# Observer Theory, Bayes Theory, and Psychophysics

**Bruce M. Bennett**

*Department of Mathematics, University of California,*
*Irvine, California 92717*

**Donald D. Hoffman**

*Department of Cognitive Science, University of California,*
*Irvine, California 92717*

**Chetan Prakash**

*Department of Mathematics, California State University,*
*San Bernardino, California 92407*

**Scott N. Richman**

*Program in Mathematical Behavioral Science, University of California,*
*Irvine, California 92717*

## 1. INTRODUCTION

The search is on for a general theory of perception. As the papers in this volume indicate, many perceptual researchers now seek a conceptual framework and a general formalism to help them solve specific problems.

One candidate framework is "observer theory" (Bennett, Hoffman, and Prakash, 1989a). This paper discusses observer theory, gives a sympathetic analysis of its candidacy, describes its relationship to standard Bayesian analysis, and uses it to develop a new account of the relationship between computational theories and psychophysical data. Observer theory provides powerful tools for the perceptual theorist, psychophysicist, and philosopher. For the theorist it provides (1) a clean distinction between competence and performance, (2) clear goals and techniques for solving specific problems, and (3) a canonical format for presenting and analyzing proposed solutions. For the psychophysicist it provides techniques for assessing the psychological plausibility of theoretical solutions in the light of psychophysical data. And for the philosopher it provides conceptual tools for investigating the relationship of sensory experience to the material world.

Observer theory relates to Bayesian approaches as follows. In Bayesian approaches to vision one is given an image (or small collection of images), and a central goal is to compute the probability of various scene interpretations for that image (or small collection of images). That is, a central goal is to compute a conditional probability measure, called the "posterior distribution," which can be written $P(Scene \mid Image)$ or, more briefly, $P(S \mid I)$. Using Bayes rule one writes

$$P(S \mid I) = \frac{P(I \mid S)P(S)}{P(I)}.$$

This formula gives a method for computing the desired posterior distribution, and is widely referred to in vision research (Geman and Geman, 1984; Marroquin, 1989; Szeliski, 1989; Bülthoff, 1991; Bülthoff and Yuille, 1991; Clark and Yuille, 1990; Geiger and Yuille, 1991; Knill and Kersten, 1991; Yuille, Geiger, and Bülthoff, 1991; Freeman, 1992; Belhumeur and Mumford, 1992; see also Nakayama and Shimojo, 1992). It provides a powerful approach to understanding and modeling human perceptual capacities. But it has a well-known limitation. For real vision problems the collection of images that might be obtained is very large. (For instance, there are about $10^{15}$ possible true-color images of 1024 by 1024 pixel

resolution.) Therefore $P(I)$ and $P(I \mid S)$ are either 0 or near 0 for most images and the form of Bayes rule given above is either undefined or unstable. We cannot remove this problem by conditioning on large collections rather than on small collections of images, because the task we typically face in image understanding is to interpret a given single image or small collection. In special cases the instability problem can be avoided by the use of energy functionals and Gibbs distributions (Marroquin, Mitter, and Poggio, 1987; Poggio and Girosi, 1989). What we need, however, is a *general* form of Bayes rule that allows conditioning on events of probability zero and that requires no special assumptions.

Of what practical use is a general form of Bayes rule? Can instabilities that would arise from conditioning on sets of very small measure be avoided by use of this formulation? In many cases, yes. For example, a general (i.e., nondiscrete) formulation frequently permits the calculation of a local value as a limit of more global statistics. The very nature of these statistics is to wash out noise. By contrast, a single local measurement is highly sensitive to noise. Thus we would expect that to calculate the local number as a limit of these statistics provides a much more robust computational strategy. In addition, the existence of general nondiscrete formulations of a theory provides a reliable foundation for the design of discrete approximations; in fact there are many ways to discretize a nondiscrete system. The optimal comparison and selection of these various discretizations is only possible to the degree that the general formulation is well understood.

Observer theory gives a general form of Bayes rule. The result in the noise-free case is a "competence observer" and in the noisy case a "performance observer." The general form of Bayes rule requires two technical tools: regular conditional probability distributions and Radon-Nikodym derivatives. One goal of this paper is to provide an intuitive understanding of these tools and their practical application to problems in vision. After using these tools to derive a new general form of Bayes rule, we apply it to (1) the derivation of posterior distributions of practical use in vision, (2) the development of a theoretical framework for interrelating psychophysical experiments with computational theories, and (3) the analysis of hierarchical and weak coupling as theories of sensor fusion.

## 2. COMPETENCE: SOME BASIC IDEAS

Perception is a process of inference, or informed guessing. From phase and amplitude differences in the acoustic waveforms at the two ears we infer the location of a sound source. From disparities in the images at the two eyes we infer the three-dimensional (3D) shapes and locations of visual objects. From temporal variations in pressure at the finger pads we infer the 3D shapes and identities of haptic objects. As Helmholtz (1925) noted, such inferences are often fast and unconscious (Ittleson, 1960; Gregory, 1973; Rock, 1977; Marr, 1982). Indeed many of these inferences may never have been conscious at any point in our ontogeny or phylogeny. They may be instantiated in low-level neuronal networks, but they are inferences nonetheless.

Perceptual inferences are not deductively valid: the conclusions of perceptual inferences typically go well beyond their premises. The examples just adduced illustrate this point, as do others. Given gradients of shading in a two-dimensional (2D) image as premises, we reach conclusions about the 3D shapes of objects. Given the activity of chemoreceptors in the mouth and nose as premises, we reach conclusions about the safety and palatability of food. These conclusions are not logically dictated by the premises. Indeed, now and then a conclusion is wrong—not due to neurological malfunctions, but due to the nondeductive character of the inferences. The result is a perceptual illusion. We see a 3D shape in a stereogram, when in fact the stereogram is flat. We hear a train rushing by, when in fact we are listening to headphones. Such illusions are not evidence that sense data are incorrigible (Quinton, 1965), but they are powerful evidence that perceptual inferences are not deductively valid (Fodor, 1975).

Perceptual inferences are statistical, but with a unique feature. The inferences typical of statistical decision theory become trivial in the absence of noise. Not so perceptual inferences. Even in the absence of noise and quantization error, perceptual inferences have a nontrivial structure; we call this structure a "competence observer." Even if our corneas and lenses had no optical scatter, our retinas had infinite resolution, and our neurons had no computational limits, still the inferences underwriting stereovision would be nontrivial, and a stereogram could fool the visual system. The role of noise in perceptual inferences must, of course, be carefully studied. But the essential structure of perceptual inferences remains unaltered in the absence of noise. Describing this structure is the proper subject

of a competence theory of perception, and a prerequisite to the proper treatment of noise and other issues of performance.

Observer theory captures this structure in its formal definition of a competence observer. A major thesis of observer theory is that the same competence structure lies at the core of every perceptual capacity, and therefore that every perceptual capacity can be described, in its noise-free essence, as an instance of a competence observer. This "observer thesis" cannot be proven, since it states a relationship between a formalism (a competence observer) and an informal concept (a perceptual capacity); but it could, in principle, be disconfirmed by a counterexample, and is therefore an empirical hypothesis. In this regard it resembles the Church-Turing thesis, which states that every effective procedure can be described as some Turing machine. This thesis also cannot be proven, since it states a relationship between a formalism (the Turing machine) and an informal concept (an effective procedure); but it could, in principle, be disconfirmed by a counterexample, and is therefore an empirical hypothesis. Competence observers are proposed to play the same role for the description and analysis of perceptual capacities that Turing machines play for the description and analysis of effective procedures.

To illustrate the intuitions underlying the definition of a competence observer, we consider first the perception of 3D structure from image motion. Experiments by Braunstein and others (Braunstein, Hoffman, Shapiro, Andersen, and Bennett, 1987; Braunstein, Hoffman, and Pollick, 1990) indicate that subjects can reliably discriminate visual displays which depict rigid motions of points from displays which depict nonrigid motions, and that subjects can do this with as few as two views of four moving points. For displays depicting rigid motions, subjects report seeing specific 3D interpretations, and can reliably indicate the interpretations they see (Liter, Braunstein, and Hoffman, 1993). For displays depicting nonrigid motions, subjects generally report seeing no rigid 3D interpretations.

A competence theory of this capacity must account for (i) the ability to discriminate between rigid and nonrigid displays and (ii) the particular 3D interpretations that are reported for the rigid displays. As mentioned before, this account will be idealized, in the sense that it ignores noise and quantization error.

To account for (i) a competence theory must assign (one or more) rigid interpretations not to every display, but only to those displays that depict rigid motions. Most displays

do not depict rigid motion. In fact one must carefully program a display to make it depict a rigid motion. Therefore the competence theory must assign no interpretations to most displays. Put more formally, a competence theory for the problem of giving rigid interpretations to motion displays must make sure that the problem is almost surely ill-posed, and must not regularize the problem (Tichonov and Arsenin, 1977; Poggio, Torre, and Koch, 1985). To regularize the problem (in the sense of Tichonov) would be to require that *every* display have one interpretation, and this would eliminate any ability to discriminate rigid from nonrigid displays. By contrast, making the problem almost surely ill-posed, by requiring that almost all displays have no interpretations, restores the ability to discriminate rigid from nonrigid displays.

To account for (ii) a competence theory must assign one or more 3D interpretations to each display depicting a rigid motion. If subjects see just one 3D interpretation of some rigid display, then the theory should assign that interpretation to the display. If subjects see more than one interpretation, then a complete theory of competence should assign a probability measure supported on these interpretations, giving the appropriate relative frequencies or strengths of the interpretations. For displays in which subjects see just one interpretation, the techniques of (Tichonov style) regularization theory can sometimes be applied. For displays in which subjects see more than one interpretation, more general techniques are required.

## 3. COMPETENCE: AN EXAMPLE

We now consider one specific computational theory of structure from motion in order to motivate the definition of competence observer and to demonstrate useful computational techniques within the framework of competence observers. There are many accounts of structure from motion that could equally well serve this purpose (e.g., Ullman, 1979; Hoffman and Bennett, 1986), but we consider, for simplicity, the account presented by Bennett, Hoffman, Nicola, and Prakash (1989b). They prove the following "Two-View Theorem":[1]

[1] For ease of search, we number all definitions, theorems, corollaries, and displayed

**Theorem 1.** *(Two-View Theorem).* Two orthographic views of four noncoplanar points (i) almost surely[2] have no rigid interpretations, but (ii) have a one-parameter family of such interpretations if they have any.

This theorem assumes that the correspondence between points in the two views is known. We can also assume, since the projection is orthographic, that one of the points is taken to be the origin in both views, so that only three points have variable coordinates in each view.

The inference of structure from motion defined by this theorem is as follows. An elementary premise for the inference is two views of three points (plus the origin, which we will no longer mention). The set of all elementary premises is therefore the set of all two views of three points, which is $\Re^{12}$ (i.e., two views $\times$ three points per view $\times$ two coordinates per point). Here, and throughout the paper, we denote the set of all elementary premises by $D$. (Here the mnemonic is $D$ for "data," since premises are the data assumed for an inference.) According to the theorem, most of these premises, in fact almost all of these premises, have no rigid 3D interpretations. However there is a small subset of premises which have rigid 3D interpretations. These correspond to displays that depict rigid motion. This subset, which we denote $D_s$ and call the "special premises," has measure zero in $\Re^{12}$. Thus an element $d \in D_s$ represents two views of three points which have a rigid interpretation. A generically chosen element $d \in D$ represents two views of three points which, almost surely, have no rigid interpretation.

If $d \in D$ is an elementary premise consisting of two views of three points, i.e., of six points in the plane, then an elementary conclusion, $c$, compatible with $d$ is obtained by adding a depth coordinate to each of these six points, thereby creating a 3D interpretation for $d$. (Hereafter, in discussing this inference, we use "elementary conclusion" and "3D interpretation" interchangeably.) The depth coordinate of each of these six points in each

equations in one sequence. Figures are numbered in a separate sequence.

[2] We refer here to Lebesgue measure. "Lebesgue almost surely" means "up to measurable sets of Lebesgue measure zero." Intuitively, a measurable set $A$ is a subset of a space $X$ for which it is meaningful to assign a volume. Lebesgue measure is the standard way to assign volumes to measurable sets of Euclidean spaces. For more on Lebesgue measure see the Appendix, A0.

view can vary independently of the others, so there is, for each $d \in D$, a six-dimensional (6D) space of such conclusions $c$. Concretely, let $d$ be the set of points $\{(x_{ij}, y_{ij})\}$. (Here, and in the following, $i = 1, 2, 3$ indexes the points and $j = 1, 2$ indexes the views.) Then $c$ is a set of points $\{(x_{ij}, y_{ij}, z_{ij})\}$. We denote the 6D space of all such $c$ by the symbol $[d]$. According to the Two-View Theorem, for almost every $d \in D$ the 6D space $[d]$ contains no rigid 3D interpretations; however, for $d \in D_s$, the 6D space $[d]$ contains a one-dimensional subset of rigid 3D interpretations. This one-dimensional subset can be parametrized by the angle between the image plane and the axis of rotation associated to each rigid 3D interpretation (Bennett et al., 1989b). This angle is called the "slant" of the axis of rotation, and its value varies in the open interval $(0, \pi/2)$.

The set of all elementary conclusions for this inference is simply the union, over all $d$ in $D$, of the elementary conclusions $[d]$. The set of all elementary conclusions is therefore $\Re^{18}$ (i.e., $\Re^{12} \times \Re^6$). Here, and throughout the paper, we denote the set of all elementary conclusions by $C$. Almost all elements of $C$ represent nonrigid 3D interpretations. However there is a small subset of elements of $C$ which represent rigid 3D interpretations. This subset, which we denote $C_s$ and call the "special conclusions" or "special interpretations," has measure zero in $C$. $C_s$ corresponds to a bias or *a priori* assumption of the observer. For purposes of Bayesian analysis, the "prior probabilities" assumed by the observer are probability measures supported in $C_s$.

The collection of all elementary premises $D$ and the collection of all elementary conclusions $C$ are related by a "rendering function." In this case the rendering function is a map $\pi: C \to D$ given by $\{(x_{ij}, y_{ij}, z_{ij})\} \mapsto \{(x_{ij}, y_{ij})\}$. From this definition it follows that every 3D interpretation $c$ that is compatible with the image data $d$ actually gets mapped to $d$ by the rendering function $\pi$. We can write, therefore, that $\pi(c) = d$ for any $c$ in $[d]$. Equivalently, we can write $\pi^{-1}(d) = [d]$. This structure, plus some extra structure we will discuss shortly, is illustrated in Figure 1. A good way to check that one understands this figure and the discussion thus far is to convince oneself that $\pi(C_s) = D_s$.

So far we have an abstract framework that permits inferences. Now we must, so to speak, breathe life into this framework so that it will in fact perform inferences. We will do this by means of "kernels."[3] Intuitively, given two sets $X$ and $Y$, a kernel $N$ from $X$

[3] For a precise discussion of kernels see the Appendix, sections A1–A4.

**Figure 1.** *A competence observer: a canonical representation of a class of perceptual inferences.*

to $Y$ is a device which associates to each element of $X$ a probability measure on $Y$. If $x$ is an element of $X$ and $A$ is a subset of $Y$, then we use the notation $N(x, A)$ to denote the probability assigned to $A$ by the probability measure associated, by $N$, to $x$. Sometimes we use the notation $N(x, \cdot)$ to denote the whole probability measure associated by $N$ to $x$.

We model our perceptual inferences by a kernel, $\mathcal{I}$. We call $\mathcal{I}$ an "interpretation kernel" or "inference kernel" because it allows us to estimate conditional probabilities of scene interpretations given images. These probabilities are estimated assuming that there is no noise, i.e., assuming that the 3D interpretation $c$ is in the 6D space $[d]$ iff $c$ is invariably rendered as the two-view display $d$.

In fact, by definition of a kernel, $\mathcal{I}$ is an infinite collection of distinct probability measures, one probability measure for each two-view display $d$ that depicts a rigid motion. The probability of a set, $A$, of 3D interpretations given the two-view display $d$ is written $\mathcal{I}(d, A)$ or $\mathcal{I}(A \mid d)$. For purposes of Bayesian analysis, this probability is a posterior probability $Pr(A \mid d)$ under the assumption of no noise. The corresponding prior is a probability measure supported in $C_s$ (the rigid 3D interpretations). The likelihood function associated to any 3D interpretation $c$, rigid or not, is the conditional probability on images given the scene $c$. As such this likelihood is the indicator function $1_{\pi(c)}$, i.e., the function which assigns the value 1 to $\pi(c)$ and 0 to all other elements of $D$. The likelihood function takes this simple form because we are still in the noise-free competence case. We later

extend this discussion to the performance case.

