Vision: Bayesian Inference and Beyond

Daniel Kersten[1,3] and Alan Yuille[2,3]
[1]Department of Psychology, University of Minnesota
[2]Departments of Statistics and Psychology, University of California, Los Angeles
[3]Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea

Introduction

Although research has provided an enormous amount of knowledge about visual brain anatomy, physiology and neural mechanisms, this knowledge is insufficient to quantitatively 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 neurobiology of visual circuitry must be augmented by understanding visual processing at a more abstract level (Figure 1). Such understanding should take into account behavioral functions or tasks, and the computational problems the organism must solve. We get insight into tasks through the study of how animals, such as ourselves, use vision (Milner & Goodale, 2006) as well as their relationships to the development of computer vision systems that are required to achieve specific goals (Ullman, 2000).

Published in: **The New Visual Neurosciences (2013)**. MIT Press, Cambridge MA. John S. Werner and Leo M. Chalupa (Editors). Comments may be sent to the author at kersten@umn.edu. D.K. and A.Y. were supported by the WCU (World Class University) program funded by the Ministry of Education, Science and Technology through the National Research Foundation of Korea (R31-10008).

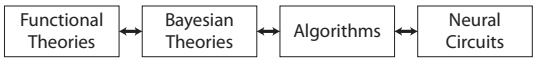

Figure 1. Probabilistic/Bayesian theories fall between the functional ("computational") and algorithmic levels of analysis in Marr's categorization of three levels of analysis for the study a complex information processing system such as vision (Marr, 1982).

This chapter describes how visual processing can be modeled and understood in terms of *probabilistic inference*, or equivalently, as a decoding problem where the goal is to determine information about the world from *image patterns* reaching the eyes. Information gathering, however, is not a passive process and depends on the goals and abilities of the *organism* performing the decoding. From this perspective, three important questions are: 1) what types of image patterns occur? 2) what information can be extracted from these patterns for a given task? and 3) how does the organism's inferences compare with optimal inference given answers to 1) and 2)? (See Fig. 2.)

At this abstract level, the primary concern is with how the required computations and algorithms for inference and learning constrain neural processing. This priority is arguably a 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" (Marr, 1982). Theories of

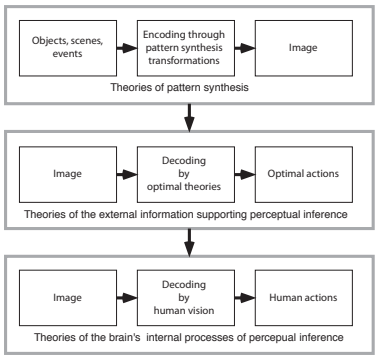

Figure 2. Three levels of study in perceptual inference.

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.

Advances in applied mathematics, probability theory, information theory, artificial neural networks, and artificial intelligence have helped to develop a *Bayesian* framework for approaching vision as well as other problems, including sensorimotor control (Körding & Wolpert, 2006; Schlicht & Schrater, 2007; Battaglia & Schrater, 2007; Franklin & Wolpert, 2011) and cognition (Griffiths et al., 2010). This framework includes a common language for describing vision problems, mathematical techniques for modeling them, and algorithms that can be applied to solve them. These advances have been facilitated by the enormous increase in computer power which has made it possible to explore increasingly complicated probability models.

The Bayesian framework reduces to standard signal detection theory as a special case (Green & Swets, 1966), but goes beyond it in its range of applicability, the power of its techniques, and the kinds of questions asked (Kersten & Schrater, 2002; Chater et al., 2006). The development of signal detection theory in the 1950s was strongly motivated by the idea that errors in psychophysical decisions varied because of noise and bias inside the observer. However, not long after computer vision began attempts to mimic human perception in the late 1960s, it became apparent that the fundamental limit to accurate and reliable perceptual decisions (by machine or organism) was not noise in the internal processing or sensory measurements, but rather the ambiguity in what local image measurements implied about the external world. This problem was exacerbated with the realization that useful information, i.e. signals like "curved line", "round shape", or "baseball player", require substantial computation to be decrypted from input image patterns. While understanding how the brain deals with noisy input and neurons is important, this chapter focuses on elements of the Bayesian program that are important for studying vision when internal noise can be neglected, and where the challenge is to discover and extract important information given the inherent ambiguities and complexities of images.

The past decade has produced a substantial body of research interpreting human behavior and neural coding from a Bayesian framework[1]. For practi-

[1] For a general introduction to Bayesian inference in cognitive science see Griffiths & Yuille (2008), for Bayesian models applied to ideal observer psychophysics and natural systems analysis, Geisler (2008, 2011), for an overview and critique applied to cognitive neuroscience, Colombo & Seriès (2012), for Bayesian decision theory, Körding (2007); Maloney & Mamassian (2009), to optimal learning Fiser et al. (2010), and for an earlier review of object perception, see Kersten et al. (2004). For examples of applications of ideal observers to learning, attention, and letter detection, see Trenti et al. (2010); Eckstein et al. (2006); Pelli et al. (2006), to saliency and eye movements (Torralba et al., 2006; Itti & Baldi, 2009; Zhang et al., 2008; Chikkerur et al., 2010), to contour and shape (Feldman, 2001; Feldman & Singh, 2005; Wilder et al., 2011), filling-in (Zhaoping & Jingling, 2008), and to cross-modal interactions see Battaglia et al. (2011); Körding et al. (2007); Shams (2010). For neural coding, see Pouget et al. (2000); Knill & Pouget (2004); Fiser et al. (2010); Ma (2010); Berkes et al. (2011).

cal experimental reasons, this work has focused on modular problems, where the tasks are limited and the stimulus descriptions simple. Yet, human vision deals with images that are enormously complex and for a wide range of tasks. Our aim in this chapter is to show how recent developments in Bayesian models can provide insight into how the visual system deals with the complexities of natural images, and its relation to the hierarchical organization of the visual cortex.

The rest of this chapter is divided into three main sections: 1) *Bayesian Modeling* describes basic probabilistic concepts and operations; 2) *Dealing with image complexity and task flexibility* develops ideas from computer vision to questions of ventral stream feature hierarchy; 3) *Behavioral and neural evidence for Bayesian computations* describes results of several experiments consistent with hierarchical, probabilistic computations in the visual system.

Bayesian Modeling

Basic ingredients. We assume that the knowledge required for perception is represented in terms of a probability distribution, $p(I, s)$, over random variables that are measurable, I, and hidden variables, s, some or all of which need to be estimated. [2] At the most basic level, we think of measurable variables as the intensity pattern at the eye, I. But measurables are often assumed to be features that are "easily" computed from the image early in processing. Deciding what is a measurable feature and what is a state to be inferred is a modeling assumption. For example, estimating the distance of an object from an image requires a "measurement" of angular size; however, determining the angular size of an object is itself a non-trivial inference in a natural image with background clutter.

Hidden variables can represent interpretable, external states of the world such as objects and events, or more abstract causes or "latent variables" that capture regularities in the image.

Using the product rule to *condition* the joint distribution on a measurement I, reduces the uncertainty about s:

$$p(s|I) = p(s, I)/p(I)$$

This distribution is called the posterior distribution of s on I, and is the basis for optimal inference.

