Kersten D. & Yuille, A.L.(2014) Inferential Models of the Visual Cortical Hierarchy. The New Cognitive Neurosciences, 5th Edition. Michael S. Gazzaniga and George R. Mangun (editors) . Chapter 34, pp. 389-404.

Inferential Models of the Visual Cortical Hierarchy*

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

Human visual object decisions are believed to be based on a hierarchical organization of stages through which image information is successively transformed from a large number of local feature measurements with a small number types (e.g. edges at many locations) to increasingly lower-dimensional representations of many types (e.g. dog, car, ...). Functional utility requires integrating a large number of local features to reduce ambiguity, while at the same time selecting task-relevant information. For example, decisions requiring object recognition involve pathways in the hierarchy in which representations become increasingly selective for specific pattern types (e.g. boundaries, textures, shapes, parts, objects), together with increased invariance to transformations such as translation, scale, and illumination. Computer vision architectures for object recognition and parsing, as well as models of the primate ventral visual stream are consistent with this hierarchical view of visual processing. The hierarchical model has been extraordinarily fruitful, providing qualitative explanations of behavioral and neurophysiological results. However, the computational processes carried out by the visual hierarchy during object perception and recognition are not well-understood. This chapter describes how a Bayesian, inferential perspective may help to understand the organization of visual knowledge, and its utilization through the feedforward and feedback flow of information.

To appear in: The New Cognitive Neurosciences, 5th Edition. Please do not cite without permission.

It takes just one quick glance at the picture in Figure 1A to see the fox, a tree trunk, some grass and background twigs. This is a remarkable achievement in which the visual system turns a massive set of highly ambiguous local measurements, (Figure 1B), into accurate, and reliable identifications. But that is just the beginning of what vision enables us to do with this picture. With a few more glances, one can see a whole lot more: the shape of the fox's legs and head, the varying properties of its fur, guess what it is doing, whether it is young or old. The ability to generate an unbounded set of descriptions from a virtually limitless number of images illustrates the extraordinary versatility of human perception.

This chapter focuses on the following question: What knowledge representations and computational processes are needed to achieve reliable and versatile object vision? Although we are far from complete answers, there has been substantial progress in the overlapping fields of perceptual psychology, computer vision/robotics, and visual neuroscience.

Comments may be sent to the author at kersten@umn.edu. D.K. and A.L. 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) and by ONR N000141210883.

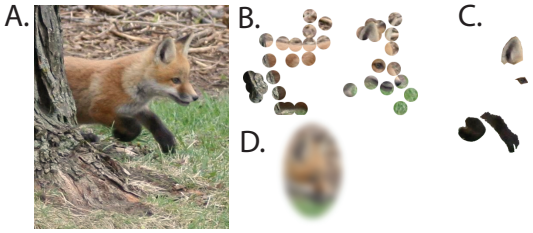

Figure 1. **A**. This figure illustrates two problems. 1) How can local measurements made from small patches (**B**), using neurons with small receptive fields, be integrated to recognize objects and patterns (e.g. fox, tree trunk, grass)? 2) How does the visual system support a limitless number of descriptions of a single scene? Answers need to account for flexible access to information of various types over a range of spatial scales, such as the various edge and textural properties of local regions **B**, the shape of parts **C**, and intermediate- and higher-level concepts such as "head" **D**, respectively. There is a bootstrapping problem in that the accurate interpretation of any local patch is ambiguous without knowledge of the rest.

In all three fields, theories of representations of visual knowledge and the processes acting on them are constrained by: 1) functional behaviors or tasks, and their priorities; 2) the statistical structure of the visual world, and consequently in images received; 3) algorithms and knowledge structures for getting from images to behaviors; and 4) neurophysiological (or hardware) limitations on what can be computed by collections of neurons (or components and circuits).

There has been considerable growth in 4), our knowledge of the neurophysiology and anatomy of the primate visual system at the level of large-scale organization of visual areas and their connections (Kourtzi & Connor, 2011; Kanwisher, 2010), and the finer scale level of cortical (Markov et al., 2013; Callaway, 1998; Lund et al., 2003) and sub-cortical neuro-circuitry (Guillery & Sherman, 2002). The larger picture is that visual processing involves processing within a visual area (both laterally and across laminae), and hierarchical – feedforward and feedback – processing between areas with various feature selectivities (Figure 2).

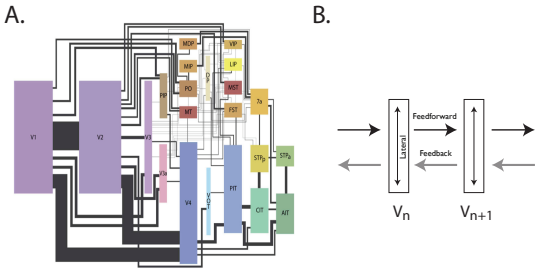

Figure 2. **A**. Schematic of macaque monkey visual cortex. The colored rectangles represent visual areas (see Felleman & Van Essen (1991)). The gray lines show the connections between areas, with the thickness proportional to estimates of the number of feedforward fibers. Areas in warm and cool tones belong to the dorsal and ventral streams, respectively. (Figure from Wallisch & Movshon (2008); see also Lennie (1998)) **B**. Feedforward and feedback connections represent transmission of feedforward and feedback signals between visual areas. Lateral (also called "horizontal") organization within areas, representing features of similar types and level of abstraction.

However, despite growth in our knowledge of the visual brain, there remains a gap in our understanding of how the biology of vision enables common behaviors.[1] An immediate problem faced when beginning such an analysis is that the large-scale systems nature of the problem makes it difficult to empirically test theories of behavior at the level of neurons. One strategic solution is to temporarily ignore the details of the neurophysiology and neurocircuitry (i.e., 4) above), and try to understand a nar-

[1] Even complete knowledge of neural network connectivity and dynamics would be insufficient to explain visual function. For example, a complete description of spatial-temporal switching of the billion plus transistors in a video game console would provide little insight into how these patterns relate to game goals, algorithms or behavior.

rower problem–what are the representations, learning principles, and types of computations required for competent visual behavior?

A key idea, inspired by both computer vision research and quantitative studies of human behavior, is that vision is fundamentally inferential. More specifically, visual perception involves processes of *statistical inference*, which can be as simple as heuristic rules, to more complex, probabilistic processes. Further, methods of statistical inference can also be applied, specifically through machine learning techniques, to understand how hierarchical representations of feature types are constrained by the statistical regularities in natural images.

In the next section, we review basic concepts of statistical inference, focusing on Bayesian decision theory. In subsequent sections, we discuss the functions of within-area (focusing on lateral representations), feedforward, and feedback visual processing from an inferential perspective, with a view towards a better understanding how the visual cortical architecture may support human visual object perception and recognition.