A good way to check that one understands the discussion of $\mathcal{I}$ thus far is to convince oneself that $\mathcal{I}(c \mid d) = 0$ if $c$ is not in $[d] \cap C_s$.

The markovian kernel $\mathcal{I}$ expresses an infinite collection of posterior probabilities as a single linear operator (which maps measures on $D$ to measures on $C$—see Appendix, A3). This is a powerful tool, as we shall see, for interpreting noisy images. And it provides a new way to view perceptual inferences: as logic morphisms between spaces of probability measures, one space of measures representing premises and another representing conclusions (Bennett, Hoffman, and Murthy, 1993). We discuss kernels later.

The Two-View Theorem determines a class of markovian kernels $\mathcal{I}$, each of which defines a possible competence observer with the given $C$, $D$, $C_s$, $D_s$, and $\pi$. For each two-view display $d \in D_s$, i.e., for each two views of three points that has a rigid interpretation, the posterior $\mathcal{I}(\cdot \mid d)$ is a probability measure on the 6D space $[d]$. But this probability measure is not supported on the entire 6D space. Instead it is supported on a one-parameter subset of that space, viz., the one-parameter family of rigid 3D interpretations. The Two-View Theorem alone does not provide the information required to select among these posteriors in a principled manner. One candidate posterior gives equal weight to each of the rigid interpretations in the one-parameter family. (A uniform probability density is possible here since the parameter of the family lies in the interval $(0, \pi/2)$.)

We have now described a particular competence theory licensed by the Two-View Theorem. Psychophysical experiments suggest that this theory is psychologically plausible as a competence theory of the ability to discriminate between rigid and nonrigid displays (Braunstein et al., 1987, 1990). But psychophysical experiments also suggest that a uniform posterior is not psychologically plausible in a competence theory of the particular 3D interpretations that are reported for the rigid displays (Liter et al., 1993). Subjects do not give equal probability to all rigid interpretations in the one-parameter family compatible with a given two-view display. Instead, some rigid interpretations are almost always seen, and others are almost never seen. This suggests that the competence theory just discussed must be (1) abandoned or (2) refined to include constraints in addition to rigidity. The second option amounts to finding a more restrictive prior (Feldman, 1992, gives strong evidence for modal priors in perception and categorization). We consider this later.

## 4. COMPETENCE OBSERVER: THE DEFINITION

With a concrete example now in hand, we are ready to consider the abstract definition of a competence observer (see Figure 1).

**Definition 2.** A *competence observer* $\mathcal{O}$ is a collection $(C, D, C_s, D_s, \pi, \mathcal{I})$ with the following properties (see Figure 2).

(a) $C$ and $D$ have measurable structures[4] $(C, \mathcal{C})$ and $(D, \mathcal{D})$, respectively. The points of $C$ and $D$ are measurable.

(b) $C_s$ and $D_s$ are measurable subsets of $C$ and $D$ respectively, i.e., $C_s \in \mathcal{C}$ and $D_s \in \mathcal{D}$. (So $C_s$ and $D_s$ inherit measurable structures $\mathcal{C}_s$ and $\mathcal{D}_s$ respectively.)

(c) $\pi: C \to D$ is a measurable surjective function with $\pi(C_s) = D_s$.

(d) $\mathcal{I}$ is a markovian kernel on $D_s \times \mathcal{C}_s$ such that for each $d$, $\mathcal{I}(d, \cdot)$ is a probability measure supported in $[d] \cap C_s$. ($[d] = \pi^{-1}(d)$ is called the fiber of $\pi$ over $d$.)

$C$ is called the *configuration space* or *conclusion space* of the competence observer. We can think of it as the space of possible scene interpretations. $D$ is called the *data space.* We can think of it as the space of possible image data. $C_s$ is the set of *special configurations.* $C_s$ represents the bias of the observer, and measures on $C_s$ correspond to the priors of Bayesian analysis. $D_s$ is the set of *special premises.* It corresponds to those image data for which the observer is willing to assert nontrivial posterior distributions on possible scene interpretations. In the terminology of Jepson and Richards (1992), points of $D_s$ correspond to "key features." $\pi$ is the *perspective.* It corresponds to a rendering function. $\mathcal{I}$ is the *interpretation kernel.* It is a collection of posterior probabilities, represented as a linear operator.

A competence observer $\mathcal{O}$ "works" as follows. $\mathcal{O}$ receives a premise $d \in D$ and makes a decision: if $d$ is in $D_s$, then $\mathcal{O}$ gives interpretations in $[d] \cap C_s$ with (posterior) distribution $\mathcal{I}(d, \cdot)$; if $d$ is not in $D_s$, then $\mathcal{O}$ gives no interpretations, i.e., it remains inert. In other

[4] A measurable structure on a space $X$ is $X$ itself together with a collection, $\mathcal{X}$, of subsets of $X$ that includes $X$ and is closed under countable union and complement. The sets in $\mathcal{X}$ are called "events" and correspond intuitively to the possible outcomes of an experiment or observation. For more on measurable structures see Halmos (1950).

words, $\mathcal{O}$ gives interpretations only for the special premises (premises in $D_s$), and the interpretations given are only special (interpretations in $C_s$). It is in this sense that $C_s$ and $D_s$ represent a bias of a perceptual capacity.

There, in a nutshell, is the competence observer. The observer thesis asserts that each competence theory of a perceptual capacity can be written as an instance of a competence observer. If this thesis is correct, the competence observer provides a canonical form for the presentation of competence theories of perceptual capacities. Several examples of perceptual capacities presented in this canonical form can be found in Bennett et al. (1989a).

## 5. COMPETENCE: THE DECISION

The first thing a competence observer must do is make a decision: act or not. Given a premise $d$, corresponding perhaps to some image data, the competence observer must decide whether or not $d$ is in $D_s$. If it is, then the competence observer must act by assigning a posterior distribution $\mathcal{I}(d, \cdot)$. If it is not, then the competence observer does nothing.

We consider this decision first in the noise-free case, and then consider the effects of noise in the next section.

To do so, we return to the Two-View Theorem. In the course of their proof, Bennett et al. (1989b) derive a polynomial $f$ on $D$ which vanishes precisely on $D_s \subset D$: for $d \in D_s$, $f(d) = 0$; for $d \in D - D_s$, $f(d) \neq 0$. They did so as follows. Denote the 3D coordinates of feature point $i$ in view $j$ by

$$\vec{a}_{i,j} = (x_{i,j}, y_{i,j}, z_{i,j}), \tag{3}$$

where $i = 1, 2, 3$ and $j = 1, 2$. Denote the 2D coordinates of the image of feature point $i$ in view $j$ by

$$\vec{b}_{i,j} = (x_{i,j}, y_{i,j}), \tag{4}$$

where again $i = 1, 2, 3$ and $j = 1, 2$. Clearly $\vec{a}_{i,j} = (\vec{b}_{i,j}, z_{i,j})$. Thus from the images we know the $x_{i,j}$ and $y_{i,j}$ coordinates of the feature points, but we do not know the six $z_{i,j}$ coordinates. A necessary condition for points to undergo a rigid motion between the frames is given by the six quadratic polynomial equations

$$\vec{a}_{m,1} \cdot \vec{a}_{n,1} = \vec{a}_{m,2} \cdot \vec{a}_{n,2}, \qquad 1 \leq m, n \leq 3, \tag{5}$$

where $\cdot$ indicates the dot (scalar) product of vectors. These equations state that the lengths of the vectors $\vec{a}_{i,j}$ and the angles between them remain constant over the two instants of time. Although there are six quadratic equations in the six unknown $z_{i,j}$'s, the system is nevertheless inconsistent and therefore almost surely has no solutions. That is, for almost any choice of image data $\vec{b}_{i,j}$, Equations 5 have no solutions, and thus rigid interpretations are impossible. Only for a measure zero set of image data $\vec{b}_{i,j}$ do Equations 5 have any solutions for rigid interpretations. With some algebraic manipulation, one can

show that a necessary and sufficient condition on the image data for Equations 5 to have a solution is that

$$f = \det \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & h_{3,2} & h_{3,3} \end{bmatrix} = 0, \tag{6}$$

where $h_{m,n} = \vec{b}_{m,2} \cdot \vec{b}_{n,2} - \vec{b}_{m,1} \cdot \vec{b}_{n,1}$. From (6) one can show that $f$ is a homogeneous polynomial of sixth degree in the $x_{i,j}$'s and $y_{i,j}$'s, i.e., in the image data (Bennett et al., 1989b).

Having derived $f$, it is now trivial in the noise-free case to decide if the image data $d$ ($= \{\vec{b}_{i,j}\}$) is in $D_s$, i.e., to decide if the image data have a rigid interpretation. One simply plugs the $\vec{b}_{i,j}$ into the determinant in Equation 6. If the result is zero, then the image data have a rigid interpretation; otherwise they do not.

## 6. PERFORMANCE: THE DECISION

If there is noise the decision is not so simple: if the image data $d$ is near $D_s$, but not in it, we might still decide that $d$ has a rigid interpretation. This decision is often amenable to the well-known tools of statistical decision theory. We will not review these tools here (see, e.g., Savage, 1972). Instead we discuss a class of computational techniques particularly suited to the decisions typical of perceptual capacities.

To make such a decision in a principled way we need (i) a measure of the distance between $d$ and $D_s$ and (ii) the receiver operating characteristic as a function of this distance.

We consider both points, beginning with the measure of distance. For the example of detecting rigid motion, $D_s$ is not a linear or quadric surface to which euclidean distance is easy to compute. Instead $D_s$ is a sixth degree surface and distance from this surface is hard to compute. However for purposes of our decision we do not need a measure of distance from $D_s$ that is well defined for all points of $D$. Instead we need a measure that is well defined only nearby $D_s$ (in the euclidean sense), since it is only for points nearby $D_s$ that we have a difficult decision to make. For this purpose one candidate approach is to use the function $f$ of Equation 6. This function is zero on $D_s$ and its absolute value increases with increasing euclidean distance from $D_s$, at least for points nearby $D_s$. Here

$f$ will have to be normalized for scale: since $f$ is homogeneous of sixth degree, its value nearby a point $d \in D_s$ grows as the fifth power of the size of the corresponding rigid object.

Thus the value of $f$, suitably normalized, provides a measure of distance from $D_s$. But this measure of distance will aid our decision only if the receiver operating characteristic (ROC) based on this measure is sufficiently sensitive (Green and Swets, 1966). Recall that the ROC is a vector-valued function. For each distance $\delta$ in the domain of the function, there are two corresponding values in its range. These two values are the probability of hits and the probability of false alarms that obtain if $\delta$ is used as the decision criterion in the following manner: if $f(d) \leq \delta$ decide that $d$ has a rigid interpretation; otherwise decide that $d$ has no rigid interpretation. Typically an ROC is displayed as a graph of hit probability versus false alarm probability, with the range variable $\delta$ left implicit. We do so in this paper.

One could in principle derive the ROC analytically given the function $f$ and a model of the noise, but in practice this is difficult. Instead we ran Monte Carlo simulations as follows. We randomly generated 10,000 nonrigid structures, projected the 3D coordinates of each structure onto two randomly chosen image planes, and computed the 10,000 resulting values of $f$. We then randomly generated 30,000 different rigid 3D structures (each consisting of three points plus the origin), projected each structure onto two randomly chosen image planes. For 10,000 of these, we randomly perturbed the coordinates of the projected points with 0.05 percent gaussian noise.[5] For another 10,000 we perturbed the coordinates with 1.25 percent gaussian noise. The remaining 10,000 we perturbed with 5 percent gaussian noise. We then computed the 30,000 resulting values of $f$. Plots of the results as standard ROC curves, shown in Figure 2, demonstrate excellent detection properties even with 5 percent noise. In this case it is still possible to pick a decision criterion for which the probability of hits is better than 0.95 and the probability of false alarms is less than 0.03.

[5] In our Monte Carlo simulations the $x$ and $y$ coordinates of points in the 10,000 nonrigid structures were uniformly distributed within a range of $\pm 10$, so that the expected absolute value of each coordinate was 5. All coordinates in the 30,000 rigid structures were also restricted to a range of $\pm 10$. Thus the phrase "5 percent gaussian noise" means gaussian noise with a standard deviation of 0.25.

**Figure 2.** *Receiver operating characteristic (ROC) for the function $f$. Three ROC curves are shown. In order from uppermost to lowest, the ROCs are shown for 0.05%, 1.25%, and 5% gaussian noise in the image coordinates.*

This example demonstrates a general approach to the decision problem: (i) Construct a competence observer whose set $D_s$ has measure zero in $D$; (ii) Find a function which vanishes precisely on $D_s$; (iii) Analytically, or via Monte Carlo simulations, determine a critical value of this function which gives acceptable hit and false alarm rates for a reasonable model of noise.[6]

Note that the function $f$ of Equation 6 can be used to recognize 3D structures. Pick any rigid 3D configuration of points $R$ and any generic view of that configuration. Insert the coordinates from that view into the first frame variables (the $\vec{b}_{i1}$'s). The result is a polynomial $f_R$ of lower degree in just the second frame variables (the $\vec{b}_{i2}$'s). Now to determine if a second image depicts a different view of $R$, insert its coordinates into $f_R$ and compute the resulting value. If the value is small enough (based on the above Monte Carlo results) then decide that the second image depicts $R$; otherwise decide it does not.

[6] More examples of this approach are contained in Bennett, Hoffman, and Prakash (1993a,b). These examples include affine motions and weak perspective projection. Another approach to detecting rigid objects is given by Thompson, Lechleider, and Stuck (1993). Their approach, however, requires an extra step: for each new image to be analyzed one must first estimate a four-dimensional vector $r$ which minimizes the median residual error in a set of linear equations.

This approach allows one to recognize any view of a (transparent) 3D structure after being given only a single view (cf. Basri and Ullman, 1991; Poggio, 1990). For an opaque object, one can recognize any view within one cell of its aspect graph after being given only one view from within that cell.

One can draw a general conclusion from this section. To be able to deal with noise, the definition of competence observer must be augmented with a function $\alpha: D \to [0,1]$ which, intuitively, is related to the degree of confidence that each premise should be assigned a special interpretation. Intuitively, we can take $\alpha(d)$ to be the hit rate at the point of the ROC curve corresponding to the parameter value $\delta = f(d)$. This motivates the following definition of performance observer (Bennett, Hoffman, Kakarala, 1993).[7]

**Definition 7.** A *performance observer* is a competence observer $(C, D, C_s, D_s, \pi, \mathcal{I})$ together with a function $\alpha: D \to [0,1]$. We call $\alpha$ the *confidence function* of the performance observer.

## 7. COMPETENCE: THE POSTERIOR IN THE DISCRETE CASE

In this section we derive a form of the Bayes posterior in the discrete case that will lead naturally to a formulation of the more general case.

Given an image $d \in D$ we want to assign probabilities to possible scene interpretations $c \in C$. Thus we want to compute the conditional probability $Pr(c \mid d)$. For simplicity, and to fix ideas, we consider first the discrete case. In this case we can use Bayes rule:

$$Pr(c \mid d) = \frac{Pr(d \mid c)\mu(c)}{\lambda(d)}. \tag{8}$$

Here $\mu$ denotes a prior measure on scenes, $Pr(d \mid c)$ is the likelihood function, and $\lambda$ is a measure on images which expresses the probability that the image $d$ will be acquired, assuming $\mu$ is the actual measure on scenes. This expression is well-defined only if $\lambda(d) > 0$.

[7] Later we give a precise definition of $\alpha$ (Definition 51) and briefly discuss its relationship to standard Bayesian decision methods. Definition 7 is different in form but equivalent in content to the definition given by Bennett, Hoffman, and Kakarala (1993).