Typically more than one cause influences a measurable image variable (Figure 3B). One component of cause s in $s = (s_1, s_2)$ may be important to estimate, whereas another is a confounding or "nuisance" variable to be discounted (Figure 3C). For example, suppose the image intensity is $I = s_1 \cdot s_2$, with s_1 and s_2 representing reflectance ("grayness" of surface) and illumination (level of light falling on a surface), respectively. The pattern of light, s_2, falling on an object is usually considered a nuisance variable, whereas s_1 is important because it is an invariant surface property useful for object recognition (e.g. "is it a white or black piece of paper?"). Like conditioning, discounting can reduce uncertainty about the remaining variables, through application of the "sum rule", called marginalization:

$$p(s_1|I) = \sum_{s_2} p(s_1, s_2|I).$$

Because the problem of confounding variables is the rule rather than the exception, marginalization can be viewed as a basic operation for inference that underlies useful decisions. Its effects could be built in, as in distinct visual pathways specialized for different functions (Milner & Goodale, 2006; Fang & He, 2005). Or it could involve dynamic neural processes that adapt to the task at hand. Freeman

[2] The term "hidden" emphasizes the fact that the values of states of the world are effectively encrypted and need to be decoded. This is the central mystery of perception. The shape of an object may *appear* to be a direct measurement (and at some higher level could be treated as such), but computational theory and experiments have shown that inferring shape from image patterns requires inferential processes.

(1994) showed how marginalizing with respect to viewpoint and illumination could substantially reduce uncertainty about shape and material. Integrating out illumination direction can also disambiguate depth from cast shadows (Kersten, 1999). Freeman (1994) further showed that marginalization could be viewed as choosing the estimate least sensitive to variations in the confounding variable, providing a Bayesian interpretation of the generic viewpoint principle (see Lowe (1987); Nakayama & Shimojo (1992)). Beck et al. (2011) show that given certain assumptions about spike train statistics, some kinds of marginalization can be achieved through divisive normalization (usually thought of in terms of gain control and attentional processes, Reynolds & Heeger (2009); Carandini & Heeger (2011)).

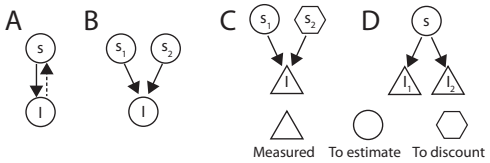

Figure 3. Simple directed graphs. The solid arrows represent conditional probability relationships between parents and descendants. **A**. "Basic Bayes", where the joint is factorized as $p(s, I) = p(I|s)p(s)$. The dashed arrow indicates the bottom-up direction of inference. **B**. More than one cause contributes to measurements, $p(I, s_1, s_2) = p(I|s_1, s_2)p(s_1)p(s_2)$. This kind of graph implies conditional dependence of the causes given a measurement, and marginal independence over measurements. Causes can effectively compete to explain a measurement, a Bayesian phenomenon called "explaining away" (Pearl, 1988). **C**. If only one parent variable is important to estimate, the other one needs to be integrated out, i.e. discounted. **D**. One cause leads to two effects, i.e. image measurements. The graph implies that the two measurements are conditionally independent given s. This generative model is the basis for tests of optimal cue integration.

Generative models. The posterior probability, $p(s|I)$, represents the information required for reliable estimates of states of the world from image input. The posterior could be as simple as a look-up table assigning probabilities to scenes for a given input, without an explicit model of input variation[3]. But visual images are complex, inferences are computationally hard, and there are big advantages to expressing the posterior in terms of generative components, i.e. a description of the regularities in the causes (state variables describing objects, scenes, events, independent of the resulting images) and how those causes create images[4]. To provide some intuition as to why generative knowledge is useful, consider the many images that could result from a pair of gardening shears. Its appearance will vary with lighting, position, 3D orientation and the opening angle. In practice, it is impossible to have lookup table that can anticipate all possible appearances. While some variations may be efficiently handled bottom-up, others such as discounting 3D orientation or opening angle can be much more efficiently handled with object-specific knowledge–i.e. that a shears can rotate in depth, and open and close Ullman (2000). Bayes rule describes the relationship between the posterior and information about how images could be formed. It says:

$$p(s \mid I) = \frac{p(I \mid s)p(s)}{p(I)} = \frac{p(I \mid s)p(s)}{\sum_{s'} p(I \mid s')p(s')}.$$

This theorem re-expresses $p(s \mid I)$, the probability of the state given the measurement, in terms of $p(I \mid s)$, the probability of the measurement given the state (called the *likelihood* of s), and $p(s)$, the probability of the state variables (called the *prior*).

[3] Decisions based on $p(s|I)$ in the absence of a generative model are called *discriminative*. The distinction between discriminative and generative models is related to the distinction made between policy and model-based learning in Bayesian reinforcement learning theory (Strens, 2000).

[4] Generative means that, in principle, one can generate observables I by first drawing a sample from $p(s)$, call it s', and then a sample of I from $p(I|s')$.

It tells us that we can look to models of how image patterns are formed by states of the world (e.g. shape, material, lighting and viewpoint) as well as models of regularities of those states (e.g. surface smoothness, pigmentation, ...).

Elaborating graphical structure. Bayes' formula makes generative knowledge explicit in terms of two basic factors represented as "Basic Bayes" in the graph of Figure 3A; however, more structure is needed to cope with the range of tasks and the complex interdependency of image intensities. It is impossible to do Bayesian inference on high dimensional joint distributions, $p(I_1, I_2, ..., s_1, s_2, ...)$, without appropriate knowledge structuring. A straightforward next step is to consider slightly more complex graphs, for example involving more than one cause (Figure 3B & C), one cause leading to more than one measurement (Figure 3D), and simple 3-level hierarchies (Figure 7).

Consider the case where one cause produces more than one measurement, providing the basis for cue integration (Figure 3D). A distinctive prediction of Bayesian observers is that decisions should be based on full knowledge of the posterior distribution, i.e. including higher-order moments. Following the work of Clark & Yuille (1990), Jacobs (1999), and Ernst & Banks (2002); Ernst & Bülthoff (2004), numerous studies have tested whether humans combine sensory information weighted by reliability (the reciprocal of the second moment, variance). The majority of these studies confirm optimality, with possible exceptions (cf. Cheng et al., 2007; Gori et al., 2008). Most studies of cue integration have been restricted to continuous-valued measurements whose precision is modeled using standard gaussian and independence assumptions, and thus a linear weighting function. Stevenson & Körding (2009) devised a Bayesian ideal observer that uses occlusion, a non-metric cue, and showed that their model accurately predicts human depth judgments.

While small graphs are useful for low-dimensional, modular analyses of visual behavior, substantially more structure is needed to understand recognition with natural images, and we return to this later when we discuss a possible relationship of cortical visual architecture to graphical representations.

Making decisions & estimates. To make a decision, one can choose values of the state variables that maximize the appropriately conditioned and marginalized distribution–this is the *maximum a posterior* (MAP) estimate, which is optimal in the sense of minimizing the average error rate. But human perception is flexible and some tasks may need several estimates, each requiring different levels of precision, and perhaps at different times. To pick up a ball, you need to know how far away and how big it is. Given a measurement of angular size, I, and a constraint $I \approx s_2/s_1$, planning the reach requires a good estimate of distance (s_1), while the actual grasping requires a good estimate of physical size (s_2). Bayesian decision theory generalizes the operation of marginalization and defines optimality for an action as minimizing the risk:

$$R(\alpha; I) = \sum_s L(\alpha, s) p(s \mid I),$$

where the loss function $L(\alpha, s)$ is the cost of deciding to take action α when the true world state is s (see Geisler & Kersten, 2002; Maloney & Mamassian, 2009). To keep things simple, the rest of this chapter assumes that conditioning and marginalization choices are sufficient to characterize different task requirements.