Vision as statistical inference

How can one begin to model vision as inference? To begin, we need to specify the task requirements: what should be estimated, and the image information to get there. The number of models to get from input to output can be very large, suggesting the strategy of first characterizing the requirements for optimal inference, and then interpreting actual performance in terms of approximations (see ideal observer analysis below). Bayesian inference theory provides a well-developed set of concepts for modeling optimal inference, including discriminative and generative probability models, and decision rules.

In its basic form, Bayesian theory provides mathematical tools for estimating hypotheses with potentially complex interdependencies (e.g. causal relationships), given varying degrees of uncertainty and importance. Bayesian inferences are based on knowledge represented by the joint probability distribution, $p(s_1, s_2, \ldots, I_1, I_2, \ldots)$ – a model of the probability of descriptions ("explanations" or "hypotheses") $s = s_1, s_2, \ldots$, together with the patterns of image measurements (or "features") $I = (I_1, I_2, \ldots)$.

The joint distribution, however, can be quite complex, reflecting causes of image patterns that are often subtle and deep. For example, the descriptions of the fox in Figure 1 included inferences of category (which influence 3D object shape, and thus measurements available in the image projection), subcategory (baby fox, which affects the size and contours of the head), material (fur properties, together with shape and lighting produce image texture), relative depths (the tree occludes part of the fox, which in turn occludes background), and pose (the image of fox's head is to the right of the body). This suggests a causal, top-down hierarchical structure, with variables representing abstract concepts at the top, to variables at the bottom representing local features shared among many objects.

Formally, the structure of images can be formulated in terms of probability distributions over structured graphs (Lauritzen & Spiegelhalter, 1988; A. Yuille, 2010). The graphical language helps capture the causal structures and the dependencies/independencies between causes. The nodes are random variables that represent hypotheses about events, objects, parts, features, and their relations. The links express the statistical dependencies between nodes. The links can be directed, representing causal influence, or undirected. Inference and task flexibility is achieved by fixing values of nodes based on local image measurements, or decisions made elsewhere in the system (e.g. through "priming"), together with integrating out variables that are unimportant for a given task (for a simple example, see Figure 3A) [2].

[2] Until the advent of computers, it was difficult to

Optimality is defined by a criterion (e.g. "minimize average error") which determines a decision rule (e.g. "pick the values of the unknowns that maximize the posterior probability") [3].

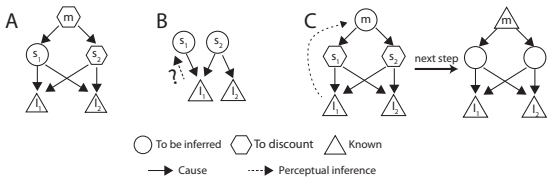

Figure 3. **A**. A simple graph illustrating the generative constraints on incoming data. See main text. **B**. More than one combination of causes s, could explain local image measurement, I_1. Optimal perception seeks an explanation, i.e. values of s_1 or s_2 that give the most probable explanation for how the image measurement could have been generated. For example, Bayes optimal calculations show that without feature I_2, s_1 takes on one value, but with a measurement of I_2, it takes on a different value. Pearl (1988) calls this "explaining away". **C**. Bayesian coarse-to-fine. Different "models", m, can be different functions of the parameters s, which in turn lead to different image features. An initial, "quick and dirty" visual inference may be at the top level (e.g. it is a "fox") ignoring shape details (but using for example features from the wooded context, fur color, "features of intermediate complexity" or "fragments", that may be sufficient). Fixing the hypothesis of "fox" can be followed by reliable inferences at a lower-level (e.g. "shape of the head of the fox").

Bayesian algorithms can be *discriminative*, based on a model of the posterior: $p(s|I) = p(s,I)/p(I)$ – the probability of a description $s = s_1, s_2, \ldots$, given a pattern of image measurements (or "features") $I = (I_1, I_2, \ldots)$. Discriminative algorithms are bottom-up, and do not incorporate explicit models of how image patterns are caused by objects. For example, in it simplest form, a discriminative algorithm could be a look-up table which maps an image pattern to the most probable hypothesis, which in neural terms is not that different from a reflex (Purves & Lotto, 2003)[4].

Bayesian algorithms can also be *generative*. Generative models rely on knowledge in the likelihood, $p(I|s)$ which specifies how an image results from causes or explanations s, and a prior $p(s)$. These probabilities are related to the posterior through Bayes rule: $p(s|I) = p(I|s)p(s)/p(I)$. Generative algorithms make explicit use of top-down *generative processes*, in which high-level hypotheses are used to simulate the values of lower-level nodes, ultimately generating a prediction of I (Mumford, 1992; A. Yuille & Kersten, 2006). Generative models provide a number of advantages. For example, by elaborating the structure of the likelihood, computational studies have shown that a generative process can improve recognition through "explaining away", useful for both learning (Hinton, 2009; Zeiler et al., 2011), and inference applied to image parsing (Tu et al., 2005). Generative algorithms predict appearances in time (e.g. Bayes-Kalman; Burgi et al., 2000), and cope more efficiently with a wider range of variability, such as the virtually unlimited ways in which objects can be composed (A. L. Yuille & Mottaghi, 2013; Chang et

handle Bayesian calculations beyond a few dimensions. Today, computer vision algorithms find Bayes optimal solutions for problems involving thousands of dimensions. Optimization methods include regression, various message-passing algorithms such as EM, and belief-propagation. It is largely an open question if and how such algorithms could be implemented in a neurally plausible fashion

[3] Bayesian decision theory generalizes "integrating out" by introducing a loss (or utility) function to allow for relative costs of imprecision in the estimation of various contributing values of s_i. Optimality is then defined as maximizing utility (or minimizing risk) (Maloney & Zhang, 2010; Geisler & Kersten, 2002)

[4] A discriminative algorithm can implement a decision rule with no explicit use of probabilities. For example, with a large number of samples, a rule to minimize empirical risk (Schölkopf & Smola, 2002) becomes equivalent to minimizing Bayes risk, as discussed in D. Kersten et al. (2004).

al., 2011), discussed more below.

Computer vision studies have shown discriminative and generative models can be combined (Tu et al., 2005)–an algorithmic strategy similar in spirit to two-stage processing accounts of human visual recognition, in which an initial, fast decision about the "gist" of a scene narrows the space of specific objects to match to the image (Bar, 2003).

Bayesian probabilistic methods have been applied in a number of quantitative studies of human visual behavior. There is a long history to studying human perception (and neural responses) using "ideal observer analysis" (Gold et al., 2009). Here one makes quantitative comparisons between what an ideal (Bayesian) observer can achieve with humans or neurons (Geisler, 2011; Trenti et al., 2010). A strategic benefit of ideal observer analysis in studies of human behavior is that it helps to distinguish perceptual limitations inherent to the information processing problem from limitations of the neural mechanisms (cf. Weiss et al., 2002; Eckstein et al., 2006).

