Vision as Bayesian Inference: A Historical Perspective

Computational Cognition, Vision, and Learning

Alan Yuille

Departments of Cognitive Science and Computer Science

Johns Hopkins University

Structure of the Talk

- Part 1. Marr's Dream. Unifying the study of Biological Vision (BV) and Computer Vision (CV).
- Part 2. Why I became a Bayesian. Brief history of Bayes in 1980's.
- Part 3. Examples.
- Part 4. Bayes and the Brain. Analysis by Synthesis.
- Part 5. Reviving Marr's dream.

Part 1: Marr's Dream of Vision

- When I started vision in 1982 there was a dream – articulated by David Marr in his book "Vision" – that Computer Vision (CV) and Biological Vision (BV) could be studied together in a complimentary manner.
- Computer Vision was a very new and disorganized field with roots in Artificial Intelligence, Image Processing, Pattern Analysis (an early version of Machine Learning), and Neural Networks.
- Biological Vision was much older. Psychophysics was established in the 19th century. Neuroscience studies of vision were more recent but had produced Nobel prize-winning work (Hubel & Wiesel).
- BV was studied at many universities, CV at only a few (e.g., MIT, Stanford, CMU). BV conferences were much bigger than CV conferences.

Marr & Poggio's Three Levels of Analysis

- Marr's dream was based on his (and Poggio's) three levels of analysis.
- 1. Computational Level
- 2. Algorithmic Level
- 3. Hardware Level.

- It was argued that the core computational task of vision was the same for BV and CV namely: how to estimate physical properties of the underlying 3D scene from images (patterns of light rays).
- So the computational level should be similar. The algorithms might be, depending on whether you believed in neural network models. The hardware of computers and brains were certainly different.

Natural Constraints and Ecological Constraints

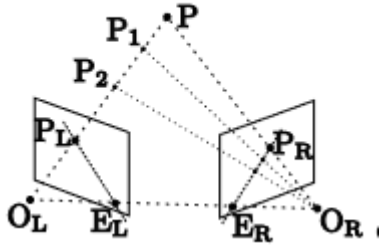
- Vision was gradually being perceived to be a very hard problem. Images were ambiguous – they requires assumptions about the real world to enable vision to be unambiguous (and well-posed).
- Marr argued that natural constraints were needed to make vision possible – surfaces are usually smooth, objects are typically rigid, etc. Gibson's "ecological constraints" captured a similar idea. *One aim of vision scientists was to identify these natural constraints.*
- Marr acknowledges that some aspects of the brain were surely due to biology/evolution and that others, like blind spot in the retina, attentional mistakes like change blindness, were properties that CV systems were unlikely to want.

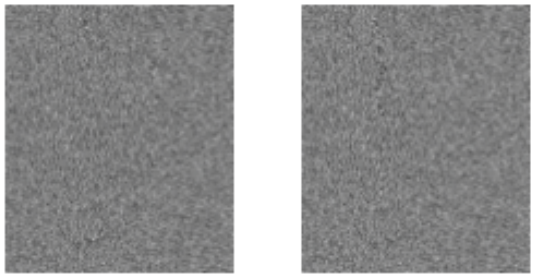
Marr's Framework for Vision

- Marr proposed a framework for vision. This consisted of constructing a series of representations.
- The primal sketch represented image properties.
- The 2 1/2D sketch represented surfaces depth/orientation – shape from X -- as seen from the viewer.
- The 3D model which represented objects in terms of 3D geometric primitives (Biederman's Geon theory).
- These representations could be roughly mapped to areas of the visual cortex.

Marr's Theory: Mind, Brains, and Machines

- The link between Marr's theory and Neuroscience (Brain) was limited, due to the challenges of performing neuroscience experiments.
- There was closer link to Psychology and Cognitive Science (Mind) because you could compare Computational models to behavioral experiments (mostly qualitative).
- Example: Marr & Poggio's model of stereo agreed with human experiments on Julesz's Random Dot Stereograms (RDS).
- Left (Stereo), Center (RDS). Right (Fox – the result of matching the RDS).





