# Vision as Bayesian Inference: Analysis by Synthesis?

Alan Yuille[1] and Daniel Kersten[2].

Departments of Statistics UCLA[1], Psychology University of Minnesota [2].
emails: yuille@stat.ucla.edu, kersten@umn.edu.

## Abstract

We argue that the study of human vision should be aimed at determining how humans perform natural tasks on natural images. Attempts to understand the phenomenology of vision from artificial stimuli, though worthwhile as a starting point, risk leading to faulty generalizations about visual systems. In view of the enormous complexity of natural images, they are similar to trying to evaluate the performance of a soldier in battle from his ability at playing with a water pistol. Dealing with this complexity is daunting, but Bayesian inference on structured probability distributions offers the ability to design theories of vision that can deal with the complexity of natural images and which use analysis by synthesis strategies with intriguing similarities to the brain.

## Introduction: Perception as inference

Experimental studies of biological vision systems have typically been performed using artificial stimuli and tasks. While much progress has been made, it is questionable how much light these studies shed on the performance of biological vision systems when faced with the complexity of natural images and natural tasks such as segmentation and object detection and recognition. For example, recent studies suggest that the responses of neurons in V1 to natural stimuli cannot be predicted from their responses to artificial stimuli [3]. It is well known to computer vision researchers that vision algorithms that work on artificial stimuli almost never generalize to natural images (e.g. there are many algorithms which solve the correspondence problem and estimate depth on Julesz random dot stereograms, but give terrible results when applied to natural images).

We argue that the major difficulty of vision arises because natural images are both complex and objectively ambiguous. Similar 3D objects can result in different images, and different objects can result in similar images. Moreover, typical images can be highly complex and consist of hundreds of objects, many of which are overlapping. Therefore theoretical and experimental attempts to understand biological systems must come to terms with the daunting complexity of images somwhat analogous to the way that molecular biologists have had to tame the complexity of the genome.

In this article, we argue that the theoretical principles described in this special issue are rich enough to deal with the complexities and ambiguities of natural images and to perform major perceptual tasks such as recognition and segmentation. The approach is based on Bayesian inference using probability distributions defined on structured representations [42, 44]. Vision is treated as an inverse inference problem, in the spirit of Helmholtz [19], where the goal is to estimate the factors that have generated the image. Two major themes follow naturally from this approach.

Firstly, vision as inverse inference presupposes a formalization of how the input is generated, and which of the causes of that input should be estimated. This probabilistic generative approach

enables us to define Bayesian Ideal Observers (BIO) based on the principles of Bayesian Decision Theory and using structured representations and distributions which subsume the standard models of signal detection theory [17] and give theories of optimal performance against which human performance can be compared (the article by Griffiths and Yuille introduces these concepts). Recent survey papers justify this approach and give an entry point to the growing body of literature on this approach [31, 16, 25]. An alternative perspective is given by Howe *et al* [20] who question both the use of Bayesian Decision Theory and whether the factors being estimated are physical variables (like reflectance, size, i.e. Helmholtz [19]) or "orderings of visual stimuli" based on past experience with all such stimuli (as advocated in [20]).

Secondly, this inverse inference perspective suggests that the inference algorithm should combine a top-down generative component with bottom-up processing. The generative component allows the system to internally simulate, or synthesize, from the probability distributions and so is known as "analysis by synthesis" and relates to the forward and backward projections in the brain [30] [34, 45, 18, 39, 14, 15] (see [8] for how this may relate to neurotransmitters). But it must be emphasized that analysis by synthesis is not necessarily required for Bayesian inference (see, for example, work on the detection of hands [6]) and so we will give arguments for it in the next section.

The themes in this paper are common to other aspects of cognitive science as described in this special issue (see introduction by Chater, Tenenbaum and Yuille). The ability to simulate by generative models also occurs naturally within visuo-motor control where they can be used to simulate the consequence of actions, cf. [23, 41, 12]. Mental imagery [27] also suggests the ability to do internal simulation, but over long time scales. At a more theoretical level, statistical inference on structured representations offers a common mathematical framework for cognitive science which leads naturally to theoretical models for coupling different sensory modalities and for integrating perception with planning. For example, recent work [10], [40] has built on theoretical studies [5],[49] to model the coupling of visual cues and haptic cues and shown good fit with experimental data.

## The need for Bottom-up and Top-down processing

We now give arguments why visual inference requires bottom-up and top-down components. This contrasts with standard textbook theories of vision which favour bottom-up processing based on computing low-level representations such as edge maps (cf. [38]) or 2-1/2D representations of depth and shape (cf. [32]).

First observe that low-level vision is ambiguous. For example, it is extremely difficult to determine whether there is an edge present in a small region of an image [33]. These findings are supported by empirical studies of the relative ineffectiveness of edge detectors on natural images [26] and the limitations of regional cues for segmentation [42].