Quantitative behavioral experiments have shown near optimality or ideal-like behavior in a variety of domains, including visual cue integration (Jacobs, 1999), visual motor control (Orban & Wolpert, 2011; Wolpert & Landy, 2012), learning (Green et al., 2010), and attention (Chikkerur et al., 2010). For reviews, see Geisler (2011); D. J. Kersten & Yuille (2013); Vilares & Körding (2011). Findings of optimal behavior have raised the question of whether neural populations within the brain explicitly represent and compute with probabilities, e.g. using information about both the mean and covariance of perceptual variables (cf. Koch et al., 1986; Ma, 2012; Ganguli & Simoncelli, 2011; Ma, 2010; Beck et al., 2011; Ma et al., 2006, 2008; T. S. Lee & Mumford, 2003; Knill & Pouget, 2004; Zemel & Pouget, 1998).

Bayesian methods applied to graphical models have provided a unified framework within which to understand generative and inverse inference, as well as statistical learning (Jordan & Weiss, 2002). And while it isn't always practical to develop a quantitative model for a complex visual function, the basic concepts provide a common language for describing how image representations with an area might be discovered from natural image regularities, how complexity is managed, and how reliable, flexible decisions may be made through the combination of feedforward and feedback flow of cortical information.

Representations and computations in visual hierarchies

In the following three sections, we discuss within-area, feedforward and feedback computations from an inferential perspective, with particular attention to how lateral/within-area and between-area (feedforward and feedback) processes may relate to primate vision. Because most relevant research has been on early retinotopic visual areas, our examples focus there. The computations and surface representations in early visual cortex may be more complex than traditionally thought, making V1 a good test-bed for ideas regarding hierarchical functions generally (T. S. Lee, 2003; Olshausen & Field, 2005; Gilbert & Sigman, 2007).

Within-area representations

Cortical maps are a fundamental, large-scale property of lateral, within-area cortical organization with a well-established empirical and theoretical history (Mountcastle, 1997; Hubel & Wiesel, 1977; Barlow, 1981). Specifically, the columnar organization within a visual area reflects the requirement that units representing similar image features should be nearby on the cortical surface (Durbin & Mitchison, 1990). This arrangement is believed to provide the basis for perceptual organization, for example to group local edges into object boundaries. The presumption is that local features of a similar

type can be more easily linked over cortical space. A given area represents spatially organized information of a similar type and level of abstraction (Connor et al., 2007; Orban, 2008). Are there natural image regularities that support the evolution, development, and adult plasticity of lateral, within-area feature representation? If so, what theoretical learning principles might help to explain the discovery and representation of regularities? How do the task requirements of object perception constrain representations?

Insight comes from computational studies that have shown how structured image knowledge can be discovered, through "unsupervised" as well as task-based learning (e.g. "supervised" learning) from collections of natural images. Such "discoveries" in an organism presumably arise through evolution and development of the visual system through exposure to natural images, as well as to their behavioral outcomes. It makes sense that early visual features would be more general-purpose, involving representations shared among many objects, and thus more strongly constrained by the statistical regularities in natural images, discoverable through unsupervised learning. As one moves up the visual hierarchy, the contingencies of primary tasks become more important. This may account for multiple parallel pathways (Nassi & Callaway, 2009; Freiwald & Tsao, 2010; Beauchamp et al., 2002), and the divergence, following V1 and V2, into multiple visual areas in which different causal contributions are discounted (integrated out) based on different task requirements. Such specialization would be constrained through adaptations based on outcomes (e.g. task-based or reinforcement learning) across phylogenetic and ontogenetic time scales.

Unsupervised learning of feature representations. An early idea was that, in its simplest form, N discrete levels (or areas, or layers of neural units) are required to detect Nth-order image regularities. With such a system in place, vision operates in a feedforward manner in which progressive conjunctions of features are detected, eventually leading to the detection of whole objects. Barlow (1990) suggested that mechanisms for learning Nth-order image regularities could rely on the detection of "suspicious coincidences" in the combinations of input features (i.e., test whether $p(s_1, s_2) >> p(s_1)p(s_2)$, and if so recode to remove this dependency). Some coding could be "hard-wired", and modulated or built during early development. At the behavioral level, it has been shown that human adults can learn, without supervision, part combinations by detecting co-occurence of features (Orbán et al., 2008; Fiser et al., 2010).

There have been a large number of computational studies aimed at explaining the neural population architecture in V1 in terms of efficient codes that exploit the regularities in natural images. Neural response properties, such as orientation and spatial frequency tuning in V1 neurons, are consistent with a sparse coding strategy adapted to the statistics of natural images (Olshausen, 1996; Hyvärinen, 2010). In addition, neurons in primary visual cortex 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 (additional) suspicious coincidences. Recently, Ziemba et al. (2013) developed a texture model based on high-order statistical dependencies in natural images that could account for selectivities in both macaque and human V2.

Purely bottom-up, unsupervised feature learning typically ignores task requirements (i.e. what to discount) and eventually the behavioral end-goal of a visual pathway needs to be taken into account[5].

[5] Discounting can be achieved through unsupervised learning. Cadieu & Olshausen (2012) show unsupervised learning of invariances of form by factoring out contributions from motion.

However, some task requirements are general, suggesting that certain kinds of information can be discounted early on.

Generic task constraints on early representations. It is believed that early vision involves both contour- and region-based linking (Grossberg & Mingolla, 1985; Lamme et al., 1998; T. S. Lee, 2003; Roe et al., 2009, 2012). For contour features, conditional probabilities, fit with natural image statistics, predict aspects of human contour perception, such as the Gestalt property of "good continuation" – nearby contour elements tend to have similar orientations (Geisler & Perry, 2009; Elder & Goldberg, 2002). Region-based grouping relies on the prior assumption of piece-wise smoothness in low- and higher-order intensive attributes (i.e. texture; Shi & Malik, 2000). The assumed function of edge- and region-based grouping is to compute surface representations that are more reliably associated with object than image properties, providing a front-end to a variety of object-based tasks, including recognition (Marr, 1982). And a first step would be to begin the process of discounting causes of image patterns that are not needed.

The accurate inference of illumination level and direction is low priority for both "what" and "how" tasks, which care primarily about objects and surfaces. This suggests that at least some components of illumination variation would be discounted early in the visual system. This is consistent with retinal lateral inhibition filtering out slow spatial gradients (presumed due to illumination), and emphasizing edges (presumed due to surface changes). However, illumination effects are complicated: slow gradients can also be caused by shape, and simple filtering neither accounts for human perception of brightness (Knill & Kersten, 1991; Kingdom, 2011), nor provides accurate reflection estimation in computer vision applied to natural images (Tappen et al., 2005).