As discussed below, Bayesian models for object recognition represent knowledge as probabilities over complex, but structured graphs. Inference on complex graphs requires message passing algorithms that update local conditional distributions at all nodes by passing information between neighbors, given that the values of some nodes are fixed and others integrated out. Belief propagation is one

such method (Pearl, 1988; MacKay, 2003). Messages that update probabilities rather than passing on decisions respect the Bayesian version of Marr's principle of least commitment (Marr, 1982). The question of whether communication between neural populations observes this principle, (cf. Ma et al., 2006), or instead involves decisions that get passed from one population to the next (cf. Lennie, 1998), is a fundamental unanswered question in brain science.

So far we have treated decisions at an abstract level without distinguishing deliberate from automatic decisions, or those at the level of subprocesses. An intriguing proposal is that perception doesn't actually commit to a decision in the usual sense. Instead a conscious percept reflects a stochastic sample drawn from a posterior distribution. The posterior is represented by a collection of exemplars that approximate the frequencies of occurrence of the perceptual hypothesis conditioned on the measurements (Lee & Mumford, 2003; Fiser et al., 2010). There are a number of behavioral results, including perceptual bistability, consistent with this idea (Sundareswara & Schrater, 2008; Gershman et al., 2012; Moreno-Bote et al., 2011; Battaglia et al., 2011).

In summary, visual knowledge can be conceptualized as a joint distribution over possible images and interpretations. The operations of conditioning and marginalizing narrow the space of possibilities through image measurements and task assumptions, respectively. Optimal decisions require computations that take full account of the knowledge available, including knowledge of uncertainty.

Dealing with image complexity and task flexibility

While Bayesian models have provided compelling accounts of how human visual behavior manages uncertainty for low-dimensional problems, a central problem is to understand how the visual system is organized to deal with the high-dimensionality of raw image data. In the next sections, we compare what is known about the neural basis of object processing and recognition in the ventral stream with Bayesian computer vision systems for recognizing objects in natural images.

The basic assumption is that the organization of cortical areas reflects the structure of natural images for efficient representation, learning, and inference.

Object recognition and ventral stream processing

A large body of results from primate neurophysiology and human neuroimaging is consistent with the idea that visual object processing is based on a hierarchical organization of stages through which image information is successively transformed from a high-dimensional set of local feature measurements with a small number of types (e.g. edges at many locations) to increasingly lower-dimensional representations of many types (e.g. dog, baseball player, ...), and that this increase in selectivity is accompanied by increased invariance to illumination, translation and scale (cf. Grill-Spector & Malach, 2004; Kourtzi & Connor, 2011; Roe et al., 2012; DiCarlo et al., 2012).

Further, activity in these cortical stages is modulated by context and task suggesting computations that operate within and between these areas. However, from a computational perspective, there is no clear consensus as to exactly what information visual cortical areas represent, why they have the form they do, or what operations act on these representations. (For various theories and analyses, see Marr (1982); Ullman (1995, 2007); Epshtein et al. (2008); Poggio (2011); Friston (2005); DiCarlo et al. (2012)).

Bayesian systems for object recognition

From a Bayesian perspective, the structure of images can be formulated in terms of probability dis-

tributions over graphs. To do this, we assume the following.

1) The visual system's structure should rely on the compositional structure of natural images (Geman et al., 2002; Jin & Geman, 2006; Yuille, 2011). Compositionality refers to the ability to construct hierarchical representations, whereby features/parts are used and shared to describe a virtually unlimited number of relational compositions. One argument is that without such a structure, we could not account for the speed with which humans can acquire and generalize visual knowledge.

2) Visual knowledge can be represented as a graphical structure in which nodes are random variables that represent hypotheses about features, parts, objects and their relations, and the links capture the statistical dependencies between nodes. It is important to note that this assumption involves explicit representations with accessible semantic interpretations, not just sequential banks of spatial filters (see below). This seems to be necessary to perform a range of visual tasks at different levels of analysis, and for the ability to transfer learning (e.g., not just detect objects like cats, but also locate their parts and spatial relations, and even learn to recognize hybrid objects, e.g. which have cat torsos and dog legs).

3) The system should be organized to be able to detect conjunctions of features that belong together as part of an object, while at the same time discounting, through disjunction, sources of variation, such as scale, illumination, position, and articulation. An illustration of this principle are AND/OR graphs (see below).

4) Inference and task flexibility is achieved by fixing values of nodes based on image measurements, priming, or attention, together with integrating out variables that are unimportant for a given task.

Motivated by stiff competitions within the computer vision community to build recognition systems that can handle the enormous variation of natural images (e.g. (Everingham et al., 2006)), algorithms have been developed that can do flexible inference as well as learn a generative structure. For a review of several Bayesian hierarchical composition models for recognition see Zhu et al. (2011). Next we describe an example to illustrate how knowledge is represented at different levels, how it performs different tasks, and how it can perform bottom-up and top-down inference. As with the primate ventral stream, higher level nodes represent hypotheses more coarsely, with increased specificity and invariance from bottom to top.

An example of a computer model for recognition. The task is to detect and parse a baseball player, given the naturally occurring wide range of variation in lighting, poses, and background clutter (Zhu, Chen, Lin, & Yuille, 2010). Figure 4 illustrates a model of a baseball player formulated as a probabilistic model defined over an AND/OR graph (Zhu, Chen, Lin, & Yuille, 2010). Selectivity and invariance are achieved through AND and OR operations. Although the logical operations are reminiscent of the "simple" and "complex" cell units in Riesenhuber & Poggio (1999), here the nodes represent random variables with semantic content and with associated distributions, and the links involve two-way interactions. The high level nodes represent body parts such as head, torso, and feet–while the lower level nodes represent the subparts. The high level nodes/parts are either compositions of lower nodes/subparts or are choices between different alternative subparts. The compositions and choices correspond to logical AND and OR operations, respectively.

The representation of a baseball player requires specifying the state variables of all the nodes in the hierarchy. The high level nodes only contain executive level descriptions – e.g., the center position, size, and orientation of the head – leaving more precise information – e.g., the precise position of the boundary of the head – to be represented by the

states of the lower-level nodes. This is similar to the gradual loss of positional specificity as one moves up the ventral stream. The existence of some representation of spatial relations even at higher levels is consistent with behavioral and with functional magnetic resonance imaging (fMRI) studies (Kravitz et al., 2010).

The edges between the different graph nodes, and the local conditional probabilities defined over them (see description of Markov Random Fields, Figure 6), capture the statistical relations about the spatial configurations of the baseball player (i.e. the likely relative positions of the body parts, and which body parts are likely to occur taking into account viewpoint and pose changes).

