, D. & Kersten, D. (1991). Ideal perceptual observers for computation, psychophysics and neural networks. In *Vision and Visual Dysfunction, Vol. 14: Pattern Recognition by Man and Machine*, ed. R. Watt. New York: MacMillan Press.

Koenderink, J.J. & van Doorn, A.J. (1975). Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer, *Optica Acta*, **22**, 773-791.

MacKay, D.J.C. (1991). A practical Bayesian framework for backpropogation networks, *Neural Computation*, 448-472.

Madarasmi, S., Kersten, D. & Pong, T. (1993). The computation of stereo disparity for transparent and opaque surfaces. In *Advances in Neural Information Processing*, ed. G.J. Giles, S.J. Hanson and J.D. Cowan. Morgan Kaufman.

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

Marr, D. & Poggio, T. (1979). Cooperative computation of stereo disparity, *Science*, **194**, 283-287.

Marroquin, J.L., Mitter, S. & Poggio, T. (1985). Probabilistic solution of ill-posed problems in computational vision. In *Proceedings Image Understanding Workshop* ed. L. Baumann, pp. 293-309. McLean, VA: Scientific Applications International Corporation.

Szeliski, R. M. (1989). *Bayesian Modeling of Uncertainty in Low-Level Vision*. Norwell, MA: Kluwer Academic Press.

Ullman, S. (1979) The interpretation of structure from motion, *Proc. R. Soc. London, B*, **23**, 405-426.

Yuille, A.L. (1989). Energy functions for early vision and analog networks, *Biological Cybernetics*, **61**, 115-123.

Yuille, A.L. & Clark, J.J. (1993). Bayesian models, deformable templates and competitive priors. To appear in *Spatial Vision in Humans and Robots*, ed. L.Harris and M. Jenkins. Cambridge University Press, Cambridge.