This problem is naturally cast in terms of Bayesian inference, where the generative knowledge is contained in the image formation model, $I = f(E, R, S)$ and spatial priors on illumination (E), reflectance (R), and shape (S)–spatial maps called "intrinsic images" (Barrow et al., 1978). Conceptually, a Bayesian model would use a posterior proportional to the product of a likelihood function $p(I - f(R, S, E))$, and priors that characterize the spatial regularities in the natural patterns of reflectance, shape, while discounting illumination through integration (see Freeman, 1994). While computing intrinsic images from natural images can be done in special cases, it nevertheless remains a challenging problem (Grosse et al., 2009; Barron & Malik, 2012).

Perceptual evidence for human computation of an intrinsic image for reflectance comes from human lightness judgments which are more strongly correlated with reflectance than image intensity or contrast. The classic Craik-O'Brien lightness illusion, shown in the upper middle panel of Figure 4A, illustrates this. Regions with identical physical intensities appear to have different lightnesses. The functional interpretation is that the illusion is due to a mechanism designed to produce an estimate of surface reflectance, based on the assumption that reflectance changes are often abrupt, and illumination changes tend to be gradual (Figure 4B).

fMRI evidence for processes involved in computing a lightness map in human V1 and V2 is shown in Figure 4A (Boyaci et al., 2007). Activity in localized regions of visual cortical areas V1 and V2 (distant from the central edge) respond to a perceived change in lightness in the absence of a physical change in intensity (see lower panels in Figure 4A). While purely lateral computations have been invoked to explain this kind of "filling-in", it has also been shown that human V1 response to lightness change is also sensitive to perceptual organization of occluded surfaces, suggesting that top-down feedback may be involved (Boyaci et al., 2010).

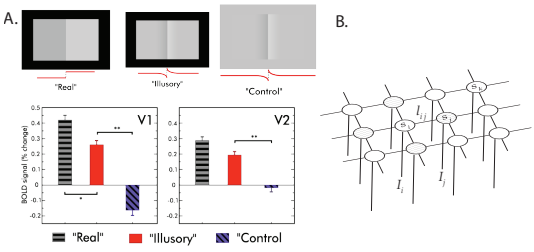

Figure 4. **A**.The upper middle panel shows a classic illusion known as the Craik-O'Brien effect. Away from the vertical border, the left and right rectangles have the same luminance, as indicated by red line which shows how light intensity varies from left to right. The interesting perceptual observation is that the left rectangle looks darker than the right. In fact, there is little difference between the appearance of a real intensity difference (upper left), and the illusory one. The lower graphs show that voxels in both V1 and V2 respond to apparent changes in lightness almost as strongly as real changes, as compared with a control. **B**. An undirected graph (Markov Random Field) can be used to formulate prior probabilities representing lateral, spatial statistical dependencies for contours and surface properties such as reflectance (cf. Marroquin et al., 1987; Kersten, 1991).

In addition to allowing for illumination variation, object recognition has the additional requirement that variations due to position and depth need to be discounted. We discuss within-area computations supporting invariant recognition in the later section on feedforward computations.

Learning hierarchically organized area representations for recognition. One can use the end-goal of object classification as a constraint on learning feature hierarchies through successive, top-down categorization of intermediate-level features. Here the invariance requirements are built into the choice of what distinguishes the top-level training classes. The basic principle is to learn diagnostic features (such as "fragments" or "features of intermediate complexity") that maximize the information for distinguishing object classes (Ullman et al., 2002). Humans and non-human primates seem to learn such features (Harel et al., 2007; Lerner et al., 2008; Hegdé et al., 2008; Kromrey et al., 2010). To build a feature hierarchy one applies this principle at the highest level to learn high-level features that optimally distinguish object classes. At the next level down the principle is again applied to learn lower-level features that distinguish the previous features learned, and so forth (Epshtein & Uliman, 2005). The task requirement of what to discount is built into the *a priori* selection of the training classes to be distinguished. Simulations have shown examples that once the features have been learned, accurate object recognition and localization can be achieved with one forward pass followed by one backward pass through the hierarchy (Epshtein et al., 2008).

Learning object compositions to manage image complexity. Compositionality refers to the human ability to construct hierarchical representations, whereby features/parts are used and shared to describe a potentially unlimited number of relational compositions (Geman et al., 2002). It is argued that without such a generative structure underlying scene and object compositions, we could not account for the efficiency and versatility with which humans can acquire and generalize visual knowledge. There is also evidence that humans exploit compositionally when learning new patterns (Barenholtz & Tarr, 2011). One aspect of compositionality is the ability to represent spatial relationships between parts, an idea with an early history (Waltz, 1972; Marr & Nishihara, 1978; Biederman, 1987; Hummel & Biederman, 1992). A second aspect, consistent with current models of primate recognition, is the idea of "reusable" features or parts, where lower levels have only a few feature types (e.g. edges), but these can be combined in many ways to make compositions of parts with increasing specificity at higher levels.

An underlying compositional structure to the visual world suggests that learning should exploit that

assumption, and computer vision work has demonstrated unsupervised learning of levels of reusable parts from natural image ensembles which they then apply to multi-class recognition (Zhu, Chen, Torralba, et al., 2011; see Figure 5).

Figure 5. **A**. Examples of the mean shapes of visual concepts automatically learned for multiple objects with part sharing between objects (Zhu, Chen, Torralba, et al., 2011; Zhu, Chen, & Yuille, 2011). The specificity and the number of types of features increases as one goes up the hierarchy, consistent in general terms, with the progression of neural selectivities as one moves up the ventral stream.

Feedforward computations

Invariant object recognition by the ventral stream requires discounting spatial position and size (Fukushima, 1988; Rolls & Foldiak, 1993; Riesenhuber & Poggio, 1999; DiCarlo et al., 2012). The basic feedforward computations are assumed to be the detection of conjunctions of features that belong together as part of an object, while at the same time discounting, through disjunction (which can be viewed as an approximation for "integrating out"), sources of variation, including position and scale.

It has been argued that a hierarchy of multiple areas is required to achieve functional invariance given the biological properties of neurons and their connections (Poggio, 2011). In this account, discounting is achieved incrementally through levels of the ventral stream, through the operation of AND-like (to detect feature conjunctions) and OR-like operations (to discount variations in position, size) over levels (Zhu et al., 2010), via simple and complex type cells respectively (Riesenhuber & Poggio, 1999).

During the first feedforward pass, information necessarily gets left behind in the race to quickly and accurately draw from a relatively small set of high priority, categorical hypotheses. But "no going back" requires strong *a priori* architectural assumptions regarding what constitutes high priority end-goals, as well as a strategic balancing of the trade-off between selectivity and invariance. Invariance is achieved at the cost of loss of information–too much loss and categories become indistinguishable; too little, and there are too many object types.