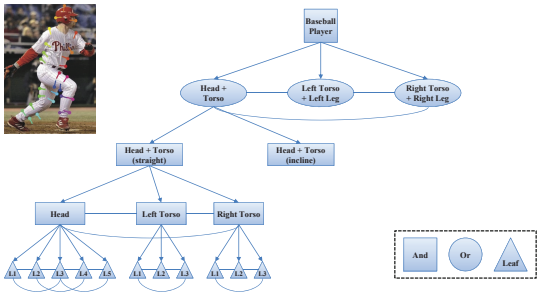

Figure 4. The AND/OR Graph Model (Zhu, Chen, Lin, & Yuille, 2010). The Baseball player is an AND of the head and torso, and left and right legs, but the head is an OR of straight head and torso or an inclined head and torso (top left).

Because the AND/OR graph is a probability model defined over a graph, one can perform inference to estimate the states of the unknown graph nodes, conditioned on the values of a subset of the nodes by message passing algorithms. This leads to a bottom-up and top-down strategy which is driven by input from the images (i.e. the model is conditioned on the states of the bottom-levels). Hypotheses are computed at the lower-levels and propagated up the hierarchy to form hypotheses for larger parts of the object and top-down processing is used to remove the "false" hypotheses at lower levels. Intuitively there will be many possible hypotheses for the small sub-parts of the object, since small features and parts of the object are easy to confuse with the background clutter. Compositions of sub-parts, however, are less likely to occur by chance in the background. So as the algorithm passes messages up the hierarchy it will tend to converge to the correct solution. Convergence speed depends on the reliability of the initial measurements.

One can also run the model in an attentional/priming mode where some of the top-level nodes are fixed (or conditioned on) – i.e. the system is primed to see a baseball player but does not know exactly where it is – while the bottom level nodes are also specified by the data (if we condition only on the top-level nodes, allowing no input from the image, then this is like imagining, or dreaming, a baseball player). This requires passing messages both top-down (from the primed nodes) and bottom-up (from the input nodes).

Learning hierarchical structure in natural images

While the brain clearly needs to be adaptive to image structures relevant for *successful behaviors*, the complexity of natural images suggests that part of the brain's solution involves the discovery of hierarchical structure in images themselves. A property of natural images is that intensities are to a first approximation, piece-wise smooth, so that one can predict pixel intensities from nearby pixels. Barlow argued that if image content is recoded to remove these and other higher-order statistical dependencies (through sparse coding), it becomes easier to compute useful information with probabilities (Barlow, 2001). Then one can detect "suspicious coincidences" (is $p(s_1, s_2) >> p(s_1)p(s_2)$?), and learn to predict them so they become unsuspicious.

The principle of detecting suspicious coincidences provides the means to build up a hierarchical model of features, parts, objects and scenes (Zhu et

al., 2008). Learning relies on a compositional prior for hierarchical world structure in which an image can be composed from meaningful, "reusable" parts as in language (Geman et al., 2002). A number of learning algorithms have been developed to learn hierarchical models from data (cf. Hinton, 2007; Zhu, Chen, Torralba, et al., 2010; Zeiler et al., 2011). We describe the results of one by Zhu, Chen, Torralba, et al. (2010) that explicitly relies on the compositional nature of collections of natural objects.

Figure 5 shows the hierarchy of features, parts, and shapes that result from learning models for the small parts first and then combining them to learn models for the larger parts. Unlike other hierarchical models (e.g., deep-belief networks), the graph structure of the probability model is learned, not just the weights. This allows the model to adapt and transfer to novel stimuli. The algorithm can rapidly learn to recognize a new object if it can be constructed by combining parts that have already been learned for other objects.

What aspects of the above Bayesian models might be shared by the primate visual system? The large proportion of our knowledge of primate vision is early-level, and the next section focuses on evidence that the visual system is well-adapted to the generative structure of images, and that these early representations provide the basis for lateral and hierarchical probabilistic computations that begin at the bottom level of the visual hierarchy. This section is followed by a description of several behavioral and neuroimaging studies that support interactions between lower- and higher-level processing stages.

Behavioral and neural evidence for Bayesian computations

Early representation & processing of images

A common theme underlying explanations of neural population architecture in V1 has been in terms of efficient codes that exploit the redundancy or regularities in natural images. For example it has been shown that neural response properties, such as orientation and spatial frequency tuning in V1 neurons, may be accounted for in terms of a sparse coding strategy adapted to the statistics of natural images (Olshausen, 1996; Hyvärinen, 2010). Neurons in primary visual cortex also show non-linear divisive-normalization behavior in which responses are inhibited by contrast variation outside the classical receptive field. Divisive normalization results in a reduction of statistical dependencies (Schwartz & Simoncelli, 2001), providing an efficient representation potentially useful for discovering *suspicious coincidences* Barlow (1990). Both sparse coding and contrast normalization have been explained in terms of an underlying generative model of natural images called a Gaussian Scale Mixture model (Wainwright & Simoncelli, 2000).

How do orientation- and spatial-frequency-selective spatial filter descriptions relate to visual behavior, to object perception, and to information and computations needed for higher-level areas? There are many answers to these questions depending on task, but the key step is to interpret neural population activity as representing a space of hypotheses about the causes of the stimulus (e.g. object boundary), rather than the stimulus itself (e.g. change in intensity). Here we focus on edges and surface regions, with a Bayesian view to understanding V1's role in object recognition.

From local spatial filters to edges. Hubel & Wiesel provided one clear functional interpretation of primary visual cortex (V1)–that a population of orientation-tuned simple and complex cells could represent an edge or line (Hubel, 1982). While there is wide consensus that information about orientation, as represented in neural populations of orientation-selective neurons, is a fundamental aspect of early visual spatial processing, exactly how such information is used to reliably determine ob-

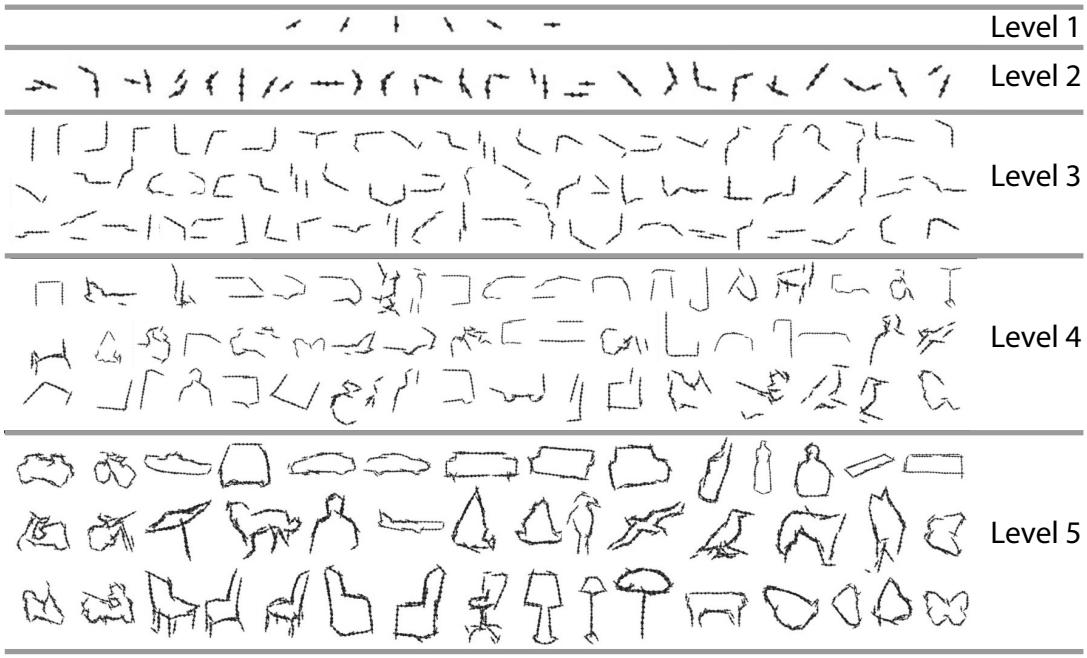

Figure 5. Examples of the mean shapes of visual concepts automatically learned for multiple objects with part sharing between objects (Zhu, Chen, Torralba, et al., 2010). The first few levels (1-3) are generic and higher levels (4 and 5) contain more specific concepts. or object parts, which occur for a limited number of objects.

ject properties, such as contour boundaries, or internal textures is still a challenging problem.