(We discuss shortly how to formulate the problem in the more realistic nondiscrete situation in which $\lambda(d) = 0$.) In the noise-free case the likelihood function $Pr(d \mid c)$ takes, as we have seen, a particularly simple form. It assigns the value 1 to $\pi(c)$, i.e., to the image that would be rendered if the scene were $c$, and it assigns the value 0 to all other images. We can write this as

$$Pr(d \mid c) = 1_{\pi(c)}(d), \tag{9}$$

where $1_{\pi(c)}(d)$ denotes the indicator function of $\pi(c)$. This, together with our assumption on $\lambda$, implies that

$$\lambda(d) = \mu\pi_*(d). \tag{10}$$

Here the measure $\mu\pi_*$ is, by definition, the distribution of $\pi$ with respect to $\mu$ defined by $\mu\pi_*(B) = \mu(\pi^{-1}(B))$, where $B$ is any event in $\mathcal{D}$.

For example, for the competence observer defined by the Two-View Theorem, if $c = \{(x_{ij}, y_{ij}, z_{ij})\}$, where $i = 1, 2, 3$ and $j = 1, 2$, then $Pr(d \mid c)$ is 1 for $d = \{(x_{ij}, y_{ij})\}$, and zero otherwise.

Substituting (9) into (8) gives

$$Pr(c \mid d) = \frac{1_{\pi(c)}(d)\mu(c)}{\lambda(d)}, \tag{11}$$

but this numerator is equal to $\mu(c)$ if $\pi(c) = d$ and is 0 otherwise, so we can write

$$= \frac{\mu(c \cap \pi^{-1}(d))}{\lambda(d)},$$

which by (10) is

$$= \frac{\mu(c \cap \pi^{-1}(d))}{\mu(\pi^{-1}(d))},$$

or

$$= \overline{\mu|_{[d]}}(c),$$

where $\overline{\mu|_{[d]}}$ indicates the normalized restriction of $\mu$ to $\pi^{-1}(d)$, i.e., $\overline{\mu|_{[d]}}(c) = \mu(c \cap \pi^{-1}(d))/\mu(\pi^{-1}(d))$. Thus the posterior probability $Pr(c \mid d)$ is the normalized restriction of the prior probability $\mu$ to the set $[d]$. Using the observer language of interpretation

kernels we can write

$$Pr(c \mid d) = \overline{\mu|_{[d]}}(c) = \mathcal{I}(d, c). \tag{12}$$

If we replace the specific scene interpretation $c$ with a measurable set $A$ of scene interpretations, this posterior probability becomes

$$Pr(A \mid d) = \sum_{c \in A} Pr(c \mid d) = \mathcal{I}(d, A). \tag{13}$$

Thus for competence observers the posterior distributions contained in the interpretation kernel $\mathcal{I}$ are normalized restrictions of a prior measure $\mu$ on $C_s$. The restriction sets are fibers $[d]$ of the rendering function $\pi$. For a competence observer compatible with the Two-View Theorem, $\mathcal{I}(d, \cdot)$ may be taken to be, for example, a uniform probability measure on the set $[d] \cap C_s$; this measure is uniform with respect to the parametrization of $[d] \cap C_s$ by the slant of the axis of rotation.

Note that in this derivation the prior measure $\mu$ might be improper, i.e., it might not be a probability measure. The derivation requires, however, that $0 < \mu([d]) < \infty$ for each $d \in D$.

## 8. COMPETENCE: THE POSTERIOR IN THE CONTINUOUS CASE

To generalize the Bayesian results to the nondiscrete case, we must use *regular conditional probability distributions* (Parthasarathy, 1968; Bennett et al., 1989a). Regular conditional probability distributions (rcpd's) are defined precisely in the Appendix, section A5. However here we give a concrete example to illustrate the definition. Let $\mu$ be the usual joint gaussian on $\Re^2$ and let $\pi$ be the projection from $\Re^2$ onto the $x$-axis. Then the rcpd of $\mu$ with respect to $\pi$ is a kernel $\eta(x, \cdot)$ which assigns to each point $x$ of the $x$-axis a gaussian measure on the vertical line in $\Re^2$ through that point $x$.

The intuitions are as follows. Suppose that we are modeling a competence observer $O = (C, D, C_s, D_s, \pi, \mathcal{I})$ and that our prior measure on scene interpretations is a probability measure $\mu$ on $C_s$. If we are given the image data $d$, our posterior probability distribution, according to Equation 12, is the normalized restriction of $\mu$ to the set $\pi^{-1}(d) = [d]$. That is, the posterior distribution assigns to a set $A$ of scene interpretations the probability $\mu(A \cap$

$[d])/\mu([d])$. But what if the probability (i.e., $\mu$ measure) of the set $[d]$ is zero, as it most often is in the continuous case? Then this formulation of the posterior distribution is undefined. Unfortunately, this is precisely the case we need most. In practical situations we are given an image $d$ and must come up with a posterior probability on scene interpretations. This image $d$, however, is usually just one of a large collection of possible images that we might have been given, and therefore its probability, and the probability of $[d]$, is zero or near zero. To get the desired posterior distributions in this case, what we need is a way to get normalized restrictions of $\mu$ to sets $[d]$ of probability zero. Or, to put it slightly differently, we need to be able to condition on sets of probability zero.

This is precisely the power of rcpd's—to condition on sets of probability zero and still get a well-defined probability measure as a result. An rcpd is a markovian kernel. In the case at hand it is denoted $m_\pi^\mu$, where the superscript indicates the prior probability and the subscript indicates the rendering function.

The application of rcpd's to competence observers is immediate. The interpretation kernel $\mathcal{I}$ just is the rcpd $m_\pi^\mu$. The probability measures $\mathcal{I}(d,\cdot) = m_\pi^\mu(d,\cdot)$ then correspond to the posterior distributions.

## 9. PERFORMANCE: THE UNDERLYING PROBABILITY SPACE

Now we consider measurement noise. A general model for this noise is a markovian kernel

$$N\colon C \times \mathcal{D} \to [0,1], \tag{14}$$

where $\mathcal{D}$ denotes the collection of all measurable subsets[8] of the space $D$ of image data. We interpret this kernel as follows. We fix a scene $c \in C$ and let $B \in \mathcal{D}$ denote a set of images. Then $N(c,B)$ is the probability that the rendered image $d$ is in the set $B$ given that the scene is $c$. (In the discrete case it is sufficient to know the probabilities when $B = \{d\}$ for all images $d \in D$; for simplicity we denote this $N(c,d)$ instead of $N(c,\{d\})$. It is the probability that $d$ is the image rendered given that the scene is $c$.) So interpreted,

[8] In the discrete case the $\sigma$-algebra is just the collection of all subsets. In the general case the $\sigma$-algebra is a nontrivial subcollection.

the kernel $N$ is a collection of likelihoods and we would like to write

$$Pr(d \mid c) = N(c, d), \tag{15}$$

where $Pr$ is some appropriate probability on perceptual events. In order to do this we must be more explicit and rigorous about the probabilistic setting.

Since each perceptual event is a pair consisting of a scene from $C$ and an image from $D$, the underlying space for perceptual events is $C \times D$. Our fundamental probability space then is a space $(C \times D, \mathcal{C} \otimes \mathcal{D}, Pr)$.[9] If we let $A \in \mathcal{C}$ denote a measurable collection of scene interpretations, and $B \in \mathcal{D}$ a measurable collection of images, then $Pr(A \times B)$ denotes the probability of the perceptual event in which the true scene is in the set $A$ and the rendered image is in the set $B$. We assume we are given a prior probability $\mu$ on $C_s \subset C$ and a conditional probability $N: C \times \mathcal{D} \to [0, 1]$ expressed as a markovian kernel. This conditional probability $N$ reflects the effects of noise as described above and is the "likelihood function" of Bayesian analysis. It follows by definition of conditional probability that in the discrete case $Pr$ assigns to a typical set $A \times B \in \mathcal{C} \otimes \mathcal{D}$ the probability

$$Pr(A \times B) = \sum_{c \in A, d \in B} Pr(c, d) = \sum_{c \in A, d \in B} \mu(c) N(c, d) \tag{16}$$

and correspondingly in the continuous case

$$Pr(A \times B) = \int_A \mu(\mathrm{d}c) \int_B N(c, \mathrm{d}d). \tag{17}$$

Thus we may say that in the discrete case

$$Pr(c, d) = \mu(c) N(c, d), \tag{18}$$

while in the continuous case

$$Pr(\mathrm{d}c, \mathrm{d}d) = \mu(\mathrm{d}c) N(c, \mathrm{d}d). \tag{19}$$

In both cases, we may use the language of kernel multiplication (Appendix, A3) to observe that

$$Pr(A \times B) = \mu(1_A N)(B). \tag{20}$$

[9] The product $\sigma$-algebra $\mathcal{C} \otimes \mathcal{D}$ is the $\sigma$-algebra generated by all subsets of the form $A \times B$, $A \in \mathcal{C}$, $B \in \mathcal{D}$. In the discrete case the product algebra is (trivially) the collection of subsets of $C \times D$, but in the continuous case $\mathcal{C} \otimes \mathcal{D}$ is nontrivially related to $\mathcal{C}$ and $\mathcal{D}$. See, e.g., Halmos (1950).

It follows that the marginal of $Pr$ on $C$ is the prior measure $\mu$ on scenes and the marginal of $Pr$ on $D$ is the measure $\mu N$ on images, i.e., $(\mu N)(B) = \int_C \mu(\mathrm{d}c)N(c, B)$. As in the noise-free discrete case, we have assumed that the prior probability on scenes is described by a measure $\mu$ on $C_s \subset C$. In the presence of noise, however, we have obtained a different measure $\lambda$ on images. Recall from (10) and (11) that in the noise-free case this measure is $\lambda(d) = \mu(\pi^{-1}(d)) = \mu\pi_*(d)$. That is, $\lambda$ is simply obtained by "pushing down" by $\pi$ onto $D$ the prior measure $\mu$ on $C_s \subset C$. But if there is noise described by a kernel $N$, then this push down, as we have just seen, must instead be of the form

$$\lambda(d) = (\mu N)(d) = \sum_{c \in C} \mu(c)N(c, d) \tag{21}$$

in the discrete case, and of the form

$$\lambda(\mathrm{d}d) = (\mu N)(\mathrm{d}d) = \int_C \mu(\mathrm{d}c)N(c, \mathrm{d}d) \tag{22}$$

in the continuous case. If $N(c, \cdot)$ is the Dirac measure at $\pi(c)$, i.e., if there is no noise, then Equations 21 and 22 reduce to Equation 10.

In the discrete case, we can now make sense of Equation 15 in terms of the above postulated probability measure $Pr$: the correct expression for $Pr(d \mid c)$ is, with our present notation, $Pr(C \times \{d\} \mid \{c\} \times D)$.

## 10. PERFORMANCE: DERIVATION OF THE POSTERIOR (STANDARD BAYESIAN ANALYSIS)

We now want the posterior probability, i.e., the conditional probability that the scene is in a set $A$ given that the rendered image is $d$. In the general case there is a question about the meaning of the posterior probability since the probability of $\{d\}$ might be zero. We will deal with the discrete case first and then use our results to motivate the derivation in the general case.

We now give the discrete posterior distribution in the presence of noise. The result (Equation 24 below) is a Bayes rule for the perceptual situation. If $A \in \mathcal{C}$ is a measurable collection of scene interpretations, and $d$ is an image, the posterior probability given that image is the conditional probability $Pr(A \times D \mid C \times \{d\})$. From now on we shall write

this as, simply, $P(A \mid d)$. In what follows we will often abuse notation in this way, e.g., by writing just $A$ instead of $A \times D$, $B$ instead of $C \times B$ etc. In particular, with this notation $Pr(A)$ is the marginal $\mu(A)$, $Pr(B)$ is the marginal $\mu N(B)$, and $P(A \mid d)$ is the posterior we seek. (The exact meaning should be clear from the context, but we caution the reader that the meaning in terms of the underlying probability space be always kept in mind so as to avoid mistakes in computation.)

**Theorem 23.** *(Discrete Posteriors).* Let the measure $\mu$ be the prior distribution on scene interpretations and let the markovian kernel $N$ be the likelihood function. Then the posterior probability that the true scene is in $A \in \mathcal{C}$ given that the rendered image is $d$, is given by

$$P(A \mid d) \quad \stackrel{\text{def}}{=} \quad Pr(A \times D \mid C \times \{d\}) = \frac{(\mu(1_A N))(d)}{(\mu N)(d)}, \tag{24}$$

whenever $\mu N(d) \neq 0$ and $P(A \mid d) = 0$ otherwise. The posterior satisfies, for any $B \in \mathcal{D}$,

$$\sum_{d \in B} \mu N(d) P(A \mid d) = Pr(A \times B). \tag{25}$$

Moreover, the posterior probability is related to a regular conditional probability distribution as follows: Let $q\colon C \times D \to D$ be the projection mapping $q(c, d) = d$. Then the rcpd $m_q^{Pr}$ exists and for any $A \in \mathcal{C}$,

$$P(A \mid d) = m_q^{Pr}(d, A \times \{d\}) = m_q^{Pr}(d, A \times D), \tag{26}$$

for $\mu N$-almost all $d$.

**Proof.** Appendix, A9.

Equation 24 is Bayes rule expressed in kernel notation, and relates the posterior to the prior $\mu$ and the noisy likelihood function $N$. This is a formulation in terms of kernels, equivalent to the usual form. In fact, (24) when $A$ is the singleton $\{c\}$ reduces to the familiar form, i.e., $P(c \mid d) = N(d \mid c)\mu(c) / \sum_c N(d \mid c)\mu(c)$. Equation 24 itself expands out to $P(A \mid d) = \sum_{c \in A} N(d \mid c)\mu(c) / \sum_c N(d \mid c)\mu(c)$. Thus the likelihood function in Bayesian analysis is a Markovian kernel. Equations 24–26 will allow us to generalize from the discrete to the continuous case. One checks that the last equality in Equation 24 shows

that $P(A \mid d)$ is a *markovian kernel* on $D \times \mathcal{C}_s$, since $\mu$ is supported in $C_s$. Of course, $P$ naturally extends to a submarkovian kernel on $D \times \mathcal{C}$.

We are now ready for the continuous case. Given the issue of zero probabilities for $d$, how should we define $P(A \mid d)$? The answer lies in generalizing a property in Theorem 23: In the discrete case, $P(A \mid d)$ is the value of a markovian kernel on $D \times \mathcal{C}_s$ (with the usual order of arguments of kernels reversed[10]), satisfying Equation 25. This motivates

**Definition 27.** *(General Posterior).* The general posterior $P(A \mid d)$ is any markovian kernel on $D \times \mathcal{C}_s$, satisfying

$$Pr(A \times B) = \int_{d \in B} \mu N(\mathrm{d}d) P(A \mid d) = \int_{c \in A, d \in B} \mu N(\mathrm{d}d) P(\mathrm{d}c \mid d). \tag{28}$$

From this and (20) it follows that $P$ is a Bayesian posterior for the likelihood $N$ given $\mu$ iff

$$(\mu 1_A N)(B) = ((\mu N) 1_B P)(A) \quad (= Pr(A \times B)). \tag{29}$$

It is equivalent to say (by Lemma A39 of the Appendix) the following: For all bounded measurable functions $f$ on $C$ and $g$ on $D$,

$$\mu f N g = (\mu N) g P f. \tag{30}$$

**Notes on Definition 27:**

(i) The notion of Bayes posterior as a relationship between kernels is only meaningful given the prior $\mu$.

(ii) (29) constrains $N$ and $P$ only a.e. $\mu$ and $\mu N$ respectively. This means that we can modify $P(d, \cdot)$ arbitrarily for $d$ in any set of $\mu N$ measure zero without affecting (29). Thus the notion of Bayes posterior is defined only a.e. $\mu$ and $\mu N$ respectively, and this is why, in Definition 27, we said "a" posterior rather than "the" posterior.

[10] Note that a kernel on $D \times \mathcal{C}_s$ is normally written with $d \in D$ as its first argument and $A \in \mathcal{C}_s$ at its second; in keeping with traditional Bayesian notation, however, we are writing $A$ first and $d$ second in $P(A \mid d)$.

By Definition 27, to say that $N$ is the Bayesian posterior for the likelihood $P$ given $\mu N$ means that the last equation holds, but with the $\mu$ on the left replaced by $\mu NP$. Thus we note that Definition 27 is symmetric provided $\mu = \mu NP$. We will prove in Theorem A below that this always holds if $P$ is the posterior of $N$ with respect to $\mu$, i.e., that *the relationship of Bayesian posterior is symmetric.* Thus if $P$ is the posterior of $N$ with respect to $\mu$ then $N$ is the posterior of $P$ with respect to $\mu N$.