Using compositions. Compositional arguments may help to answer the question of why a hierarchical visual architecture desirable. A. L. Yuille & Mottaghi (2013) conjecture that the key problem of vision is complexity. The visual system needs to be organized in such a way that it can represent a very large number of objects and be able to rapidly detect which ones are present in an image. They demonstrate by mathematical analysis that this can be achieved using compositional models which are constructed in terms of hierarchical dictionaries of parts (see Figure 5). There are two key issues. Firstly, this visual architecture exploits part sharing between different objects which leads to great efficiency in representation and speed of detection. The lower-level parts are small and are shared between many objects. The high-level parts are larger (are composed from lower-level parts) and are shared less because they are more specific to objects. Secondly, objects are represented in a distributed hierarchical manner where the positions, and other properties, of the high-level parts are specified coarsely while the low-level parts are specified to higher-precision. This "executive-summary principle", combined with part-sharing,

can lead to exponential gains in the number of objects that can be represented, and the speed of recognition. For these types of models (based on Zhu, Chen, Torralba, et al., 2011) recognition is performed by propagating up hypotheses about which low-level parts are present to obtain an unambiguous high-level interpretation. Top-down processing, discussed in the next section, can be used to remove false low-level hypotheses (using a high-level context).

We noted at the beginning the extraordinary reliability and versatility of human vision, in its ability to respond both to challenging input (partially hidden objects, confusing background clutter, camouflage) and diverse task demands, such as the fox description example. What if the information for a low-level hypothesis (e.g. precise object boundary location, or the direction of movement of a local edge) is not sufficiently reliable from a single forward pass? What if a task needs information not present or easily computable within top-levels of the hierarchy? Earlier we noted some of the computational advantages of generative models in resolving residual ambiguity. The next section discusses human behavioral and neuroimaging experiments, based primarily on the effects of context on local decisions, that are consistent with cortical feedback computations.

Feedback computations

Most interpretations of top-down visual processes have focused on selective attention, which is viewed as feedback that improves sensitivity at attended locations and/or features (Desimone & Duncan, 1995; Noudoost et al., 2010; Petersen & Posner, 2011). Top-down (or "endogenous") visual attention is typically interpreted as selective tuning in which information is routed through the visual processing hierarchy to amplify some features relative to others. In particular, Tsotsos et al. (1995) argues that attention acts to optimize visual search for features through a top-down hierarchy of winner-take-all processes. A Bayesian perspective emphasizes preservation of information about uncertainty about hypotheses, and its sequential reduction by message-passing between units and areas (T. S. Lee & Mumford, 2003). In addition, the diversity of visual descriptions suggests flexible access to hierarchically organized information. While there is no direct evidence, at this time, for neural populations representing hypotheses rather than decisions, or for probabilistic computations (as in message passing) (Lochmann & Deneve, 2011), there are behavioral and neuroimaging results that are suggestive of Bayesian top-down computations down the cortical hierarchy. We briefly describe some of them.

Coarse-to-fine inferences. A basic lesson learned from computer vision is: to be certain about a local region of a natural image requires knowledge of the whole (Figure 1B). Local perceptual decisions can be automatic, constrained by spatial or temporal context (as in priming or prior learning, cf. Hsieh et al. (2010)) or be consciously task driven and specified by a higher-level "executive".

Automatic (and executive) coarse-to-fine inference can be modeled as an initial high-level decision which "fixes" the value in the upper level of a hierarchical model, constraining subsequent lower-level decisions (Figure 3C). An optimal decision restricted to a high level requires integrating out intermediate-level parameters. Several behavioral results are consistent with Bayesian coarse-to-fine computations over a simple hierarchical graph structure (Knill, 2003; Körding et al., 2007; Wozny et al., 2010; Wu et al., 2008; Stocker & Simoncelli, 2008). For example, Wu et al. (2008) have shown that human velocity discrimination performance is consistent with an initial classification of motion type (rotation, expansion, translation).

Does feedback enhance or suppress feature representations?. There are several ways in which top-down signals could change the neural representation of the probability distributions. Top-down

processes may enhance or suppress low-level features consistent with a descriptions or hypotheses at higher levels (Mumford, 1992; Rao & Ballard, 1999; T. S. Lee & Mumford, 2003; A. Yuille & Kersten, 2006; Friston & Kiebel, 2009; Spratling, 2012). Enhancement is consistent with neurophysiological and brain imaging studies that have demonstrated that perceptual grouping is correlated with the amplification of neural responses throughout the visual hierarchy (Kourtzi et al., 2003; Roelfsema, 2006). Enhancement is also consistent with the compositional models described earlier, in which information about a given object is represented and bound hierarchically. In principle and depending on the task, feature enhancement could either be automatic, or correspond to executive, top-down ("endogenous") attention. There is also evidence for suppression of lower-level features which are consistent with a high-level hypothesis. Such a mechanism, sometimes referred to as "predictive coding", could support detecting and subsequently processing image information that does not fit with the current interpretation. Such a bottom-up signal would provide the basis for exogenous attention, but in contrast to a saliency computation (Li, 1997; Rao & Ballard, 2013; Itti & Baldi, 2009; Chen et al., 2013; L. Zhang et al., 2008; X. Zhang et al., 2012), which could be accomplished laterally, the signal increase is the result of a top-down prediction that fails.

Figure 6 shows behavioral evidence consistent with a predictive coding interpretation of "explaining away", in which occlusion cues provide an explanation for the missing vertices of the diamond (see Kersten et al. 2004). When the diamond is seen during an adaptation period (Figure 6C), there was an increase in the strength of adaptation to shape (e.g. adapting to a skinny diamond results in seeing a standard comparison diamond as fatter); at the same time, there was a decrease in the strength of adaptation to the local orientation of comparison gratings. The converse was found when the occlusion cues were inconsistent with a diamond (Figure 6D). The interpretation, consistent with other research, rests on the assumption that the sites of orientation and shape adaptation are in early and higher-level cortical areas, respectively.

There is also evidence from human fMRI studies for context-dependent suppression of neural activity in earlier areas in some cases (Murray et al., 2002; Fang, Kersten, & Murray, 2008; Alink et al., 2010; Rauss et al., 2011; Cardin et al., 2011), but not all (Mannion et al., 2013). And suppression measured using fMRI activity does not necessarily show the spatial specificity suggested by the above adaptation study or by theory (Wit et al., 2012).

In the language of signal detection theory, the suppression of false and true positives through feedback could both be computationally useful. Suppression of false positives and/or enhancement of true positives in one population of neurons could serve to bind object representations with parts and features at lower levels, as in the above compositional model. At the same time, increased activity in another neural population could signal false positives, i.e. inconsistent features that need to be resolved with other hypotheses (Rao & Ballard, 1999; Friston, 2005; Clark, 2012). Ultra-high field fMRI with sub-millimeter resolution has 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 prediction errors.