By contrast, high-level vision is rarely ambiguous. The patterns of objects, such as faces or other objects, are complex and rarely occur by chance. Moreover, they are often easy to resolve when they do. For example, patterns in the bark of a tree may occasionally look like a face, see Figure 5, but can easily be disambiguated by the alternative explanations for these patterns. This lack of ambiguity for high-level vision is highlighted by the recent successes of computer vision algorithms for detecting high-level objects such as faces and text reliably from natural images [48, 46, 4].

From this perspective, a major problem for vision systems is how to use low-level cues to rapidly

access the correct high-level models so as to quickly resolve the low-level ambiguities. The difficulty is knowing which high-level model(s) to use. Consider the text-book example of the black and white dalmation dog [21]. Low-level cues for this image contain little evidence to activate a high-level dog model, and so naive subjects take a long time to detect the dog. But subjects who have seen the image before, and know that there is a dog in it, can perceive it instantaneously.

This suggest a visual system where low-level cues make bottom-up proposals which are validated by high-level models. This bottom-up processing should include standard methods such as edge detection and grouping by regional properties – but it must also include special processes for detecting important objects such as faces and text [48, 4], and for rapidly classifying the scene [37]. These high-level models access the image, or a filtered version of it, in a top-down process to ensure consistency of the image interpretation. The early visual areas such as V1/V2 are the most natural candidates for such an image representation [34, 28]. In certain cases, the bottom-up cues are sufficiently unambiguous and so the object, or scene structure, can be detected without high-level feedback.

## Generative Models, Image Parsing and Analysis by Synthesis

We now describe how this visual system can be implemented. We first give an introductory example which illustrates the main points of the approach on simple examples. Then we describe a more advanced theory that applies to natural images.

### Introduction to Generative Models and Analysis by Synthesis

First, we consider models for generating an image. Suppose we start with a simple vocabulary of shapes and patterns which contains the letters $A, B, C, ....$ We can define a simple probability model for generative images built out of this vocabulary by using templates for each letter and allow the letter to be placed randomly at any position in the image. We can also give an ordering on the letters so that one letter can completely, or partially, occlude another letter. Sampling from this distribution will yield images as shown in Figure 1A.

Next, we can make this generative model richer by expanding the vocabulary to include additional objects such as rectangular bars and fragments of letters, Figure 1B. The addition of the elements allows us to generate more complex images but the richness of the vocabulary makes some images potentially ambiguous. The same image can be generated in two distinct ways, see Figure 1B.

This ambiguity in generation leads to potential ambiguities of the inverse process of inference/interpretation. This is resolved by having probabilities on the different ways that images can be generated. The chance alignment of fragments is more likely than a $B$ with an invisible (white) occluder (see Figure 1B). But a $B$ with a visible (black) occluding bar is more likely than a set of fragments which are accidentally aligned with the bar.

We now extend the vocabulary of the generative model in several ways, see Figure 1C. Firstly, we we can allow the letters to have properties such as size, font, and shading pattern. Secondly, we can put probabilities on the spatial relations between letters so that, for example, they line up to form words. This extension of the vocabulary leads to further ambiguities.

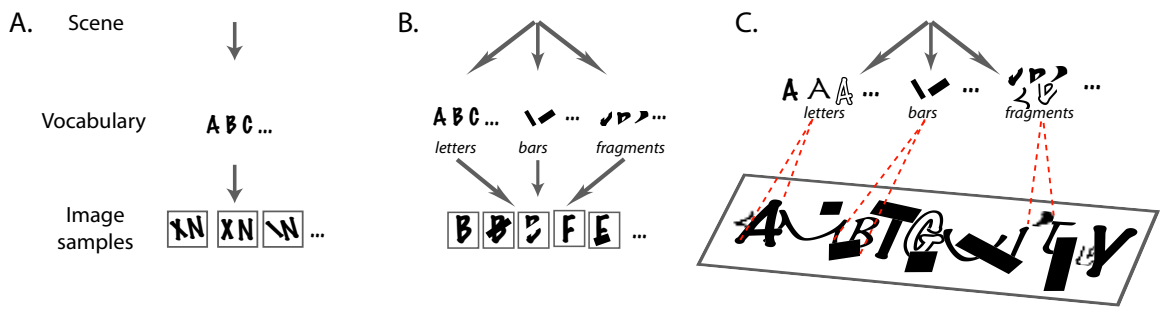


Figure 1: Left Panel (A). A simple vocabulary for generating the image. There is little, or no, ambiguity in interpreting images. At worst, the letter $X$ may be confused with a slanted $I$ partially occluding a vertical $I$. Centre Panel (B). A richer vocabulary. A given cause, such as a particular letter, can be manifest in many different images. But there are now multiple ways to generate identical images, see text. Right Panel (C). The richer the vocabulary, the greater the image ambiguity, and the harder it is to interpret the image. This leads to a formidable inference problem.

But now we turn to a different issue – how fast can we perform inference to determine the most likely way the image was generated? Our explanations, so far, have assumed that we can check all possible ways to generate the image and decide on the most probable. But this becomes completely impractical when the size of the vocabulary becomes very large. So how can we hope to solve the inverse inference problem of estimating which configuration of objects is most likely to have generated the observed image?