Marr's Theory: Mind, Brains, and Machines

- Other Examples:
- Ullman's computational models for estimating 2D and 3D motion.
- S. Ullman. The Interpretation of Visual Motion. 1977.
- Related work. Grossberg's dynamical system models for spatial grouping.
- All these models gave qualitative agreement with the perceptual findings. Very few quantitative results.
- The computational models were designed to work on simplified artificial stimuli. But some models could be extended to real images.

The Computational Theories

- The theories were largely formulated in terms of minimizing energy, or cost, functions.
- These energy functions could often be expressed in terms of a *data term* plus a *natural constraint term.*
- The natural constraints were fairly simple – e.g., surfaces are spatially smooth, objects tend to move rigidly.
- At the same there was similar CV work by B. Horn who used energy models for shape from shading and structure from motion. Again energy = data term + smoothness term.
- The models were beginning to work on real world images.

Why I became a Bayesian.

- I was strongly influenced by the work of S. Geman and D. Geman "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images" PAMI. 1984.
- It offered a mathematically coherent framework for addressing all vision problems. By defining a Gibbs distribution – the exponential of the negative energy – it could subsume all "energy function" models of vision. The *data terms* became *the likelihood function* and the *natural constraints* became *priors*.
- *This probabilistic formulation also suggested algorithms for performing inference and learning. Geman & Geman used Gibbs sampling and simulated annealing.*
- As an ex-Physicist, I was very attracted by this theoretical framework.

Advantages of Bayes.

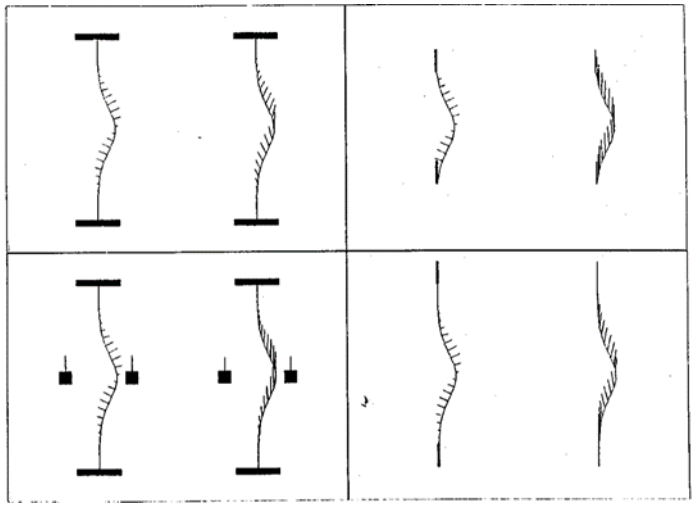
- Advantages of the Bayesian Formulation were exploited and developed in subsequent papers.
- (1) Using Gibbs distributions – almost all the energy function models could be reinterpreted as Bayesian models. Natural constraints as priors.
- (2) This probabilistic formulation naturally suggested inference algorithms. In particular we developed new algorithms by adapting *mean field theory (MFT)* from Statistical Physics. These have since been generalized and re-branded as *variational inference.* Belief propagation is closely related.
- (3) For certain classes of models, the MFT algorithms became identical to neural network models (Hopfield) and several of the dynamical systems models (Grossberg) could also be re-derived in this manner.
- (4) The probabilities of Bayes meant that you could combine visual cues (e.g., for estimating shape) in a principled way by taking into account their statistical dependencies. This led to distinguishing between “weak coupling” and “strong coupling”.

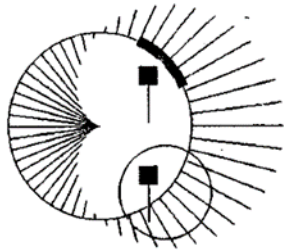
Advantages of Bayes.