Hierarchically organized expertise. In the race to make high priority decisions quickly, as in "core" or basic-level recognition (DiCarlo et al., 2012), detailed information about position, size, shape, material and illumination direction is left behind, but not necessarily discarded. We know that human vision can discriminate subtle differences in shape and material, and even see gradients of illumination, suggesting that it has the ability to access low-level information, or recover transformations discounted earlier (Grimes & Rao, 2005; Tenenbaum & Free-

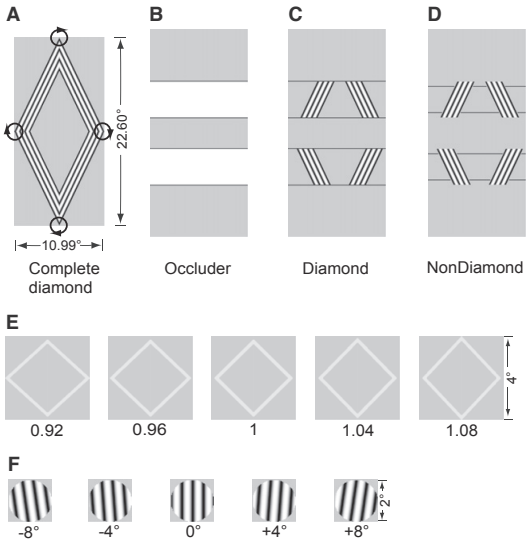

Figure 6. In studies with human subjects, He et al. (2012) showed that perceptual grouping amplifies the effect of adaptation to a whole shape, while reducing the strength of adaptation to local tilt. Thus perceptual grouping is consistent with enhancement of high-level shape representation and attenuation of the low-level feature representation, possibly the result of top-down predictive coding. **A** The diamond corners undergo tight rotations during adaptation. When covered by occluder, shown in **B**, the diamond can still be perceived as shown in **C**. The diamond percept can be disrupted by the occlusion relationships shown in **D**. **E**, **F** show test stimuli for measuring the aftereffects of shape, and tilt, respectively. Figure adapted from He et al. (2012).

man, 2000; Olshausen et al., 1993).

The ability of vision to extract information of different types and across multiple spatial scales raises the possibility that feedback signals in visual hierarchies have a richer computational function than so far discussed. Neuronal activity and receptive fields as early as primary visual cortex appear to be modulated by task requirements (Gilbert & Sigman, 2007; McManus et al., 2011). The interesting possibility is that the representation of information across levels of the visual hierarchy is accessible for a range of tasks. But for what functions, representations, and operations?

One possibility is that the optimal machinery, representations, or coordinate frames for the task exists at a lower level. T. Lee et al. (1998) suggested that higher level computations that involve fine-grain spatial and orientation information would necessarily involve V1. There are a number of results consistent with this idea. For example, Harrison & Tong (2009) analyzed patterns of fMRI voxel activity to show that visual areas from V4 down to V1 can retain orientation information held in working memory over many seconds. Variations in perceptual learning and its transfer may be understood in terms of whether the learning task requires the "expertise" of a lower- vs. higher-level of processing (Hochstein & Ahissar, 2002). In another study, Williams et al. (2008) found that the measured patterns of fMRI activity near foveal retinotopic cortex could discriminate which object category the observers had been seeing with their peripheral vision. It has been known for some time that visual imagery involving fine spatial discrimination, and even orientation-specific tactile tasks may activate representations in early visual areas (Kosslyn et al., 1993; Kosslyn & Thompson, 2003; Zangaladze et al., 1999; Lucan et al., 2010).

Consider the everyday task of inferring an object's physical size from its image. This is a nontrivial computation with no current computer vision solution. The visual system has to decide which features form the boundary of the object's image, i.e. a challenging segmentation and grouping problem, that could require feedback to retinotopic areas. The locations of these features are needed to summarize the average diameter, or angular size. Then to estimate physical from angular size, the system needs to process the larger context in order to take the object's depth into account. Further, size perception often involves comparisons with other objects, raising the question of where to make those.

The complexity of the analysis suggests an interplay between high-level representations, and early retinotopic areas, particularly V1 for its high spatial precision. Studies by Murray et al. (2006) and Fang, Boyaci, et al. (2008) used a classic depth illusion to show that the pattern of spatial activity in V1 activity is indeed modulated by 3D depth context (Figure 7). When an object (a ring) appeared bigger, its "neural image" on V1 was bigger (i.e. activation shifts to a more eccentric representation of the visual field). This effect was significantly stronger when observers attended to the object, consistent with feedback from higher-level areas that process depth in the larger context of the scene. Psychophysical data is also consistent with a top-down influence of depth on orientation-selective, and putatively early cortical regions (Arnold et al., 2008).

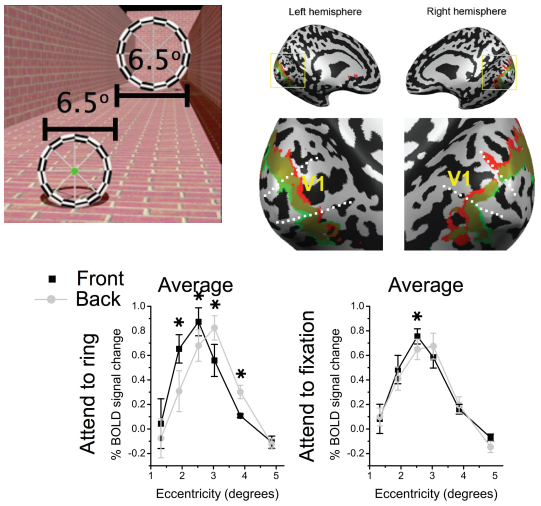

Figure 7. This figure illustrates how global, contextual information for 3D depth can shift the spatial extent of activity in human V1.

The longer you look, the more you see.

Not many decades ago "perception" seemed to be not much more than a screen, admittedly with some puzzling distortions, viewed by a high-level executive agent. Then retinal and cortical studies showed that neurons were doing much more than transmitting image information: they were emphasizing certain kinds of information, such as edges, at the expense of others (smooth gradients). This led to the idea of the retina and early visual cortical areas as spatio-temporal filter banks. But still, the emphasis was on early perceptual processing as a set of filtering stages, effectively passing decisions forward from one stage to the next (Lennie, 1998).

Computer vision has provided the perspective that in order to produce useful behavioral outcomes, the human visual system is solving a decoding problem whose understanding requires concepts and a level of analysis beyond traditional neural network filtering. The past decade has seen substantial progress in both the computational and neural understanding of how vision could be solving the problems of object perception. We have discussed potential limitations on the robustness and versatility of vision with strictly feedforward processing and have reviewed arguments and results suggesting that both automatic and executive processes access built-in image knowledge at several levels of abstraction. We conjecture that the brain's ability to solve the problems of local uncertainty and task versatility rests on deep generative knowledge of the structure of images. A major challenge for the future is to better understand the way the brain represents and controls the top-down utilization of this knowledge (cf. Ullman, 1984; Blanchard & Geman, 2005), eventually explaining how the brain enables us to see so much in just one picture of a fox.