We propose an "analysis by synthesis" strategy where low-level cues, combined with spatial grouping rules (similar to Gestalt laws), make bottom-up proposals which activate hypotheses about objects and scene structures. These hypotheses are accepted, or rejected, by direct comparison to the image (or a filtered version of it) in a top-down process. These bottom-up proposals come with probabilities, which are a measure of their strength. If bottom-up proposals are sufficiently strong (i.e. the low-level cues are sufficiently unambiguous) then they may be accepted without any need for verification at a lower level.

For example, consider an image containing letters, rectangles, and letter fragments, Figure 2. We can obtain bottom-up cues in several ways. Firstly, we can run an edge detector to discard the shading information (which is often variable) and then spatially group the edges into segments by using the principles of continuity, parallelism and colinearity. This will enable us to get cues for the positions of rectangles and the identity of some letters. These features are used to make bottom-up proposals regarding possible objects. Remaining ambiguity is eliminated in a synthesis stage that tests how well the object models explain the image features. Other more sophisticated cues can be used for the proposals. For example, we can treat text as a type of texture and design detectors which respond to characteristic patterns of text (e.g. [4]). We stress that this is an illustration of this model only, ignoring shading information may be a mistake for other more realistic stimuli.

## Natural Image Parsing

The acid test for generative models and analysis by synthesis algorithms is whether they can be extended to deal with the complexities of natural images. We now describe recent work which

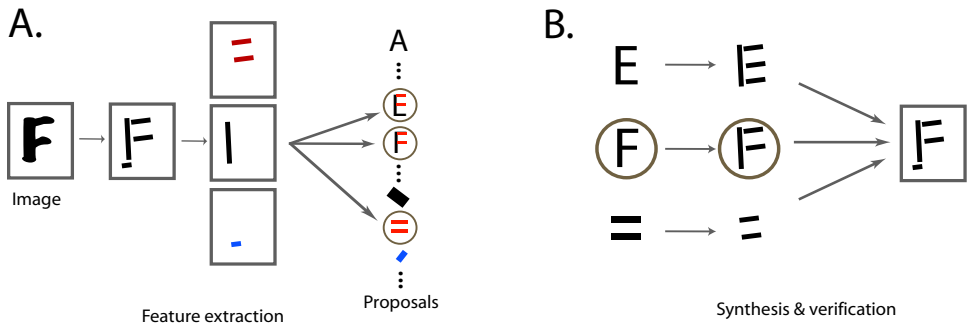


Figure 2: Analysis by synthesis. A. Low-level processing (left panel) can extract edge features, such as bars, and use conjunctions of these features to make bottom-up proposals to access the higher-level models of objects. B. The high-level objects access the image top-down to validate or reject the bottom-up proposals (right panel). In this example, the low-level cues propose that the image can be interpreted as an $E$, an $F$, or a set of parallel bars. But interpreting it as an $F$ explains almost all the features in the image and is preferred.

suggests that they can.

This work uses a generative model for images [42, 44] which is similar to a two level probabilistic context free grammar (PCFG) (see article by Griffiths and Yuille) and hence we refer to the approach as image parsing. This model is illustrated in Figure (3) (left panel) where the root node represents the entire image scene. The first level corresponds to the non-terminal nodes representing parameterized models of objects, such as faces or letters, or of generic regions such as textures or shaded patterns. A non-terminal node $i$ has attribute variables $(\zeta_i, L_i, \Theta_i)$ where $\zeta_i$ labels the type of the model (e.g. face, letter, texture or shading), $\Theta_i$ denotes the model parameters (e.g. the parameters that determine the shape of the letter), and $R(L_i)$ denotes the region in the image that the model generates. These regions are non-overlapping and the discontinuities across the boundaries are not explicity modeled. Formally, we can summarize the non-terminal nodes by $W = \{(\zeta_i, L_i, \Theta_i) : i = 1, ..., N\}$ where the number $N$ of nodes is a random variable. These non-terminal nodes are obtained by sampling from a distribution $P(W) = p(N) \prod_{i=1}^{N} p(L_i)p(\zeta_i|L_i)p(\Theta_i|\zeta_i)$ (this is conceptually similar to applying production rules to the root node). In turn, the observed intensity values on the image lattice (the terminal nodes of the graph) are obtained by sampling from generative models $p(I_{R(L)}|\zeta, L, \Theta)$ for the specific regions which depend on their model type and their parameters (similar to applying production rules to the non-terminal nodes in a PCFG). This includes models for generating the appearance of faces and letters, see samples in Figure (3) (right panel). Overall, this second level gives a model $p(I|W) = \prod_{i=1}^{N} p(I_{R(L_i)}|\zeta_i, L_i, \Theta_i)$ [44]. This combines with the first stage to have a full generative model $p(I|W)p(W)$ for the image.

There are many ways to extend this model by augmenting the number of pattern types, by including Gestalt laws and other principles of spatial organization [43], and by having hierarchical models [1] [11]. In particular, the pattern types can be expanded to include material properties which are not explicit objects.