**Proposition 31.** Let $\pi\colon X \to Y$ be a measurable mapping and $\mu$ a measure on $X$. Then the kernels $\mathcal{I}$ from $Y$ to $X$ and $\pi_*$ from $X$ to $Y$ are Bayesian posteriors of each other iff $\mathcal{I}$ is the rcpd of $\mu$ with respect to $\pi$ (see Appendix A5 for definition of rcpd, and Appendix A4 for how $\pi_*$ can be viewed as a kernel). In other words, in the competence case $\mathcal{I}$ and $\pi_*$ are Bayes posteriors of each other.

**Proof.** Appendix A36.

Do the markovian kernels which are the posteriors of Definition 27 exist in general in the continuous case? Indeed so; the appropriate generalization of (24) is exhibited in (35) below. To see this we need *Radon-Nikodym derivatives,* which we now briefly review.[11]

**Definition 32.** *(Radon-Nikodym Derivative).* If $\nu$ and $\rho$ are measures on some measurable space $(W, \mathcal{W})$ and if, for all $D \in \mathcal{W}$, $\nu(D) = 0$ implies $\rho(D) = 0$, then $\rho$ is said to be *absolutely continuous* with respect to $\nu$, and we write $\rho << \nu$. If $\rho << \nu$, the Radon-Nikodym Theorem states that there exists a measurable function $f$ such that

$$\rho(B) = \int_B f \mathrm{d}\nu, \qquad \forall B \in \mathcal{W}, \tag{33}$$

as long as $\nu$ is a $\sigma$-finite measure and $\rho$ is finite. Any function $g$ which differs from $f$ only on a set of $\nu$ measure zero also satisfies (33), and is also called a Radon-Nikodym derivative of $\rho$ with respect to $\nu$, and denoted $d\rho/\mathrm{d}\nu$. Thus the Radon-Nikodym derivative with respect to $\nu$ is defined only up to sets of $\nu$-measure zero. The Radon-Nikodym derivative $f$ is sometimes called the *density of $\rho$ with respect to $\nu$.*

[11] For a detailed discussion of the relationship of Radon-Nikodym derivatives to conditional expectation see, e.g., Chung (1974). See also Grenander (1981) for an application of Radon-Nikodym derivatives to abstract inference.

The perceptual Bayes rule in the general case now follows:

**Theorem 34.** *(General Posterior).* Let the measure $\mu$ be the prior distribution on scene interpretations and let the markovian kernel $N$ be the likelihood function. Then the posterior probability that the true scene is in $A \in \mathcal{C}$ given that the rendered image is $d$ is given by

$$P(A \mid d) = \frac{\mathrm{d}(\mu(1_A N))}{\mathrm{d}(\mu N)}(d), \qquad \text{a.e. } d\ (\mu N). \tag{35}$$

In fact, by definition of Radon-Nikodym derivative (Definition 32), this means that for any $B \in \mathcal{D}$,

$$\int_{d \in B} \mu N(\mathrm{d}d) P(A \mid d) = P(A \times B). \tag{36}$$

Let $q\colon C \times D \to D$ be the projection mapping $q(c,d) = d$. Then we can write $P(A \mid d)$ as a regular conditional probability distribution as follows:

$$P(A \mid d) \quad \stackrel{\text{def}}{=} \quad Pr(A \times \{d\} \mid C \times \{d\}) = m_q^{Pr}(d, A \times D). \tag{37}$$

**Proof.** Appendix, A22.

An important special case of this theorem is when the prior and likelihood have densities. This means that there are measures $\lambda$ on $C$ and $\rho$ on $D$, with $\mu(\mathrm{d}c) = f(c)\lambda(\mathrm{d}c)$ and $N(c, \mathrm{d}d) = n(c,d)\rho(\mathrm{d}d)$; we say that $f(c)$ is the "density" of $\mu$ (with respect to $\lambda$) and $n(c,d)$ is the density of $N(c, \mathrm{d}d)$ (with respect to $\rho$). (The case most familiar in the perception literature is where $C$ and $D$ are $\Re^n$ and $\Re^m$ for some $n$ and $m$ and $\lambda$ and $\rho$ are the ususal euclidean volumes.) Then the posterior kernel $P(\mathrm{d}c \mid d)$ also has a density, say $p(c,d)$, with respect to $\lambda$, i.e., $P(\mathrm{d}c \mid d) = p(c,d)\lambda(\mathrm{d}d)$. In terms of these densities Bayes Rule says:

$$p(c,d) = \frac{n(c,d)f(c)}{Z(d)}, \tag{38}$$

where $Z(d) = \int n(c,d)f(c)\lambda(\mathrm{d}c)$ is a normalization factor. (Again, the discrete case (25) follows upon requiring that $\lambda$ be counting measure on a finite set.)

Equation 35 is of practical importance in computing posteriors in the perceptual situation, as we shall see in the next section. This form of Bayes rule is general. It applies to arbitrary prior probability measures $\mu$. It applies to arbitrary forms of noise

$N$; in particular, it is not restricted to an assumption of gaussian noise. It allows some improper, i.e., infinite, prior measures.[12] And it allows one to condition on images that have probability zero. Thus this form of Bayes rule can be applied to many problems in vision in which a posterior distribution is sought. In section 12 we use it to compute a special class of posteriors that often arise in vision research.

Theorem 34 implies that $P(A \mid d) = P(A \cap C_s \mid d)$. One can see this directly from (35) by noting that $\mu$ is supported on $C_s$, so that $\mu(1_A N) = \mu(1_{A \cap C_s} N)$.

If the measures involved are all discrete, Theorem 34 specializes to Theorem 23. For in that instance, integrals become sums and Radon-Nikodym derivatives become simply ratios ((25) is the analogue of (36) in the discrete case at hand). Moreover, Theorem 34 also confirms the noise-free or competence case. Here, $N(c, \mathrm{d}d) = \epsilon_{\pi(c)}(\mathrm{d}d)$, i.e., the Dirac kernel at $\pi(c)$. In the proof of Proposition 31 we show that then $\mu N = \pi_* \mu$ and that $\mu 1_A N(B) = \int_B \mu N(\mathrm{d}d) \mathcal{I}(d, A)$. Thus the Radon-Nikodym derivative $P(A \mid d)$ is just the interpretation kernel $\mathcal{I}(d, A)$, as expected.

By Definition 27, $P(\mathrm{d}c \mid d)$ is a markovian kernel. It will be of use later to evaluate integrals of the form $\int P(\mathrm{d}c \mid d) f(c)$, where $f$ is a bounded, measurable function on $C$. Equation 35 tells us what this integral is for $f = 1_A$. We have

**Corollary 39.** For any bounded, measurable function $f$ on $C$,

$$\int_{c \in C} P(\mathrm{d}c \mid d) f(c) = \frac{\mathrm{d}(\mu(fN))}{\mathrm{d}(\mu N)}(d), \qquad \text{a.e. } d\ (\mu N).$$

**Proof.** Appendix, A24.

We reiterate a key consequence of the foregoing derivations: *In the general case, Bayesian posteriors and likelihoods are kernels.* This is useful for several reasons. First,

[12] Improper priors are an issue, e.g., when the measure $\mu N$ assigns the value infinity to some singleton. The general form of Bayes rule (35) will accept any (possibly improper) prior measure $\mu$ as long as $\mu N$ is $\sigma$-finite, i.e., as long as there is a measurable partition $\{A_n\}$ of $D$ such that $N(c, A_n)$ is integrable with respect to $\mu$ for all $n$. Note that in this case (35) is still defined as a markovian kernel. For more on improper priors, see Hartigan (1983).

kernels are well understood mathematical objects whose properties have been extensively investigated. Second, kernels provide a convenient computational tool in the general (i.e., nondiscrete) case. In particular they allow us to compute probabilities of scenes as expectations of functions on the space of images; they also allow us to compute probabilities of images as expectations of functions on the space of scenes.

With the foregoing definitions and characterization of the Bayesian posterior we can prove several important results on the general behavior of these posteriors. For this purpose we will introduce terminology and notation which is simplifying and suggestive.

**Definition 40.** Let $X$ and $Y$ be measurable spaces. Let $\mu$ be a probability measure on $X$. Let $H$ from $X$ to $Y$ be a markovian kernel. We will denote by $H^\dagger_\mu$ (or just by $H^\dagger$ when there is no confusion about $\mu$) the Bayesian posterior of $H$ for the prior probability $\mu$. We will also refer to $H^\dagger_\mu$ as the *Bayes adjoint* of $H$ for $\mu$, or just the *adjoint* for short when $\mu$ is unambiguous. (We remind the reader that the Bayes adjoint is well defined only a.e. $\mu H$ in the sense of note (ii) after Definition 27.) According to (30), $H^\dagger$ is defined by the relation

$$\mu f H g = (\mu H) g H^\dagger f,$$

for all bounded measurable functions $f$ on $X$ and $g$ on $Y$. It is very suggestive to use the following notation. In general, on a measure space $X$ with measure $\mu$ and functions $f$ and $h$, we denote $< f, h >_\mu = \int_X f(x)h(x)\mathrm{d}\mu(x)$. The definition of the Bayes adjoint $H^\dagger$ of $H$ with respect to $\mu$ is

$$< f, Hg >_\mu = < H^\dagger f, g >_{\mu H}, \tag{41}$$

which is a restatement of (30).

**Notes on Definition 40**

(i) We recall that the existence of $H^\dagger$ is guaranteed by Theorem 34

(ii) The adjointness terminology is further justified below by Theorems 42, 43, and 44.

(iii) If $f \in L^2(X, \mu)$ and $g \in L^2(Y, \mu H)$, with $\|f\| = \|g\| = 1$, it is natural to view $f$ and $g$ as *states* on $X$ and $Y$. The quantity $< f, Hg >_\mu = < H^\dagger f, g >_{\mu H}$ may then be interpreted as a measure of the compatibility of the states $f$ and $g$, or as the probability

of simultaneous occurrence of the two states. A compatibility of numerical value 1 (perfect compatibility) occurs when $f = Hg$, or equivalently, when $g = H^\dagger f$.

**Theorem 42.** Given $X$, $\mu$ on $X$, and a kernel $H$ from $X$ to $Y$,

$$\mu H H^\dagger = \mu.$$

**Proof. Appendix A37.**

Intuitively, $H^\dagger$ reverses the effect of $H$ on the prior $\mu$.

**Theorem 43.** With $X$, $\mu$, and $H$ as above,

$$(H_\mu^\dagger)_{\mu H}^\dagger = H.$$

**Proof.** We refer to (41) as the definition of the Bayes posterior. Applying this definition first to $H$ and then to $H_\mu^\dagger$, we see that

$$< f, Hg >_\mu = < H^\dagger f, g >_{\mu H} = < f, (H_\mu^\dagger)_{\mu H}^\dagger g >_{\mu H H^\dagger} .$$

The right-hand side equals $< f, (H_\mu^\dagger)_{\mu H}^\dagger g >_\mu$ by theorem A. Comparing this with the left-hand side we are done. ∎

The following theorem is very useful for the study of composite Bayesian inferences.

$$\begin{array}{ccccccc} & \mu & & & \nu=\mu H & & \lambda=\nu K \\ X & & \underset{H_\mu^\dagger}{\overset{H}{\rightleftarrows}} & Y & & \underset{K_\nu^\dagger}{\overset{K}{\rightleftarrows}} & Z \end{array}$$

$$(HK)_\mu^\dagger$$

**Theorem 44.** Let $X$, $Y$, and $Z$ be measurable spaces. Let $H$ be a markovian kernel from $X$ to $Y$ and $K$ a markovian kernel from $Y$ to $Z$. Let $\mu$ be a probability measure on $X$. Let $\nu = \mu H$ and $\lambda = \nu K = \mu H K$ on $Y$ and $Z$ respectively. Then

$$K_\lambda^\dagger H_\nu^\dagger = (HK)_\mu^\dagger.$$

(This equation holds only a.e. $\lambda$, $\nu$, and $\mu$. See note (ii) after Definition 27.)

**Proof.**

$$\begin{aligned} < f, HKg >_{\mu} &= < f, (HK)g >_{\mu} \\ &= < H^{\dagger}_{\mu} f, Kg >_{\mu H} \\ &= < K^{\dagger}_{\mu H} H^{\dagger}_{\mu} f, g >_{\lambda = \mu HK} . \end{aligned}$$

By Definition 41, this means

$$K^{\dagger} H^{\dagger} = (HK)^{\dagger}. \tag{45}$$

## 11. COMPUTING POSTERIORS: A CLASS OF EXAMPLES

We now use Theorem 34 and the assumption of simple noise to derive an explicit expression for the posterior distributions in a special class of cases that includes the Two-View Theorem observer and many other observers of interest in computer vision.

**Theorem 46.** *(Some Useful Posteriors).* Let the set of scene interpretations, $C$, be a euclidean space $\Re^j$ and let the set of special scene interpretations, $C_s$, be a measurable subset of $C$. Let the set of images, $D$, be a euclidean space $\Re^k$ and let the special images, $D_s$, be a measurable subset of $D$.[13] Let the rendering function, $\pi\colon C \to D$, be measurable. Let the prior measure on scene interpretations be $\mu$, and the measure on images be $\lambda$ $(= \mu N)$. Let the noise kernel $N$ be constant on fibers of $\pi$, and in fact be modeled by independent, identically distributed gaussian random variables with mean zero and standard deviation $\sigma$. Thus, for any scene interpretation $c$ and set $D$ of images,

$$N(c, D) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \int_{D \in \mathcal{D}} \exp\left(\frac{-\|\pi(c) - d\|^2}{2\sigma^2}\right) \lambda(\mathrm{d}d), \tag{47}$$

where $m$ is the dimension of $D$. Then the posterior probability that the scene interpretation

[13] This implies that each of these sets is a standard Borel space and that therefore rcpd's exist (see Appendix A8).

is in the set $A$ given that the image is $d$ is the following:

$$P(A \mid d) = \frac{\mathrm{d}(\mu(1_A N))}{\mathrm{d}(\mu N)}(d) = \frac{\int_{C_s \cap A} \exp\left(\frac{-\|\pi(c)-d\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c)}{\int_{C_s} \exp\left(\frac{-\|\pi(c)-d\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c)}. \tag{48}$$

**Proof.** Appendix, A25.

We now have an explicit form for the posterior distribution $P(d, A)$ for a large class of observers of interest to researchers in computer vision. This class includes the rigid motion observer based on the Two-View Theorem (Theorem 1). If we wish to pick a "best" interpretation on the basis of the posterior distribution we have several standard options (see, e.g., Papoulis, 1984; Gelb, 1974). For instance, we can choose the interpretation with the maximum posterior probability, the so-called MAP (maximum *a posteriori*) estimator; if the prior is uniform, this is called the maximum likelihood estimate (see Witkin, 1981, for an early use of maximum likelihood estimates in vision).

## 12. THE BAYESIAN OBSERVER DIAGRAM

So far we have developed several different threads: competence observers, performance observers, and standard Bayesian estimation. Each has its own meaning and applications in the study of perception. Each is also associated with a compound structure embodying various mathematical objects such as measures, mappings, and kernels. It is the purpose of this section to tie these constructs together. We will introduce one diagram, the "Bayesian Observer Diagram" (henceforth "BOD"), which displays their relevant mathematical structures and the precise relationships between them.

The BOD consists of probability measures related by arrows representing kernels. The point here is that kernels operate on measures to produce other measures (Appendix A3). Thus we can move around the diagram by successively applying kernels to a given measure. The BOD diagram is shown in Figure 3.

The objects at the corners of the diagram are, as we have noted, probability measures. The measures at the upper corners are probabilities on the conclusion space $C$.

**Figure 3.** *Bayesian observer diagram showing the relationship of observer theory to standard Bayesian analysis.*

The measures at the lower corners are probabilities on the data space $D$. Each edge of the diagram consists of arrows in both directions. There are also two arrows in opposite directions along a diagonal of the diagram. These arrows, as mentioned above, represent kernels. The left vertical edge corresponds to a competence observer (Definition 2). The measure $\mu$ is a probability measure on conclusions which expresses the prior probability on scenes adopted by the competence observer. $\mathcal{I}$ is the interpretation kernel of the competence observer (i.e., the *noise-free* posterior used by the competence observer). We can also think of $\mathcal{I}$ as describing the conditional probabilities induced by $\mu$, given points of $D$. For each noise-free image, $\mathcal{I}$ gives a probability measure on scene interpretations that are compatible, via the function $\pi$, with that image.

