Computational Vision: Principles of Perceptual Inference*

Alan L. Yuille[†], James M. Coughlan,[†] and Daniel Kersten[‡]

November 25, 1998

*This is the first chapter from a book in preparation by the authors.

[†]Smith-Kettlewell Eye Research Institute, San Francisco, CA

[‡]Department of Psychology, University of Minnesota, Minneapolis, MN
email: kersten@eye.psych.umn.edu

1 Introduction

It is generally accepted that understanding intelligence and brain function is an important scientific goal of interest to scientists, medical practitioners, and engineers. The visual system, in particular, is a very big component of the brain and it is estimated that at least fifty percent of neurons in the human cortex are involved in visual processing (the proportion rises to sixty percent for macaque monkeys). It is generally believed that the cortex is the seat of intelligence (the rise of the human species in the past few million years correlates strongly with increase in cortex size). Moreover, the uniformity of the cortex suggests that if one understands part of it, such as the visual system, then this knowledge would generalize to help understand the rest.

How then, can one study the brain and intelligence? One direct approach is to treat it as a physical system and study it in the same way that physicists study elementary particles or complex systems. However, the brain is extremely complicated and is, perhaps, the most complex known physical system in the universe. It contains at least 10^{12} neurons all of which have many components such as axons, the soma, and dendritic trees. Although scientists have learnt an enormous amount about brain anatomy, physiology and neural mechanisms, this knowledge is still not nearly enough to determine analytic equations which describe neural dynamics of large systems of neurons (except in certain restricted regimes). Moreover, even precise knowledge of neural dynamics would only yield partial understanding of brain function (knowing the electronic dynamics of a computer does not give much insight into the algorithm that the computer is running). This suggests that direct approaches based on studying the biology of the brain must be augmented by understanding visual processing at a more abstract level. Such understanding must take into account the tasks the biological, or artificial, vision system should achieve, what domains it must function in and what computational resources it has.

The approach described in this book, is to study vision (and hence intelligence and the brain) in terms of information processing. This enables us to formulate vision in terms of *probabilistic inference*, or equivalently, as a decoding problem where the goal is to determine information about the world from *intensity patterns* reaching the eye or a camera. Information gathering, however, is not a passive process and will depend on the goals and abilities of the *agent* performing the decoding. From this perspective, the three important issues are what types of image patterns occur, how can information be extracted from these patterns, and what information should be extracted. These three issues will be discussed in later sections of this chapter.

At this abstract level, we are not concerned with how the algorithms for learning and inference are implemented by hardware or wetware. For biological systems it is plausible that certain visual abilities, such as the ability to perceive symmetry, may be learnt during evolution and be encoded in genes and developmental processes (which may be sensitive to the early environment). Such abilities may appear to be "hardwired" in the sense that they are fixed by an early stage of an animal's development. Other abilities, such as the

ability to recognize cars, are presumably learnt at a later stage and can be modelled in terms of neuronal plasticity. From our perspective, it is important to understand what visual abilities can be learnt either during early development or by a mature adult. But the corresponding biophysical mechanisms are beyond the scope of this book.

Our holy grail is a theory of visual perception which is as quantifiable as physics or engineering. Such a theory could be applied to solve real world computer vision problems (in the sense that a mature engineering field, such as control theory, contains techniques for solving most control problems) and which can also provide a theory for human vision which makes quantitive predictions which can be tested by rigourous experiments. A theory of this type does exist for the first stage of visual and speech processing, and is known as signal detection theory (Green and Swets [45]). Signal detection theory (SDT) assumes that an ideal observer would make optimal use of the information in the stimulus to perform specific visual tasks (for example, the ideal uses the optimal statistical criteria to distinguish signal from noise.) It thus provides a rigourous quantitive comparison to evaluate human performance. The problem is that, in its current form, the theory is hard to apply except to simple stimulus types. For complex real world stimuli, it is difficult to determine how to represent and quantify the statistical properties of real world images, or shapes, and estimate the performance of an ideal observer. For example, it seems impossible to apply signal detection theory to high level vision tasks such as object recognition or scene understanding from natural images.

Recently, however, advances in a variety of disciplines – applied mathematics, probability theory, information theory, speech recognition, artificial neural networks, artificial intelligence – have helped develop a framework for describing vision and other perceptual and inference problems. This framework includes a common language for describing vision problems, mathematical techniques for modeling them, and algorithms that can be applied to solve them. The framework reduces to standard signal detection theory as a special case but goes a long way beyond it both in its range of applicability and the power of its techniques. (These advances have been facilitated by the enormous increase in computer power which has made it possible to explore increasingly complicated probability models. For example, one pragmatic reason for the immense popularity of Gaussian probability models was the ease of computing results for them by hand in the days before computers.)

The goal of this book is to give an introduction to these recent advances and provide pointers to the literature where further information can be obtained. Our book will concentrate on the basic mathematical techniques and illustrate them by numerous examples from the psychophysics literature. We hope that our work can act as a bridge between artificial and biological vision and will help stimulate further advances in these areas. Understanding the "inner space" of the brain has been compared to exploring the universe. Within this analogy, our aim is to provide techniques for creating Saturn 5 rockets and Apollo spacecraft to do the exploration.

In this chapter we will describe the basic concepts of this approach. We start by a brief historical review of signal detection theory (SDT). The next three sections introduce pattern theory (PT), bayesian inference, and agents interacting with the world. The fourth section reviews the basic perceptual tasks performed by both artificial and biological vision systems.

1.1 Signal Detection Theory and Ideal Observers

Signal detection theory was developed in the 1950's to model and analyze human sensory decisions in the face of signal uncertainty due to background noise, both internal and external [95]. The theory was the result of combining earlier work in statistical decision theory [36], [87], [46] with communication systems theory [108],[100]. The mathematical results of these theoretical developments and their application to human sensory systems were brought together in the classic book of Green & Swets [45]. Signal detection theory's contribution to human perception was two-fold: 1) statistical decision theory showed how to analyze the internal processing of sensory decisions, and; 2) communication theory showed how the external stimulus and task limit reliable decisions. Let's consider each of these in turn.

1.1.1 Modeling internal processing of perceptual decisions

The application of statistical decision theory to psychophysics showed that sensory decisions were determined by two experimentally separable factors: sensitivity (related to an inferred internal signal-to-noise ratio, called d', which we will discuss extensively in the next chapter) and the decision criterion. This distinction rested, in part, on understanding the significance of the fact that there are two ways of being correct in a detection task–an observer can say "yes I detected the signal" and be right (a "hit") or can say "no, I didn't" and also be right (a "correct rejection"), depending on whether the signal was really transmitted or not[1]. Thanks to signal detection theory, this distinction is now "obvious". For a fixed sensitivity, an observer can only increase the hit rate at the expense of the correct rejection rate. The hit and correct rejection rates are interpreted as the result of: 1) an internal decision variable having two (usually Gaussian) probability densities depending (i.e. conditional) on whether the signal was present or not; 2) basing a decision on whether the internal variable was bigger or less than a criterion value. The most striking success of the decision theory aspect of signal detection theory was the large body of empirical results to emerge showing a remarkable invariance of sensitivity over different tasks (e.g. over various criteria of the observer). On the methodological side, the use of two-alternative forced-choice tasks (as a means to efficiently eliminate criterion contributions from measures of sensitivity) in psychophysics became a de facto standard.

[1]The two ways of being wrong are often called "false alarms" and "misses". This distinction corresponds to the standard definition of Type I and Type II errors in statistical significance testing.

Although the derived measure of sensitivity (d') used in psychophysical experiments has its roots in signal detection theory, in a number of cases, however, these roots have been cut and the measure has taken on a life of its own. This can be dangerous. d' is a signal to noise measure appropriate when both signal and noise can be modelled by univariate Gaussian densities with identical variances. As we will describe in this book, there are other ways to capture the concepts of signal and noise, based on Bayesian probability theory, which are more appropriate for many situations.

1.1.2 Ideal Observer: Modeling external limits to reliable decisions

The second major contribution of signal detection theory grew out of results from communication theory which showed that there were inherent physical limits to the reliability of information transmission, and thus detection, independent of the specific implementation of the detector, i.e. whether it be physical or biological. These limits can be modeled by a mathematically defined ideal observer for the task. For the ideal observer, the signal-to-noise ratio can be obtained from direct measurements of the variations in the transmitted signal. As we will see below, the ideal observer is of particular relevance to the pattern theory (PT) approach we take in this book and introduced in detail later in this chapter. The ideal observer presaged Marr's ideas of a computational theory for an information processing task, as distinct from the algorithm and implementation to carry it out [80]. The ideal provides a *quantitative* computational theory.

An important precursor of ideal observer analysis is the classic experiment of Hecht, Schlaer and Pirenne (1942) [49], which showed that, under certain viewing conditions, the retinal absorption of something on the order of ten photons or so was sufficient for human light detection. As a consequence, it was concluded that a rod photoreceptor could transduce a single photon (see [26] for an excellent account of this experiment). Their conclusions rested on the Poisson theory of fluctuations in photon emission. However, Hecht et al. did not analyze both hit and false alarm rates to determine sensitivity, and a more complete analysis of light sensitivity had to wait for the development of signal detection theory and the ability to compare estimates of the internal and external signal-to-noise ratios.

1.1.3 Statistical efficiency: Comparing measurements of internal and external signal-to-noise ratios

As an information processing organ, the brain is limited both by the objective or external information available to make perceptual decisions, and its ability to process that information. Experimental studies of human perceptual behavior are often left with a crucial, but unanswered question: To what extent is the measured performance limited by the information in the task rather than by the perceptual system itself? Answers to this question are critical for understanding the relationship between perceptual behavior and its underlying biological mechanisms. One way in which ideal observer analysis provided

a means of addressing this question was through measurements of statistical efficiency.

The application of the ideal observer in vision had to wait almost two decades after Hecht, Schlaer & Pirenne, when in 1962, Barlow determined the optimal quantum efficiency of human light discrimination [9]. By considering both the external and internal sources of variability, Barlow showed that an ideal photon detector could get by with about one tenth the number of photons as a human for the same combination of hit and correct rejection rates. Almost all of this quantal inefficiency was accounted for by optical light and receptor absorption losses (see [92]). The key idea was to calculate the external signal-to-noise measure d'_i determined by the physical stimuli and compare this to the inferred internal d'_h (e.g. as determined by the human observers hit and correct rejection rates). Statistical efficiency (equal to the quantum efficiency) can be defined as: $(d'_h/d'_i)^2$, and is always less than or equal to 1. In succeeding chapters, we will encounter other applications of ideal observers to human vision.

The above example of ideal observer analysis of light detection illustrates our fundamental strategy for studying perception, consisting of three modeling domains (figure 1). First, how does the signal (i.e. light switch set to "bright" or "dim") get encoded into intensity changes in the image? For light discrimination, the answer must deal with variations due to quantal fluctuations. Second, what are the limits to optimal theories for decoding the light changes in the received image to infer which signal was transmitted? Answers to this question rely on theories of ideal observers, or more generally of optimal inference. Third, how does a comparison of human and ideal behavior inform our theories of human vision?

The philosophy of this book follows the general strategy of ideal observer analysis illustrated in figure 1; but our goal is much more ambitious–to provide the tools and principles applicable to natural images and the full range of perceptual tasks. One of the crucial differences between classical SDT and the aims of this book is our emphasis on modeling the external limits to inference, including both synthesis and optimal decoding. This, we believe, is a necessary consequence of the inherent complexity of natural image patterns. Marr wrote in 1982: "...the nature of the computations that underlie perception depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented." Theories of human perceptual inference require an understanding of the limits of perceptual inference through optimal decoding theories. These theories, in turn, require an understanding of the transformations and variations introduced in pattern formation. With these aims in mind, let's take a closer look at how signal detection theory needs to be extended.