The advantages of a generative model for the entire image include the ability to "explain away". Submodels corresponding to different objects, or processes, compete and cooperate to explain dif-

ferent parts of the image (e.g. the letter $B$ plus bar competes with the interpretation of accidentally aligned fragments in Figure 1B). A face model might hallucinate a face in the trunk of a tree; but a tree model can overule this and provide the correct interpretation of the tree trunk, see Figure (5). In addition, full generative models enforce consistency of the interpretation of the image.

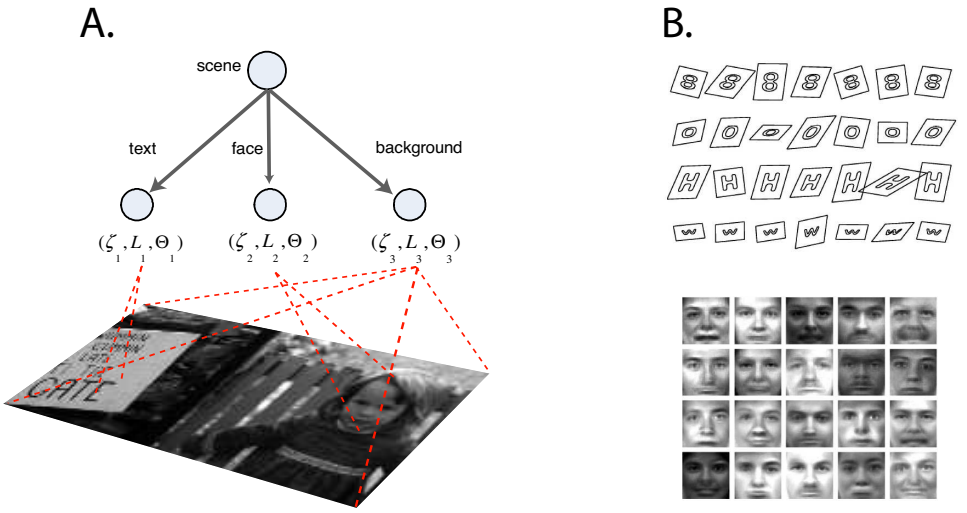


Figure 3: A. The image is generated (left panel) by a probabilistic context free grammar shown by a two layer graph with nodes with properties $(\zeta, l, \theta)$ corresponding to regions $L_i$ in the image. B. The right panel shows samples from the face model and the letter model – i.e. from $p(I_{R(L)}|\zeta, L, \Theta)$.

We now switch to the task of performing inference on this generative model to estimate $W^* = \arg\max_W P(W|I)$. This requires a sophisticated inference algorithm that can perform operations such as creating nodes, deleting nodes, diffusing the boundaries, and altering the node attributes.

The strategy used in [44] is to perform analysis by synthesis by a data-driven Markov Chain Monte Carlo (DDMCMC) algorithm. This algorithm is guaranteed to converge by standard properties of MCMC. Informally, low-level cues are used to make hypotheses about the scene which can be verified or rejected by sampling from the models. For example, low-level cues [48, 46] can be used to hypothesize that there is a face in a region of the image. This hypothesis can be validated or rejected by sampling from a generative face model. The bottom-up cues propose that there are faces in the tree bark, but this proposal is rejected by the top-down generative model, see Figure (5). Inference is performed by applying a set of operators which change the structure of the parse graph, see Figure (4). These operators are implemented by transition kernels $K$, see Box 1 for a more technical description of the algorithm. The bottom-up cues are based on *discriminative models* which are described in Box 2.

## Implications for Cognitive Science

We claim that the above model for image parsing shares key elements with human visual processing. This claim raises a number of important questions.

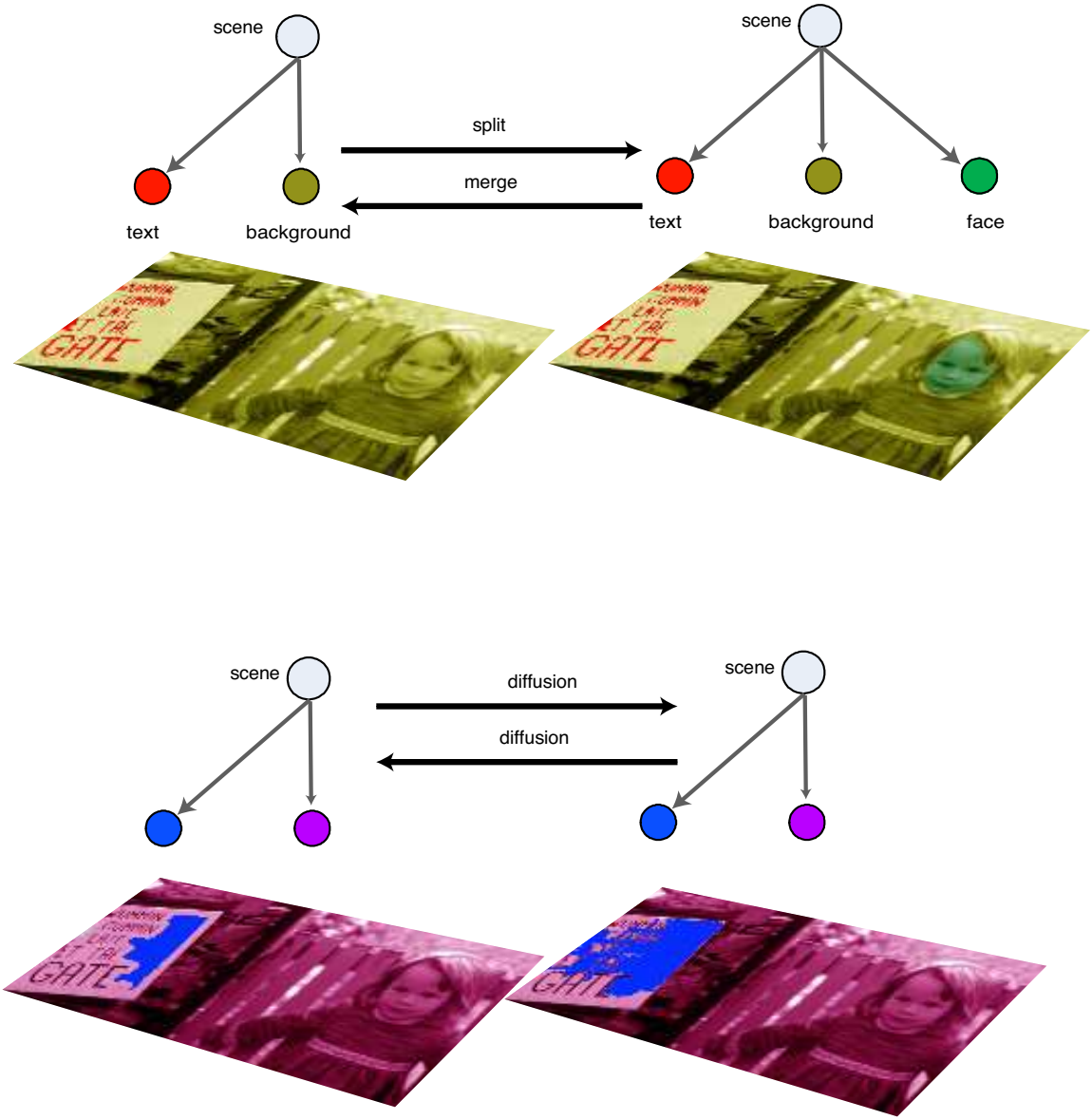


Figure 4: Examples of Markov chain dynamics that change the graph structure or the node attributes of the graph giving rise to different ways to parse the image. Different dynamics, for example creating or deleting nodes, are performed by different kernels $K$.

### "Isn't feedback inconsistent with fast processing in human object recognition?"

We argue that the bottom-up proposals are consistent with fast feedforward processing. If these proposals are strong, then the high-level percept can occur before top-down validation has begun. There is evidence that reliable diagnostic information for certain categories is available from very simple image measurements [37, 46], and that humans make certain categorical decisions sufficiently fast to preclude a verification loop [47](but see [13] and [22]).

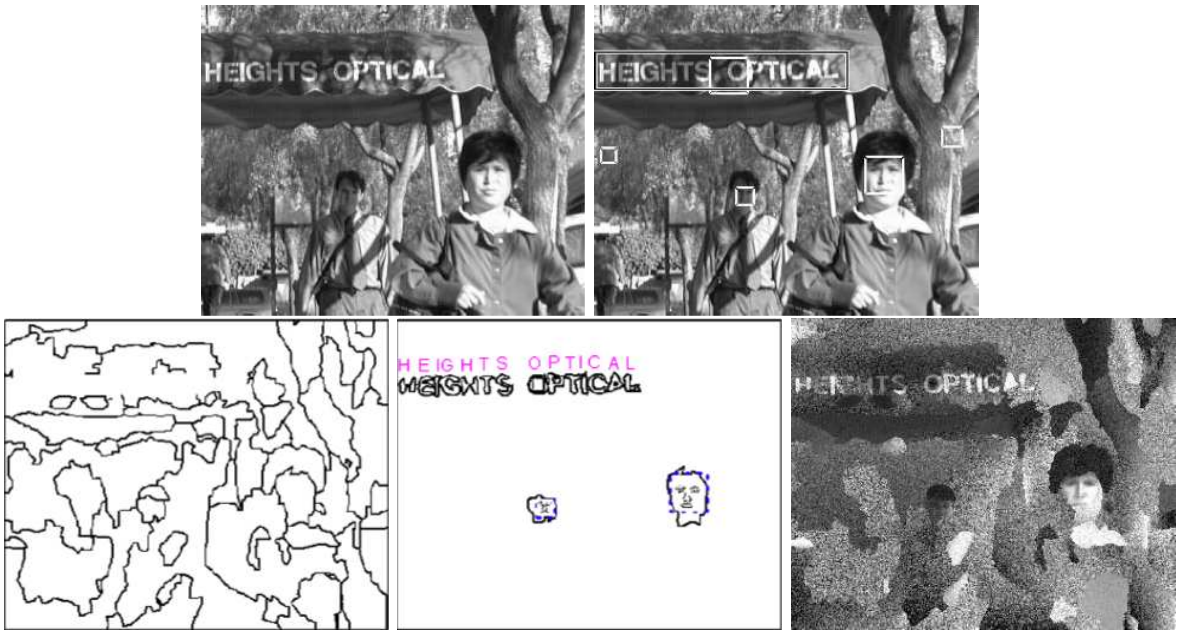


Figure 5: Top left: Input image. Top right: Bottom-up proposals for text and faces are shown by boxes. A face is "hallucinated" in a tree. Bottom centre: Overall segmentation (bottom left), Detection of letters and faces. Bottom right: Synthesised image

### "Where do the generative models come from?"

Ideally the generative models, the discriminative models, and the stochastic grammar would all be learnt from natural images. This is not difficult in principle because, as discussed in Griffiths and Yuille, learning the model from data is simply another example of statistical inference. The Helmholtz machine [7] gives an illustration of how a generative model, and an inference algorithm, can be learnt. This approach, however, has been applied only to simple visual stimuli. Similarly Friston [15] suggests learning models using the Expectation-Maximization algorithm. Although this is a useful metaphor, the challenge is to see whether this idea can be translated to algorithms that can deal with the complexities of natural images.

Learning generative and discriminative models is an extremely difficult problem in practice due to the large dimensionality of natural images. There has recently, however, been dramatic progress on the similar, but arguably simpler, problem of learning a stochastic grammar for natural languages (see article by Chater and Manning). At present, different components of the image parsing model are learnt individually. For example, the discriminative models for text and faces are trained using labelled examples of "face", "text", and "non-face", "non-text". Similarly the generative models for faces and text are learnt from labelled examples of faces and text. This, however, is far easier than unsupervised learning of the full stochastic grammar.

### "Where is attention in the image parsing model?"

The current DDMCMC algorithm contains no explicit attention mechanism. Instead bottom-up cues act everywhere in parallel to activate the generative models. But there are some mechanisms in DDMCMC which are similar to attention – for example, proposals which have high confidence are processed quickly. Moreover, it is possible to add an attentional mechanism whereby processing is directed to maximize criteria such as the expected gain in information. In other words, uncertainty in the residual signal helps prioritize where DDMCMC should process the image. This could be part of the mechanisms for driving extrinsic attention. It is unclear how this relates to standard theories of attention and the pre/post attentive distinction [38]. But it is also unclear how these types of theories, developed for artificial idealized stimuli, relate to the performance of human observers on realistic natural images. For example, Li *et al* [29] show that human subjects can pre-attentively detect complex objects, such as animals, in natural images.

### "How does this relate to neural mechanisms?"

There have long been speculations about the relationship between analysis by synthesis and forward and backward pathways in the cortex (e.g. see [34]). For a recent review of the ascending and descending pathways between visual areas, see [14].

Although it is impractical at present to test the details of analysis by synthesis models there are some relevent findings using functional magnetic resonance imaging (fMRI) and high-density electrical mapping of evoked response potentials (ERP). Earlier fMRI work by a number of groups has shown that the human lateral occipital complex (LOC) increases activity during the perception of object completion. Murray et al (2002) used fMRI to show that when local visual information is perceptually organized into whole objects, activity in human primary visual cortex (V1) decreases over the same approximate period that activity in higher, lateral occipital areas (LOC) increases. The authors interpret the activity changes in terms of high-level hypotheses that compete to explain away the incoming retinal data. This interpretation is also consistent with findings of another group, Murray et al (2004), who combined ERP and fMRI. Based on timing measurements, they concluded that dorsal regions provide input to LOC, where signatures of illusory contour completion first appear, then followed by activity in V1/V2 . Activity in fusiform gyrus has also been associated with object recognition. Bar et al. [2] used magnetoencephalography (MEG) and fMRI in an object recognition task to show that low spatial frequency specific activity appeared in left orbitofrontal cortex 50 msec earlier than than in fusiform gyrus of temporal cortex. They interpreted their results as consistent with a fast, spatially coarse analysis that selects probable object interpretations which are subsequently integrated with bottom-up information.

## Conclusion

We have argued that the major goal of vision science should be to determine how biological systems work under natural conditions and when performing natural tasks. This requires understanding and modeling the complexity of images. We argue that studies of perceptual abilities on simple synthetic stimuli can be misleading and unrepresentative of how perceptual systems function under realistic conditions.

By using recent examples from the computer vision literature, we showed that probability distributions defined on structured representations offer the promise to model natural images. Current research is extending this model including more sophisticated representations, better bottom-up cues, and extensions to greater ranges of objects and scenes. One advantage of this type of theory is that is readily "extensible" in the sense that by designing increasingly sophisticated generative models of this form we can in principle develop artificial visual systems of arbitrary effectiveness. These generative models can be used as Bayesian Ideal Observers.

The inference algorithm for this generative model is intriguing because it naturally follows the "analysis by synthesis" strategy that may correspond to the forward and backward pathways in the cortex. This algorithm combines bottom-up and top-down processing by using low-level cues to activate high-level models which are compared to the image in a top-down process. It is hoped that as techniques like fMRI, EEG, MEG and multi-unit recording continue to develop it will be possible to make direct experimental predictions from these models.

**Box 1: Data Driven Markov Chain Monte Carlo**

The DDMCMC algorithm requires designing transition kernels $K_i(W, W')$ for the graph operations illustrated in Figure (4). These kernels give a probability to transition from state $W$ to state $W'$ and obey the normalization condition $\sum_{W'} K_i(W, W') = 1, \ \forall W, i$. They are also designed to obey the detail balance condition $P(W|I)K_i(W, W') = P(W')K_i(W', W)$, which ensures that repeatedly sampling from these kernels will give samples from the posterior distribution $P(W|I)$ (plus some technical conditions). The full system combines all these kernels into a single kernel $K(W, W') = \sum_i \alpha_i K_i(W, W')$. The $\alpha_i$ ($\sum_i \alpha_i = 1$) are probabilities, so at each time-step one kernel (i.e. type of transition) is selected with probability $\alpha_i$. The kernels are designed to be of Metropolis-Hastings form $K_i(W, W') = q_i(W, W')a_i(W, W')$, where a transition from $W$ to $W'$ is proposed by $q_i(W, W')$ and accepted, or rejected, by $a_i(W, W')$. The proposals $q_i(W, W')$ are designed to be bottom-up proposals which are designed using discriminative models $Q(W|I) \propto \prod_i Q(w_i|\phi_i(I))$ which give easily computable cues to determine the components $w_i$ of the representation $W$ in terms of features $\phi_i(I)$ computed from the image (see Box 2). The acceptance probabilities $a_i(W, W')$ are based on the high-level models, for details see [44].

**Box 2: Generative and Discriminative Models**

Originally discriminative methods were defined by decision rules $\alpha(I)$ which can be described in terms of Bayesian Decision Theory, see box in Griffiths and Yuille's article. These decision rules output discrete values (e.g. "face" or "non-face") and there was no attempt to model the probability distribution $P(I, W)$. Discriminative methods of this type include classic techniques like the perceptron and more recent methods such as AdaBoost and Support Vector Machines [9]. More recently, discriminative methods have been generalized to include any method that approximates the posterior distribution $P(W|I)$. Intuitively, these methods make decisions but, by including probabilities, they give a measure of confidence in the decision. This is the sense in which we used discriminative methods in this article. Discriminative methods can be applied to learn approximate distributions $Q(w_1|\phi(I))$ for components $w_1$ of the full interpretation $W$, where $\phi(I)$ is a set of features extracted from the image. The key idea, to ensure speed of discriminative proposals, is that the feature $\phi(I)$ can be rapidly extracted and the approximate distribution $Q(w_1|\phi(I))$ is rapid to compute. For example, AdaBoost learning can be used to learn discriminative probabilities for the presence, or absence, of a face at a specific scale, orientation, and location in the image.

# ACKNOWLEDGMENTS

We would like to acknowledge helpfull feedback from the reviewers and the TICS editors. This work was supported by ONR N00014-05-1-0124 and NIH RO1 EY015261.

## References


[1] Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci, 8(10), 457-464.

[2] Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., Hamalainen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R. & Halgren, E. (2006). Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A, 103(2), 449-454.

[3] Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L. & Rust, N. C. (2005). Do we know what the early visual system does? J Neurosci, 25(46), 10577-10597.

[4] Chen X & Yuille AL. "Detecting and Reading Text in Natural Scenes". CVPR (2) (2004): 366-373

[5] Clark JJ, & Yuille AL. 1990. *Data Fusion for Sensory Information Processing*. Boston: Kluwer Academic Publishers

[6] Coughlan JM, Yuille AL, English C, and Snow D. "Efficient deformable template detection and localization without user initialization". *Computer Vision and Image Understanding* Volume 78 , Issue 3. pp 303-319. 2000.

[7] Dayan P, Hinton GE, Neal RM, & Zemel RS. (1995). The Helmholtz machine. *Neural Comput* 7: 889-904

[8] Dayan, P. and Yu, A. "Norepinephrine and Neural Interrupts" NIPS Advances in Neural Information Processing Systems 18. 2005.

[9] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). Pattern Classification. John Wiley and Sons, Inc.

[10] Ernst MO, & Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415: 429-33

[11] Feng Han, & Zhu, S.C. "Bottom-up/Top-Down Image Parsing by Attribute Graph Grammar". *International Conference on Computer Vision*. Beijing, China. (2005). 1778-1785.

[12] Flanagan, J. R., Vetter, P., Johansson, R. S., & Wolpert, D. M. (2003). Prediction precedes control in motor learning. Curr Biol, 13(2), 146-150.

[13] Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. Exp Brain Res, 142(1), 139-150.

[14] Friston, K. (2003). Learning and inference in the brain. Neural Netw, 16(9), 1325-1352.

[15] Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci, 360(1456), 815-836.

[16] Geisler WS, & Kersten D. 2002. Illusions, perception and Bayes. *Nat Neurosci* 5: 508-10

[17] Green DM, & Swets JA. 1974. *Signal Detection Theory and Psychophysics*. Huntington, New York: Robert E. Krieger Publishing Company

[18] Grenander U. 1996. *Elements of pattern theory*. Baltimore: Johns Hopkins University Press. xiii, 222 pp.

[19] Helmholtz H. 1867. *Handbuch der physiologischen Optik*. Leipzig: L. Voss. (tranlasted in English by JPC Southall as *Treatise on Physiological Optics*)

[20] Howe, C. Q., Lotto, R.B., & Purves, D. (2006). Comparison of Bayesian and empirical ranking approaches to visual perception. J Theor Biol.

[21] Dalmatian photo by R. C. James, in R. L. Gregory, The Intelligent Eye (McGraw-Hill, New York, 1973), p. 14.

[22] Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. J Vis, 3(7), 499-512.

[23] Kawato, M. (1999). Internal models for motor control and trajectory planning. Curr Opin Neurobiol, 9(6), 718-727.

[24] Kersten D, & Schrater PW. 2002. Pattern Inference Theory: A Probabilistic Approach to Vision. In *Perception and the Physical World*, ed. R Mausfeld, & D Heyer. Chichester: John Wiley & Sons, Ltd.

[25] Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian Inference. Annual Review of Psychology, 55, 271-304.

[26] Konishi SM, Yuille AL, Coughlan JM, & Zhu SC. "Statistical Edge Detection: Learning and Evaluating Edge Cues". IEEE Trans. Pattern Anal. Mach. Intell. 25(1): 57-74 (2003)

[27] Klein, I., Paradis, A. L., Poline, J. B., Kosslyn, S. M., & Le Bihan, D. (2000). Transient activity in the human calcarine cortex during visual-mental imagery: an event-related fMRI study. J Cogn Neurosci, 12 Suppl 2, 15-23.

[28] Lee TS, & Mumford D. 2003. Hierarchical Bayesian Inference in the Visual Cortex. *Journal of the Optical Society of America A*, 1434-1448. Issue 7.

[29] Li FF, VanRullen R, Koch C, & Perona P. "Rapid natural scene categorization in the near absence of attention". *Proc. Natl. Acad. Sci.* 99, 8378 - 8383, 2002.

[30] MacKay, D. M. (1956). Towards an information-flow model of human behavior. British Journal of Psychology, 47, 30-43.

[31] Mamassian P, Landy MS, & Maloney LT. (2002). Bayesian modelling of visual perception. In *Probabilistic Models of the Brain*. R. Rao, B. Olshausen and M. Lewicki (Eds), pp. 13-36. Cambridge, MA: MIT Press

[32] Marr D. **Vision**. W.H. Freeman. 1982.

[33] McDermott, J. (2004). Psychophysics with junctions in real images. Perception, 33(9), 1101-1127.

[34] Mumford D. 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern* 66: 241-51

[35] Murray SO, Kersten D, Olshausen BA, Schrater P, & Woods DL. 2002. Shape perception reduces activity in human primary visual cortex. *Proc Natl Acad Sci U S A* 99: 15164-9

[36] Murray, M. M., Foxe, D. M., Javitt, D. C., & Foxe, J. J. (2004). Setting boundaries: brain dynamics of modal and amodal illusory shape completion in humans. J Neurosci, 24(31), 6898-6903.

[37] Oliva A & Torralba A. "Modeling the shape of the scene: A holistic representation of the spatial envelope". *International Journal of Computer Vision,* 42(3):145–175, 2001.

[38] Palmer SE. **Vision Science: Photons to Phenomenology** pBradford Books, MIT Press, Cambridge, MA. 1999.

[39] Rao RP, & Ballard DH. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2: 79-87

[40] Schrater PR, & Kersten D. 2000. How optimal depth cue integration depends on the task. *International Journal of Computer Vision* 40: 73-91

[41] Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. Nat Neurosci, 5(11), 1226-1235.

[42] Tu Z, & Zhu S-C. 2002. Image Segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24

[43] Tu Z, & Zhu S-C. 2002. *Parsing Images into Region and Curve Processes*. Proc. of the 7th European Conference on Computer Vision, pp 393-407.

[44] Tu Z, Chen X, Yuille AL, & Zhu S-C. Image Parsing: Unifying Segmentation, Detection and Recognition. *International Journal of Computer Vision*. (63) 2, pp 113-140. 2005.

[45] Ullman S. 1996. *High-level Vision: Object Recognition and Visual Cognition*. Cambridge, Massachusetts: MIT Press

[46] Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. Nat Neurosci, 5(7), 682-687.

[47] VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. J Cogn Neurosci, 13(4), 454-461.

[48] Viola P, & Jones MJ. 2001. Robust real-time object detection. *Proc. of IEEE Workshop on Statistical and Computational Theories of Vision*

[49] Yuille AL, & Bülthoff HH. 1996. Bayesian decision theory and psychophysics. In *Perception as Bayesian Inference*, ed. DC Knill & W Richards, pp. 123-161. Cambridge, UK: Cambridge University Press