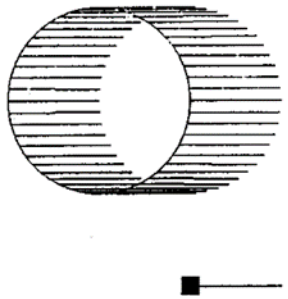
- (5) Bayes Decision Theory. This gives a direct link to Signal Detection Theory, Ideal Observer Theory, and Control Theory. (And to versions of Machine Learning with empirical risk). A unified framework.
- (6) Bayes defined probabilities over problem instances. This enabled performance bounds (e.g., Bayes risk) but even convergence rates of algorithms, in some cases. This could also be used for learning algorithms, (Smirnakis and Yuille 1993), recently rediscovered as unsupervised learning
- (7) Analysis of Bayesian models showed close relationships between models that appeared very different (by integrating out variables).
- (8) Bayes and the Brain. Analysis by Synthesis. See later section.

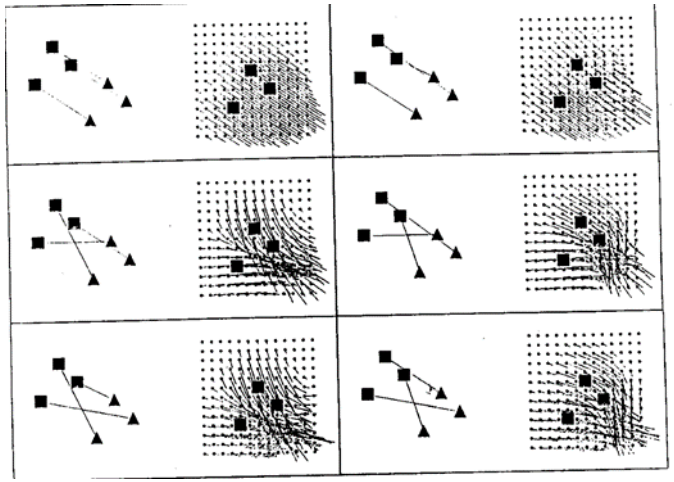
Examples of Bayes.

- The Motion Coherence Theory. A.L. Yuille & N.M. Grzywacz. 1988.
- We proposed that motion perception used a slow+smooth prior. This accounted qualitatively for a range of perceptual phenomena: motion capture, motion coherence. For short- and long-range motion.









Example.The Motion Coherence Theory.

- It also technically derived the solution as linear combinations of kernels which were the eigenfunctions of differential operators.

$$\vec{v}(\vec{r}) = \sum_i \frac{\vec{\beta}_i}{2\pi\sigma^2} \exp \frac{-(\vec{r} - \vec{r}_i)^2}{2\sigma^2}$$

where the β_i are solutions of

$$(\lambda\delta_{ij} + G_{ij})\vec{\beta}_j = \vec{U}_i$$

where

$$G_{ij} = \frac{1}{2\pi\sigma^2} \exp \frac{-(\vec{r}_j - \vec{r}_i)^2}{2\sigma^2}$$

$$\sum_{m=0}^{\infty} \frac{\sigma^{2m}}{(m!2^m)} \nabla^{2m} G(\vec{x},\sigma) = \delta(\vec{x})$$

- This helped inspire Poggio's work on Radial Basis functions for learning.
- Later studies gave some quantitative support for this model (Watamaniuk et al, H. Lu & A.L. Yuille). Very nice demonstrations of a similar model by Y. Weiss et al.
- Methods like these are used for state-of-the-art matching algorithms. E.g., Jiayi Ma et al.

Analog Neural Networks on Early Vision

- Koch, Marroquin, Yuille. Analog Neural Networks in Early Vision. PNAS. 1986.
- This work used mean field theory to give a neutrally plausible implementation for Geman & Geman's model.
- This may be consistent with properties of the visual cortex. See T-S Lee's research group at CMU.
- Note that T-S Lee has several neuroscience findings that give evidence for some of these computational models, e.g., support for Marr Poggio's theory of stereopsis.