1.2 Pattern Theory

The term "Pattern Theory" was introduced by Ulf Grenander in the 70's [48] as a name for a field of applied mathematics which gave a theoretical setting for a large number of related ideas, techniques and results from fields such as computer vision, speech recognition, image

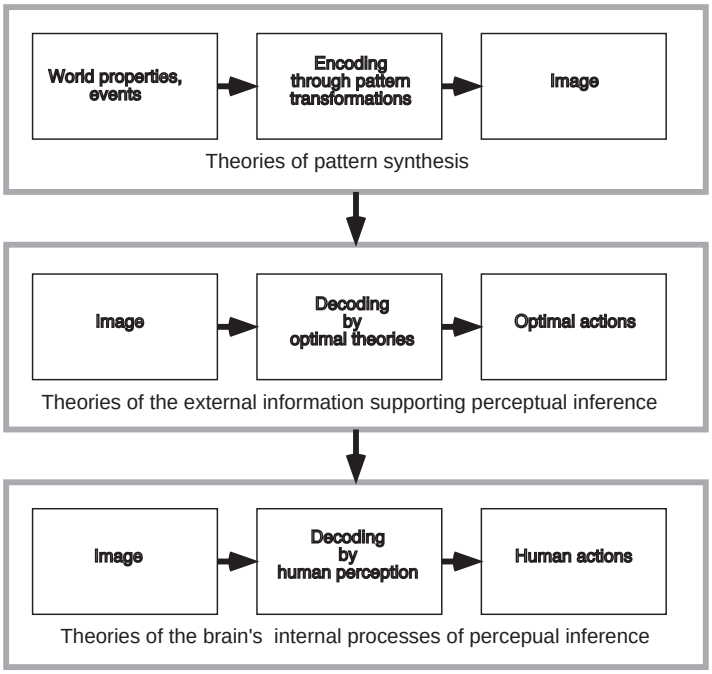

Figure 1 Three levels of study in perceptual inference. Pattern synthesis: How do sources of information about objects and events in the world get encoded into the images received? Optimal decoding: What are the optimal ways of decoding the image to uncover the sources? Human perception: What theories of human perception can explain how human performance differs from that of optimal observers?

and acoustic signal processing, pattern recognition and its statistical side, neural nets and parts of artificial intelligence [2],[101],[50], [106],[91],[74],[3].

In mathematical terms, pattern theory corresponds to formulating problems in terms of Bayesian probability theory. This requires having theories which can both synthesize patterns and also analyze and interpret them. This can be summed up with the slogan "analysis by synthesis" and will be discussed in the next section. More technically, it leads to theories defined in terms of probabilities on graphs [91],[74].

Image patterns correspond to spatio-temporal intensity arrays on a two-dimensional lattice. But similar concepts can be applied to other perceptual modalities and even to more abstract inference systems such as medical expert systems. Such patterns are typically generated by partially hidden processes and the task of the observer may be to decode the signal to determine these causes. For medical expert systems the hidden processes can be the diseases themselves which are unobservable and can only be determined by examining the symptons of the patient or by investigating causal factors for the disease (such as whether the patient smoked). As for vision, the complexities of these medical domains means that knowledge can typically only be expressed as probabilities

rather than deterministic laws.

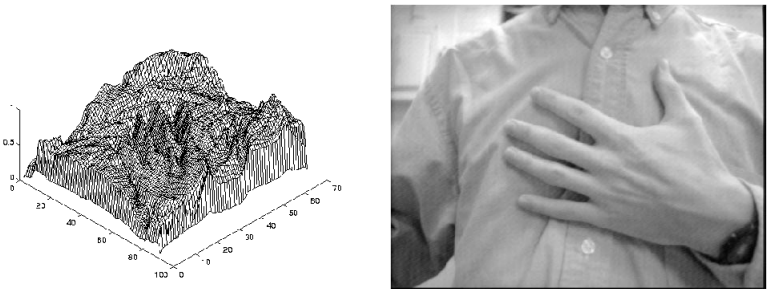

Figure 2 One way to see the difficulty of interpreting images is by representing them in non-standard form. Left, an image of a hand represented by a grid plot where the z axis plots the intensity. This is an accurate representation of the light that reaches the camera yet it is very hard to interpret. Right, the orginal image in conventional form. Our visual skills enable us to interpret this image very easily.

The ability to synthesize patterns is a key element of pattern theory. Where do image patterns and their transformations come from? Image formation is a complex process which depends on several independent factors. These factors include the shape of the objects being viewed, the texture or albedo of the objects, the light sources, and the spatial layout. In addition, there are a variety of complications such as mutual illumination between the different objects. The following parable, derived from Adelson and Pentland [1], illustrates how different generating factors interact when forming an image and how hard it may be to distinguish between them.

A theatre director wishes to design a set. He employs a carpenter, a lightsman and a painter. He explains how the set must appear to viewers in the Royal Box and asks each employee in turn whether he can configure the set to obtain this appearance. All his employees assure him that it is easy. The carpenter describes how by altering the geometry of the theatre scenery he can obtain the desired effect. The lightsman says there is no need to move the scenery, he can ensure the desired image by simply putting in a few lights at strategic positions. The painter says it is even easier, just give him a blank screen and he can paint any image the director likes. The point is that the albedos of objects, their geometry, and the scene lighting are the three most important factors that generate the image. To obtain any desired image one can simply adjust one factor keeping the other two constant. More generally, by manipulating these factors we can come up with an infinite number of possible scenes that cause the identical image. Inverting this process to interpret the image is therefore impossible unless we make assumptions about the likely structure of the scene. In other words to determine that certain image structures are more likely to be due to albedo changes, others to lighting, and so on.

In this parable it is assumed that it is possible to specify simple physical models for combining the different generative factors to determine the image. Complex real world

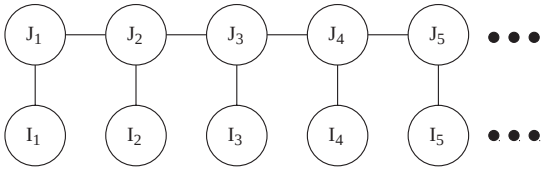

Figure 3 Graph structure for restoring a one-dimensional image. $J_1, J_2, \ldots J_N$ represent the underlying image intensities without noise, while $I_1, I_2, \ldots I_N$ are the noisy pixel values that are actually measured. The goal is to find the MAP estimate of $\{J_i\}$ given only the $\{I_i\}$, i.e. the $\{J_i\}$ that maximizes $P(\{J_i\}|\{I_i\})$, which is equivalent to maximizing $P(\{J_i\})P(\{I_i\}|\{J_i\})$. The prior $P(\{J_i\})$ enforces smoothness of neighboring J_i's (as expressed by the lateral connections in the graph) but allows for occasional discontinuities such as occur at object boundaries in the image. The likelihood function $P(\{I_i\}|\{J_i\})$ enforces the fact that $\{I_i\}$ is a noisy sample of $\{J_i\}$. See Chapter 3 for details on this model.

scenes, however, will often be too complex to determine such models (barring major advances in computer graphics). Fortunately in many cases it may suffice to approximate the image formation process probabilistically. Such an approach has been succesfully used in hidden markov models of speech where modelers ignored the physical processes which generated elementary sound units, such as phonemes, and instead trained markov models to represent them probabilistically. If possible, however, it is preferable to have transparent models which explicitly take into account all the factors which generate the image.

These complexities and interacting factors require sophisticated probability distributions defined over many random variables. A typical probabilistic model for an images, described in detail in Chapter 3, is shown in figure (3). The large number of random variables required, which may be thousands or more, is significantly different from the standard models used in signal detection theory, see figure (4), where there are typically only two random variables. In order to define probability distributions on a large number of variables we first have to know which variables directly influence each other. This "influence" can be represented by a graph, see figure (3), where the random variables are represented by nodes of the graph and two variables directly influence each other only if the nodes are connected, see figure (11) for another example. Finally, we need to define probability distributions on this graph, and the distributions we use, such as $P(\{I_i\}|\{J_i\})$ in figure (3) are learnt by statistical analysis of datasets. This means that they will often not be Gaussians[2].

[2]Gaussians are very popular models partially because the typically lead to straightforward linear analysis. In addition, the Central Limit theorem states that Gaussian distributions will arise if a number of independent processes are combined. On the other hand, the Central Limit theorem only applies asymp-

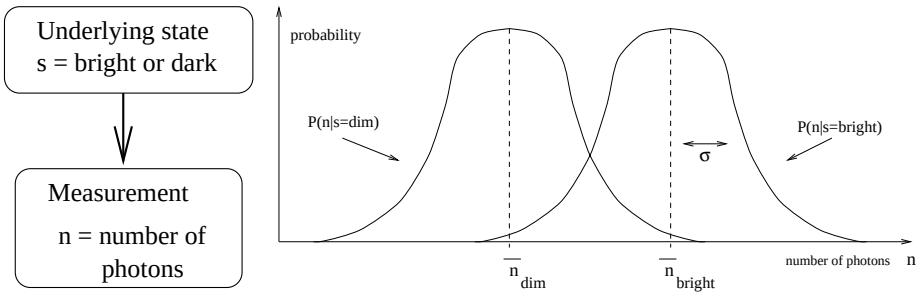

Figure 4 A standard problem from signal detection theory expressed as a probability on a graph. The problem is to infer whether a bright or dim light was flashed based on the number of photons measured. The bright light produces more photons on average, but the actual number of photons fluctuates from trial to trial, making the inference subject to error. Left, the Bayes net specifies the structure of the problem: the state s causes the measurement of a certain number of photons n. The prior is $P(s)$ and the likelihood is $P(n|s)$, and the posterior probability $P(s|n) = P(s)P(n|s)/P(n)$ is specified once a measurement n is made. The difficulty of making an inference depends on how similar $P(n|s = bright)$ and $P(n|s = dim)$ are. For the case in which $P(n|s = bright)$ and $P(n|s = dim)$ are Gaussians of equal variance, as shown on right, this difficulty may be summarized by the signal-to-noise ratio $d' = (\overline{n}_{bright} - \overline{n}_{dim})/\sigma$. See Chapter 2 for more details of this model.

Consider, for example, the theatre workshop. It is straightforward to model it by probabilities defined on graphs. The image is generated by a number of interacting factors and the influences between these factors can be captured by a graph on which probability distributions can be defined. By contrast, it does not fit easily into the paradigm of target stimulus plus additive Gaussian noise. Instead it is more helpful to think of a *transformations* that can act on patterns (see Chapter 1, by D.B. Mumford in [72]). These transformations include spatial warping, occlusion, and superposition. We illustrate these transformations in figure (5). As we will show in this book, it is possible to model these types of transformations by probabilities on graphs.

Such a paradigm, when applicable, leads to powerful techniques for modeling a variety of physical systems, particularly those occuring in electrical engineering (where the assumption can often be justified by the physics of the situation). But when many factors are involved it is hard, and often impossible, to characterize them simply as target or noise. In the example above, is the albedo the noise? Or the lighting? Or the geometry? In situations where noise can be unambiguously defined it is still far from clear that additive Gaussian models are appropriate. In certain cases the central limit theorem can be used to justify the Gaussian approximation but it should be stressed that the central limit theorem is only valid for a restricted class of situations (the basic theorem requires a large

totically and only for restricted ways of combining processes. Moreover, as emphasized by Huber [51], and repeatedly rediscovered by workers in computer vision, Gaussian distributions are very *non-robust* and sensitive to contaminants in the data. The Central Limit Theorem and Robust Statistics are described in Chapter 2.

Figure 5 An example of the diffrent types of operations that are performed on image patterns. Left, a distinguished scientist. Left center, a spatial warp of the scientist. Center right, the warped scientist behind bars. Right, the warped scientist beyind bars viewed through a window on which a cat's image is reflected.

number of independently, identically distributed additive samples – though the results can be generalized to include limited interactions between the samples). Moreover, it has long been known that probabilistic inferences derived from assuming Gaussian models are highly sensitive to contamination from non-Gaussian sources and are *non-robust* in the sense that a single outlier in a data sample can cause enormous changes in the estimated parameters of the Gaussian, such as its mean and variance [51]. Many examples from computer vision illustrate the importance of robustness and the dangers of incautiously assuming Gaussian models. Our view is that the central limit theorem is too often used as "the valium of the intellectuals" and that appeals to it should be justified by mathematical reasoning or careful experimental analysis.

1.2.1 Pattern theory as a generalization of ideal observer theory

It is now helpful to contrast pattern theory, and probabilities on graphs, with signal detection theory and other ways, such as linear systems analysis, for modeling vision.

Although pattern theory is in the tradition and spirit of signal detection theory, PT involves major conceptual and theoretical advances. For both SDT and PT, the problem is one of decoding: the input pattern is the result of multiple causes, but only a subset of the causes, the signals, are useful for a perceptual decision or action. Let's consider six ways in which PT goes considerably beyond this common foundation.

Signals are not simple functions image intensities. In SDT and classical ideal observer analysis, the input, the noise, and the signal, are all the same "stuff": a physical quantity (e.g. luminance or sound pressure level) as a function of time and/or space. Perceptual decisions are often limited to information which is explicit in the decoded signal–e.g. does the signal image have more light intensity than another? The decoder which answers this question is simple–it merely measures whether the image intensity is bigger. In pattern

theory, the signals are any properties of the world useful for the visual agent; for example, estimates of object shape and surface motion are crucial for actions such as recognition and navigation, but are not simple functions of light intensity. Natural images are much more complex functions of useful signals, and finding an inverse function relating image measurements to useful signals is a major theoretical challenge.