Given certain assumptions (Poisson-like spiking behavior), the pattern of activity across a population of orientation-tuned neurons can be assumed to represent the likelihood of orientation, and thus provide the basis for Bayesian decoding of edge orientation (Pouget et al., 2000). However, determining whether such an edge is part of an object boundary requires additional inference. Natural images are full of local edges, many due to accidents of illumination, occlusion, and background clutter. Distinguishing those which belong together as part of an object of interest likely requires a combination of lateral (within a cortical area) and top-down (between cortical areas) interactions. Local computation could be based on natural, local smoothness constraints (i.e. priors) on object boundaries (Geisler & Perry, 2009) to link nearby edges of similar orientation (Figure 6) (for relationship to independent coding see Garrigues & Olshausen (2008)). However, several decades of computer vision studies have shown that local, lateral constraints are insufficient, and top-down processes that incorporate intermediate-level, gestalt constraints (parallels, symmetry), and object-specific information are required for robust segmentation (see the intermediate-level generic, and later shape-specific visual concepts in Figure 5). Neurophysiological results are increasingly consistent with this view (cf. McManus et al., 2011). We return to the question of hierarchical representations of object knowledge, and how feedback may interact with edge representations later.

From edges to surface regions. The visual system is sensitive to shape as well as to surface properties of color, reflectance, and texture. An open

question has been whether regional surface properties are represented in a dense retinotopic format. An image-like representation of surface properties, shape and reflectance, could support more robust matching of input to higher-level templates. Could such a representation exist in early topographic visual areas, such as V1 or V2?

Neuroimaging experiments have shown that human V1 responds to lightness change, a correlate of reflectance, almost as strongly as it does to luminance change, and in V2 just as strongly (Boyaci et al., 2007). Using the same kind of stimulus, work in anesthetized monkeys has found luminance responses to temporal changes in spatially uniform regions to neurons in cytochrome oxidase blobs of primary visual cortex (V1), and responses to illusory lightness in cells in the color-activated regions (thin stripes) of V2 (Roe et al., 2005).

To represent reflectance, however, the visual system would have to solve a non-trivial inference problem. The first is that reflectance, shape, and illumination ("intrinsic images") are all confounded in the (luminance) image. Solving this problem requires prior knowledge of the spatial regularities of the intrinsic images. For example, reflectance could be roughly approximated as piece-wise constant, shape as piece-wise smooth, and illumination as smooth. One way to characterize these properties is through measurements of statistics on intrinsic images. One recent model factors out all three intrinsic images by using priors from objectively measured intrinsic images (Barron & Malik, 2012).

There is evidence for mechanisms that smoothly "fill-in" surface color between boundaries (Komatsu, 2006; Lee & Yuille, 2006), suggesting neural processes that assume smoothness to construct a map of surface color. Graph structures such as illustrated in Figure 6A can be used to describe smoothness priors on reflectance. Given some nodes initialized with reliable measurements (e.g at edges), missing or noisy values can be smoothed out by interpolation. There are a number of algorithms to do this, (cf. Gibbs sampling and MCMC methods in MacKay, 2003). Edges can be made explicit as random variables, called "line-processes", which have their own 1D smoothness priors, and when coupled with the regional surface process, serve to break neighborhood relations (see Figure 6B and (Koch et al., 1986; Kersten, 1991)). Whether and how such algorithms are neurally implemented is an open question. Distinct populations for boundaries and regions together with linking mechanisms have been proposed (Grossberg & Hong, 2006; Roe et al., 2005), however further experiments are needed to determine if and how such neural populations interact within and across visual areas. MRFs can also be used to model the strength of spatial grouping of features and parts at more abstract levels in a visual hierarchy. This is discussed in a later section.

A final problem is that the perception of lightness is sensitive to 3D layout of surfaces, (Gilchrist, 1977), suggesting that reflectance computations require non-local computations. For example, occlusion can interpose one surface over another, spatially separating image regions with the same underlying reflectance. There is evidence that, as early as V1, regions spatially disconnected by occlusion, with similar but different luminances are grouped together as a surface with a common lightness (Boyaci et al., 2010). It is possible that both lateral MRF-like, and top-down mechanisms are involved in early lightness computations, suggesting interactions in the cortical hierarchy.

Top-down, bottom-up interactions

What is the evidence for top-down/bottom-up Bayesian computations over hierarchical representations in human perception? Let's consider what this question might mean. An advantage of generative models is that they allow for flexible inference. Consider the three-level hierarchical graphical model shown in Figure 7A. Even for this simple model, there are a large number of possible inferences depending on which variables are known

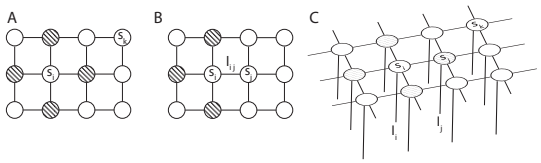

Figure 6. Markov Random Fields (MRF) can be used to model smoothness and piece-wise smoothness in image intensities or intrinsic images. The nodes represent random variables and the links the relationships. **A.** An undirected graph representing a Markov Random Field. The probability distribution of s_i depends only on variables in a predefined neighborhood (e.g. in this example, the four nearest neighbors, dashed nodes). Prior probabilities on smoothness and piece-wise smoothness constraints can be represented with this kind of graph. **B.** An MRF with explicit line processes that when switched on break the neighborhood relations. **C.** An MRF where the values to be estimated are linked to the image measurements, I_i.

and which are integrated out. For example, suppose the task requires estimating the top-level value m representing an object category. The MAP solution would be to find that value of m which maximizes $p(m|I_1, I_2)$. This could be achieved with a purely bottom-up algorithm with the middle variables, s_1, s_2, integrated out from the posterior, $p(s_1, s_2, m|I_1, I_2)$. Further, as noted earlier in the discussion of cue integration, a distinctive aspect of a Bayesian solution is to integrate the image measurements based on full knowledge of the uncertainty represented by the posterior. On the other hand, suppose the category m is known (from some other source of knowledge), and one wants to estimate the middle parameters, s_1, s_2, representing possible shapes within a specific category m'. The MAP solution would be to find s_1 and s_2 that maximize $p(s_1, s_2|I_1, I_2, m')$. An algorithm could use generative knowledge, that $p(s_1, s_2|I_1, I_2, m') \propto p(I_1, I_2|s_1, s_2)p(s_1, s_2|m')$. Note that the factors on the right hand size use knowledge both of how the category influences shape, and of how shape influences the image measurements. Below, we discuss a related third possibility, "Bayesian coarse-to-fine", in which the first priority is to make an accurate high-level decision, followed by estimation of lower-level parameters. "Explaining-away" is another kind of inference which uses top-down generative knowledge, also discussed below (Figure 3B).