Example: Cue Coupling

- Psychophysical studies of Shape from X by Buelthoff and Mallot.
- These involved cues such as texture, shading, and stereo.
- Their finding supported Bayesian theories and were inconsistent with linear weighted averaging.
- *Are cues independent or not?*

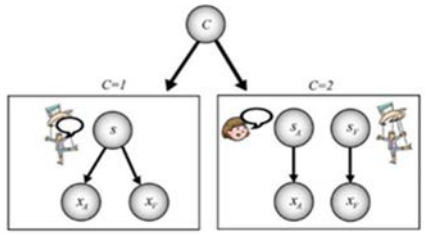


Fig. 40. The subject is asked to estimate the position of the cues and to judge whether the cues are from a common cause – i.e. at the same location – or not. In Bayesian terms the task of judging whether the cause is common can be formulated as model selection – are the auditory and visual cues more likely to generated from a single cause (left) or by two independent causes (right). Figure reprinted with permission from [?].

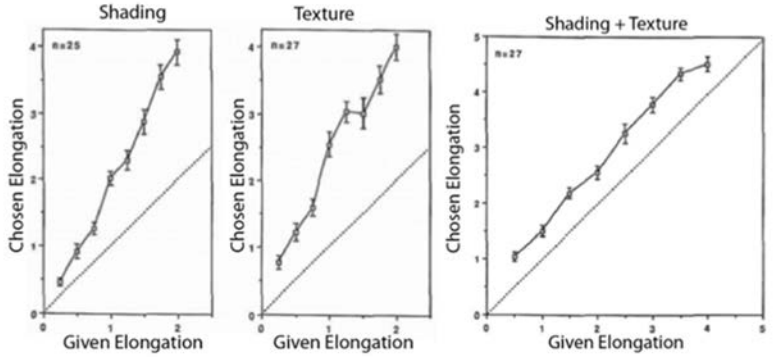


Fig. 34. Cue coupling results which are inconsistent with linear weighted average [?]. Left Panel: If depth is estimated using shading cues only then humans underestimate the perceived orientation (i.e. they see a flatter surface). Center Panel: Humans also underestimate the orientation if only texture cues are present. Right Panel: But if both shading and texture cues are available then humans perceive the orientation correctly. This is inconsistent with taking the linear weighted average of the results for each cue separately. Figure reprinted with permission from [?].

Example: Strong Coupling.

- Other studies suggested strong coupling and, in particular, model selection between competing explanations of the data. Blake and Buelthoff. Kersten et al. Buelthoff and Mallot.

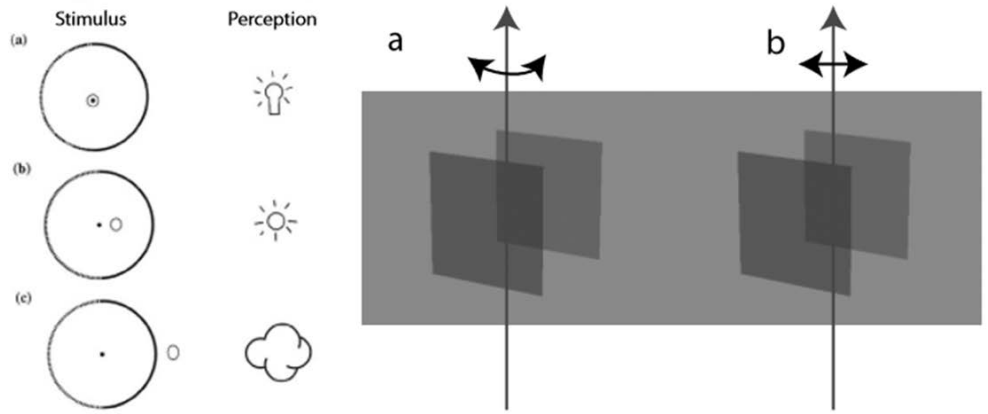


Fig. 39. Examples of strong coupling with competitive priors. A sphere is viewed binocularly (left) and small changes in the position of the specularity lead to very different percepts (Blake and Bülthoff 1990). Similarly altering the transparency of the moving surfaces (right) can make the two surfaces appear to rotate either rigidly together or independently.