Useful information is confounded by more than noise. In signal detection theory, the other causes of the input pattern are called noise; in pattern theory, the other causes include noise as well as unneeded variables of different stuff. In terms of pattern theory, one can distinguish between *primary* and *secondary* variables.[3] The primary variables are those which the system is designed to estimate. By contrast, the secondary variables are not estimated but are instead integrated or summed out. It should be emphasized that the distinction between primary and secondary depends on the specific task the system is designed to solve. Variables which are secondary for one task may be primary for another. For example, estimating the illumination is often unimportant for visual tasks and so the illumination variable(s) is often secondary. But there will be occasional tasks when one wishes to determine the illumination, in which case the variable becomes primary. Illumination is just one of several secondary variables (depending on the task) which play the role of noise in SDT. This leads to a third major distinction.

Natural images are not linear combinations of relevant signals. Signal detection theory is primarily linear: the input is the sum of the signal and the noise. Except in rare instances (e.g. contrast detection limited by photon fluctuations at high light levels), natural perceptual tasks involve inputs which are non-linear functions of the signals and the noise (or secondary variables). For example, light intensity is a non-linear function of object shape, reflectance, and illumination. It should be stressed that much psychophysics up to now has been investigating the measurement stage – i.e. the retina. There has been more concern with photoreceptor noise and other sources of early noise than with the more general "noise", or variability, caused by the viewed scene. It is generally believed that human visual perception is fairly accurate after the imaging stage and so there is little, or no, "internal noise" after the photoreceptors (see [10] for an estimate of the residual high-level "noise" in light detection). (Though individual neurons may be noisy the entire system appears to have a lot of redundancy built in).

This brings us to an important distinction between artificial and biological vision modelers about where the uncertainty, or noise, should be modelled. In computer vision it is always assumed that the noise is due to the outside world and the cameras which capture the photons and digitize the image. There is assumed to be no noise in the computations performed when processing the image (unless it is explicitly inserted as part of a stochastic

[3]Primary and secondary variables have also been referred to as explicit and generic variables, respectively. The theory of generic views treats viewpoint as a secondary variable, enabling resolution of ambiguities in shape perception [86]. Light direction as a secondary variable can be used to obtain a unique estimate of depth from cast shadows [67].

sampling algorithm). By contrast, many models of biological systems assume that there is "noise" in the visual system in addition to the noise in the photoreceptors whereas noise in the external stimuli is discounted. Computer vision is thus in the spirit of the external noise component of SDT; whereas, most psychophysics has been influenced by the internal noise contribution of SDT.

Variables of interest are rarely Gaussian. Signal detection theory approximates noise variations as Gaussian processes. A Gaussian approximation works very well in certain domains (as an approximation to Poisson light emission), but as we will see, fails in most others. Both the linear and Gaussian assumptions have had a striking success in the general problem of modeling human perceptual and cognitive decisions, where the variability is inside the observer [45],[112]. But, when modelling external variability the Gaussian assumption fails miserably.

Perception involves more than classification. Not surprisingly, for signal detection theory, the primary focus is on signal detection–was the signal sent or not? Pattern theory subsumes SDT by considering all perceptual tasks: classification, estimation, control, and learning. SDT has not precluded abstract signals (e.g. "is any one of 100 signals there or not?"); but we also require a framework that can handle continuous estimations (e.g. distance or shape) as well as more complex categorical decisions: e.g. is the input pattern due to a cat, a dog, or "my cat"? Pattern theory provides the tools for formulating the generalization of ideal observers for the complex tasks of natural perception.

Most of the interesting perceptual knowledge on priors and utility is implicit. Both signal detection and pattern theory rest on Bayesian decision theory. As is introduced below, two important components of decision theory are the specification of prior probabilities of scene properties or signals and the costs and benefits of actions, through a loss function. In most applications of SDT, it has been the experimenter that manipulates the priors and the loss functions. The human observer is often aware of the changes, and can adopt a conscious strategy to take these into account. For us, the most important priors are largely determined by the structure of the environment and can, in principle, be modeled independently of perceptual inference (i.e. in the synthesis phase of study). As we will see later, modeling priors (e.g. through density estimation) is a hard theoretical problem in and of itself, largely because of the large dimensionality. In classical SDT, probabilities are typically specified over small dimensional spaces. The costs and benefits are inherent to the type of perceptual task, and determine the primary and secondary variables. Thus, to elaborate on Helmholtz's definition of perception: we are largely concerned with perception as unconscious inference involving unconscious priors, and unconscious loss functions.

Although there is a growing literature of ideal observer analysis in human vision, the application of the ideal observer has been limited to simple statistical models and tasks. Past studies have been confined to low-dimensional natural tasks (e.g. Poisson fluctuation models for resolution [39]), artificial laboratory manipulations (e.g. additive

Gaussian noise, [92]), and (overly) simple approximations of natural image statistics (e.g. fractal image ensembles, [69]). We now introduce the recent theoretical advances which allow us to generalize signal detection theory and ideal observers to almost all perceptual tasks.

2 Bayes Probability Theory: Pattern Analysis and Pattern Synthesis

A key element of this book is the use of Bayesian probability theory. An advantage of the Bayesian approach, compared with other vision theories, is that they separate the models and their inferences from the precise algorithms which compute these inferences and, even further, from the hardware/wetware on which the algorithms are implemented[4]. This allows us to determine limits of performance for visual tasks which are *independent* of the specific algorithm used and lies at the heart of the *ideal observer* concept. This also distinguishes pattern theory from neural network approaches such as multi-layer perceptrons [50]). From our perspective, these neural network theories try to solve two difficult tasks — both modeling and computation at once. In a Bayesian approach, as in Hidden Markov models and Bayes Nets, we first learn the probability distributions and verify them explicitly by stochastic sampling, see figure (9), *and then* determine algorithms for applying the models to practical problems. We believe that learning models and algorithms separately will lead to more tractable problems. Moreover, the explicit nature of the representations is more conducive to understanding the internal workings of an algorithm, and to knowing what problems they will generalize to.

We now give a brief introduction to the techniques of Bayesian probability theory. In general, we wish to infer the state of the world $\mathbf{S}$ given some measurement $\mathbf{I}$. Thus the variables $\mathbf{S}$ would correspond to the variables in our representations of the world, for example the variables representing the shape of a face, while the measurement $\mathbf{I}$ would correspond to the observed images. Within the Bayesian framework, one infers $\mathbf{S}$ by considering $P(\mathbf{S} \mid \mathbf{I})$, the *a posteriori* probability of the state of world given the measurement. Note that by definition of conditional probabilities, we have

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

Dividing by $P(\mathbf{I})$, we obtain Bayes' theorem

$$P(\mathbf{S} \mid \mathbf{I}) = \frac{P(\mathbf{I} \mid \mathbf{S})P(\mathbf{S})}{P(\mathbf{I})} = \frac{P(\mathbf{I} \mid \mathbf{S})P(\mathbf{S})}{\sum_{\mathbf{S}'} P(\mathbf{I} \mid \mathbf{S}')P(\mathbf{S}')}. \tag{1}$$

This simple theorem re-expresses $P(\mathbf{S} \mid \mathbf{I})$, the probability of the state given the measurement, in terms of $P(\mathbf{I} \mid \mathbf{S})$, the probability of observing measurement given the state, and

[4]See Marr 1982 for a discussion of the importance of these types of distinctions.

$P(\mathbf{S})$, the probability of the state. Each of the terms on the right-handside (RHS) of the above equation has an intuitive interpretation.

Diversion: Conditional probabilities. Let's take a moment to illustrate conditional probabilities. The probability of rain in Seattle is high; but the probability of being in Seattle, given rain is much lower. The probability of some red pixels in an image given a tomato in a scene is fairly high. But the probability of a tomato in a scene, conditional on some red pixels in an image, is low.

This next one is a bit more subtle[5]. A murder suspect tests positive in a DNA test and it is reported that the chances of a match with an innocent person are 1 in a million. So is it a good guess that the guy is guilty? Not necessarily, we don't have the right conditional probability yet. We've been told the false positive rate, $p(DNA\ match|Not\ Guilty)$; but because of the high moral cost of convicting the wrong person, the more crucial conditional probability is $p(Not\ Guilty|DNA\ match)$. Suppose that, due to variability in DNA testing, there are 3 people out of several million in the city that would show a positive test. Then $p(Not\ Guilty|DNA\ match) = 2/3$–evidence, which by itself would constitute reasonable doubt. Of course there is always more than one piece of evidence, and ultimately, what one wants is $p(Not\ Guilty|All\ Evidence)$.

Many people, including experts, get wrong conditional probabilities in this next, even subtler, example. It is sometimes called the "Monty Hall problem". A prize is behind one of three doors, called A, B, and C. The contestant picks a door, say A, but it is left closed. Monty Hall opens up C to show the contestant that the prize is not there. Should the contestant change his mind and pick B? Well, the probability of the prize being behind door A hasn't changed: $p(prize\ behind A) = p(prize\ behind A|C\ shown\ empty) = 1/3$. But $p(prize\ behind\ B|C\ shown\ empty) = 2/3$, so yes, the contestant should change his mind! The information gain, on observing Monty to have opened door C, is reflected by the change in the conditional probabilities. The correct answer isn't intuitively obvious to most people[6].

The expression $P(\mathbf{I} \mid \mathbf{S})$, often termed the *likelihood function*, is a measure of how likely an image is given what we know of the state of the world. To see this, note that

---

[5]Adapted from an example in "A Mathematician Reads the Newspaper" by John Allen Paulos.

[6]The Monty Hall problem gained some notoriety through Marilyn Vos Savant's explanation of this non-intuitive answer in Parade Magazine. Even experts disagreed with Vos Savant, leading to some professional embarassment. Vos Savant's explanation was insightful and went along the following lines. Imagine there are a 1000 doors and the prize is behind one. You pick one and then Monty Hall opens all the doors *except* for the one you picked and one other (so 998 are shown to be empty). Should you switch? The answer is now obviously "yes". It is interesting that intution gets the right conditional probability for 1000 doors, but not for only 3. This may be because people incorrectly perceive Monty Hall to open one of the two doors at random–whereas he obviously avoided one particular door when there are a 1000. This also raises a point that recurs in this book. We generally assume that cognitive or perceptual decisions are non-optimal; however, a reasonable strategy is to try to discover tasks for which we are. We do make the right decision for the 1000 door case.

assuming we know the state of the world, e.g. the light sources, the objects, and the reflectance properties of the surfaces of the objects, then we can re-create, as an image, our particular view of the world. Yet, due to noise in our imaging system and imprecision of our models, this re-creation will have an implicit degree of variability. Thus, $P(\mathbf{I} \mid \mathbf{S})$ probabilistically models this variability.

The expression $P(\mathbf{S})$, referred to as the *prior model*, models our prior knowledge about the world. Because of the ambiguities inherent in vision, the prior is necessary to make vision "well-posed." To illustrate, consider the well-known Necker cube illusion, see figure (6). This illusion is often used to illustrate the ambiguity of vision by observing that there are two stable percepts, both of a cube (i.e. the internal vertices are seen in either front or in back). From our perspective, however, the real interest of this illusion is why there are only two percepts. After all, there are an infinite number of wire-frame objects which could have generated the image. Why do we only perceive cubes? In Bayesian terms, one explains this phenomenon by asserting that the prior probability for symmetric shapes such as cubes, which occur frequently in our world, is greater than for other irregular shapes. This prior may correspond to low level cues, such as symmetry and the high probability of right angles, or to a specific prior for cube shapes (consider cubist paintings!) Thus, even though the values of the likelihood functions for a cube shaped wire-frame and a irregular shaped wire-frame may be equal, i.e. $P(I|\ cube\}) = P(I|\ not-cube\})$, the prior term creates the observed bias toward cubes. Figure 7 illustrates how information from the likelihood and prior probabilities can uniquely constrain the perceptual interpretation of a cube. (This illlusion can also be "explained" by the concept of generic views, or accidental alignments, which takes into account the *phase space* of different interpretations and which we will describe from a Bayesian perspective in Chapter 2.)

Another important example are the Mooney images described by Moore and Cavanagh [84]. Such images, (see figure (8), are readily identified as faces despite their severe degradation. The bias for humans to detect faces is so strong that they ignore other conflicting sources of information. As can be verified, by visiting the San Francisco Exploratorium[7], an inverted (i.e. convex) face mask will be perceived as a normal face. This perception persists even as the observer walks past the mask. The visual system prefers to interpret the image as a distorting real face rather than a static inverted face.

As with the Necker cube and Mooney examples, in general the image alone is not sufficient to determine the scene and, consequently, the choice of priors becomes critically important. They embody the assumptions about the world that the visual system must make to achieve valid percepts. Such assumptions have been proposed by workers in biological vision and include Gibson's *ecological constraints* and Marr's *natural constraints*.

We therefore need likelihood functions and priors sufficiently rich to model important aspects of the world. These probability distributions should, if possible, be learnt from the

[7]In Europe, go to the London Museum of Science.