References

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

Arnold, D. H., Birt, A., & Wallis, T. S. A. (2008,

June). Perceived Size and Spatial Coding. *Journal of Neuroscience*, *28*(23), 5954–5958.

Bar, M. (2003, May). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600–609.

Barenholtz, E., & Tarr, M. J. (2011, April). Visual learning of statistical relations among nonadjacent features: Evidence for structural encoding. *Visual Cognition*, *19*(4), 469–482.

Barlow, H. (1981, May). The Ferrier Lecture, 1980: Critical Limiting Factors in the Design of the Eye and Visual Cortex. *Proceedings of the Royal Society B: Biological Sciences*, *212*(1186), 1–34.

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

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

Barrow, H., Tenenbaum, J., & SRI International. Artificial Intelligence Center. Computer Science and Technology Division. (1978). Recovering intrinsic scene characteristics from images.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*, *34*(1), 149–159.

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

Biederman, I. (1987, April). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Blanchard, G., & Geman, D. (2005, June). Hierarchical testing designs for pattern recognition. *The Annals of Statistics*, *33*(3), 1155–1202.

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.

Burgi, P. Y., Yuille, A., & Grzywacz, N. M. (2000, August). Probabilistic motion estimation based on temporal coherence. *Neural Computation*, *12*(8), 1839–1867.

Cadieu, C. F., & Olshausen, B. A. (2012, April). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation; Neural Computation*, *24*(4), 827–866.

Callaway, E. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, *21*, 47–74.

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.

Chang, L., Jin, Y., Zhang, W., Borenstein, E., & Geman, S. (2011). Context, computation, and optimal roc performance in hierarchical models. *International Journal of Computer Vision*, *93*(2), 117–140.

Chen, C., Zhang, X., Wang, Y., & Fang, F. (2013). Measuring the Attentional Effect of the Bottom-Up Saliency Map of Natural Images. *Intelligent Science and Intelligent ...*.

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, A. (2012). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science.

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007, April). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, *17*(2), 140–147.

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). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Durbin, R., & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, *343*(6259), 644–647.

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

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, *2*(4), 324–353.

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.

Epshtein, B., & Uliman, S. (2005, October). Feature hierarchies for object classification. , *1*, 220–227.

Fang, F., Boyaci, H., Kersten, D., & Murray, S. O. (2008, November). Attention-dependent representation of a size illusion in human V1. *Current biology : CB*, *18*(21), 1707–1712.

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.

Felleman, D., & Van Essen, D. (1991, January). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

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

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

Freiwald, W. A., & Tsao, D. Y. (2010, November). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, *330*(6005), 845–851.

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.

Fukushima, K. (1988). Neocognitron - a Hierarchical Neural Network Capable of Visual-Pattern Recognition. *Neural Networks*, *1*(2), 119–130.

Ganguli, D., & Simoncelli, E. P. (2011, November). *Neural implementation of Bayesian inference using efficient population codes* (Tech. Rep.).

Geisler, W. S. (2011, April). 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.

Gilbert, C. D., & Sigman, M. (2007, June). Brain States: Top-Down Influences in Sensory Processing. *Neuron*, *54*(5), 677–696.

Gold, J. M., Abbey, C., Tjan, B. S., & Kersten, D. (2009, November). Ideal Observers and Efficiency: Commemorating 50 Years of Tanner and Birdsall: Introduction. *Journal of the Optical Society of America A, Optics, Image Science, and Vision*, *26*(11), IO1–IO2.

Green, C. S., Pouget, A., & Bavelier, D. (2010, September). Improved probabilistic inference as a general learning mechanism with action video games. *Current biology : CB*, *20*(17), 1573–1579.

Grimes, D., & Rao, R. P. (2005). Bilinear sparse coding for invariant vision. *Neural Computation*, *17*(1), 47–73.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Attention, Perception, & Psychophysics*, *38*(2), 141–171.

Grosse, R., Johnson, M., Adelson, E., & Freeman, W. (2009). Ground truth dataset and baseline evaluations for intrinsic image algorithms. *Computer Vision, 2009 IEEE 12th International Conference on*, 2335–2342.

Guillery, R. W., & Sherman, S. M. (2002, January). Thalamic relay functions and their role in corticocortical communication: generalizations from the visual system. *Neuron*, *33*(2), 163–175.

Harel, A., Ullman, S., Epshtein, B., & Bentin, S. (2007, July). Mutual information of image fragments predicts categorization in humans: Electrophysiological and behavioral evidence. *Vision Research*, *47*(15), 2010–2020.

Harrison, S. A., & Tong, F. (2009, February). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635.

He, D., Kersten, D., & Fang, F. (2012, June). Opposite modulation of high- and low-level visual aftereffects by perceptual grouping. *Current biology : CB*, *22*(11), 1040–1045.

Hegdé, J., Bart, E., & Kersten, D. (2008, April). Fragment-based learning of visual object categories. *Current biology : CB*, *18*(8), 597–601.

Hinton, G. (2009, November). Learning to represent visual input. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1537), 177–184.

Hochstein, S., & Ahissar, M. (2002). View from the Top:: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, *36*(5), 791–804.

Hsieh, P. J., Vul, E., & Kanwisher, N. (2010). Recognition Alters the Spatial Pattern of fMRI Activation in Early Retinotopic Cortex. *Journal of Neurophysiology*.

Hubel, D., & Wiesel, T. (1977). Ferrier lecture: Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 1–59.

Hummel, J. E., & Biederman, I. (1992, July). Dynamic binding in a neural network for shape

recognition. *Psychological Review*, *99*(3), 480–517.

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.

Jordan, M., & Weiss, Y. (2002). Probabilistic inference in graphical models. *Handbook of neural networks and brain theory*.

Kanwisher, N. (2010, May). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(25), 11163.

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

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

Kersten, D. J., & Yuille, A. L. (2013, April). Vision: Bayesian Inference and Beyond. (88), 1–16.

Kingdom, F. A. A. (2011, April). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, *51*(7), 652–673.

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

Knill, D. C., & Kersten, D. (1991, May). Apparent surface curvature affects lightness perception. *Nature*, *351*(6323), 228–230.

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.

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.

Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Maljkovic, V., Weise, S. B., Chabris, C. F., et al. (1993, July). Visual Mental Imagery Activates Topographically Organized Visual Cortex: PET Investigations. *Journal of Cognitive Neuroscience*, *5*(3), 263–287.

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin*, *129*(5), 723–746.

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.

Kourtzi, Z., Tolias, A. S., Altmann, C. F., Augath, M., & Logothetis, N. K. (2003, January). Integration of local features into global shapes-monkey and human fMRI studies. *Neuron*, *37*(2), 333–346.

Kromrey, S., Maestri, M., Hauffen, K., Bart, E., & Hegdé, J. (2010, November). Fragment-Based Learning of Visual Object Categories in Non-Human Primates. *PLoS ONE*, *5*(11), e15444.