Currently there is no direct evidence for neural populations representing hypotheses rather than decisions, or for probabilistic computations (as in message passing). However, there are behavioral and neuroimaging results that are suggestive of Bayesian computations at different levels of abstraction, and between cortical areas. We briefly mention some of them.

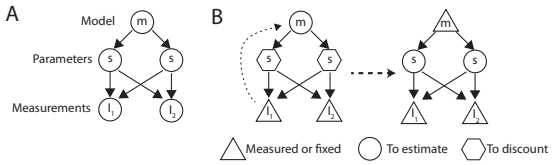

Figure 7. **A**. A simple hierarchical graphical model. Values of a discrete random variable at the top level of the hierarchy represent "model" choices. Various models represent various categories of lower-level parameters. Depending on the task, some nodes are fixed (either through measurement at the lowest level, or through prior decisions, attention, or priming at the highest, model level), others get integrated out, and others are estimated. For example, the network can be run in a generative fashion where samples, I, are drawn conditional on model choices, m. **B**. "Conditioned perception" (Stocker & Simoncelli, 2008), is the particular sequence of marginalization where: 1) intermediate-level parameters are integrated out, and the high-level model values estimated (the dashed arrow shows the direction of inference), and then 2) the value of the inferred model is held fixed and the parameters are estimated.

Bayesian coarse-to-fine. A hierarchical architecture allows for "Bayesian coarse-to-fine" processing in which an initial high-level inference, at a coarse level of abstraction, constrains a subsequent finer-grained analysis of lower-level features. It is well-established that certain visual decisions ("animal present or not?") can be made extremely fast (Rous-

selet et al., 2004). This is consistent with the above flexible Bayesian-style architecture, in which decisions are effectively confidence-driven. For example in the AND/OR baseball player models, hypotheses can be rapidly propagated bottom-up, and if there is little uncertainty at the lower levels, there may be no need for top-down removal of false hypotheses. However, in the process the top-down information can help to specify the precise positions of the baseball player's boundaries (and sub-parts). More generally, Bayesian hierarchical models can allow reductions in uncertainty at a high-level of abstraction to later affect decisions at lower levels.

Perceptual decisions can be automatic, driven by strong, reliable context evidence (either spatial information elsewhere in the image, or temporal as in priming) or consciously task driven and specified by some even higher-level "executive". Both strategies are consistent with a coarse-to-fine computation in which a high-level decision "fixes" the value in the upper level of a hierarchical model, constraining or biasing subsequent lower-level decisions. An optimal decision restricted to a high level requires marginalization over intermediate-level parameters. We brief describe several behavioral results which are consistent with Bayesian coarse-to-fine computations over a simple hierarchical graph structure (see Figure 7).

Knill (2003) showed that, when estimating surface orientation from texture cues, the visual system uses different texture models depending on evidence in the image or non-linearity weighting of cues. The model types, texture orientation parameters, and image measurements (cues) can be represented at the top, middle, and bottom levels of the hierarchy, respectively. The system usually interprets a 2D texture as caused by an underlying isotropic 3D texture. However, textures may also be anisotropic. Human surface judgments were well-modeled by a Bayesian observer that uses a texture cue to make a coarse inference to decide the texture model type, and then applies this model to interpret the texture cues. Other work has shown that humans infer causal structure (i.e. "common or independent sources?") when integrating sound and visual stimuli to estimate direction, consistent with Bayesian estimation (Körding et al., 2007; Shams, 2010). Human perception of complex motion fields is also consistent with Bayesian selection of motion type (rotation, expansion, translation) influencing discrimination performance (Wu et al., 2008).

Human biases in motion direction estimation following a conscious classification decision (Jazayeri & Movshon, 2007), have been also been explained in terms of model selection (Stocker & Simoncelli, 2008), Figure 7B.

Top-down prediction. Top-down hypotheses can compete to explain data. The Bayesian graphical interpretation is shown in Figure 3B. A number of perceptual phenomena, in which contextual changes drastically change local visual interpretations, are consistent with "explaining away" (Kersten et al., 2004). There are several ways in which the representation of the probability distributions could change. In the earlier computer vision baseball example, top-down processes suppressed false hypotheses at a lower level, keeping those consistent with the larger picture emerging as a consequence of message passing. Another strategy is "predictive coding" in which s^{n+1} is a provisional guess at stage $n+1$ that can use generative knowledge, f, to predict the input, s^n, of an earlier stage n. If the prediction error, $|s^n - f(s^{n+1})|$ is small (represented by low neural activity), the likelihood is high and the hypothesis is a good fit; if not, another hypothesis needs to be tried (Rao & Ballard, 1999; Lee & Mumford, 2003; Yuille & Kersten, 2006; Friston & Kiebel, 2009).

There is growing evidence from human fMRI studies for context-dependent suppression of neural activity in an earlier area in certain cases (Murray et al., 2002; Summerfield et al., 2006; Fang et al., 2008; Alink et al., 2010; Rauss et al., 2011; Cardin

et al., 2011). However, the spatial resolution of fMRI is too coarse to tell whether reduced activity may be due to suppression of false Grossberg (1999) or true hypotheses Rao & Ballard (1999). In fact, because of different functional requirements, the suppression of lower-level false and true hypotheses could both be occurring. A top-down attentional task might suppress competing lower-level ambiguous hypotheses in order to select true hypotheses relevant to the task Grossberg (2007). But the system may also rely on automatic processes to detect new information that does not fit with the current interpretation. The first strategy corresponds to executive, top-down ("endogenous") attention, and the second to stimulus-driven, bottom-up ("exogenous") attention (Desimone & Duncan, 1995; Petersen & Posner, 2011). Thus there may be two types of neural populations, those whose activity codes errors, and those whose increased activity represents increased confidence in local feature hypotheses coded in that area (Friston, 2005). Thus, enhancement of consistent features could be accompanied by suppression of false positives in one population, but consistent features suppressed in another population. The laminar pattern of feedback connections raises the possibility that these representations may lie in distinct cortical layers. A recent study used ultra-high field fMRI with sub-millimeter resolution and found stronger fMRI response in middle cortical layers of V1 during the presentation of scrambled objects as compared with intact objects (Olman et al., 2012), similar to what one might expect from error units.

Egner et al. (2010) proposed that distinct populations of "representational" units (feature detectors) and "error" units could be activated independently by behavioral manipulations of expectation and surprise, respectively. They manipulated expectation by training with cues that were diagnostic of the appearance of either a face or a house on a given trial. The subject's task was to detect an occasional and thus surprising, upside-down face or house. The authors were able to account for the pattern of results in human fusiform face area in terms of a sum of activity due to separate populations of units whose activity reflected expectation and surprise, rather than by face-specific responses predicted by bottom-up selectivity alone.