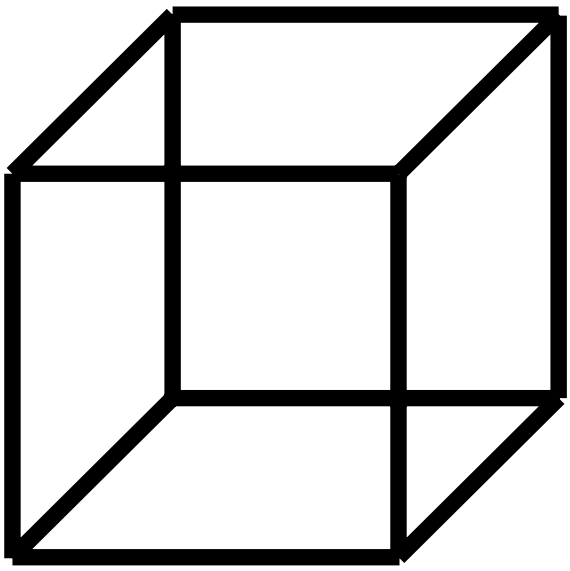

Figure 6 In this figure of the Necker Cube, the observer always perceives a cube, even though there is an infinite number of possible 3-D wire-frames which could produce this image.

application domain and are expected to relate to the types of fundamental transformations desctribed by pattern theory. If the signals can be modelled by low dimensional probability distributions (i.e. with a small number of random variables) then standard statistical methods are sufficient to learn these models. Such techniques are not sufficient for higher dimensional distributions, such as intensity images with up to $250,000$ dimensions, and novel approaches are required such as Minimax Entropy Learning theory [131] which has been demonstrated to learn probability distributions for generic images [132], and image textures [131].

An important advantage of Bayesian models of patterns is that it allows us to stochastically sample from the model, possibly fixing some of the world variables $\mathbf{S}$, using this distribution to construct sample signals $\mathbf{I}$ generated by various classes of objects or events. The basic idea of sampling is as simple as tossing a coin many times to determine if it is biased (We will discuss sampling algorithms in detail in later chapters). A good test of whether the prior has captured all the patterns in some class of signals is to see if these samples are good imitations of life. From a pattern theory perspective, the analysis of the patterns in a signal and the synthesis of these signals are inseparable problems and use a common probabilistic model: computer vision should not be separated from computer graphics, nor speech recognition from speech generation.

Another example of sampling comes from modelling language [109] in terms of the probability of a letter conditional on the N previous letters. The order of these Markov models indicate the amount of "history" encoded within them. For example, a zeroth-order model has no history at all and each letter is sampled independently from the

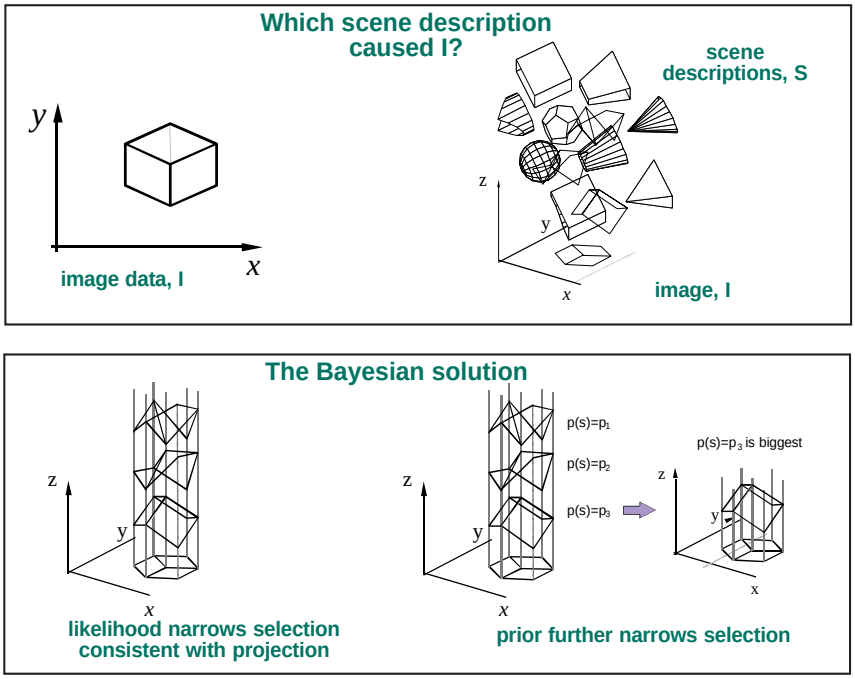

Figure 7 Bayes Cube. What 3D object caused the image of a cube in the upper left panel? The likelihood constrains the possible set of objects (or scene descriptions) to those consistent with the image data, but even this is an infinite set. The prior probability constrains the consistent set of 3D objects to those that are more probable in the world. If the prior probability has two equal valued modes, there would be two equally probable interpretations, as with the Necker cube. Although this particular example suggests a sequential application of likelihood and prior constraints, Bayes does not specify a procedure, and many Bayesian implementations simultaneously enforce both constraints. (Adapted from Sinha & Adelson, 1993

same distribution. By comparison, a second order model samples its letters based on the previous two letters. For a zero-order Markov model, the letters are independent and although their frequency matches English text, there is little apparent regularity in the samples produced. For a first-order Markov model, the frequency of pairs matches English text and samples begin to have recognizable letter groupings, such as "ON", or "ARE" and near-hits such as "DEAMY". Even more structure emerges in higher-order Markov samples. One can generate samples based on Nth-order Markov models of words, instead of letters and generate short phrases that often seem to make sense[8]

Once a suitable Bayesian theory has been specified we will need to perform proba-

[8]It is well-known that the combinatorics become unmanageable long before meaningful sentences begin to emerge with any grammatical consistency. This was one of the observations that led to the development of formal theories of grammar [24].

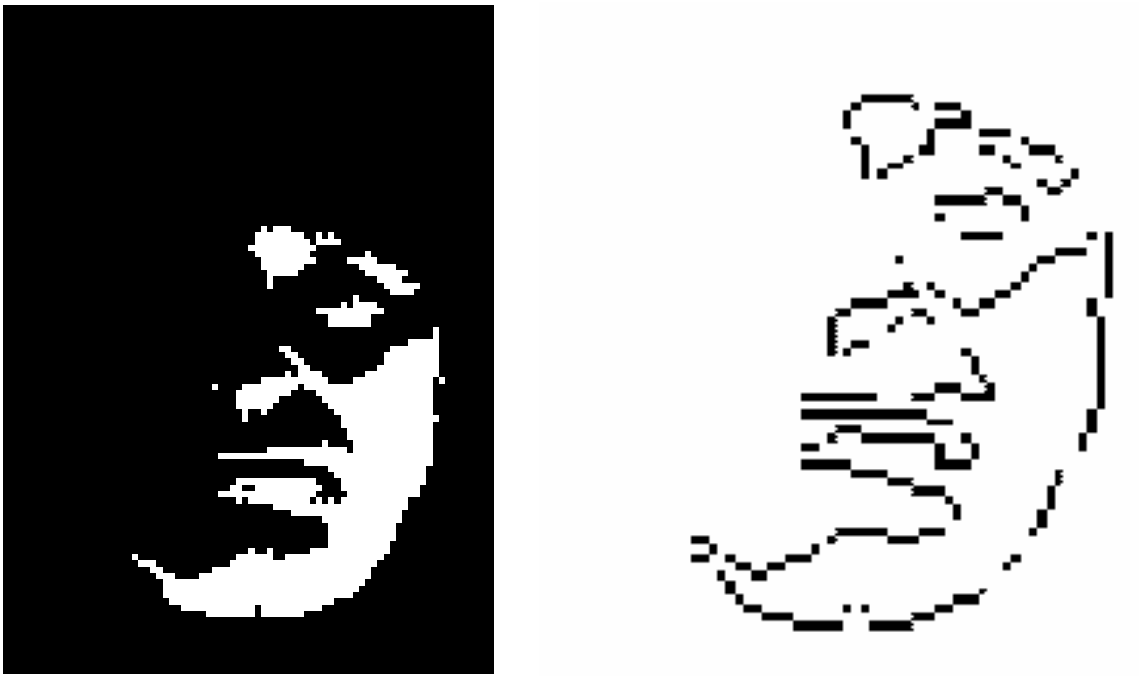

Figure 8 Left, a binarized "Mooney" image of a face. It can be perceived as a face despite the severe degradation of the image. Right, the edge map corresponding to the binary image. Observe how ambiguous this is.

Figure 9 Samples which test three different probability distributions for curves (courtesy of Song Chun Zhu). Observe how these three samples bring out the different aspects of the curves distributions. Left, this model biases towards jagged curves. Center, this model gives smoother curves and has multiple parts. Right, this model has elongated smooth curves with multiple parts.

bilistic inference such as finding the most probable (MAP) estimate of the state of the world. To illustrate how one might formulate a problem using Bayesian reasoning, we have constructed a toy example. In this example, an observer must determine whether or

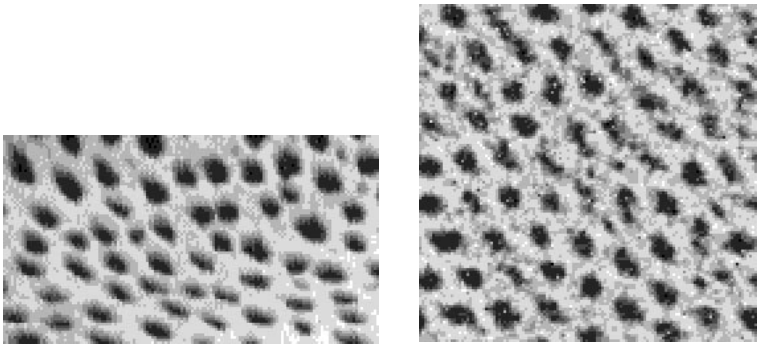

Figure 10 Texture (from Zhu, Wu, Mumford 1997). (Left) the observed texture image from which the Bayesian model was learnt. (Right) an image synthesized by stochastic sampling from the learnt model.

---

not a family member, seen at an odd angle, is his grandmother.

Problem:

"I see a member of my family - definitely a women and definitely wearing a bikini - sunbathing on the front deck of our sailboat. Because of her position I can't make out her face, but it appears her hair is gray."

"Is this my grandmother?"

Bayesian Analysis:

$$\begin{aligned} \texttt{Obs} &= \text{Hair appears gray} \\ \texttt{Gran} &= \text{Grandmother in a bikini} \\ \texttt{W or D} &= \text{Wife or Daughter in a bikini} \end{aligned}$$

$$\begin{aligned} P(\texttt{Gran}) &= 0.01 \quad \text{– unlikely she would wear a bikini} \\ P(\texttt{W or D}) &= 0.99 \\ P(\texttt{Obs} \mid \texttt{Gran}) &= 0.5 \\ P(\texttt{Obs} \mid \texttt{W or D}) &= 0.05 \quad \text{– odd lighting could make hair seem gray} \end{aligned}$$

hence by Bayes' Theorem

$$P(\texttt{Gran} \mid \texttt{Obs}) = \frac{P(\texttt{Obs} \mid \texttt{Gran})P(\texttt{Gran})}{P(\texttt{Obs})} = \frac{0.005}{P(\texttt{Obs})}$$

similarly

$$P(\texttt{W or D} \mid \texttt{Obs}) = \frac{P(\texttt{Obs} \mid \texttt{W or D})P(\texttt{W or D})}{P(\texttt{Obs})} = \frac{0.0495}{P(\texttt{Obs})}$$

hence

$$P(\texttt{Obs}) = 0.005 + 0.0495 = 0.0545$$

$$P(\texttt{Gran} \mid \texttt{Obs}) = \frac{10}{99} \; !$$

In this granny example, the random variables, the distributions, and the inference strategy are straightforward. We will now introduce a more powerful graphical way of representing these probabilities and performing inference on them. This approach is known as *probablities on graphs* and it will be used extensively in this book. It has been built out of mathematical developments in probability, neural networks, probabilistic AI, speech, etc. as a way for overcoming earlier limitations of probabilty theory and information theory.

Firstly, we represent the system in terms of the *joint distribution* $P(\mathbf{S}, \mathbf{I})$. This distribution is defined over the set of random variables representing $\mathbf{S}$ and $\mathbf{I}$. The graph structure represents explicitly whether specific variables are dependent, independent, or conditionally independent.

A classic way of illustrating conditional independence of variables is by considering the well known phenomenon that eating ice-cream is correlated with drowning. But this correlation does not indicate a causal relationship, or direct influence, between these two events. Instead there is a more basic event of being at a beach which increases the probabilities of both eating ice-cream and drowning. In probabilistic terms we can represent being at a beach, drowning and/or eating ice cream as random variables $B = Beach,\ D = Drown,\ I = Ice-Cream$. The situation is completely specified by the joint distribution $P(B, D, I)$. If we ignore the beach variable then we must consider the joint distribution $P(D, I) = \sum_B P(B, D, I)$. In this case we find that $P(D, I) \neq P(D)P(I)$ and so the variables are not independent and hence are correlated. (Two variables X and Y are independent if $P(X, Y) = P(X)P(Y)$ for all X, Y – see Appendix 1).

Now we consider the full problem. By the definition of conditional probability, see Appendix 1, this can always be expressed as $P(B, D, I) = P(I, D|B)P(B)$. But our knowledge of the *influence relationships* between the variables means we can express $P(I, D|B) = P(I|B)P(D|B)$, or equivalently we state that I and D are *conditionally independent* if B is specific (i.e. as being at beach or not). We can therefore express the full probability distribution as:

$$P(B, D, I) = P(D|B)P(I|B)P(B), \tag{2}$$

which can be represented by a graph with arrows connecting the random variable B to D

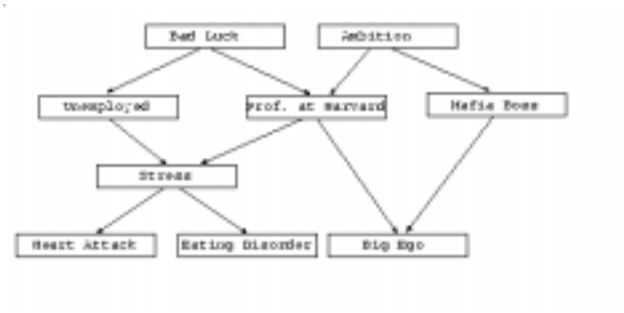

Figure 11 A Bayes Net is a directed graph with probabilities. The arrors derive links between nodes which directly influence each other. Imagine a gameshow where the players have to identify the professions of people but only by asking indirect questions. The jobs are "unemployed", "Harvard professor" and "Mafia Boss". The players are not allowed direct questions but they can ask about causal factors – e.g. "bad luck" or "ambition" – or about symptoms – "heart attack", "eating disorder", "big ego".

and I. This graph makes clear the dependencies between the variables and, in particular, that D and I are conditionally independent given that we are on a beach.

The key point here is the distinction between direct and indirect influences. Being at a beach *directly* influences drowning and eating ice-cream. But ice-cream and drowning, in themselves, are only *indirectly* related. The concept of influence is very general and subsumes the notion of causality. It also allows us to represent probability distributions in many variables by graphs where the random variables are represented by nodes and direct influence between random variables is denoted by links joining the appropriate nodes. The nature of the problem will determine the complexity of the graph needed to represent it. For a problem with many independent variables only a few links will be needed and the graph structure will be simple. For more complex situations with direct influences between many variables we will require complicated graphs. In fact one, abstract, way to consider the structure of this book is in terms of increasing complexity of the graph structures which we need for modelling increasingly complex vision problems.

The graphical representation makes explicit the dependencies between random variables and makes probabilistic inference more straightforward by basic operations such as *marginalization* and *conditioning*. For example, both the prior and likelihood function can be obtained by standard manipulations known as *marginalization* and *conditioning*. The prior on $\mathbf{S}$ can be obtained directly by summing out the $\mathbf{I}$ variables, $P(\mathbf{S}) = \sum_{\mathbf{I}} P(\mathbf{S}, \mathbf{I})$, which is marginalization. To determine the likelihood function we condition to obtain $P(\mathbf{I}|\mathbf{S}) = P(\mathbf{S}, \mathbf{I})/P(\mathbf{S})$. Marginalization and conditioning are two of the most important ways of performing inference and we will continually use them in this book. We note that after conditioning on a variable we obtain a *conditional distribution*.

It is helpful to also consider the problem from the perspective of information theory [28]. This approach has its roots in work of Barlow [9] (see also Rissanen [102]). The close link between the Bayesian and the information-theoretic approaches which comes from

Shannon's optimal coding theorem. This theorem states that given a class of signals $\mathbf{I}$, the coding scheme for such signals for which a random signal has the smallest expected length satisfies:

$$\text{length}(\text{code}(\mathbf{I})) = -\log_2 p(\mathbf{I}), \tag{3}$$

where, following Shannon, we assume that the signals are encoded by a binary alphabet (i.e. the alphabet consists of $\{0, 1\}$) though other codings are possible. (We also gloss over technical issues involving fractional code length – see [28]). The intuition that we should use shorter length codes to represent commonly occuring signal elements is at the basis of most code designs, such as Morse code.

Given an input image $\mathbf{I}$ we may choose to describe it in terms of variables $\mathbf{S}$ which do not correspond to direct observables but which represent the structure of the image better. (One can even think of $\mathbf{S}$ as representing the objects in the scene and $\mathbf{I}$ being the image of them.) From this perspective, finding the most probable interpretation of a signal – finding $\mathbf{S}^*$ which maximizes $P(\mathbf{S}|\mathbf{I})$ – is equivalent to finding the *shortest way to encode* $\mathbf{S}$ given that we have observed signal $\mathbf{I}$. In other words, we seek the simplest explanation of the input signal. This principle of economy dates back to William of Occam and is referred to as Occam's razor and lies at the heart of mimimum description length theory [102].

Although this connection is attractive opinions differ about its utility. Our view is that it offers a new perspective on probability theory which may be a useful guide for defining probability distributions on complex events. The theatre workshop analogy, described earlier, is an example where minimum description length is a natural criterion. Indeed, the original example was formulated along these lines with costs for painting, costs for lighting, costs for carpentry.

Many of the basic concepts of information theory are directly related to perceptual processing and will be introduced in detail in later chapters of this book. Indeed, as we stated earlier, perceptual processing can be thought of as decoding input signals by extracting information. In addition, a visual system will need to encode information efficiently in order to transfer it from one part of the system to another. For such problems, concepts such as the *entropy* of a distribution and the capacity of channels play crucial roles. Roughly, the entropy measures the uncertainty of the distribution: highly peaked distributions have low entropy and spread-out ones have high entropy.

One can also ask questions about the entropy of images and what aspects of images convey important information. One way to approach this is based on Shannon's guessing game for estimating the entropy of English: the entropy can be bounded by observing how good humans are at predicting sequences of words. Experiments based on this idea have been applied to images [65].

Finally, we are aware that the word "Bayes" comes with several associations which are irrelevant to this book. In particular, it has been associated with philosphical arguments

about the difference between "subjective" interpretations of probabilities (which involve degrees of belief) versus so-called "objectivist", or "frequentist" approaches. We do not wish to get involved in this controversy and stress that the probability distributions we use are intended to correspond as much as possible to the objective reality of the domain and should, if possible, be measured or learnt. Nevertheless we stress that objectivist approaches to probability theory often smuggle in unstated assumptions and we have sympathy with Good's remark that "the subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science". Many theories of computer vision can be criticized in this light. A big advantage of the Bayesian approach is precisely that it forces all the assumptions of the model to be clearly stated so that they are then capable of empirical investigation.

From the perspective of the book the basic questions about Bayes are technical issues. Can we represent the problems of interest by Bayesian networks? And can we learn these probabilities accurately? If so, can we perform probabilistic inference efficiently? (When using this approach to model biological systems one must keep in mind that the biophysics of neurons and the anatomical structure of the brain may put restrictions on the types of theories which can be implemented.)

3 Vision for an Agent in the World

Up to this point we have been talking about vision in a passive unteachable agent. By this we mean that the agent passively observes and never acts on the basis of what it sees; and that it analyzes its sensory data using a fixed set of expectations which are never updated. The complete agent does neither.

The complete agent uses the input from the visual system to help decide on which actions to perform in the world. The agent's real interest is not in merely guessing the state of the world, but taking those actions which optimize its chances of achieving its goals. Thus, certain incorrect actions are more dangerous than others, and this may bias your guesses about the state of the world, given incomplete knowledge. The trade-offs affecting such a choice can be formalized using decision theory which introduces the concept of a loss function which depends on the action performed and the state of the world. In our Granny example, the required action might be to call out the name of the wearer of the bikini and tell her that you have finished cooking lunch. The loss function will then depend on the name you call and on the true identity of the bikini wearer. Granny may be flattered to be mistaken for your wife or daughter yet your daughter may be unhappy if she is mistaken for Granny. Bayesian decision theory suggests that you take the action which minimizes the expected loss of your action or, if necessary, obtain additional information before making a decision.

More formally, suppose the action $\mathbf{A}$ is one of the possible actions that can be taken after interpreting the scene $\mathbf{S}$. A *loss function* $L(\mathbf{A}, \mathbf{S})$ specifies the penalty for performing

action $\mathbf{A}$ if the scene is $\mathbf{S}$. The *risk* $R(\mathbf{A};\mathbf{I})$ of taking action $\mathbf{A}$ when the input is $\mathbf{I}$ is defined to be the expected loss:

$$R(\mathbf{A};\mathbf{I}) = \sum_{\mathbf{S}} L(\mathbf{A},\mathbf{S})P(\mathbf{S} \mid \mathbf{I}),$$

with respect to the a posterior probability, $P(\mathbf{S}|\mathbf{I})$ of $\mathbf{S}$. An important special case is when the action $\mathbf{A}$ is to guess the state $\mathbf{S}$. In this situation, one possible loss function is that you pay penalty 0 for making any wrong interpretations and receive the reward +1 for all of the correct interpretations. In other words, $L(\mathbf{A},\mathbf{S}) = -1$ if $\mathbf{A} = \mathbf{S}$ and $= 0$ otherwise. In this case the risk reduces to

$$R(\mathbf{A};\mathbf{I}) = -P(\mathbf{A} \mid \mathbf{I}),$$

and hence the best strategy is to pick the most likely interpretation. This is called *MAP estimation* (Maximum a posteriori). As we will see in later chapters other estimation strategies are sometimes better. (Note that decision theory is used in Signal Detection theory and adjusting payoff rewards can be used to affect the observers' decisions).

After finding the action $\mathbf{A}^*$ which minimizes the expected loss the agent can decide whether the anticipated risk, $R(\mathbf{A}^*;\mathbf{I})$, is too great. If so, the agent can postpone taking action until it has gathered further information. Actively gathering more data until the objects in the world are better identified is one of the main principles of so-called *active vision*. The simplest procedure may be moving your head to get a better view, or more simply to make the foreground pop out from the background by giving it relative motion. The idea that passive vision, especially of still images, was an unnatural degenerate case of vision for an agent goes back to J.J. Gibson.

More formally, the agent can take another measurement $\mathbf{J}$ and determine the modified a posteriori distribution:

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

The expected risk, and hence the optimal action, can now be determined using the modified a posteriori function. Hopefully the additional information will make the correct interpretation more probable and hence decrease the anticipated risk. Additional measurements can be taken until the anticipated risk is considered acceptable. Of course, this analysis has assumed that measurements are freely available. If not, the cost of making them must be taken into account. It may be better to chance missidentifying Granny than to take your eye off the stove and risk spoiling the lunch.

To give another example where using MAP estimates of the world without considering risk leads to suboptimal actions, consider a frog placed in a box into which flies are very occasionally placed. Suppose the frog correctly learns its priors from the statistics of the past, so that it gives a very low prior probability to the occurrence of flies. Thus even

when a fly is present the *a posteriori* probability of seeing a fly will be very small and the frog may miss all potential meals. This problem can be avoided by adopting the correct loss function, so that the frog decides to act as though it has detected a fly even when the a posteriori probability is very small.

This example also illustrates the fact that agents operate within specific domains and their perceptual systems may break down in novel domains or if characteristics of the domains are altered. For example, it is reported that certain birds are unable to distinguish between their eggs and beer bottles inserted into their nests by inquisitive scientists. Beer bottles were presumably not a major factor during the many years when these bird's visual systems developed so there would have been little need to waste valuable neurons to represent them. Of course, the visual systems of humans and the higher mammals display great plasticity and ability to learn novel stimuli. If a westerner travels to the far east, he quickly begins to learn distinctive facial characteristics of easterners and thus acquire better abilities to recognize people he meets. Moreover, many psychophysical experiments show that human observers can learn to perform visual tasks at which they are initially poor and which seem to have little in common with those tasks that they would perform in their normal environments. Furthermore experiments on adult monkeys have shown dramatic changes in brain representations followed by exposure to specific stimuli for an extended period.

This domain specificity should be kept in mind when considering ideal observer experiments. A human observer may be ideal when performing everyday visual tasks in a known environment. But under experimental conditions, it is possible that an observer needs extended exposure to the task and domain in order to learn the appropriate Bayesian model. In some situations strong hints might be helpful. Moreover, there is definitely a limit to the plasticity of the brain and hence a limit to the tasks that an observer can be trained to do ideally.

This leads us to another objection to Bayesian models which is sometimes raised. Experiments on human reasoning have demonstrated that humans can be poor at estimating and computing with probabilities. Would this prevent humans from behaving as ideal observers? We argue that the answer is no. On the one hand, it is very dubious to draw analogies between conscious reasoning and low-level unconscious perceptual inference. On the other, we suspect that humans are perfectly able to compute probabilities for the tasks which matter to them. A bookmaker, or an insurance agent, who ignored probabilities would rapidly go broke. In summary, an intelligent agent needs only perform some tasks optimally and this will depend on its goals, resources, and environment.

4 Tasks for Experiments and Artificial Systems

Although, functional visual tasks are diverse, both computer vision and experimental psychophysics concentrate on a small number of types of inferences. This section introduces the basic types. As we will see throughout the book, Bayesian probability theory, together

	Object perception		**Spatial layout**		
	Object-centered (object recognition)		World-centered	Observer-centered (hand action)	
	Entry-level	*Subordinate-level*	*Planning*	*Reach*	*Grasp*
Shape	P	P	S	S	S
Material	S	P	S	S	S
Articulation	S	P	S	S	P
Viewpoint	S	S	S	P	S
Relative position	S	S	P	S	S
Illumination	S	S	S	S	S

Table 1 Table illustrating how the visual task determines the primary (or explict) variables to be estimated (P) , and secondary (or generic) scene variables to be discounted (S), e.g. through marginalization. The pattern of image intensities is determined by all of the scene variables, object shape, material, object articulation (e.g. body limb movements or facial expression), viewpoint, relative position between objects, and illumination. Our classification into primary and secondary is by no means definitive. For example, viewpoint may be secondary in some types of spatial planning, such as putting a puzzle together, but primary for planning a trajectory.

with rules such as conditioning and marginalization, gives a calculus for performing these inferences.

4.1 Functional Visual Tasks

The two fundamental tasks of vision are object *localization* and *recognition*. But there are numerous variations depending on the task. One can localize an object, track an object, and localize oneself. The visual requirements for reaching to an object are quite different than those for judging relative position between objects. One can see the color, shape and relative position of an object, as well as determine whether it is an animal, a dog, or "my dog, Snuggles". A given task requires certain scene variables to be estimated more accurately than others, depending on the loss function.

Recognition and localization can be broken down even further. Research in human object recognition has distinguished entry-level from subordinate-level object recognition [104]. Entry-level recognition is the intial fast-acess to object categories typically at an intermediate level of abstraction (e.g. "Is this a dog?"). Subordinate-level recognition is finer grain (e.g. "Is this a Doberman?"). It has been argued that entry-level recognition is determined primarily by object shape [18], whereas subordinate-level recognition requires fine-grain shape, material, or articulation estimates. Spatial layout tasks include scene recognition, determining the relative location between objects, and path planning. Picking up an object requires reach and grasp, which each have their own requirements for primary variables. The distinction between secondary and primary variables for a common set of visual tasks is illustrated in Table (1).

4.2 Types of inference for visual tasks

Suppose we identify various nodes in the graphical structure of a problem as signals or causes ($\mathbf{S}$), intermediate variables (α), and image patterns or effects ($\mathbf{I}$). Our knowledge of a problem is completely characterized by the joint probabiity, $p(\mathbf{I}, \alpha, \mathbf{S})$. Perceptual inference is deciding on some hypothesis, $H(\mathbf{S})$, regarding the causes given the image pattern. This requires knowledge characterized by $p(\mathbf{I}, \mathbf{S})$, obtained by marginalizing over the intermediate variables, α. On the other hand, learning the relationship between images and signals requires knowledge of intermediate parameters, α, via $p(\alpha)$, obtained through marginalization over causes and effects. In this section we will describe the standard visual tasks and illustrate how we can model them by Bayesian inference. In later chapters, we will show how similar concepts apply to learning and control tasks. Many computer vision systems are designed to perform actions on the world. This can be modelled by introducing techniques from Decision Theory and Control theory to augment pattern theory.

4.2.1 Types of tasks

There are a relatively small number of types of inference task in psychophysics and computer vision, most of which are variations on classification or estimation.

In *classification*, we have a discrete set of hypotheses $\{H_i\}$, and the task is to observe $\mathbf{I}$, and choose its category i:

$$H_i : \mathbf{I} \in S_i, i = 1 \ldots n$$

The rule for pattern synthesis can be quite abstract. For example, one may have to decide whether image data $\mathbf{I}$ is a picture of "Elias, Jess, or Simeon". If $n = 2$, the task is *discrimination*. The term *detection* is often used if $n = 2$, and one of the hypotheses is: $H_0 : \mathbf{I} \notin S_1$ (e.g. "Was there a light flash or not?") This is the "yes/no" task in psychophysics.

In computer vision and experimental psychophysics, the rule for image formation is often a simple function, $F(\mathbf{S}, \alpha)$, of the signal ($\mathbf{S}$) and secondary (α) variables (e.g. $I = S + \alpha$). It is common in psychophysics for the difficulty of the task to be controlled by small changes of a single signal parameter, $\Delta\mathbf{S}$ which can correspond to an image parameter such as contrast, wavelength, duration, noise or image position. For example, the discrimination task is determined by:

$$H_1 : \mathbf{I} = F(\mathbf{S}_0, \alpha)$$
$$H_2 : \mathbf{I} = F(\mathbf{S}_0 + \Delta\mathbf{S}, \alpha)$$

where α is the noise. Barlow's task, described earlier, was of this sort: "Was the flash of light from the bright or the dim source?". *Threshold* is defined as, $\Delta\mathbf{S}_c$, for a criterion

performance level (e.g. percent correct). For a continuous signal parameter, detection corresponds to $\mathbf{S}_0 = 0$. The independent experimental variable, $\mathbf{S}$, can be a scene variable whose effects in the image may be quite complex. For example, $\mathbf{S}$ could be the distance of an object from the viewpoint. Performance is summarized by percent correct, or hit and false alarm rates in a yes/no task, or by a confusion matrix in a classification task[9].

The *estimation task* requires us to assign a continuously valued number, $\hat{\mathbf{S}}$ to some continuously valued target parameter, $\mathbf{S}$. Formally, the task is, given $\mathbf{I}$, estimate $\mathbf{S}$, where $\mathbf{S}$ is a continuous index on the space of hypotheses:

$$H_{\mathbf{S}} : \mathbf{I} = F(\mathbf{S}, \alpha)$$

Estimation is a common computer vision task. In human perception, there is a long tradition, outside the signal detection theory framework, of doing "magnitude estimation" for sensations. Human observers can show surprising lawfulness when assigning numbers to sound pressure level, contrast, and so forth [111]. Some psychophysicists have been suspicious of magnitude estimation, and instead prefer "matching" tasks.

Continuous matching requires us to adjust a continuous control parameter, θ, to match some aspect of the target:

$$H_\theta : \mathbf{S} = \mathbf{S}(\theta), \ \textit{where } \mathbf{I}_\theta = F(\mathbf{S}(\theta), \alpha)$$

In a *discrete matching task* or "match-to-sample", $\mathbf{I}$, corresponding to $\mathbf{S}$ is observed, and a collection of hypotheses are represented by their observed outputs, $\mathbf{I}_i$. The task is to choose i that gives the best match between the target and comparison:

$$H_i : \mathbf{S} = \mathbf{S}_i, \ \textit{where } \mathbf{I}_i = F(\mathbf{S}_i, \alpha)$$

.

Matching has been used with great success to discover the mechanisms of early color vision ([81],[53]).

The "same/different" task is closely related, but checks for a match or not between two samples: $\mathbf{S}_2 \neq \mathbf{S}_1$?. Like color matching, the observer is usually required to ignore some aspect of information. For example, one might have to decide whether $\mathbf{I}_2$ is from the same object, $\mathbf{S}$, as is $\mathbf{I}_1$, given some unknown change in viewpoint α. Psychophysical results from such experiments can provide tests of hypotheses regarding the mechanisms vision has to allow for variations in the image due to secondary variables.

[9]For both detection and discrimination, a rating scale method can be used in which the confidence of a response to a fixed signal strength $\mathbf{S}$ is broken up into discrete bins in which one assigns numbers answering the question: "Was the signal present?" (E.g. 1=no chance; 2=probably not ; 3 = can't tell; 4=probabily; 5=certainly). These data are used to construct an *Receiver Operating Characteristic* or ROC curve (see next chapter). There are well-known theoretical relationships between the yes/no, 2AFC and rating scale measurements of sensitivity ([76] [45]).

Given a vector variable, $\mathbf{S}$, one can classify, discriminate or estimate with respect to some projection of $\mathbf{S}$. In color experiments, an observer may be asked to match or discriminate along the hue dimension. This adds the complication of how the human task is affected by the requirement to do projection: E.g. color matches with 3 "knobs" for hue, saturation, and brightness aren't necessarily the same as matches with only one "knob", say for hue.

We mentioned earlier that the *two-alternative forced-choice* (2AFC) task has become a standard tool in threshold measurements. Here, both $\mathbf{I}_1$ and $\mathbf{I}_2$ are observed, and the task is to decide which was due to, say, case $i = 1$. So, for the continuum case, one hypothesis is:

$$H_1, \mathbf{S}_2 > \mathbf{S}_1, \ \textit{where } \mathbf{I}_i = F(\mathbf{S}_i, \alpha)$$

and the other hypothesis is the null, $H_0, \mathbf{S}_2 = \mathbf{S}_1$.

5 Bayesian estimation: Putting the pieces together

It should be stressed that different tasks require us to perform different probabilistic inferences. In many of these cases we are only concerned with a subset of the random variables which specify the problem. The input image I may be generated by two factors F, C by $P(I|F, C)$ and with priors $P(F, C)$. For example, the variable F might correspond to the identity of the target and the variable C to its color in the image.

An example in which coupling between two factors plays an important role is provided in the following simple Bayes net. This net (see figure (12) represents the problem of determining the identity and color of a fruit based on noisy shape and color measurements derived from a color image of the fruit. The fruit F assumes value a for apple or t for tomato, and its color C is either r for red or g for green. The shape measurement I_S yields an apple or tomato response, denoted a or t, and the color measurement I_C yields a red or green response, r or g. Because the measurements are noisy their responses may not reflect the true identity and color of the fruit.

The prior probabilities are specified as follows. $P(F = a) = 9/16$ and $P(F = t) = 7/16$. The conditional probabilities of color given fruit identity are: $P(C = r|F = a) = 5/9$ and $P(C = g|F = a) = 4/9$ for an apple and $P(C = r|F = t) = 6/7$ and $P(C = g|F = t) = 1/7$ for a tomato. The joint prior probability of F and C are then given by $P(F, C) = P(F)P(C|F)$, which equals 5/16 for (a,r), 4/16 for (a,g), 6/16 for (t,r) and 1/16 for (t,g). Thus the prior probability on color alone is $P(C = r) = 11/16$ and $P(C = g) = 5/16$, and we notice that F and C are not independent, i.e. $P(F, C) \neq P(F)P(C)$, but are coupled.

The imaging probabilities $P(I_S|F)$ and $P(I_C|C)$ relate identity and color to shape and color measurements and are listed here (in pairs for conciseness):

$P(I_S = a, t|F = a) = 11/16, 5/16$

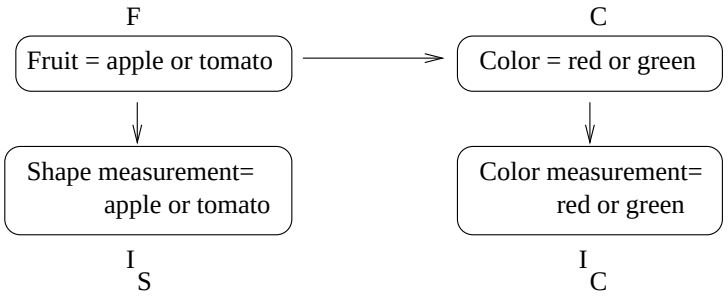

Figure 12 Simple Bayes net relating a fruit and its color to shape and color measurements. The arrows in the net express the form of the joint distribution on all the variables: $P(F, C, I_S, I_C) = P(F)P(C|F)P(I_S|F)P(I_C|C)$. Given the measurements I_S and I_C, estimates of the identity F and color C of the fruit depend on whether F and C are estimated jointly or separately.

$P(I_S = a, t|F = t) = 5/8, 3/8$

$P(I_C = r, g|C = r) = 9/16, 7/16$

$P(I_C = r, g|C = g) = 1/2, 1/2$

Let us now assume that for one particular fruit sample the shape measurement is $I_S = a$ and the color measurement is $I_C = r$. (The measurements alone suggest "red apple" as the most likely interpretation, but we must also taken to account the prior information $P(F, C)$.) The posterior probability $P(F, C|I_S = a, I_C = r)$ is then proportional to $P(F, C, I_S = a, I_C = r) = P(F)P(C|F)P(I_S = a|F)P(I_C = r|C)$, which equals $495/16^3$ for (a, r), $352/16^3$ for (a, g), $540/16^3$ for (t, r) and $80/16^3$ for (t, g). Note that the constant of proportionality is independent of F and C, so we can normalize to obtain $P(F, C|I_S = a, I_C = r) = 495/1467$ for (a, r), $352/1467$ for (a, g), $540/1467$ for (t, r) and $80/1467$ for (t, g).

F and C may be estimated jointly as the MAP (maximum a posterior) estimate of $P(F, C|I_S = a, I_C = r)$, i.e. $(F^*_{joint}, C^*_{joint}) = (t, r)$ – the most likely interpretation of the data is that the true fruit is a red tomato. However, if we estimate F and C separately we will *not* obtain $F^*_{separate} = t$ and $C^*_{separate} = r$. Instead, $P(F|I_S = a, I_C = r) = \sum_C P(F, C|I_S = a, I_C = r)$, yielding $P(F = a, t|I_S = a, I_C = r) = 847/1467, 620/1467$ and hence $F^*_{separate} = a$. Similarly, $P(C|I_S = a, I_C = r) = \sum_F P(F, C|I_S = a, I_C = r)$, yielding $P(C = r, g|I_S = a, I_C = r) = 1035/1467, 432/1467$ and hence $C^*_{separate} = r$. The coupling between identity and color, combined with the unreliability of the measurements, means that, although the most likely single interpretation of the data is "red tomato", the most likely identity overall is "apple" when we allow for the possibility of it being either red *or* green.

If we want to estimate both target identity and color then we must apply Bayes rule to get a *posterior* probability for both F, C. More precisely, we must estimate F, C from:

$$P(F, C|I) = \frac{P(I|F, C)P(F, C)}{\sum_{F',C'} P(I|F', C')P(F', C')}. \tag{4}$$

If, however, we only care about the target identity F then we must sum over all possible colors C of the target and then estimate F from:

$$P(F|I) = \sum_{C} P(F, C|I). \tag{5}$$

In general, the value F^*_{joint} we which obtain from $P(F, C|I)$ when estimating F, C together will differ from the value $F^*_{separate}$ obtained by estimating F on its own. This result sometimes goes against our intuitions. Surely detecting whether a target is present in an image must first require us to determine its color? From a Bayesian perspective, performing the detection optimally requires us to sum out over all the colors which the target might have in the image. In mathematical terms this requires us to *marginalize* our distribution by summing out the random variables whose values we are not concerned with–the secondary variables. In many situations, the contribution to this sum will be strongly dominated by colors of the target which are close to the correct color. The result will then be almost identical to estimating the color and the target identity simultaneously. But there will be situations where the results will differ in important ways. An important case, which we will discuss in the next chapter, concerns the *accidental viewpoint* assumption whose interpretation, from a probabilistic perspective, critically depends on whether variables are summed out or not.

Marginalization over secondary or unneeded variables plays an important role in Bayesian inference as we will illustrate throughout this book. Indeed it is perhaps the most important inference tool and the computational complexities of many models depend critically on whether marginal distributions can be calculated in reasonable time.

Key points:

- Vision, and other forms of perceptual inference, can be thought of in terms of decoding input image signals in order to extract information and determine appropriate actions.
- Realistic images are generated in complex ways involving multiple factors. It is usually difficult to model such problems within the standard target/noise task. In particular, typical images require an immense number of random variables to model them.
- Images and the World consist of patterns. These patterns must be described and understood. Although patterns are complex there are regularities and, in particular, a limited number of transformations which constantly appear.
- In Bayesian models the objects of interest, both in the image and in the scene, are represented by random variables. These probability distributions should represent the important properties of the domain and should be learnt or estimated if possible. Stochastic sampling can be used to judge the realism of the distributions.
- Visual inference about the world would be impossible if it were not for regularities occuring in scenes and images. The Bayesian approach gives a way of encoding these assumptions probabilistically. This can be interpreted in terms of obtaining the simplest description of the input signal and relates to the idea of vision as information processing.
- The Bayesian approach separates the probability models from the algorithms required to make inferences from these models. This makes it possible to define ideal observers and put fundamental bounds on the ability to perform visual tasks *independently* of the specific algorithms used.
- Various forms of inference can be performed on these probability distributions. The basic elements of inference are marginalization and conditioning.
- Probability distributions on many random variables can be represented by graph structures with direct influences between variables represented by links. The more complex the vision problem, in the sense of the greater direct influence between random variables, the more complicated the graph structure.
- The purpose of vision is to enable an agent to interact with the world. The decisions and actions taken by the agent, such as detecting the presence of certain objects or moving to take a closer look, must depend on the importance of these objects to the agent. This can be formalized using concepts from decision theory and control theory.

- Computer vision modelers assume that the uncertainty lies in the scene and pay less attention to the image capturing process. By contrast, biological vision modelers have paid a lot of attention to modeling the noise in the photoreceptors and retina – and less on the scene.
- There are certain precise mathematical tasks which we need to compute: Detectability, Discrimination, Localization, Identification/Classification/Estimation. These involve various forms of inference.

References

[1] Adelson, E. H., & Pentland, A. P. (1996). The Perception of Shading and Reflectance. In Knill, D. C., & Richards, W. (Ed.), *Perception as Bayesian Inference*(pp. 409-423). Cambridge: Cambridge University Press.

[2] D.J. Amit. **Modeling Brain Function**. Cambridge Unniversity Press. 1989.

[3] M.A. Arbib (Ed.). **The Handbook of Brain Theory and Neural Networks.** A Bradford Book. The MIT Press. 1995.

[4] Atick, J. J., & Redlich, A. N. "What does the retina know about natural scenes?". *Neural Computation*. 4,(2), 196-210. 1992.

[5] Attneave, F. "Some Informational Aspects of Visual Perception.". *Psychological Review*. 61,((3)), 183-193. 1954.

[6] R. Balboa. PhD Thesis. Department of Computer Science. University of Alicante. Spain. 1997.

[7] Barkow, J. H., Cosmides, L., & Tooby, J. (1992). The Adapted Mind. Oxford: Oxford University Press.

[8] Barlow, H. B. (1959). "Sensory mechanisms, the reduction of redundancy, and intelligence". National Physical Laboratory: HMSO, London, 1959.

[9] Barlow, H. B. "A method of determining the overall quantum efficiency of visual discriminations.". *J. Physiol. (Lond.)*. 160, 155-168. 1962.

[10] Barlow, H. B. (1977). Retinal and central factors in human vision limited by noise. In B., B. H.,& P., F. (Ed.), *Photoreception in Vertebrates*(pp. Academic Press.

[11] Barlow, H. B., & Reeves, B. C. "The versatility and absolute efficiency of detecting mirror symmetry in random dot displays.". *Vision Research*. 19, 783-793. 1979.

[12] Barlow, H. B. "The absolute efficiency of perceptual decisions". *Phil. Trans. R. Soc. Lond. B*. 290, 71-82. 1980.

[13] Barlow, H. B., & Foldiak, P. "Adaptation and decorrelation in the cortex". In Miall, C., Durban, R. M., & Mitchison, G. J. (Ed.), *The Computing Neuron*(pp. Addison-Wesley. 1989.

[14] M. Barzohar and D. B. Cooper, "Automatic Finding of Main Roads in Aerial Images by Using Geometric-Stochastic Models and Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 459-464, 1993.

[15] Beck, J., Prazdny, K., & Rosenfeld, A. "A Theory of Textural Segmentation.". In Beck, J., Hope, B.,& Rosenfeld, A. (Ed.), *Human and Machine Vision*(pp. 1-38). New York: Academic Press. 1983.

[16] P. N. Belhumeur, "A binocular stereo algorithm for reconstructing sloping, creased and broken surfaces, in the presence of half-occlusion", *Proc. ICCV*, Berlin, 1993.

[17] R. E. Bellman, *Applied Dynamic Programming.* Princeton University Press, 1962.

[18] Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**,115-147.

[19] A. Blake and A. Zisserman. Visual Reconstruction. MIT Press. 1987.

[20] Blake, A., Bulthoff, H., & Sheinberg, D. "Shape from Texture: Ideal Observers and Human Psychophysics", 1992.

[21] Bossomaier, T., & Snyder, A. W. "Why Spatial Frequency Processing in the Visual Cortex?". *Vision Research.* 26,(8), 1307-1309. 1986.

[22] Burgess, A. E., & Abbey, C. K. "Visual signal detectability with two noise components: anomalous masking effects". *J. Opt. Soc. Am. A.* 14, 2420 -2442. 1997.

[23] Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. "Efficiency of human visual signal discrimination.". *Science.* 214, 93-94. 1981.

[24] Chomsky, N. (1956). "Three models for the description of language." *IRE Transactions on Information Theory*, 2 (3), 113-124.

[25] J. Coughlan, D. Snow, C. English, and A.L. Yuille. "Efficient Optimization of a Deformable Template Using Dynamic Programming". In *Proceedings Computer Vision and Pattern Recognition. CVPR'98.* Santa Barbara. California. 1998.

[26] Cornsweet, T. N. (1970). **Visual Perception**. New York: Academic Press.

[27] Cosmides, L. (1989). "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task." *Cognition*, **31**,187-276.

[28] T.M. Cover and J.A. Thomas. **Elements of Information Theory**. Wiley Interscience Press. New York. 1991.

[29] Cross, G. R. and Jain, A. K. "Markov random field texture models." *IEEE, PAMI*, **5**, 25-39. 1983.

[30] M.H. DeGroot. **Optimal Statistical Decisions**. McGraw-Hill. 1970.

[31] Egan, J. P. (1975). **Signal detection theory and ROC-analysis**. New York: Academic Press.

[32] Elder, J., & Zucker, S. "Evidence for Boundary-Specific Grouping.". *Vision Research.* 38,(1), 143-152. 1998.

[33] Field, D. J. "Relations between the statistics of natural images and the response properties of cortical cells". *Journal of the Optical Society*4,(12), 2379-2394. 1987.

[34] Field, D. J., Hayes, A., & Hess, R. F. "Contour integration by the human visual system: evidence for a local "association field"". *Vision Research.* 33,173-193. 1993.

[35] Field D.J, Hayes, A. and Hess, R.F. "Contour Integration by the Human Visual System: Evidence for a Local Association Field". *Vision Research.* 33, pp 173-193. 1993.

[36] Fisher, R. A. (1925). **Statistical Methods for Research Workers**, Edinburgh: Oliver and Boyd.

[37] D. Geiger, B. Ladendorf, and A.L. Yuille. "Occlusions and Binocular Stereo". In *Proc. ECCV.* Genoa, Italy. 1992.

[38] D. Geiger and T-L Liu. "Top-Down Recognition and Bottom-Up Integration for Recognizing Articulated Objects". In *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition.* Ed. M. Pellilo and E. Hancock. Venice, Italy. Springer-Verlag. May. 1997.

[39] Geisler, W. "Sequential Ideal-Observer analysis of visual discriminations". *Psychological Review.* 96,(2), 267-314. 1989.

[40] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". IEEE Trans. on PAMI 9(7), pp 721-741. 1984.

[41] D. Geman and G. Reynoids, "Constrained restoration and the recover of discontinuities", *IEEE Trans. PAMI*, vol.14, pp.367-383, 1992.

[42] D. Geman, S. Geman, C. Graffigne and P. Dong, "Boundary detection by constrained optimization", *IEEE trans on PAMI*, vol.12, No.7, July, 1990.

[43] D. Geman. and B. Jedynak. "An active testing model for tracking roads in satellite images". *IEEE Trans. Patt. Anal. and Machine Intel.* Vol. 18. No. 1, pp 1-14. January. 1996.

[44] Goldstein, E. B. (1995). "Sensation and Perception", (4th ed.). Belmont, California: Wadsworth Publishing Company. 1995.

[45] Green, D. M., & Swets, J. A. (1974). "Signal Detection Theory and Psychophysics". Huntington, New York: Robert E. Krieger Publishing Company. 1974.

[46] Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv. Mathematik*, **1**, (17), 195.

[47] U. Grenander, Y. Chow, and D.M. Keenan. *Hands. A Pattern Theoretic Study of Biological Shapes*. Springer:New York. 1991.

[48] U. Grenander. General Pattern Theory, Oxford Univ Press. 1993

[49] Hecht, S., Shlaer, S., & Pirenne, M. H. (1942). Energy, quanta, and vision. *Journal of General Physiology*, **25**, 819-840.