Lamme, V. A., Sup, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing cortex. *Current Opinion in Neurobiology*, *8*, 529–535.

Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 157–224.

Lee, T., Mumford, D., Romero, R., & Lamme, V. A. (1998, June). The role of the primary visual cortex in higher level vision. *Vision Research*, *38*(15-16), 2429–2454.

Lee, T. S. (2003, March). Computations in the early visual cortex. *Journal of Physiology-Paris*, *97*(2-3), 121–139.

Lee, T. S., & 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.

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

Lerner, Y., Epshtein, B., Ullman, S., & Malach, R. (2008, July). Class information predicts activation by object fragments in human object areas. *Journal of Cognitive Neuroscience*, *20*(7), 1189–1206.

Li, Z. (1997). Primary cortical dynamics for visual grouping.

Lochmann, T., & Deneve, S. (2011, October). Neural processing as causal inference. *Current Opinion in Neurobiology*, *21*(5), 774–781.

Lucan, J. N., Foxe, J. J., Gomez-Ramirez, M., Sathian, K., & Molholm, S. (2010). Tactile shape discrimination recruits human lateral occipital complex during early perceptual processing. *Human Brain Mapping*, NA–NA.

Lund, J., Angelucci, A., & Bressloff, P. C. (2003). Anatomical substrates for functional columns in macaque monkey primary visual cortex. *Cerebral Cortex*, *13*(1), 15–24.

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

Ma, W. J. (2012, October). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, *16*(10), 511–518.

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.

Ma, W. J., Beck, J. M., & Pouget, A. (2008, April). Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology*, *18*(2), 217–222.

Maloney, L. T., & Zhang, H. (2010, November). Decision-theoretic models of visual perception and action. *Vision Research*, *50*(23), 2362–2374.

Mannion, D. J., Kersten, D. J., & Olman, C. A. (2013, September). Consequences of polar form coherence for fMRI responses in human visual cortex. *NeuroImage*, *78*(C), 152–158.

Markov, N. T., Vezoli, J., Chameau, P., Falchier, A., Quilodran, R., Huissoud, C., et al. (2013, August). The anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J Comp Neurol*, n/a–n/a.

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.

Marr, D., & Nishihara, H. K. (1978, February). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society B: Biological Sciences*, *200*(1140), 269–294.

Marroquin, J., Mitter, S., & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 76–89.

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.

Mountcastle, V. B. (1997, April). The columnar organization of the neocortex. *Brain*, *120 (Pt 4)*, 701–722.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Murray, S. O., Boyaci, H., & Kersten, D. (2006, February). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience*, *9*(3), 429–434.

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.

Nassi, J. J., & Callaway, E. M. (2009, April). Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, *10*(5), 360–372.

Noudoost, B., Chang, M. H., Steinmetz, N. A., & Moore, T. (2010, April). Top-down control of visual attention. *Current Opinion in Neurobiology*, *20*(2), 183–190.

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.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. (1993, November). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*(11), 4700–4719.

Olshausen, B. A., & Field, D. J. (2005, August). How close are we to understanding v1? *Neural Computation*, *17*(8), 1665–1699.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(7), 2745.

Orban, G., & Wolpert, D. M. (2011, August). Representations of uncertainty in sensorimotor control. *Current Opinion in Neurobiology*, *21*(4), 629–635.

Orban, G. A. (2008, January). Higher order visual processing in macaque extrastriate cortex. *Physiological reviews*, *88*(1), 59–89.

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

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*.

Purves, D., & Lotto, R. (2003). *Why we see what we do: An empirical theory of vision.* Sunderland, Mass., U.S.A. : Sinauer Associates.

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

Rao, R. P., & Ballard, D. H. (2013, April). Probabilistic Models of Attention based on Iconic Representations and Predictive Coding. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention* (pp. 1–16). Academic Press.

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.

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., Chen, G., & Lu, H. (2009, May). Visual System: Functional Architecture of Area V2. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 331–349). Elsevier.

Roelfsema, P. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, *29*, 203–227.

Rolls, E., & Foldiak, P. (1993). Learning invariant responses to the natural transformations of objects. *Neural Networks*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Cambridge, Mass. : MIT Press.

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

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(8), 888–905.

Spratling, M. W. (2012, December). Distinguishing Theory from Implementation in Predictive Coding Accounts of Brain Function. , 1–3.

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

Tappen, M., Freeman, W., & Adelson, E. (2005). Recovering intrinsic images from a single image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *27*(9), 1459–1472.

Tenenbaum, J. B., & Freeman, W. (2000). Separating style and content with bilinear models. *Neural Computation*, *12*(6), 1247–1283.

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

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, *78*(1), 507–545.

Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image parsing: Unifying segmentation, detection, and recognition. In *International journal of computer vision* (pp. 113–140).

Ullman, S. (1984). Visual routines. *COGNITION*, *18*(1-3), 97–159.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002, June). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*.

Vilares, I., & Körding, K. P. (2011, April). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22–39.

Wallisch, P., & Movshon, J. A. (2008, October). Structure and Function Come Unglued in the Visual Cortex. *Neuron*, *60*(2), 194–197.

Waltz, D. L. (1972). *Generating semantic descriptions from drawings of scenes with shadows* (Tech. Rep.).

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002, May). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Williams, M. A., Baker, C. I., Beeck, H. P. Op de, Shim, W. M., Dang, S., Triantafyllou, C., et al. (2008). Feedback of visual object information to foveal retinotopic cortex. *Nature Neuroscience*, *11*(12), 1439–1445.

Wit, L. H. de, Kubilius, J., Wagemans, J., & Beeck, H. P. Op de. (2012, October). Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. *Journal of Vision*, *12*(11), 12–12.

Wolpert, D. M., & Landy, M. S. (2012, December). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*(6), 996–1003.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology*, *6*(8), e1000871.

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. (2010, August). An information theory perspective on computational vision. *Frontiers of Electrical and Electronic Engineering in China*, *5*(3), 329–346.

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

Yuille, A. L., & Mottaghi, R. (2013). Complexity of Representation and Inference in Compositional Models with Part Sharing. *arXiv preprint arXiv:1301.3560*.

Zangaladze, A., Epstein, C. M., Grafton, S. T., & Sathian, K. (1999, October). Involvement of visual cortex in tactile discrimination of orientation. *Nature*, *401*(6753), 587–590.

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.

Zemel, R. S., & Pouget, A. (1998, February). Probabilistic interpretation of population codes. *Neural Computation*, *10*(2), 403–430.

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.

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012, January). Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron*, *73*(1), 183–192.

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. (2011, January). 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.

Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., Movshon, J. A., & Freeman, J. (2013, May). A functional and perceptual signature of the second visual area in primates. *Nature Publishing Group*, 1–12.