Conclusions and future directions

Although it was observed by Helmholtz more than 100 years ago that perception is a process of "unconscious inference," developing and testing quantitative models that embrace this idea has occurred only recently. From this body of work we have gained new insights about inferential processes occurring in perception, and new ways of thinking about neural computation in the brain. However, achieving a full understanding of perception and its neural basis will remain an important and challenging problem for the future. A major aspect of the problem is that there is no consensus on what the overall neural or psychological constraints on a generative or inferential model should look like. We have presented the case that hierarchical, compositional models are necessary (Geman et al., 2002; Feldman, 2009); however, it is unclear what the levels should represent (truth values vs. distributions on hypotheses) and what kinds of computations are done over them (decision sequences vs. propagating hypotheses and updating distributions). Variables in a level may represent higher-order image statistics, or they could (also) have accessible semantic interpretations, such as "edge", "curved line", "limbs", and "baseball player".

Models of image generation must eventually capture the flexibility of humans to interpret images that deviate substantially from natural, real-world experience. Human perception is almost never stumped–it always comes up with an interpretation, even if bizarre, wrong, or self-contradictory (consider a Dali painting, a cartoon, or a video game). While developing probabilistic models on graphical representations is an important direction, the need

for an even richer structure has been recognized in proposals for image grammars (Mumford & Desolneux, 2010) and probabilistic programs (Goodman et al., 2008).

References

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. *Journal of Neuroscience*, *30*(8), 2960–2966.

Barlow, H. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, *30*(11), 1561–1571.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 1–13.

Barron, J. T., & Malik, J. (2012, March). Shape, Albedo, and Illumination from a Single Image of an Unknown Object. *CVPR*, 1–8.

Battaglia, P. W., Kersten, D., & Schrater, P. R. (2011, June). How haptic size sensations improve distance perception. *PLoS Computational Biology*, *7*(6), e1002080.

Battaglia, P. W., & Schrater, P. (2007, June). Humans Trade Off Viewing Time and Movement Duration to Improve Visuomotor Accuracy in a Fast Reaching Task. *Journal of Neuroscience*, *27*(26), 6984–6994.

Beck, J. M., Latham, P. E., & Pouget, A. (2011, October). Marginalization in Neural Circuits with Divisive Normalization. *Journal of Neuroscience*, *31*(43), 15310–15319.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011, January). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83–87.

Boyaci, H., Fang, F., Murray, S. O., & Kersten, D. (2007, June). Responses to Lightness Variations in Early Human Visual Cortex. *Current Biology*, *17*(11), 989–993.

Boyaci, H., Fang, F., Murray, S. O., & Kersten, D. (2010). Perceptual grouping-dependent lightness processing in human early visual cortex. *Journal of Vision*, *10*(9), 1–12.

Carandini, M., & Heeger, D. J. (2011, November). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*.

Cardin, V., Friston, K. J., & Zeki, S. (2011, February). Top-down Modulations in the Visual Form Pathway Revealed with Dynamic Causal Modeling. *Cerebral Cortex*, *21*(3), 550–562.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007, July). Bayesian integration of spatial information. *Psychological Bulletin*, *133*(4), 625–637.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010, October). What and where: A Bayesian inference theory of attention. *Vision Research*, *50*(22), 2233–2247.

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

Colombo, M., & Seriès, P. (2012, February). Bayes in the Brain–On Bayesian Modelling in Neuroscience. *The British Journal for the Philosophy of Science*.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012, February). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434.

Eckstein, M. P., Drescher, B., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973.

Egner, T., Monti, J. M., & Summerfield, C. (2010, December). Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. *Journal of Neuroscience*, *30*(49), 16601–16608.

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14298.

Ernst, M. O., & Banks, M. S. (2002, January). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Ernst, M. O., & Bülthoff, H. H. (2004, April). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169.

Everingham, M., Zisserman, A., Williams, C., Van Gool, L., Allan, M., Bishop, C., et al. (2006). The 2005 pascal visual object classes challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 117–176.

Fang, F., & He, S. (2005, September). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, *8*(10), 1380–1385.

Fang, F., Kersten, D., & Murray, S. O. (2008). Perceptual grouping and inverse fMRI activity patterns in human visual cortex. *Journal of Vision*, *8*(7), 2.1–9.

Feldman, J. (2001). Bayesian contour integration. *Attention, Perception, & Psychophysics*, *63*(7), 1171–1182.

Feldman, J. (2009, October). Bayes and the simplicity principle in perception. *Psychological Review*, *116*(4), 875–887.

Feldman, J., & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review*, *112*(1), 243.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010, March). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

Franklin, D. W., & Wolpert, D. M. (2011, November). Computational Mechanisms of Sensorimotor Control. *Neuron*, *72*(3), 425–442.

Freeman, W. (1994, April). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*(6471), 542–545.

Friston, K. (2005, April). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K., & Kiebel, S. (2009, March). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Garrigues, P., & Olshausen, B. A. (2008). Learning horizontal connections in a sparse coding model of natural images. *Advances in neural information processing systems*, *20*, 505–512.

Geisler, W. S. (2008, January). Visual Perception and the Statistical Properties of Natural Scenes. *Annual review of psychology*, *59*(1), 167–192.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781.

Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508–510.

Geisler, W. S., & Perry, J. (2009). Contour statistics in natural images: Grouping across occlusions. *Visual Neuroscience*, *26*(01), 109–121.

Geman, S., Potter, D., & Chi, Z. (2002). Composition systems. *Quarterly of Applied Mathematics*, *60*(4), 707–736.

Gershman, S., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, 1–24.

Gilchrist, A. L. (1977, January). Perceived lightness depends on perceived spatial arrangement. *Science*, *195*(4274), 185–187.

Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. *Uncertainty in Artificial Intelligence*, *22*, 23.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008, May). Young Children Do Not Integrate Visual and Haptic Form Information. *Current Biology*, *18*(9), 694–698.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, Wiley.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T., & Yuille, A. (2008). A primer on probabilistic inference. *The probabilistic mind: Prospects for Bayesian cognitive science*, 33–57.

Grill-Spector, K., & Malach, R. (2004, July). The human visual cortex. *Annual Review of Neuroscience*, *27*(1), 649–677.

Grossberg, S. (1999, March). The link between brain learning, attention, and consciousness. *Consciousness and Cognition*, *8*(1), 1–44.

Grossberg, S. (2007). Towards a unified theory of neocortex: laminar cortical circuits for vision and cognition. *Progress in Brain Research*, *165*, 79–104.

Grossberg, S., & Hong, S. (2006, April). A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision*, *19*(2), 263–321.

Hinton, G. E. (2007, October). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.

Hubel, D. (1982). Evolution of ideas on the primary visual cortex, 1955–1978: A biased historical account. *Bioscience reports*, *2*(7), 435–469.

Hyvärinen, A. (2010, April). Statistical Models of Natural Images and Cortical Visual Representation. *Topics in Cognitive Science*, *2*(2), 251–264.

Itti, L., & Baldi, P. (2009, June). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Jacobs, R. (1999, October). Optimal integration of texture and motion cues to depth. *Vision Research*, *39*(21), 3621–3629.

Jazayeri, M., & Movshon, J. A. (2007, April). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, *446*(7138), 912–915.

Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, *2*, 2145–2152.

Kersten. (1991). Transparency and the cooperative computation of scene attributes. In M. S. Landy (Ed.), *Computational models of visual processing* (pp. 209–228). The MIT Press.

Kersten. (1999). High-level vision as statistical inference. In M. Gazzaniga & Editor (Eds.), *The new cognitive neurosciences* (pp. 353–364). Cambridge, MA: MIT Press.