[50] J. Hertz, A. Krogh, and R.G. Palmer. "Introduction to the Theory of Neural Computation". Add-son-Wesley. 1991.

[51] P.Huber, 1981, Robust Statistics, Wiley and Sons.

[52] M. Isard and A. Blake. "Contour tracking by stochastic propagation of conditional density". *Proc. European Conf. Comput. Vision*, pp. 343-356, Cambridge, UK. 1996.

[53] Hurvich, L. M. (1981). **Color vision**. Sunderland, Mass.: Sinauer Associates.

[54] D.W. Jacobs. "Robust and Efficient Detection of Salient Convex Groups". *IEEE Trans. Patt. Anal. and Machine Intel.* Vol. 18. No. 1, pp 23-37. January. 1996.

[55] E. T. Jaynes. "Information theory and statistical mechanics". *Physical Review.* **106**, pp620-630, 1957.

[56] Julesz, B. "Experiments in the Visual Perception of Texture.". *Scientific American.* 232,(4), 34-43. 1975.

[57] Julesz, B. "A brief outline of the texton theory of human vision". *Trends in Neuroscience.* 7, 41-45. 1984.

[58] Julesz, B., Gilbert, E. N., & Victor, J. D. "Visual Discrimination of Textures with Identical Third-Order Statistics.". *Biological Cybernetics.* 31,137-140. 1978.

[59] B. Julesz. **Dialogues on Perecption**. 1995.

[60] Kanizsa, G. (1979). "Organization in Vision". New York: Praeger. 1979.

[61] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active Contour models". In *Proc. 1st Int. Conf. on Computer Vision.* 259-268. 1987.

[62] D. Kendall, "A survey of the statistical theory of shape", Statistical Science, Vol.4, No.2, pp 87-120, 1989.

[63] Kersten, D. "Spatial summation in visual noise". *Vision Research.* 24,1977-1990. 1984.

[64] Kersten, D. "Statistical efficiency for the detection of visual noise". *Vision Research*. 27,(6), 1029-1040. 1987.

[65] Kersten, D. J. (1987). Predictability and Redundancy of Natural Images. *Journal of the Optical Society of America*, **4**, 2395-2400.

[66] Kersten, D. "Statistical limits to image understanding". In Blakemore, C. (Ed.), *Vision: Coding and Efficiency*(pp. 32-44). Cambridge, UK: Cambridge University Press. 1990.

[67] Kersten, D. (1998). High-level vision as statistical inference. In Gazzaniga, M. (Ed.), *The Cognitive Neurosciences* Cambridge, MA: MIT Press.

[68] Knill, D. C., Field, D., & Kersten, D. "Human discrimination of fractal images". *Journal of the Optical Society of America, A,*. 7, 1113-1123. 1990.

[69] Knill, D. C., D. Field, and D. Kersten, 1990. Human discrimination of fractal images. Journal of the Optical Society of America, A, 7: 1113-1123.

[70] D. C. Knill, "Discrimination of planar surface slant from texture: Human and ideal observers compared". *Vision Research*. In press.

[71] D. C. Knill, "Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues". *Vision Research*. In press.

[72] D.C. Knill and W. Richards. (Eds). **Perception as Bayesian Inference**. Cambridge University Press. 1996.

[73] Kovcs, I., & Julesz, B. "A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation". *Proc. Natl. Acad. Sci. USA*. 90, 7495-7497. 1993.

[74] S. Lauritzen and D.J. Spiegelhalter. "Local Computations with probabilities on graphical structures and their application to expert systems". *Journal of the Royal Statistical Society*, B, pp 157-224. 1988.

[75] Liu, Z., Knill, D. C., & Kersten, D. "Object Classification for Human and Ideal Observers". *Vision Research*. 35,(4), 549-568. 1995.

[76] Macmillan, N. A., & Creelman, C. D. (1991). **Detection theory : a user's guide**, . Cambridge [England] ; New York: Cambridge University Press.

[77] Malik, J., & Perona, P. "Preattentive texture discrimination with early vision mechanisms". *Journal of the Optical Society of America A*. 7,(5), 923-932. 1990.

[78] J. Shi and J. Mailk. "Normalized cuts and image segmentation". In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. pp 731-737. 1997.

[79] S. Mallat. "Multi-resolution approximations and wavelet orthonormal bases of $L^2(R)$. Trans. Amer. Math. Soc. 315, pp 69-87. 1989.

[80] Marr, D. (1982). **Vision: A Computational Investigation into the Human Representation and Processing of Visual Information**, San Francisco: Freeman.

[81] Maxwell, J. C. (1855). Experiments on colour, as perceived by the eye, with remarks on colour-blindness. *21*, (2), 275-298.

[82] Mohan, R., & Nevatia, R. "Using Perceptual Organization to Extract 3-D Structures". *IEEE PAMI*. 11, 1121–1139. 1989.

[83] U. Montanari. "On the optimal detection of curves in noisy pictures." *Communications of the ACM*, pages 335–345, 1971.

[84] Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition*, **in press**.

[85] D. Mumford and J. Shah. "Optimal approximations by piecewise smooth functions and associated variational problems". Comm. Pure Appl. Math. 42, pp 577-684. 1989.

[86] Nakayama, K., & Shimojo, S. "Experiencing and perceiving visual surfaces". *Science*. 257,1357-1363. 1992.

[87] Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. London, Series A*, , 289.

[88] Olshausen, B. A., & Field, D. J. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". *Nature*. 381, 607-609. 1996.

[89] D.H. Paul. "Speech Recognition using Hidden Markov Models." The Lincoln Laboratory Journal, Volume 3, Number 1. 1990.

[90] J. Pearl. **Heuristics**. Addison-Wesley. 1984.

[91] J. Pearl. **Probabilistic Reasoning in Intelligent Systems.** Morgan Kaufmann Publishers. Inc. San Mateo, California. 1988.

[92] Pelli, D. G. "The quantum efficiency of vision". In Blakemore, C. (Ed.), *Vision:Coding and Efficiency*(pp. Cambridge: Cambridge University Press. 1990.

[93] P. Perona and J. Malik. "Scale Space and edge detection using anisotropic diffusion." *Proc. 5th IEEE Work. Computer Vision.* Miami. 1987.

[94] P. Perona and J. Malik. "Detecting and localizing edges composed of steps, peaks and roofs". Proceedings CVPR. 1990.

[95] Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Trans. IRE Professional Group on Information Theory*, **PGIT-4**, 171-212.

[96] M. Pettet, S. McKee and N.M. Grzywacz. "Constraints on Long Range Interactions Mediating Contour Detection". Vision Research. **38**, pp 865-880. 1998.

[97] S.B. Pollard, J. Mayhew and J. Frisby. "PMF: A stereo correspondence algorithm using a disparity gradient limit". *Perception*, 23. pp 449-470. 1985.

[98] T. Poggio and K. Sung. Example-based learning for view-based human face detection. In *Proc. Image Understanding Workshop*, pages II:843–850, 1994.

[99] T. Poggio and F. Girosi. "Regularization algorithms for learning that are equivalent to multilayer networks". *Science*. 247, pp 978-982. 1990.

[100] Rice, S. O. (1944). Mathematical Analysis of Random Noise. *Bell System Technical Journal*, **23**, , 282-332.

[101] B. Ripley. "Pattern Recognition and Neural Networks". Cambridge University Press. 1996.

[102] J. Rissanen. **Stochastic Complexity in Statistical Inquiry.** Singapore: World Scientific Publishing Co. 1989.

[103] Rock, I., & Palmer, S. "The legacy of Gestalt Psychology". *Scientific American*. 263,(6), 84-90. 1990.

[104] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

[105] Ruderman, D., & Bialek, W. "Statistics of Natural Images: Scaling in the Woods". *Physical Review Letters*. 73, Number 6,(8 August 1994), 814-817. 1994.

[106] S. Russell and P. Norvig. "Artificial Intelligence: A Modern Approach. Prentice-Hall. 1995.

[107] Srinivasan, M. V., Laughlin, S. B., & Dubs, A. "Predictive coding: A fresh view of inhibition in the retina". *Proc. Roy. Soc. Lond. B*. 216, 427-459. 1982.

[108] Shannon, C. E., & Weaver, W. (1949). "The mathematical theory of communication", Champaign, IL: U. Illinois Press. 1949.

[109] Shannon, C. (1951). Prediction and entropy of printed English. *Bell. Sys. Tech. J.*, **30**, 50-64.

[110] inha, P. & Adelson, E. (1993). Recovering reflectance and illumination in a world of painted polyhedra. *Proceedings of Fourth International Conference on Computer Vision* Berlin: 156-163.

[111] Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, **54**, 377-411.

[112] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, (4857), 1285-93.

[113] Tanner, W. P. J., & Birdsall, T. G. (1958). Definitions of d' and n as psychophysical measures. *Journal of the Acoustical Society of America*, **30**, 922-928.

[114] K.K. Thornber and L.R. Williams. "Characterizing the Distribution of Completion Shapes with Corners Using a Mixture of Random Processes". In *Energy Minimization Methods in Computer Vision and Pattern Recognition.* Eds. M. Pelillo and E.R. Hancock. Springer-Verlag Lecture Notes in Computer Science. 1997.

[115] Tjan, B., Braje, W., Legge, G. E., & Kersten, D. "Human efficiency for recognizing 3-D objects in luminance noise". *Vision Research*. 35,(21), 3053-3069. 1995.

[116] Treisman, A. M., & Gelade, G. "A feature-integration theory of attention". *Cognitive Psychology*. 12,97-136. 1980.

[117] M. Turk and A. Pentland. " Eigenfaces for recognition." *J. of Cognitive Neuroscience*, 3(1), 1991.

[118] . C.W. Tyler (Ed.). **Human Symmetry Perception and its Computational Analysis.** VSP. Utrecht. 1996.

[119] Victor, J. D. "Images, statistics, and textures: implications of triple correlation uniqueness for texture statistics and the Julesz conjecture: comment". *Journal of Optical Society of America, A*. 11,(5), 1680-1684. 1994.

[120] Watson, A. B., H. B. Barlow, and J. G. Robson, 1983. What does the eye see best? Nature 31,: 419-422.

[121] Watson, A. B., Barlow, H. B., & Robson, J. G. "What does the eye see best?". *Nature*. 31, 419-422. 1983.

[122] Wolfe, J. M. ""Effortless" texture segmentation and "parallel" visual search are not the same thing". *Vision Research*. 32,(4), 757-763. 1992.

[123] A.L. Yuille. "Deformable Templates for Face Recognition". *Journal of Cognitive Neuroscience*. Vol. 3, No. 1. 1991.

[124] A.L. Yuille and J. Coughlan. "An A* perspective on deterministic optimization for deformable templates". To appear. Pattern Recognition Letters. 1997.

[125] A.L. Yuille and S.C. Zhu. "Geometrical Interpretation of Minimax and Discrimination". Preprint. Smith-Kettlewell Eye Research Institute. 1997.

[126] A.L. Yuille and J. Coughlan. "Twenty Questions, Focus of Attention, and A*": A theoretical comparison of optimization stratregies". In *Proc. International Workshop on Energy Minimization Methods in Computer Vision*. Venice, Italy. 1997.

[127] S.C. Zhu and A.L. Yuille. "A Unified Theory for Image Segmentation: Region Competition and Its Analysis". Harvard Robotics Laboratory Technical Report 95-3. 1995.

[128] S.C. Zhu. *Statistical and Computational Theories for Image Segmentation, Texture Modeling, and Object Recognition*, PhD Thesis. Division of Applied Sciences. Harvard University. 1996.

[129] S.C. Zhu and A.L. Yuille. "A Flexible Object Recognition and Modelling System". International Journal of Computer Vision. 20(3). pp 187-212. 1996.

[130] S.C. Zhu and A.L. Yuille. "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation". IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 18, No. 9. pp 884-900. 1996.

[131] S.C. Zhu, Y. Wu, and D. Mumford. "Minimax Entropy Principle and Its Application to Texture Modeling". Neural Computation. Vol. 9. no. 8. Nov. 1997.

[132] S.C. Zhu and D. Mumford. "Prior Learning and Gibbs Reaction-Diffusion". IEEE Trans. on PAMI vol. 19, no. 11. Nov. 1997.

[133] S.C. Zhu and D. Mumford. "GRADE: A framework for pattern synthesis, denoising, image enhancement, and clutter removal." In Proceedings of International Conference on Computer Vision. Bombay. India. 1998.

[134] S-C Zhu, Y-N Wu and D. Mumford. FRAME: Filters, Random field And Maximum Entropy: — Towards a Unified Theory for Texture Modeling. Int'l Journal of Computer Vision 27(2) 1-20, March/April. 1998.

[135] S.C. Zhu. "Embedding Gestalt Laws in Markov Random Fields". Submitted to *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*.