Recall that $\pi$ is the "image rendering function" of the observer. It assigns to each conclusion (or "scene") $c$ in $C$ the unique image data $\pi(c) = d$, in $D$, which is compatible with $c$. The downward arrows denoted $\pi_*$ on each edge of the BOD are kernels. The kernel $\pi_*$ assigns to any probability measure $\nu$ on $C$ the probability measure $\nu\pi_*$ on $D$, which is defined as follows. The function $\pi$ can be thought of as a random variable on $C$ taking values in $D$; the probability $\nu\pi_*$ is then the usual "distribution" of this random variable.

The kernel $F$ expresses the effect of noise at the level of the conclusion space $C$. In this sense it associates to any noise-free prior $\mu$ of the competence observer a probability

measure $\mu'$ on $C$. We can think of $\mu'$ as the "fuzzy" analog of $\mu$, for the particular noisy circumstances described by $F$.

If we view $F$ as the likelihood function for the prior $\mu$, then $R$ is the corresponding Bayesian posterior. Given the meaning of $F$ as a fuzzing kernel, we can think of $R$ as "retracting" noisy interpretations back onto noise-free interpretations. $F_*$ and $R_*$ express the effects of $F$ and $R$ at the level of the data space $D$.

The right edge of the BOD corresponds to a performance observer (Definition 7). $\mathcal{I}'$ is its interpretation kernel. As in the competence observer of the left edge, we can think of $\mathcal{I}'$ as describing the conditional probabilities induced by $\mu'$, given points of $D$. For each noisy image, $\mathcal{I}'$ gives a probability measure on scene interpretations that are compatible, via the rendering function $\pi$, with that image.

The kernel $N$ models noise in the rendering process. Intuitively, $N(c, d)$ gives the probability that the image $d$ is rendered given that the scene is $c$. Because of noise this probability will in general be nonzero even if $d$ is not $\pi(c)$. In many concrete situations where we can express probability measures as densities with respect to some underlying measure, the density function associated to $N$ is commonly known as the "likelihood function." The kernel $P$ is the Bayesian posterior for the prior $\mu$ and the likelihood $N$. In most cases, the likelihoods and posteriors referred to in the Bayesian literature correspond to $N$ and $P$. This means that traditional Bayesian analyses refer to just the diagonal of the BOD.

Note that $\pi$, $F$, and $N$ can be specified independently of $\mu$, whereas the definitions of the kernels $R$, $P$, $R_*$, $\mathcal{I}$, and $\mathcal{I}'$ are all contingent on $\mu$.

The BOD is convenient in that it collects and displays all the relevant mathematical structures that figure in both an observer-theoretic and Bayesian analysis of perception. However, the real benefit of the diagram for purposes of conceptualization and analysis follows from the mathematical properties of the diagram itself. In particular, we will prove shortly that one can construct a *consistent* BOD and that every pair of opposite arrows represents Bayes adjoints (with respect to the measures indicated at the appropriate vertices). As a matter of general terminology, any diagram is said to be "consistent" if, whenever it is possible to move an object indicated at a vertex of the diagram by means of two different sequences of arrows which end at the same vertex, then the two results

are equal. (Note that consistency is weaker than the general mathematical notion of commutativity of diagrams.)

We now consider the BOD in detail. There are several entities we have already encountered:

1. *Competence Observer*
   a. $\mu$ on $C$ is the prior probability on scene interpretations.
   b. $\pi\colon C \to D$ is the rendering function.
   c. $\mu\pi_*$ is the distribution on images in the noise-free case.
   d. $\mathcal{I}\colon D\times\mathcal{C} \to [0,1]$ is a markovian kernel describing noise-free posterior probabilities. In fact $\mathcal{I}$ is an rcpd of $\mu$.

Thus the left side of the BOD is simply a competence observer with prior probability $\mu$. As we have discussed before, this observer is a canonical description of a perceptual capacity under the assumption that there is no noise. The diagram contains three other entities we have already encountered:

2. *Standard Bayesian Analysis*
   a. $N\colon C\times\mathcal{D} \to [0,1]$ is a markovian kernel describing the effect of noise on the image rendering process; it models the likelihoods of standard Bayesian analyses.
   b. $\mu'\pi_* = \mu N$ is the noisy distribution on images.
   c. $P\colon D \times \mathcal{C} \to [0,1]$ is a markovian kernel describing the Bayesian posterior probabilities when there is noise. According to Theorem 34, $P(A \mid d) = [\mathrm{d}(\mu(1_A N))/\mathrm{d}(\mu N)](d)$.

Thus the diagonal of this diagram—$\mu$, $N$, $\mu'\pi_*$, and $P$—describe the Bayesian analysis of perception when there is noise. The prior measure $\mu$ on scene interpretations gets pushed down via the likelihood $N$ to a measure $\mu'\pi_*$ $(= \mu N)$ on images. The posterior $P$ assigns scene interpretations to these images in a manner consistent with the prior measure $\mu$. The diagram also contains a performance observer, which we now describe:

3. *Performance Observer*
   a. $\mu'$ is a probability measure on $C$ such that $\mu'\pi_* = \mu N$, and is not, in general,

**Figure 4.** *The two observers in the Bayesian observer diagram.*

supported in $C_s$. This measure describes the actual distribution of scene interpretations perceived by subjects.

b. $\mathcal{I}'$ is an rcpd of $\mu'$ with respect to the rendering function $\pi$.

This right side of the BOD—$\mu'$, $\pi$, $\mu'\pi_*$, $\mathcal{I}'$—together with a significance function $\alpha$, is a performance observer (Definition 7). The final new entities are the following:

4. *Fuzzing and Retraction*
   a. $F\colon C \times \mathcal{C} \to [0,1]$ is a markovian kernel given by $F = N\mathcal{I}'$. This kernel "fuzzes up" the ideal prior $\mu$ in a way which matches the scene interpretations perceived by subjects in actual noisy circumstances. $N$ and $F$ are both models of noisy circumstances, but $F$ is at the level of scenes and $N$ relates scenes and images. Thus $N$ and $F$ must be compatible in the sense that $N = F\pi_*$.
   b. $F_*\colon D \times \mathcal{D} \to [0,1]$ is a markovian kernel which is the "pushdown" of $F$. Thus $F_* = \mathcal{I}F\pi_* = \mathcal{I}N$. This kernel fuzzes up the ideal distribution $\mu\pi_*$ on images to match the actual distribution, $\mu'\pi_*$.
   c. $R\colon C \times \mathcal{C} \to [0,1]$ is a markovian kernel given by

$$R = F_\mu^\dagger,$$

i.e., $R$ is the Bayes adjoint of $F$ for the prior probability $\mu$. (The dagger notation is as given in Definition 40.) By Theorem 42 we then have $\mu' R = \mu$, i.e., $R$ "retracts" the noisy measure $\mu'$ back to the Bayesian prior $\mu$.

d. $R_*: D \times \mathcal{D} \to [0,1]$ is a markovian kernel which is the "pushdown" of $R$. Thus $R_* = \mathcal{I}' R \pi_*$. This kernel "retracts" the noisy distribution on images back to the noise-free distribution, $\mu\pi_*$. In fact we will prove, as part of Theorem 49 below, that $R_*$ is the Bayesian posterior for the likelihood $F_*$ with respect to the prior measure $\mu\pi_*$.

**Figure 5.** *The Bayesian likelihoods and posteriors in the BOD.*

Thus the BOD ties together two observers—competence and performance—and displays their relationship with standard Bayesian estimation. A competence observer is a competence theory of a perceptual capacity in the noise free case. A performance observer is a point at which empirical data makes close contact with the formalism, allowing the competence observer to be constrained by experiments. These two observers are displayed in Figure 4. The likelihoods and posteriors of standard Bayesian estimation are highlighted in Figure 5.

What are the minimal data that are required to construct a consistent BOD? The

minimal data must include the measurable mapping $\pi$ from $C$ to $D$, and hence the kernel $\pi_*$. We also must have a prior measure on the configuration space. Because of the symmetry of the BOD, this measure can be either $\mu$ or $\mu'$. If we choose $\mu$ we need a model of "noise" $F$ that allows us to construct $\mu'$. If instead we choose $\mu'$, then we need a model $R$ for retracting $\mu'$ to $\mu$. Thus there are two possible sets of mathematically minimal data: One is $\pi_*$, $\mu$, and $F$; the other is $\pi_*$, $\mu'$ and $R$. This is the essence of the following theorem.

**Theorem 49.** If the spaces $C$ and $D$ are standard Borel spaces, then given $\{\pi_*, \mu, F\}$ or given $\{\pi_*, \mu', R\}$ one can canonically construct a consistent BOD in which each pair of opposite pointing arrows corresponds to a pair of kernels which are Bayes adjoints of each other.

**Proof.** See Appendix A40.

This theorem establishes the existence of consistent BODs. One consequence is that we now see from the BOD that, assuming consistency, $\mu N = \mu\pi_* F_*$. Thus, assuming consistency and that the prior $\mu$ is a correct model of the situation, we can always view the noise $N$ as "simple noise": $N$ is noise-free projection $\pi_*$ followed by a noise kernel $F_*$ acting *only* on the image space $D$. Hence in order to undo the noise $N$ we need only clean up noise in the image space $D$ via the kernel $R_*$. We can then use $\mathcal{I}$, the noise-free interpretation kernel, to produce our scene interpretations. This observation justifies the noise-free competence observer as an essential aspect of standard Bayesian analysis applied to perception.

## 13. PSYCHOPHYSICAL TESTS OF COMPETENCE THEORIES

Suppose we have a computational model for a visual process. And suppose we wish to test whether this model is psychologically plausible. Then we can first write a description of the model as a competence observer. This observer is the theoretical entity we will evaluate on the basis of psychophysical data. But here we face a problem: the competence observer cannot handle noise, whereas the psychophysical data we collect are doubtless

tainted with noise. We must therefore construct a model of performance based on the competence theory we wish to test and the best knowledge we have about the noise likely to obtain. This noise can occur at several levels: (1) in the generation and presentation of stimuli to a subject, (2) in the perceptual processes internal to that subject, (3) in the response processes internal to the subject, and (4) in the response apparatus used by the subject. It is the performance model that we directly test by our psychophysical data and which, in turn, allows us to accept or reject our proposed competence theory. The BOD discussed in the last section deals with the first two levels of noise: in the stimuli and in the subject's perceptual processing. The last two levels, response noise internal and external to the subject, can together be modeled by a "response kernel," $M$, that maps perceptual conclusions to possible experimental outcomes. Here we shall not concern ourselves with this kernel except to note that it is markovian and does not respect the fibres of the rendering function $\pi$. Instead we focus on the BOD and its relation to psychophysical data.

First we note that the Bayesian posterior $P$ does not model what subjects *actually perceive* in noisy displays. The reason is straightforward. $P$ gives positive probabilities only to sets of scene interpretations which have positive probability in the prior probability used to define the competence observer. If the prior probability is supported in $C_s$ then the Bayesian posterior, as we have seen, only leads to interpretations in $C_s$ (see Definition 27 and the remark after Theorem 34). Intuitively, this stipulates that, even when there is noise in the stimuli, one can only see scene interpretations that are strictly compatible with the special interpretations of the competence theory in question. But this is too restrictive. We sometimes perceive scene interpretations that are close to, but not strictly compatible with, the prior of our competence theory. If, for instance, our prior is restricted entirely to rigid motions, and we view a display depicting a nearly, but not exactly, rigid motion, then we should not be straitjacketed into seeing a rigid motion or seeing nothing. We should see a nearly rigid motion. And in fact we do. Thus the Bayesian posterior is too restrictive for the job.

The interpretation kernel $\mathcal{I}'$ of a performance observer, on the other hand, is not too restrictive. It has the flexibility to assign positive probabilities to scene interpretations outside of $C_s$.

What then is the relation between the Bayesian posterior $P$ and the performance observer's $\mathcal{I}'$ in modeling perceptions under noise? When we perceive a nearly rigid motion we can, in many cases, also visualize a rigid motion that is "close to" the perceived motion. This idealized rigid interpretation is modeled by the Bayesian posterior $P$, and the "close to" relation by the retraction kernel $R$. Hoewever, the nonrigid motion actually perceived is modeled by $\mathcal{I}'$ of a performance observer. Thus the upper left of the BOD deals with idealized perceptions, the upper right with actual perceptions. This suggests that psychophysical experiments which test subjects' actual perceptions constrain $\mathcal{I}'$ of a performance observer; psychophysical experiments which test subjects' idealized perceptions constrain $P$ of standard Bayesian analysis. The design of the experiment and the instructions to subjects will determine whether the data collected bear most directly on the actual or idealized perceptions of subjects. Here, as in all aspects of experimental design, care must be taken to assure that the appropriate type of data is collected.

How do detection experiments fit into the BOD? Detection experiments constrain the confidence function $\alpha$ of the performance observer on the right of the BOD (see Definition 7 and the immediately preceding discussion). We summarize our discussion in the following hypothesis (see Figure 6):

**Hypothesis 50.** Psychophysical experiments which test subjects actual perceptions constrain the distributions $\mathcal{I}'(d, \cdot)$ of a performance observer. Psychophysical experiments which test subjects idealized perceptions constrain the distributions $P(d, \cdot)$ of standard Bayesian analysis. Detection experiments constrain the significance function $\alpha$ of a performance observer.

Which performance observer shall we use to test our competence observer via the BOD? It should be a performance observer that is compatible with the competence observer in the sense that it has the same spaces $C$, $D$ and rendering function $\pi$. Moreover it should be related to the competence observer by the models of noise appropriate to the experimental situation. That is, referring to Figure 6, it should be related to the competence observer by maps $N$ and $F$ appropriate to the experimental situation. And finally its significance function $\alpha$ should be appropriate to the experimental situation. A performance observer which is properly related to a competence observer by these criteria

**Figure 6.** *Psychophysical experiments which test subjects' actual perceptions constrain $\mathcal{I}'$ of a performance observer; tests of idealized perceptions constrain $P$ of standard Bayesian analysis; detection experiments constrain the significance function $\alpha$.*

we call a *performance extension* of that competence observer. A change in the noise, i.e., a change in $N$ and $F$, leads to a change in the performance extension. This leads to the following definition.

**Definition 51.** A performance observer $O'$ is an *extension* of a competence observer $O$ if (1) $O$ and $O'$ satisfy[14] the BOD, with $O$ being the competence observer and $O'$ the performance observer, (2) $N$ and $F$ are proper models of the noise in the experimental situation, and (3) $\alpha(d) = R_*(d, D_s)$, where $R_*$ is the Bayesian posterior for $F_*$ of the BOD, as discussed in section 12. (Note that $\alpha$ is the standard Bayesian posterior used for Bayesian classification in the one category case.)

This definition extends and improves a similar definition given by Bennett, Hoffman,

[14] We say that an object or map satisfies a diagram if it permits the diagram to be consistent.

and Kakarala (1993). We use this definition in practice as follows:

1. Construct a performance extension of the competence theory.
2. Collect psychophysical data regarding subjects' (a) detection abilities or (b) actual perceptions, or (c) idealized perceptions.
3. Compare the detection data with the ROC's predicted by the significance function $\alpha$. Compare psychophysical data on the actual perceptions with the relevant measures $\mathcal{I}'(d, \cdot)$ of the performance extension. Compare psychophysical data on the idealized perceptions with the relevant measures $P(d, \cdot)$ of the standard Bayesian analysis.
4. Carry out standard statistical tests to decide if the comparisons in step 3 are satisfactory.
5. Conclude that the psychophysical data confirm (disconfirm) the competence theory if the statistical tests in 4 are satisfactory (dissatisfactory).

In many cases our knowledge of the experimental situation leads us to a unique $F$, and the procedure outlined above can be used without modification. If our knowledge of the experimental situation does not lead us to a unique $F$ then we cannot determine a unique performance extension for the competence observer.

## 14. AN EXAMPLE TEST

We have outlined how, in principle, psychophysical data can be used to test competence theories of perceptual capacities. One practical example of this method applied to a theory of visual surface interpolation is given in Bennett, Hoffman, and Kakarala (1993). (This example tests $\mathcal{I}'$ of a performance extension.) For another brief practical example, which tests $P$ of standard Bayesian analysis, we return again to the competence theory motivated by the Two-View Theorem (described in section 3). Recall that this competence observer takes two distinct orthographic views of four points as a premise. If the two views have no rigid interpretations the competence observer does nothing. If they have a rigid interpretation then they have, in fact, a one-parameter family of interpretations. The Two-View Theorem gives no basis for choosing among the rigid interpretations in the one-parameter family, so we constructed the competence observer to give all the interpretations

equal probability.