Kersten, Masmassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual review of psychology*, *55*, 271–304.

Kersten, D., & Schrater, P. (2002). Pattern inference theory: A probabilistic approach to vision. *Perception and the Physical World*, 191–228.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research*, *43*(7), 831–854.

Knill, D. C., & Pouget, A. (2004, December). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Koch, C., Marroquin, J., & Yuille, A. (1986, June). Analog "neuronal" networks in early vision. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(12), 4263–4267.

Komatsu, H. (2006, March). The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, *7*(3), 220–231.

Körding, K. P. (2007, October). Decision Theory: What "Should" the Nervous System Do? *Science*, *318*(5850), 606–610.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007, September). Causal Inference in Multisensory Perception. *PLoS ONE*, *2*(9), e943.

Körding, K. P., & Wolpert, D. M. (2006, July). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10*(7), 319–326.

Kourtzi, Z., & Connor, C. E. (2011, July). Neural Representations for Object Perception: Structure, Category, and Adaptive Coding. *Annual Review of Neuroscience*, *34*(1), 45–67.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010, November). High-Level Visual Object Representations Are Constrained by Position. *Cerebral Cortex*, *20*(12), 2916–2925.

Lee, T., & Mumford, D. (2003, July). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, *20*(7), 1434–1448.

Lee, T., & Yuille, A. (2006). Efficient coding of visual scenes by grouping and segmentation. *Bayesian Brain: Probabilistic Approaches to Neural Coding, MIT Press, Cambridge, MA*, 145–188.

Lennie, P. (1998). Single units and visual cortical organization. *Perception*, *27*, 889–936.

Lowe, D. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision*, *1*(1), 57–72.

Ma, W. J. (2010, October). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, *50*(22), 2308–2319.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006, November). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms* (First Edition ed.). Cambridge University Press.

Maloney, L. T., & Mamassian, P. (2009, February). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(01), 147.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

McManus, J. N. J., Li, W., & Gilbert, C. D. (2011, June). Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences*, *108*(24), 9739–9746.

Milner, D., & Goodale, M. (2006). *The Visual Brain in Action (Oxford Psychology Series)* (2nd ed.). Oxford University Press, USA.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011, July). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496.

Mumford, D., & Desolneux, A. (2010). *Pattern Theory: The Stochastic Analysis of Real-World Signals (Applying Mathematics)*. A K Peters Ltd.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002, November). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15164–15169.

Nakayama, K., & Shimojo, S. (1992, September). Experiencing and perceiving visual surfaces. *Science*, *257*(5075), 1357–1363.

Olman, C. A., Harel, N., Feinberg, D. A., He, S., Zhang, P., Ugurbil, K., et al. (2012, March). Layer-Specific fMRI Reflects Different Neuronal Computations at Different Depths in Human V1. *PLoS ONE*, *7*(3), e32536.

Olshausen, B. A. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (1st ed.). Morgan Kaufmann.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006, December). Feature detection and letter identification. *Vision Research*, *46*(28), 4646–4674.

Petersen, S. E., & Posner, M. I. (2011, July). The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, *35*(1), 120518152625006.

Poggio, T. (2011, September). The Computational Magic of the Ventral Stream: Towards a Theory. *Nature Precedings*.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2).

Rao, R. P. N., & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

Rauss, K., Schwartz, S., & Pourtois, G. (2011, April). Top-down effects on early visual processing in humans: A predictive coding framework. *Neuroscience and Biobehavioral Reviews*, *35*(5), 1237–1253.

Reynolds, J. H., & Heeger, D. J. (2009, January). The Normalization Model of Attention. *Neuron*, *61*(2), 168–185.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., et al. (2012, April). Toward a Unified Theory of Visual Area V4. *Neuron*, *74*(1), 12–29.

Roe, A. W., Lu, H. D., Hung, C. P., & Kaas, J. H. (2005, March). Cortical Processing of a Brightness Illusion. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(10), 3869–3874.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2004, August). How parallel is visual processing in the ventral pathway? *Trends in Cognitive Sciences*, *8*(8), 363–370.

Schlicht, E. J., & Schrater, P. R. (2007, June). Impact of coordinate transformation uncertainty on human sensorimotor control. *Journal of Neurophysiology*, *97*(6), 4203–4214.

Schwartz, O., & Simoncelli, E. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Shams, L. (2010, August). Probability Matching as a Computational Strategy Used in Perception. *PLoS Computational Biology*, *6*(8), e1000871.

Stevenson, I., & Körding, K. P. (2009). Structural inference affects depth perception in the context of potential occlusion. *Advances in Neural Informational Processing Systems*, 1777–1784.

Stocker, A. A., & Simoncelli, E. (2008). A Bayesian model of conditioned perception. *Advances in neural information processing systems*, *20*, 1409–1416.

Strens, M. (2000). A Bayesian framework for reinforcement learning. *Proceeedings of the Seventeenth International Conference on Machine Learning (ICML-2000),*, 943–950.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006, November). Predictive Codes for Forthcoming Perception in the Frontal Cortex. *Science*, *314*(5803), 1311–1314.

Sundareswara, R., & Schrater, P. (2008, May). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, *8*(5), 12–12.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Trenti, E. J., Barraza, J. F., & Eckstein, M. P. (2010, February). Learning motion: Human vs. optimal Bayesian learner. *Vision Research*, *50*(4), 460–472.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, *5*(1), 1–11.

Ullman, S. (2000). *High-Level Vision: Object Recognition and Visual Cognition* (1st ed.). A Bradford Book.

Ullman, S. (2007, February). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, *11*(2), 58–64.

Wainwright, M. J., & Simoncelli, E. (2000). Scale mixtures of Gaussians and the statistics of natural images. *Advances in neural information processing systems*, *12*(1), 855–861.

Wilder, J., Feldman, J., & Singh, M. (2011, June). Superordinate shape classification using natural shape statistics. *COGNITION*, *119*(3), 325–340.

Wu, S., Lu, H., & Yuille, A. (2008). Model selection and velocity estimation using novel priors for motion patterns. In D. Koller, D. Schuurmans, &

Y. B. L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 1793–1800). Cambridge, MA: MIT Press.

Yuille, A. (2011). Towards a theory of compositional learning and encoding of objects. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 1448–1455.

Yuille, A., & Kersten, D. (2006, July). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zeiler, M., Taylor, G., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2018–2025.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008, May). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 32–32.

Zhaoping, L., & Jingling, L. (2008). Filling-In and Suppression of Visual Perception from Context: A Bayesian Account of Perceptual Biases by Contextual Influences. *PLoS Computational Biology*, *4*(2), e14.

Zhu, L., Chen, Y., Lin, C., & Yuille, A. (2010, August). Max Margin Learning of Hierarchical Configural Deformable Templates (HCDTs) for Efficient Object Parsing and Pose Estimation. *International Journal of Computer Vision*, *93*(1), 1–21.

Zhu, L., Chen, Y., Torralba, A., Freeman, W., & Yuille, A. (2010). Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1919–1926.

Zhu, L., Chen, Y., & Yuille, A. (2011, April). Recursive Compositional Models for Vision: Description and Review of Recent Work. *Journal of Mathematical Imaging and Vision*, *41*(1-2), 122–146.

Zhu, L., Lin, C., Huang, H., Chen, Y., & Yuille, A. (2008). Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion.