Special Issue: Probabilistic models of cognition

Vision as Bayesian inference: analysis by synthesis?

Alan Yuille¹ and Daniel Kersten²

¹Departments of Statistics, UCLA, San Francisco, CA 94115, USA
²Department of Psychology, University of Minnesota, Minneapolis, MN 55455-0344, USA

We argue that the study of human vision should be aimed at determining how humans perform natural tasks with natural images. Attempts to understand the phenomenology of vision from artificial stimuli, although worthwhile as a starting point, can lead to faulty generalizations about visual systems, because of the enormous complexity of natural images. 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 that use ‘analysis by synthesis’ strategies with intriguing similarities to the brain. We examine these strategies using recent examples from computer vision, and outline some important implications for cognitive science.
systems is how to use low-level cues to rapidly access the algorithms for detecting high-level objects such as faces and highlighted by the recent successes of computer vision these patterns. This lack of ambiguity for high-level vision is easily be disambiguated by the alternative explanations for the bark of a tree may occasionally look like a face, but can often easy to resolve when they do. For example, patterns in complex and rarely occur by chance. Moreover, they are patterns of objects, such as faces or other objects, are the limitations of regional cues for segmentation[2].

We now give arguments why visual inference requires the need for bottom-up and top-down processing. This contrasts with standard textbook theories of vision which favour bottom-up and top-down components. This suggests a visual system where low-level cues make up processing based on computing low-level representations such as edge maps (cf.[27]).

Figure 1. (a) A simple vocabulary for generating the image. There is little or no ambiguity in interpreting images. At worst, the letter X might be confused with a slanted l partially occulting a vertical l. (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 for details). (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.

The need for bottom-up and top-down processing

We first give an introductory example implemented. We first give an introductory example that one letter can completely, or partially, occlude the image. We can also give an ordering on the letters so allow the letter to be placed randomly at any position in a simple probability model for generative images built out of templates for each letter and a simple probability model for generative images built out of templates for each letter and

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.
proposals are sufficiently strong (i.e. the low-level cues are
which are a measure of their strength. If bottom-up
process. These bottom-up proposals come with probabilities,
with the image (or a filtered version of it) in a top-down
hypotheses are accepted, or rejected, by direct comparison
hypotheses about objects and scene structures. These
Gestalt laws), make bottom-up proposals which activate
level cues, combined with spatial grouping rules (similar to
configuration of objects is most likely to have generated
solve the inverse inference problem of estimating which
image was generated? Our explanations, so far, have
assumed that we can check all possible ways to generate
image and decide on the most probable. But this
perform inference to determine the most likely way the
image can be interpreted as an
The high-level objects access the image top-down to validate or reject the bottom-up proposals. In this example, the low-level
features propose that the image can be interpreted as an
(a) (b)
We now extend the vocabulary of the generative model
synthesis algorithms is whether they can be extended to
other more realistic stimuli.
We stress that this is an illustration of this model
examples, we can treat text as a type of texture and design
sophisticated cues can be used for the proposals. For
image parsing
Natural image parsing
root node represents the entire image
structure. The first level corresponds to the non-terminal
nodes by
variables (e.g. the parameters that determine the shape
of nodes is a random variable. These non-
labels the type of the model
has attribute
labels the model
\(z_i\) where the
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. 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.
terminal nodes are obtained by sampling from a distribution $P(W)$.

In turn, the observed intensity values on the image lattice (the terminal nodes of the graph) are obtained by sampling from generative models $p(I_{RI}(L|z); z, L, Q)$ 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 3b. Overall, this second level gives a model $p(I|W)$.

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 [37], and by having hierarchical models [38,39]. 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 different 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 override this and provide the correct.

Figure 3. (a) The image is generated by a probabilistic context-free grammar, shown by a two-layer graph with nodes that have properties $(z, L, q)$ corresponding to regions $(z, L, q)$ in the image. (b) Samples from the face model and the letter model, that is, from $p(I_{RI}(L|z); z, L, Q)$.

Figure 4. (a) Left: input image. Right: bottom-up proposals for text and faces are shown by boxes. A face is "hallucinated" in a tree. (b) Left: overall segmentation. Centre: detection of letters and faces. Right: synthesised image.
interpretation of the tree trunk (see Figure 4). In addition, full generative models enforce consistency of the interpretation of the image.

We now switch to the task of performing inference on this generative model to estimate $W_0$. 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 [3] 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 [31,32] 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 (Figure 4). Inference is performed by applying a set of operators which change the structure of the parse graph (see Figure 5). 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' (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.

**Box 1. Data driven Markov Chain Monte Carlo algorithm**

The DDMCMC algorithm requires designing transition kernels $K_i(W,W')$ for the graph operations illustrated in Figure 5 (main text). These kernels give a probability to transition from state $W$ to state $W'$ and obey the normalization condition $\sum_W K_i(W,W') = 1$. They are also designed to obey the detailed balance condition $P(W'|W)K_i(W,W') = P(W|W')K_i(W',W)$, 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 $K(W,W') = \sum_i a_i K_i(W,W')$. The $a_i (\sum_i a_i = 1)$ are probabilities, so at each time-step one kernel (i.e. type of transition) is selected with probability $a_i$. 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 $Q(W|f) \propto \sum_i Q(W|\phi_i(f))$ which give easily computable cues to determine the components $w_i$ of the representation $W$ in terms of features $f$ computed from the image (see Box 2). The acceptance probabilities $a_i(W,W')$ are based on the high-level models (for details, see Ref. [3]).
Box 2. Generative and discriminative models

Originally discriminative methods were defined by decision rules $a(l)$ 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 $P(l|w)$. Discriminative methods of this type include classic techniques like the perceptron and more recent methods such as AdaBoost and Support Vector Machines [49]. More recently, discriminative methods have been generalized to include any method that approximates the posterior distribution $P(l|w)$. 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 $Q(w|l)$ for components $w_i$ of the full interpretation $W$, where $\phi(l)$ is a set of features extracted from the image. The key idea, to ensure speed of discriminative proposals, is that the feature $\phi(l)$ can be rapidly extracted and the approximate distribution $Q(W|\phi(l))$ 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.

Isn’t feedback inconsistent with fast processing in human object recognition?

We argue that the bottom-up proposals are consistent with a fast, spatially coarse analysis results as consistent with a fast, spatially coarse analysis

Where do the generative models come from?

Ideally the generative models, the discriminative models, and the stochastic grammar would all be learnt from very simple image measurements [35,32], and that validation has begun. There is evidence that reliable diagnostic information for certain categories is available then the high-level percept can occur before top-down fusion.

How does this relate to neural mechanisms?

We used discriminative methods in this article. Discriminative methods (MRI) can be applied to learn approximate distributions $Q(w|l)$ for components $w_i$ of the full interpretation $W$, where $\phi(l)$ is a set of features extracted from the image. The key idea, to ensure speed of discriminative proposals, is that the feature $\phi(l)$ can be rapidly extracted and the approximate distribution $Q(W|\phi(l))$ 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.

Isn’t feedback inconsistent with fast processing in human object recognition?

We argue that the bottom-up proposals are consistent with a fast, spatially coarse analysis results as consistent with a fast, spatially coarse analysis

Where do the generative models come from?

Ideally the generative models, the discriminative models, and the stochastic grammar would all be learnt from very simple image measurements [35,32], and that validation has begun. There is evidence that reliable diagnostic information for certain categories is available then the high-level percept can occur before top-down fusion.

How does this relate to neural mechanisms?

We used discriminative methods in this article. Discriminative methods (MRI) can be applied to learn approximate distributions $Q(w|l)$ for components $w_i$ of the full interpretation $W$, where $\phi(l)$ is a set of features extracted from the image. The key idea, to ensure speed of discriminative proposals, is that the feature $\phi(l)$ can be rapidly extracted and the approximate distribution $Q(W|\phi(l))$ 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.

Isn’t feedback inconsistent with fast processing in human object recognition?

We argue that the bottom-up proposals are consistent with a fast, spatially coarse analysis results as consistent with a fast, spatially coarse analysis

Where do the generative models come from?

Ideally the generative models, the discriminative models, and the stochastic grammar would all be learnt from very simple image measurements [35,32], and that validation has begun. There is evidence that reliable diagnostic information for certain categories is available then the high-level percept can occur before top-down fusion.

How does this relate to neural mechanisms?

We used discriminative methods in this article. Discriminative methods (MRI) can be applied to learn approximate distributions $Q(w|l)$ for components $w_i$ of the full interpretation $W$, where $\phi(l)$ is a set of features extracted from the image. The key idea, to ensure speed of discriminative proposals, is that the feature $\phi(l)$ can be rapidly extracted and the approximate distribution $Q(W|\phi(l))$ 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.
Conclusion

We are able to predict and explain many aspects of human perception. The models described in this paper can be used as Bayesian Ideal Observers.

Acknowledgements

We are grateful to the NIH for supporting this work.

Supplementary data


References

1 Caad M. et al. (2005) J. Vision. 5, 789

Have you contributed to an Elsevier publication?
Did you know that you are entitled to a 30% discount on books?
A 30% discount is available to ALL Elsevier book and journal contributors when ordering books or stand-alone CD-ROMs directly from us.
To take advantage of your discount:
1. Choose your book(s) from www.elsevier.com or www.books.elsevier.com
2. Place your order
   Americas:
   TEL: +1 800 782 4927 for US customers
   TEL: +1 800 460 3110 for Canada, South & Central America customers
   FAX: +1 314 453 4898
   E-MAIL: author.contributor@elsevier.com
   All other countries:
   TEL: +44 1865 474 010
   FAX: +44 1865 474 011
   E-MAIL: directorders@elsevier.com
   You’ll need to provide the name of the Elsevier book or journal to which you have contributed. Shipping is FREE on pre-paid orders within the US, Canada, and the UK.
   If you are faxing your order, please enclose a copy of this page.
3. Make your payment
   This discount is only available on prepaid orders. Please note that this offer does not apply to multi-volume reference works or Elsevier Health Sciences products.
   For more information, visit www.books.elsevier.com