A psychophysical test of this competence observer was conducted in a series of experiments by Liter, Braunstein, and Hoffman (1993). In one experiment they showed subjects a two-view display depicting dots in rigid motion. Subjects were simultaneously shown rigid 3D interpretations uniformly sampled from the entire one-parameter family compatible with the two-view display. Subjects selected that 3D interpretation which, up to the resolution of the monitor, best matched their idealized perception of the two-view display. In this way Liter et al. obtained psychophysical data relevant to testing hypotheses about the distribution $P(d, \cdot)$ for various two-view displays $d$.

The noise in the experimental situation was primarily due to roundoff and quantization errors in the displays. The uniform nature of this noise $N$, together with the assumption of a uniform prior $\mu$ on the one-parameter family, led Liter et al. to conclude that the posterior $P$ of the subject should also give all interpretations in the one-parameter family equal probability. Thus Liter et al. concluded, in effect, that the distributions $P(d, \cdot)$ should be uniform.

They then tested the hypothesis that the subjects' choices reflected a uniform distribution on the one-parameter family. Their data led them to reject this hypothesis and therefore to reject $P$ and, in consequence, to reject the psychological plausibility of both this competence observer and its performance extension. They found instead that subjects' choices were heavily biased towards certain of the rigid interpretations in the one-parameter family and away from others. For instance, subjects seemed to prefer "compact" rigid interpretations, i.e., interpretations in which the 3D structures were about as deep as they were wide. This led Liter et al. to suggest that subjects were using constraints in addition to rigidity to guide the interpretation process. Exactly what these constraints are is not yet known, but as hypotheses about the constraints are formulated we can construct competence observers to formalize them and performance observers to submit them to further psychophysical tests. And through a repeated cycling of theory and experiment we can hope eventually to converge on a psychologically plausible competence observer.

The psychophysical studies of Liter et al. investigate the posterior $P$ but not the significance function $\alpha$. This function has been investigated by Braunstein, Hoffman, and Pollick (1990). They had subjects observe two-view and multi-view displays, and judge

whether or not the displays depicted rigid motion. They found that the ability of subjects to detect rigid motion, as measured by their $d'$ scores, was well above chance. However subjects' performance fell well short of that predicted by the ROC curves developed in section 6. This once again suggests that further work is needed to arrive at psychologically plausible competence observers and performance extensions for the perception of structure from motion.

## 17. SUMMARY

The tools of Bayesian estimation provide a powerful approach to understanding and modeling human perceptual capacities. The discrete formulation of Bayes rule is now widely used for this purpose. However in many situations of practical interest to vision researchers the discrete formulation of Bayes rule is inappropriate because it does not allow one to condition on a measure zero event (such as obtaining a specific image out of a continuum of possible images). This paper remedies this defect by deriving a general form of Bayes rule that allows one to compute posterior distributions even when the conditioning event has measure zero. In the noise-free case this general form of Bayes rule is equivalent to the competence observers of observer theory (Bennett et al., 1989a). The consideration of noise leads to the development of performance observers. The relationship between these observers, standard Bayesian estimation, and psychophysical data can be summarized in a single commutative diagram, called the BOD. The BOD provides a useful framework for interrelating psychophysical experiments with computational theories.

## ACKNOWLEDGEMENTS

We thank M. Albert, M. Braunstein, A. Jepson, R. Kakarala, D. Knill, and W. Richards for useful discussions. We thank D. Knill and W. Richards for helpful comments on an earlier draft of this paper. Supported by NSF grant DIR-9014278 and by ONR contract N00014-88-K-0354.

## POSTLOGUE 1: MODULARITY AND COUPLING

To describe human vision by competence and performance observers is to impose a modularity. We carve the visual system into interacting components and use observer theory to describe each component and its interactions. Observer theory does not tell us a priori how to carve things up. It only provides a language for describing the units that result from such a carving. It is an empirical issue how best things should be carved.

This section gives a brief idea about how observers might be coupled or made to interact. Suppose that we have two competence observers $O_1 = (C_1,\ D_1,\ C_{s1},\ D_{s1},\ \pi_1,\ \mathcal{I}_1)$ and $O_2 = (C_2,\ D_2,\ C_{s2},\ D_{s2},\ \pi_2,\ \mathcal{I}_2)$. One way that $O_1$ and $O_2$ might be coupled is hierarchically. That is, if the conclusions of $O_1$ are used as the premises for $O_2$, then we can create a new observer $O$ which goes from the premises of $O_1$ directly to the conclusions of $O_2$. If $O_1$ infers the 3D positions of feature points given their image motion, and $O_2$ infers a 3D interpolating surface given the 3D position of feature points, then $O$ would directly infer a 3D interpolating surface given the image motion of the feature points. Formally, if $C_1$ and $D_2$ are the same space, then we can use the definition of kernel product to define a new interpretation kernel $\mathcal{I} = \mathcal{I}_1\mathcal{I}_2$ on $\pi_1(\pi_2(C_{s2})) \cap D_{s1} \times \mathcal{C}_2$, where

$$\mathcal{I}(d, A) = (\mathcal{I}_1\mathcal{I}_2)(d, A) = \int_{c \in C_1} \mathcal{I}_1(d, \mathrm{d}c)\mathcal{I}_2(c, A). \tag{52}$$

In this case we can write a new competence observer $O$ which is the *hierarchical coupling* of $O_1$ and $O_2$ by

$$O = (C_2, D_1, C_{s2} \cap \pi_2^{-1}(\pi_1^{-1}(\pi_1(\pi_2(C_{s2})) \cap D_{s1})), \pi_1(\pi_2(C_{s2})) \cap D_{s1}, \pi_1 \circ \pi_2, \mathcal{I}), \tag{53}$$

where $\mathcal{I}$ is given by (52). This observer has the same effect as first executing observer $O_1$ and then executing observer $O_2$.

Another way we might connect $O_1$ and $O_2$ is via *weak coupling* (Clark and Yuille, 1990; Bülthoff and Yuille, 1991; Knill and Kersten, 1991). We first create the *product observer*

$$O = O_1 \times O_2 = (C_1 \times C_2, D_1 \times D_2, C_{s1} \times C_{s2}, D_{s1} \times D_{s2}, \pi_1 \times \pi_2, \mathcal{I}_1 \times \mathcal{I}_2) \tag{54}$$

where, for $A_1 \in \mathcal{C}_1$ and $A_2 \in \mathcal{C}_2$, $\pi_1 \times \pi_2(A_1, A_2) = (\pi_1(A_1), \pi_2(A_2))$, and where, for $d_1 \in D_1$ and $d_2 \in D_2$, $\mathcal{I}_1 \times \mathcal{I}_2(d_1, d_2;\ A_1 \times A_2) = \mathcal{I}_1(d_1, A_1)\mathcal{I}_2(d_2, A_2)$. We describe the

noise affecting $O_1$ by a kernel $N_1$ on $C_1 \times \mathcal{D}_1$ and the noise affecting $O_2$ by a kernel $N_2$ on $C_2 \times \mathcal{D}_2$. We denote the prior distribution on scene interpretations for $O_1$ by $\mu_1$ and the prior distribution for $O_2$ by $\mu_2$. We denote their respective posterior distributions by $P_1$ and $P_2$. Using this notation we have the following definition and theorem.

**Definition 55.** *Weak Coupling.* Let $O = O_1 \times O_2$ be a product observer. Let the prior measures for $O_1$ and $O_2$ be $\mu_1$ and $\mu_2$ respectively. Let their likelihoods be given by the kernels $N_1$ and $N_2$ respectively. Then $O$ is said to be a *weak coupling* of $O_1$ and $O_2$ if its prior $\mu$ and likelihood $N$ satisfy

$$\mu N(\mathrm{d}d_1, \mathrm{d}d_2) = \mu_1 N_1(\mathrm{d}d_1)\mu_2 N_2(\mathrm{d}d_2). \tag{56}$$

**Theorem 57.** *Posterior For Weak Coupling.* Let $O$ be a weak coupling of $O_1$ and $O_2$. Let the posterior distributions for $O_1$ and $O_2$ be given by the kernels $P_1$ and $P_2$. Then $P$, the posterior distribution of $O$, is given by

$$P(A_1 \times A_2 \mid d_1, d_2) = P_1(A_1 \mid d_1)P_2(A_2 \mid d_2), \tag{58}$$

where $A_1 \times A_2 \in \mathcal{C}_1 \times \mathcal{C}_2$.

**Proof.** Appendix A41.

This theorem states that if the sources of noise are independent for two observers, then the posterior distribution associated to the product of the two observers is in fact the product of their individual posteriors. This sometimes, though not always, leads to a linear combination of cues from the two observers. There is evidence that a linear combination rule is sometimes used in human vision (Dosher, Sperling, and Wurst, 1986; Bruno and Cutting, 1988; Maloney and Landy, 1989). Nonlinear combination rules have also been formulated (Bülthoff and Yuille, 1991; Bennett, Hoffman, and Murthy, 1993).

## POSTLOGUE 2: RELATION TO REGULARIZATION AND GIBBS

To illustrate our general Bayesian formalism, we briefly describe a "Gibbsian random field" analysis of shape from shading given by Bülthoff and Yuille (1991). They adopt the Bayesian approach as a way of imposing constraints via a prior probability. Observer theory also calls, as we have seen, for the introduction of prior assumptions via probability measures on a set of special scene interpretations. In Bülthoff and Yuille's paper, as here, the particular model of noise in the perceptual modality determines the likelihood function and Bayes theorem then provides the posterior distribution. We will see that our approach helps to identify some of the issues involved in their analysis and indeed we pose a number of questions regarding the mathematical meaning of their work. We do not present any answers to the questions raised here, preferring to leave that to later research. We conclude here that in order to put their Gibbsian random field theories on a rigorous footing, yet another level of generalization seems to be required – a level beyond that to continuous systems discussed in the present paper.

Even in its nondiscrete form Bayes theorem has not, to our knowledge, been discussed or applied in the vision literature, in spite of the fact that the underlying competence theories are (usually) inherently continuous. Bennett, Hoffman, and Prakash (1989a) showed that there is a sense in which continuous systems cannot be modeled to arbitrary precision by discretized versions; a study of the continuous system itself is necessary for its proper understanding(15) The Bülthoff-Yuille analysis of shape from shading is about such an inherently continuous system and so deserves the approach of this paper.

We now present the Bülthoff-Yuille argument, modifying the language (though not the content) somewhat so as to accord with the notation of this paper. We proceed heuristically, as they and others have, interspersing comments on points of rigor.

From a given light source $\vec{s}$, they wish to infer a surface normal vector field $\vec{n}$, assuming Lambertian reflectance. Here $\vec{n} = \vec{n}(\vec{x})$ is a function of the image position $\vec{x}$ in the image rectangle $R$. The constraint they impose is smoothness of the vector field $\vec{n}$. Their prior measure thus assigns higher probabilities to smoother surfaces. They use the prior "density" $c_S^{-1} e^{-\beta E_S(\vec{n})}$ where $\beta$ is a positive real number, and $E_S$ is an "energy" or "cost" functional on the surface field[16] – the subscript $S$ refers to the prior constraint of

(15) This is a consequence of the fact that discretized approximations need not converge, in some sense, to the required values unless controlled by a knowledge of the actual system.

[16] The denotation is, respectively, from statistical mechanics or operations research.

smoothness: $E_S(\vec{n})$ is smaller, and the probability is therefore larger, for smoother fields $\vec{n}(\vec{x})$. Bülthoff and Yuille present their analysis in terms of densities; a rigorously correct theory would display full measures. Thus a prior measure on the surface normal vector fields would be of the form

$$Pr(d\vec{n}) = c_S^{-1} e^{-\beta E_S(\vec{n})} d\vec{n}, \tag{59}$$

where $d\vec{n}$ is some underlying measure on the set of vector fields. Following Horn, they take

$$E_S(\vec{n}) = \lambda \int |S\vec{n}(\vec{x})|^2 d\vec{x}, \tag{60}$$

where $\lambda$ is another real parameter and $S$ is some appropriate differential operator on the space of vector fields, such that smoother surfaces have lower energy. The quantity $c_S$ is a normalization constant. The prior (59), then, is meant to be a measure on the space of *vector fields on R*. This space is infinite-dimensional. Therefore one needs to first ascertain what reasonable measurable structures can be imposed on it, and what bona fide (i.e., $\sigma$-finite) measures exist on it. In particular, we need to be able to assert the existence of a measure $d\vec{n}$ which is concentrated on smooth vector fields, in order that a density of the sort they propose makes sense.

Next, they assume a likelihood function that represents Gaussian simple noise applied to the competence theory: Let $I(\vec{x})$ be the image intensity at $\vec{x} \in R^2$. A Lambertian competence theory requires that $I(\vec{x}) = \vec{s} \cdot \vec{n}(\vec{x})$. In observer language, the rendering function $\pi$ takes $\vec{n}$ to $\pi(\vec{n}) = \vec{s} \cdot \vec{n}$, a function on the image rectangle $R$. Define

$$E_D(I, \vec{n}) = -\int \big(I(\vec{x}) - \vec{s} \cdot \vec{n}(\vec{x})\big)^2 d\vec{x}. \tag{61}$$

The subscript $D$ stands for data. The likelihood is then the "probability" of the data $I$, given the scene $\vec{n}$:

$$Pr(dI \mid \vec{n}) = \frac{e^{-\beta E_D(I,\vec{n})} dI}{c_D(\vec{n})}. \tag{62}$$

Here $\beta$ is the same parameter as before and $c_D(\vec{n})$ is the normalization. Note that this noise is not quite the Gaussian noise we discussed in section 12—the underlying space is a now an infinite-dimensional space of *functions* and is not some $\Re^n$. But Bülthoff and Yuille assume that this noise behaves similarly to a Gaussian, in that the normalization $c_D(\vec{n})$ is independent of the particular surface field $\vec{n}$. Thus we will write simply $c_D$ for the normalization.

Bayes theorem (34) then tells us that the posterior probability that the surface field belongs to some collection $A$ of surface fields, given the image scalar field $I$, is

$$Pr(A \mid I) = \frac{\int Pr(d\vec{n}) 1_A(\vec{n}) Pr(dI \mid \vec{n})}{\int Pr(d\vec{n}) Pr(dI \mid \vec{n})}(I) \tag{63}$$

$$= \frac{\left[\int_A d\vec{n}\; e^{-\beta(E_D(I,\vec{n})+\lambda E_S(\vec{n}))}\right] dI}{\left[\int d\vec{n}\; e^{-\beta(E_D(I,\vec{n})+\lambda E_S(\vec{n}))}\right] dI}(I), \tag{64}$$

where the normalizations cancel out and the right hand sides are to be interpreted as Radon-Nikodym derivatives. Now, proceding as in section 11, it is clear that

$$Pr(A \mid I) = \frac{\int_A d\vec{n}\; e^{-\beta(E_D(I,\vec{n})+\lambda E_S(\vec{n}))}}{Z}, \tag{65}$$

where the normalization $Z$ is

$$Z = \int d\vec{n}\; e^{-\beta(E_D(I,\vec{n})+\lambda E_S(\vec{n}))}, \tag{66}$$

an integral over all possible surface fields.

This is the posterior proposed by Bülthoff and Yuille. The discretized versions of the above expressions seem straightforward to implement. In the discrete case the image rectangle $R$ is a finite set of points in $\Re^2$ and there is a finite set of possible unit normal vectors and image intensities at each point. In this case the measures $d\vec{n}$ and $dI$ can be defined as the usual (Lebesgue) uniform measures on finite-dimensional spaces. Let us denote the particular level of discretization by the subscript $k$, with $k \to \infty$ indicating the limiting passage to the continuum. Then, at level $k$, the underlying measures are $d\vec{n}_k$ and $dI_k$. The likelihood is then indeed a Gaussian. However, even in this case it is not clear how the normalizations $c_{D(\vec{n}),k}$ can be independent of the fields $\vec{n}$, considering that the image fields $I$ can never be negative. This assumption is clearly an approximation. How controlled is this approximation, as the discretization gets finer?

The question of controlling approximations becomes yet more vexed as we consider the definability of the measures in the limit of the continuum. There is, in fact, no pair of measures that the discretized uniform measures $d\vec{n}_k$ and $dI_k$ can converge to – there is no nontrivial $\sigma$-finite *uniform* measure on infinite-dimensional spaces of vector or scalar fields. Perhaps a way out of this impass is to define the discretized densities in terms of other, possibly *nonuniform,* measures $d\vec{n}_k$ and $dI_k$. For then the hope is that these measures may

be chosen so that the result (65) (which indeed holds true at discrete level $k$), converges as $k$ goes to infinity, to some appropriate kernel $Pr(A \mid I)$, i.e.,

$$\frac{\int_A d\vec{n}_k \; e^{-\beta(E_{D,k}(I,\vec{n})+\lambda E_{S,k}(\vec{n}))}}{Z_k} \longrightarrow Pr(A \mid I). \tag{67}$$

It is, after all, the existence of this kernel that we want. But the convergence process needs to be well enough controlled so that $Pr(A \mid I)$ is concentrated on *smooth* vector fields. How do we do this? Are there other ways to define prior measures, likelihoods, and posteriors in this infinite-dimensional situation? What is the structure of Bayesian analysis for random fields? We do not answer these questions here, only referring the reader interested in such questions involving infinite-dimensional integration to Simon (1979).

We make a final remark about regularization and Bayesian observer analysis. The motivation for the Gibbsian random field prior density, given above, was to regularize the situation so as to obtain a unique answer.[16] Probabilistic approaches, such as the Bayes-observer theoretic approach, allow for a consideration of multistability in a most natural fashion: instead of the maximum a posteriori estimate based on a unimodal prior, which leads to a unimodal posterior kernel, choose a *multimodal* prior (or measure on the space of conclusions); this is intended to make manifest the multistability in the posterior.

[16] Actually, the question of whether the prior of (59) is indeed unimodal (in some reasonable sense) in the continuous limit is moot.

## APPENDIX

This appendix contains technical definitions and proofs of theorems whose statements appear in the body of the paper.

**Definition A0.** *(Lebesgue Measure).* Lebesgue measure on $\Re^n$ is a translation invariant measure which assigns to each measurable set $A \subset \Re^n$ a real number equal to its $n$-volume. Any measurable set of positive codimension in $\Re^n$, i.e., any measurable set whose dimension is strictly less than $n$, has Lebesgue measure zero in $\Re^n$. The phrase "Lebesgue almost surely" means "except for a set of cases whose total Lebesgue measure is zero." Unless otherwise indicated, the phrases "almost surely," "almost all," "almost every," and "a.e.," mean, in this paper, "except for a set of Lebesgue measure zero." The phrase "a.e. $\mu$" means "except for a set of $\mu$ measure zero."

**Definition A1.** *(Kernels).* Let $(U, \mathcal{U})$, $(V, \mathcal{V})$ be measurable spaces. A *kernel on $U$ relative to $V$* or a *kernel on $V \times \mathcal{U}$* is a mapping $N: V \times \mathcal{U} \to \Re \cup \{\infty\}$, such that

(i) for every $v$ in $V$, the mapping $A \to N(v, A)$ is a measure on $U$, denoted by $N(v, \mathrm{d}u)$;

(ii) for every $A$ in $\mathcal{U}$, the mapping $v \to N(v, A)$ is a measurable function on $V$, denoted by $N(\cdot, A)$.

$N$ is called *positive* if its range is in $[0, \infty]$ and *markovian* if it is positive and, for all $v \in V$, $N(v, U) = 1$; $N$ is *submarkovian* if $N(v, U) < 1$. If $U = V$ we simply say that $N$ is a *kernel on $U$*. In the text, *all kernels are positive* unless otherwise stated.

**Definition A2.** *(Kernel Products).* If $N$ is a kernel on $V \times \mathcal{U}$ and $M$ is a kernel on $U \times \mathcal{V}$, then the *product* $NM(v, D) = \int_U N(v, \mathrm{d}u) M(u, D)$ is also a kernel on $V \times \mathcal{V}$.

**Definition A3.** *(Kernels As Linear Operators On Measures).* Let $\nu$ be a measure on $U$, and let $M$ be a kernel from $U$ to $V$. Then we define $\nu M$ to be the measure on $V$ given by $\nu M(D) = \int_U \nu(\mathrm{d}u) M(u, D)$ for measurable sets $D \subset V$.

**Remark A4.** *(Remarks On Kernels).* A measurable function $g$ on $U$ yields a new kernel $gM$ by means of $gM(u, D) = g(u)M(u, D)$. A special kind of kernel is the one on $V$

relative to $V$ called the *Dirac kernel at* $v$. This kernel is denoted $\epsilon(v, \mathrm{d}v')$ and defined by $\epsilon(v, B) = 1$ if $v \in B$ and 0 otherwise. For the Dirac kernel the measure in (i) above is the usual Dirac measure $\epsilon_v(\mathrm{d}v')$, while the mapping in (ii) above is the "indicator function" $1_B(v)$, which equals 1 if $v \in B$ and 0 otherwise.

As an example of kernels, we consider a measurable map $\pi: U \to V$. If $\mu$ is a measure on $U$, we have defined $\mu\pi_*$ to be the distribution of $\pi$ with respect to $\mu$, i.e., for a measurable set $B \subset V$, $(\mu\pi_*)(B) = \mu(\pi^{-1}(B))$. However, we can also think of this as the result of a kernel called $\pi_*$ from $U$ to $V$ operating on $\mu$. In fact, $\pi_*(u, B)$ is $1_{\pi^{-1}(B)}(u)$.

**Definition A5.** *(Regular Conditional Probability Distribution).* Let $(U, \mathcal{U})$ and $(V, \mathcal{V})$ be measurable spaces. Let $p: U \to V$ be a measurable function and $\nu$ a positive measure on $(U, \mathcal{U})$. A *regular conditional probability distribution* (abbreviated *rcpd*) of $\nu$ with respect to $p$ is a kernel $m_p^\nu: V \times \mathcal{U} \to [0, 1]$ satisfying the following conditions:

(i) $m_p^\nu$ is markovian;

(ii) $m_p^\nu(v, \cdot)$ is supported on $p^{-1}\{v\}$ for $p_*\nu$-almost all $v \in V$;[16]

(iii) If $g \in L^1(U, \nu)$, then

$$\int_U g\mathrm{d}\nu = \int_V (p_*\nu)(\mathrm{d}v) \int_{p^{-1}\{v\}} m_p^\nu(v, \mathrm{d}u)g(u). \tag{A6}$$

In view of (A1)–(A4), condition (iii) may be simply written as the kernel product formula

$$\nu = (p_*\nu)m_p^\nu.$$

In practice it is sufficient to verify (iii) in the case when $g$ is the characteristic function of a measurable subset $A \subset U$, i.e., we may replace (iii) by

(iii') If for all $A \in \mathcal{U}$

$$\nu 1_A = (p_*\nu)m_P^\nu 1_A.$$

In the special case that $U$ and $V$ are discrete spaces, integrals become sums and measures become weight functions. Thus, for the discrete situation, the appropriate version of (iii) above is

$$\sum_U g(u)\nu(u) = \sum_V (p_*\nu)(v) \sum_{p^{-1}\{v\}} m_p^\nu(v, u)g(u). \tag{A7}$$

[16] Again, $p_*\nu(\mathrm{d}v)$ is defined by $p_*\nu(B) = \nu(p^{-1}(B))$, $B \in \mathcal{V}$.

It is a theorem that if $(U, \mathcal{U})$ and $(V, \mathcal{V})$ are standard Borel spaces then an rcpd $m_p^\nu$ exists for any probability measure $\nu$ (Parthasarathy, 1968). In general there will be many choices for $m_p^\nu$ any two of which will agree a.e. $p_*\nu$ on $V$ (that is, for almost all values of the first argument). If $p\colon U \to V$ is a continuous map of topological spaces which are also given their corresponding standard Borel structures then one can show that there is a canonical choice of $m_p^\nu$ defined everywhere. This is the case typical of image understanding.

**Remark A8.** *(Some Properties Of RCPD's).* It is a theorem that if $(U, \mathcal{U})$ and $(V, \mathcal{V})$ are standard Borel spaces (e.g., euclidean spaces) then an rcpd $m_p^\nu$ exists for any probability measure $\nu$ (Parthasarathy, 1968). In general the choice for $m_p^\nu$ is not unique: any two choices will agree a.e. $p_*\nu$ on $V$ (that is, for almost all values of the first argument of the kernel). If $p\colon U \to V$ is a continuous map of topological spaces which are also given their corresponding standard Borel structures one can show that there is a canonical choice of $m_p^\nu$ defined everywhere.

**Proof A9.** *(Proof of Theorem 23).* By the definition of conditional probability,

$$P(A \mid d) = Pr(A \times \{d\} \mid C \times \{d\}) = \frac{Pr(A \times \{d\})}{Pr(C \times \{d\})} \tag{A10}$$

$$= \sum_{c \in A} \frac{\mu(c)N(c,d)}{(\mu N)(d)}$$

$$= \sum_{c \in A} \frac{\mu(c)1_A(c)N(c,d)}{(\mu N)(d)} = \frac{(\mu(1_A N))(d)}{(\mu N)(d)}, \tag{A11}$$

where we are assuming that $\mu N(d) \neq 0$. Equation (25) is now a matter of straightforward computation, left as an exercise.

Next, recall Definition A5 of rcpd's. In the present instance, the set $U$ of Definition A5 is $C \times D$, $V$ is $D$, $\nu$ there is now $P$, and $\pi$ is $q$. Now consider Equation A7. Take $g$ to be the function $1_A(c)1_B(d)$. With these identifications, Equation A7 yields

$$Pr(A \times B) = \sum_D \mu N(d) \sum_{C \times \{d\}} m_q^P(d, (c, d'))1_A(c)1_B(d) \tag{A12}$$

$$= \sum_{d \in B} \mu N(d) m_q^P(d, A \times \{d\}.) \tag{A13}$$

On the other hand, for fixed $A$, we have

$$Pr(A \times B) = \sum_{d \in B} Pr(A \times \{d\}) \tag{A14}$$

$$= \sum_{d \in B} Pr(C \times \{d\}) Pr(A \times \{d\} \mid C \times \{d\}) \tag{A15}$$

$$= \sum_{d \in B} \mu N(d) P(A \mid d), \tag{A16}$$

where we have used Equation 16 and the definition, Equation 24, of $P$. Thus the two functions of $d$, $P(A \mid d)$ on the one hand and $m_q^P(d, A \times \{d\})$ on the other, possess the same sums with respect to the measure $\mu N$. It is an elementary fact of measure theory that, since the sums over arbitrary $B$ are the same, the two functions must be equal, up to sets of $\mu N$-measure 0.

Finally, Definition A5(ii) shows that $m_q^P(d, A \times \{d\}) = m_q^P(d, A \times D)$. ∎

**Remark A17.** *(Variation of Equation 24).* Equation 24 can also be written as

$$P(A \mid d) = \frac{(\mu(1_A N)\mathcal{I})(\pi^{-1}(d))}{(\mu N \mathcal{I})(\pi^{-1}(d))} \tag{A18}$$

We can do this since, for any $d \in D$, $\mathcal{I}(d, \pi^{-1}(B))$ is 1 if $d \in B$ and is 0 otherwise, i.e., $\mathcal{I}(d, \pi^{-1}(B)) = 1_B(d)$. This means that for any measure $\nu$ on $D$ we can write

$$\nu(B) = \int \nu(\mathrm{d}d) \mathcal{I}(d, \pi^{-1}(B)) \tag{A19}$$

$$= \nu \mathcal{I}(\pi^{-1}(B)) \tag{A20}$$

Incidentally, a computation similar to that for Equation 24 shows that for any subset $B$ of $D$,

$$P(A \mid B) = \frac{(\mu(1_A N))(B)}{(\mu N)(B)} = \frac{(\mu(1_A N)\mathcal{I})(\pi^{-1}(B))}{(\mu N \mathcal{I})(\pi^{-1}(B))} \tag{A21}$$

**Proof A22.** *(Proof of Theorem 34).* We need to show first that the quantity in the right-hand side of (35) exists. Suppose $\mu N(B) = 0$. Then the positivity of $\mu$ and of $N$ show that $\mu(1_A N)(B) \leq \mu N(B) = 0$, so the derivative does exist. Next, does the proposed quantity satisfy Definition 27? Yes – immediate, given the definition of Radon-Nikodym derivatives. Finally, is the quantity a markovian kernel? That it is, for fixed $A$, measurability in $d$ is part of the definition of derivative here. Moreover, the dominated

convergence theorem shows that, for any $d$ for which it is defined, $A \to \frac{\mathrm{d}(\mu(1_A N))}{\mathrm{d}(\mu N)}(d)$ is indeed countably additive. This kernel is markovian since $\frac{\mathrm{d}(\mu(1_C N))}{\mathrm{d}(\mu N)}(d) = \frac{\mathrm{d}\mu N}{\mathrm{d}\mu N}(d) = 1$, for all $d$. The proof of (34) is just a transcription of that for the discrete case, with sums over weight functions replaced by integrals over measures. ∎

**Remark A23.** *(Remark on Theorem 34).* It follows from Theorem 34 that Equation A21 still holds in the continuous case.

**Proof A24.** *(Proof of Corollary 39).* First, observe that the displayed equation of Corollary 39 holds whenever $f$ is a linear combination of indicator functions. Every non-negative bounded measurable function is an increasing limit of linear combinations of indicator functions. By the Monotone Convergence Theorem, the displayed equation of Corollary 39 holds for non-negative functions, and since every function is a difference of two non-negative functions, it holds for all functions. ∎

**Proof A25.** *(Proof of Theorem 46).* Given the assumptions of Theorem 46 we can write, for $B \in \mathcal{D}$,

$$\mu N(B) = \int_C \mu(\mathrm{d}c) N(c, B) \tag{A26}$$

$$= \int_{C_s} \left( \frac{1}{(\sqrt{2\pi\sigma^2})^m} \int_B \exp\left( \frac{-\|\pi(c) - d\|^2}{2\sigma^2} \right) \lambda(\mathrm{d}d) \right) \mu(\mathrm{d}c) \tag{A27}$$

since $\mu$ is supported in $C_s$. Thus

$$\mu N(\mathrm{d}d) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \Big[ \int_{C_s} \exp\left( \frac{-\|\pi(c) - d\|^2}{2\sigma^2} \right) \mu(\mathrm{d}c) \Big] \, \lambda(\mathrm{d}d) \tag{A28}$$

by Fubini's theorem. A similar derivation shows that

$$\mu(1_A N)(\mathrm{d}d) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \Big[ \int_{C_s \cap A} \exp\left( \frac{-\|\pi(c) - d\|^2}{2\sigma^2} \right) \mu(\mathrm{d}c) \Big] \, \lambda(\mathrm{d}d). \tag{A29}$$

$\mu$ and $\lambda$ are of course chosen so that the integrals exist. According to Theorem 34, to find $P(A \mid d)$ we now need to compute the Radon-Nikodym derivative

$$\frac{\mathrm{d}(\mu(1_A N))}{\mathrm{d}(\mu N)}(d). \tag{A30}$$

Letting

$$F(d) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \int_{C_s} \exp\left(\frac{-\|\pi(c) - d\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c) \tag{A31}$$

and

$$F_A(d) = \frac{1}{(\sqrt{2\pi\sigma^2})^m} \int_{C_s \cap A} \exp\left(\frac{-\|\pi(c) - d\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c). \tag{A32}$$

Then (A28) and (A29) say

$$\mu 1_A N(\mathrm{d}d) = F_A(d)\lambda(\mathrm{d}d);\ \mu N(\mathrm{d}d) = F(d)\lambda(\mathrm{d}d). \tag{A33}$$

Notice that for nontrivial priors $\mu$, $F(d) > 0\ \forall d$, so that

$$\mu 1_A N(\mathrm{d}d) = \frac{F_A(d)}{F(d)} \cdot \mu N(\mathrm{d}d). \tag{A34}$$

Hence, recalling Definition 32 of Radon-Nikodym derivative, we have finally

$$\frac{\mathrm{d}(\mu(1_A N))}{\mathrm{d}(\mu N)}(d) = \frac{\int_{C_s \cap A} \exp\left(\frac{-\|\pi(c)-d'\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c)}{\int_{C_s} \exp\left(\frac{-\|\pi(c)-d'\|^2}{2\sigma^2}\right) \mu(\mathrm{d}c)}. \tag{A35}$$

This expression for $P(A \mid d)$ exists generically since the denominator is generically not zero. We will also write this posterior in the form of a kernel, $P(d, A)$. Note that in this derivation we do not have to use a gaussian. We could use any kernel which descends. ∎

**Proof A36.** *(Proof of Proposition 31).* Let us assume that $\mathcal{I}$ is the posterior for $\pi_*$. We now prove that $\mathcal{I}$ is the rcpd of $\mu$ with respect to $\pi$. From Definition 27 we have

$$(\mu 1_A \pi_*)(B) = ((\mu\pi_*)1_B\mathcal{I})(A).$$

If we let $B = Y$ we get: for any set $A$ in $X$

$$\mu 1_A \pi_* 1_Y = \mu\pi_*\mathcal{I}1_A.$$

But $\pi_* 1_Y = 1_{\pi^{-1}(Y)} = 1_X$, hence

$$\mu 1_A = \mu\pi_*\mathcal{I}1_A$$

which shows that $\mathcal{I}$ is the rcpd of $\mu$ with respect to $\pi$.

Now let assume that $\mathcal{I}$ is the rcpd of $\mu$ with respect to $\pi$. We now show that $\mathcal{I}$ is the posterior of $\pi_*$ with respect to $\mu$. We let $A$ be a measurable subset of $X$ and $B$ a measurable subset of $Y$. By Definition 27, we want to show that

$$\mu 1_A \pi_* 1_B = \mu \pi_* 1_B \mathcal{I} 1_A.$$

The right hand side is

$$\begin{aligned}
\text{RHS} &= \int_{y\in B} \mu\pi_*(dy) \int_X \mathcal{I}(y,dx) 1_A(x) \\
&= \int_Y \mu\pi_*(dy) \int_X \mathcal{I}(y,dx) 1_{A\cap\pi^{-1}(B)}(x) \\
&= \mu(A\cap\pi^{-1}(B)).
\end{aligned}$$

The left hand side is

$$\begin{aligned}
\text{LHS} &= \int_X \mu(dx) 1_A(x) \pi_*(x,B) \\
&= \int_X \mu(dx) 1_A(x) 1_{\pi^{-1}(B)}(x) \\
&= \mu(A\cap\pi^{-1}(B)).
\end{aligned}$$

We must also show that under these hypotheses $\pi_*$ is the posterior of $\mathcal{I}$ with respect to $\mu\pi_*$. But we have remarked after Definition 27 that that definition is symmetric provided that $\mu = \mu NP$, i.e., in this case $\mu = \mu\pi_*\mathcal{I}$. This follows from the definition of rcpd. ∎

**Proof A37.** *(Proof of Theorem 42.)* For any $A \in \mathcal{X}$ we have

$$\begin{aligned}
\mu H H^\dagger(A) &= \int_{y\in Y} \mu H(\mathrm{d}y) H^\dagger(y,A) \\
&= \int_{y\in Y} \mu H(\mathrm{d}y) \frac{\mathrm{d}\mu(1_A H)}{\mathrm{d}\mu H}(y) \\
&= \int_{y\in Y} \mu 1_A H(\mathrm{d}y) \qquad \text{a.e. } \nu \text{ s.t. } \pi_*\nu = \mu H \\
&= \mu 1_A H(Y) \\
&= \int_{x\in X} \mu(\mathrm{d}x) 1_A(x) H(x,Y) \\
&= \int_{x\in X} \mu(\mathrm{d}x) 1_A(x) \\
&= \int_A \mu(\mathrm{d}x) \\
&= \mu(A). \qquad ∎
\end{aligned}$$

**Proof A38.** *(Proof of Theorem 44.)* We first prove the following lemma.

**Lemma A39.** Let $X$ and $Y$ be measurable spaces, $\mu$ a probability measure on $X$, and $H$ a markovian kernel from $X$ to $Y$. The condition

$$(\mu 1_A H)(B) = ((\mu H)1_B H^\dagger)(A)$$

for all measurable $A \subset X$ and $B \subset Y$ (which defines $H^\dagger$) is equivalent to

$$\mu f H g = (\mu H) g H^\dagger f$$

for all bounded measurable functions $f\colon X \to \Re$ and $g\colon Y \to \Re$.

**Proof.** Because kernels act linearly on functions, the first equation immediately implies that the second holds for the case where $f$ and $g$ are simple functions, i.e., finite linear combinations of characteristic functions of sets. Then since every measurable function is a limit of simple functions and since our operator $H$, being markovian, is bounded we can conclude by, e.g., the Lebesgue dominated convergence theorem. ∎

We want

$$K^\dagger H^\dagger = (HK)^\dagger,$$

where we have suppressed the subscript measures since there is no ambiguity. By Lemma A39 above this means that for every bounded measurable $f\colon X \to \Re$ and $g\colon Y \to \Re$

$$\mu f (HK) g = \lambda g K^\dagger H^\dagger f.$$

We now work on the right hand side. Let us temporarily denote $h = H^\dagger f$. Then, noting

that $\lambda = \nu K$ the right hand side is

$RHS = (\nu K) g K^\dagger h$

which by Lemma A39 is

$= \nu h K g$

Here $h$ and $Kg$ are functions on $Y$, and $\nu$, a measure on $Y$, operates on their product.

Interchanging the order of multiplication, we write this as

$= \nu (Kg) h$

which we recall is

$= \nu (Kg) H^\dagger f$

$= (\mu H)(Kg) H^\dagger f$

which by Lemma A39 is

$= \mu f H (Kg)$

$= \mu f (HK) g$

which is the left hand side. ∎

**Proof A40.** *(Proof of Theorem 49).* If we are given $\{\pi_*, \mu, F\}$, then we let $\mathcal{I}$ be the canonical rcpd of $\mu$ given $\pi$. We let $N = F\pi_*$. We let $P = N_\mu^\dagger$. We let $R = F_\mu^\dagger$. We let $\mu' = \mu F$. We let $\mathcal{I}'$ be the canonical rcpd of $\mu'$ given $\pi$. We let $F_* = \mathcal{I} F \pi_*$. We let $R_* = \mathcal{I}' R \pi_*$. The constructions of $\mu\pi_*$ and $\mu'\pi_*$ are obvious from their names. This is a canonical construction of the BOD. We now show that each pair of opposite pointing arrows corresponds to a pair of kernels which are Bayes adjoints of each other. First, since $\mathcal{I}$ and $\mathcal{I}'$ are defined to be rcpd's of $\mu$ and $\mu'$ with respect to $\pi$, it follows from Proposition 31 that $\mathcal{I}$ and $\pi_*$ are Bayes adjoints with respect to $\mu$ and that $\mathcal{I}'$ and $\pi_*$ are Bayes adjoints with respect to $\mu'$. $(F, R)$ and $(N, P)$ are, by definition, pairs of Bayes adjoints with respect to $\mu$. Now $N = F\pi_*$ by definition; hence, by Theorem 44, $P = \mathcal{I}' R$. Thus the upper triangle of the BOD is consistent. By definition, $F_* = \mathcal{I} F \pi_* = \mathcal{I} N$. Also $R_* = \mathcal{I}' R \pi_*$ by Definition, and $R_* = P\pi_*$. Hence by Theorem 44, $R_*$ is the Bayes adjoint of $F_*$ with respect to $\mu\pi_*$. Thus all pairs of opposite arrows in the BOD correspond to Bayes adjoints (in either direction, thanks to Theorem 43). For consistency of the BOD we note that $\mu\pi_* F_* = \mu\pi_* \mathcal{I} N = \mu N$, and that $\mu N R_* \mathcal{I} = \mu N P \pi_* \mathcal{I} = \mu$ by two applications

of Theorem 42. All other consistency relations follow from the Bayes adjointness of pairs of opposite arrows by Theorem 42 or by the definitions of the various kernels as the appropriate composites.

The construction of the BOD given the minimal data $\{\pi_*, \mu', R\}$ and the proof of its consistency and adjointness is similarly straightforward. ∎

**Proof A41.** *(Proof of Theorem 57).* It suffices to show that

$$\int_{B_1 \times B_2} P_1(A_1 \mid d_1) P_2(A_2 \mid d_2) \mu N(\mathrm{d}d_1, \mathrm{d}d_2) = \int_{B_1 \times B_2} P(A_1 \times A_2 \mid d_1, d_2) \mu N(\mathrm{d}d_1, \mathrm{d}d_2), \tag{A42}$$

for any $B_1 \in \mathcal{D}_1$ and $B_2 \in \mathcal{D}_2$. By Definition 55 of weak coupling, the left hand side of (A42) becomes

$$LHS = \int_{B_1} P_1(A_1 \mid d_1) \mu_1 N_1(\mathrm{d}d_1) \int_{B_2} P_2(A_2 \mid d_2) \mu_2 N_2(\mathrm{d}d_2) \tag{A43}$$

which by Theorem 34 becomes

$$= \mu_1 1_{A_1} N_1(B_1) \mu_2 1_{A_2} N_2(B_2). \tag{A44}$$

But by Theorem 34, the right hand side of (A42) can be written

$$RHS = \mu 1_{A_1 \times A_2} N(B_1 \times B_2) \tag{A45}$$

$$= \int_{C_1 \times C_2} \mu_1(\mathrm{d}c_1) \mu_2(\mathrm{d}c_2) 1_{A_1}(c_1) 1_{A_2}(c_2) N_1(c_1, B_1) N_2(c_2, B_2) \tag{A46}$$

which by Fubini's Theorem becomes

$$= \int_{C_1} \mu_1(\mathrm{d}c_1) 1_{A_1}(c_1) N_1(c_1, B_1) \int_{C_2} \mu_2(\mathrm{d}c_2) 1_{A_2} N_2(c_2, B_2) \tag{A47}$$

$$= \mu_1 1_{A_1} N_1(B_1) \mu_2 1_{A_2} N_2(B_2) \tag{A48}$$

which equals (A44) and we are done. ∎

## REFERENCES


1. P.N. Belhumeur and D. Mumford. (1992). A bayesian treatment of the stereo correspondence problem using half-occluded regions. *Proceedings of the 1992 IEEE Conference on Computer Vision and Pattern Recognition,* 506–512. Los Alamitos, CA: IEEE Computer Society Press.
2. B.M. Bennett, D.D. Hoffman, C. Prakash. (1989a). *Observer Mechanics: A Formal Theory of Perception.* New York: Academic Press.
3. B.M. Bennett, D.D. Hoffman, J.E. Nicola, C. Prakash. (1989b). Structure from two orthographic views of rigid motion. *Journal of the Optical Society of America, A,* 6, 1052–1069.
4. B.M. Bennett, D.D. Hoffman, C. Prakash. (1993a). Theory of recognition polynomials, (in preparation).
5. B.M. Bennett, D.D. Hoffman, C. Prakash. (1993b). Recognition polynomials. *Journal of the Optical Society of America, A,* (in press).
6. B.M. Bennett, D.D. Hoffman, and P. Murthy. (1993). Lebesgue logic for probabilistic reasoning and some applications to perception. *Journal of Mathematical Psychology,* 37, 1, 63–103.
7. B.M. Bennett, D.D. Hoffman, and R. Kakarala. (1993). Modeling performance in observer theory. *Journal of Mathematical Psychology,* 37, 2, 220–240.
8. M.L. Braunstein, D.D. Hoffman, L. Shapiro, G.J. Andersen, B.M. Bennett. (1987). Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception and Performance,* **13,** 335–343.
9. M.L. Braunstein, D.D. Hoffman, and F. Pollick. (1990). Discriminating rigid from nonrigid motion: Minimum points and views. *Perception & Psychophysics,* **47,** 3, 205–214.
10. N. Bruno and J.E. Cutting. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General,* 117, 161–170.
11. H.H. Bülthoff and A. Yuille. (1991). Bayesian models for seeing shapes and depth. *Comments Theoretical Biology,* 2, 4, 283–314.
12. H.H. Bülthoff. (1991). Shape from X: Psychophysics and computation. In M.S.

Landy and J.A. Movshon (Eds) *Computational Models of Visual Processing,* 305–330. Cambridge, MA: MIT Press.

13. K.L. Chung. (1974). *A Course in Probability Theory.* New York: Academic Press.
14. J.J. Clark and A.L. Yuille. (1990). *Data Fusion for Sensory Information Processing Systems.* New York: Kluwer Academic Press.
15. B.A. Dosher, G. Sperling, and S. Wurst. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. *Vision Research,* 26, 973–990.
16. J. Fodor. (1975). *The Language of Thought.* New York: Thomas Y. Crowell.
17. W.T. Freeman. (1992). Exploiting the generic view assumption to estimate scene parameters. *MIT Media Lab Vision and Modeling TR–196.*
18. D. Geiger and A. Yuille. (1991). A common framework for image segmentation. *International Journal of Computer Vision,* 6, 3, 227–243.
19. A. Gelb. (1974). *Applied Optimal Estimation.* Cambridge, MA: MIT Press.
20. S. Geman and D. Geman. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741.
21. D.M. Green and J.A. Swets. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.
22. U. Grenander. (1981). *Abstract Inference.* New York: Wiley.
23. P.R. Halmos. (1950). *Measure Theory.* New York: Van Nostrand.
24. J.A. Hartigan. (1983). *Bayes Theory.* New York: Springer.
25. H. von Helmholtz. (1925). *Treatise on Physiological Optics.* New York: Dover.
26. A. Jepson and W.A. Richards. (1992). What makes a good feature? *MIT AI Memo 1356.*.
27. D.C. Knill and D. Kersten. (1991). Ideal perceptual observers for computation, psychophysics, and neural networks. In Roger J. Watt (Ed.) *Pattern Recognition by Man and Machine.* Boca Raton: CRC Press.
28. J.C. Liter, M.L. Braunstein, and D.D. Hoffman. (1993). Inferring structure from motion in two-view and multi-view displays. (under review).
29. L.T. Maloney and M.S. Landy. (1989). A statistical framework for robust fusion

of depth information. In W.A. Pearlman (Ed.), *Visual Communications and Image Processing IV. Proceedings of the SPIE, 1199,* 1154–1163.

30. J.L. Marroquin. (1989). A probabilistic approach to computational vision. In S. Ullman and W. Richards (Eds.) *Image Understanding 1989.* Norwood, New Jersey: Ablex Publishing.
31. J.L. Marroquin, S. Mitter, and T. Poggio. (1987). Probabilistic solution of ill-posed problems in computational vision. *MIT AI Memo 897.*
32. K. Nakayama and S. Shimojo. (1992). Experiencing and perceiving visual surfaces. *Science,* 257, 1357–1363.
33. A. Papoulis. (1984). *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill.
34. Parthasarathy, K. (1968). *Introduction to Probability and Measure.* New Dehli: Macmillan.
35. T. Poggio and F. Girosi. (1989). A theory of networks for approximation and learning. *MIT AI Memo 1140.*
36. T. Poggio. (1990). 3D object recognition: On a result of Basri and Ullman, *Technical Report IRST 9005-03,* MIT.
37. A.M. Quinton. (1965). The problem of perception. In *Perceiving, Sensing, and Knowing,* R.J. Schwartz (Ed.), Berkeley, Ca: University of California Press, 497–526.
38. L.J. Savage. (1972). *The Foundations of Statistics.* New York: Dover.
39. B. Simon. (1979). *Functional Integration and Quantum Physics.* New York: Academic Press.
40. R. Szeliski. (1989). *Bayesian Modeling of Uncertainty in Low-level Vision.* Boston: Kluwer Academic.
41. W.B. Thompson, P. Lechleider, and E.R. Stuck. (1993). Detecting moving objects using the rigidity constraint. *IEEE PAMI,* 15, 2, 162–166.
42. A.N. Tikhonov and V.Y. Arsenin. (1977). *Solutions of Ill-Posed Problems.* Washington, D.C.: W.H. Winston.
43. S. Ullman and R. Basri. (1991). Recognition by linear combinations of models, *IEEE Trans. Pattern Anal. Mach. Intelligence* **13,** 992–1006.
44. A.P. Witkin. (1981). Recovering surface shape and orientation from texture. *Artificial*

*Intelligence,* 17, 17–47.

45. A.L. Yuille, D. Geiger, and H.H. Bülthoff. (1991). Stereo integration, mean field theory and psychophysics. *Network,* 2, 423–442.