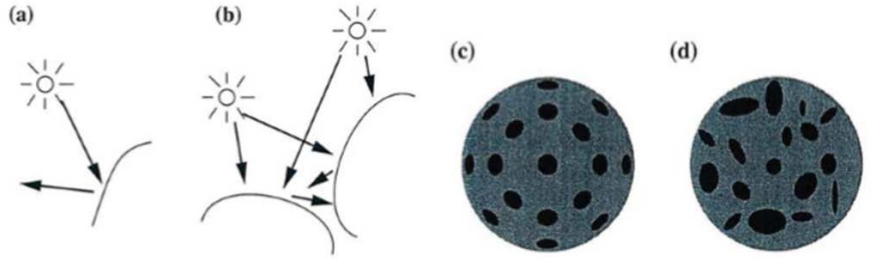
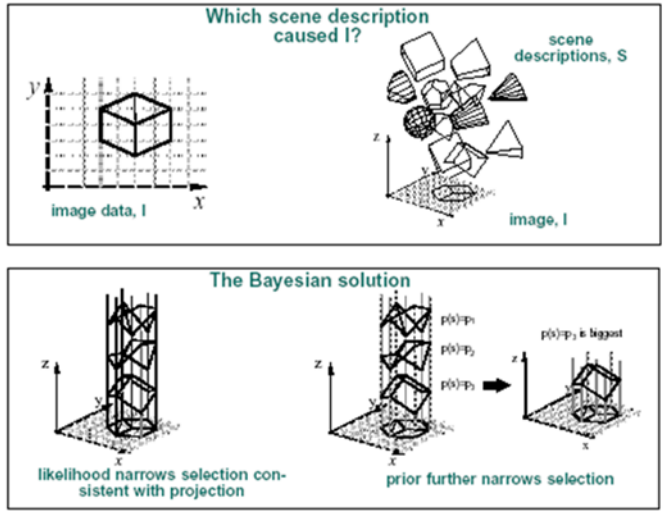


Fig. 38. Model selection may need to be applied in order to decide is a cue can be used. Shape from shading cues will work for case (a) because the shading pattern is simply due to a smooth convex surface illuminated by a single source. But for case (b) the shading pattern is complex – due to mutual reflection between the two surfaces – and so shape from shading cues will be almost impossible to use. Similarly, shape from texture is possible for case (c) because the surface contains a regular texture pattern but is much harder for case (d) because the texture is irregular.
Figure reprinted with permission from [?].

Bayes and the Brain: Analysis by Synthesis

- Helmholtz (1880's) proposed that vision could be studied as inverse inference. This requires inverting the process that generates the image.
- Inverse inference requires priors.
- There are an infinite number of ways that images can be formed.
- Why do we see a cube?
- The likelihood P(I|S) rules
- out some interpretations.
- The prior P(S) argues that cubes
- are more likely than other shapes.



Bayes and the Brain: Analysis by Synthesis

- Richard Gregory
- "Perception (vision) as hypotheses“.
- Perception is not just a passive acceptance of stimuli, but an active process involving memory and other internal processes.
- Humans have internal representations – we see images when we dream, we can imagine what animals and people will do, we can hallucinate.
- In more modern terms: “You have a physics simulator in your head”. J.B. Tenenbaum.

Vision as Inverse Inference.

Part 4: Analysis by Synthesis

- Inverse inference: optical illusions caused by incorrect inference.
- Think that the shadow is cast by the beach towel (left) or a levitating man (right).
- Ball in the box (D. Kersten).

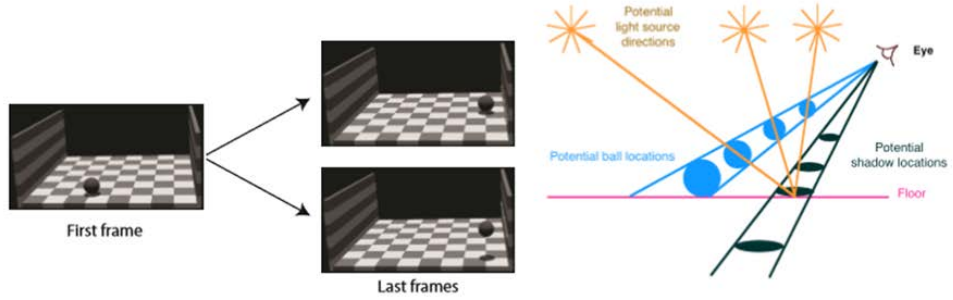
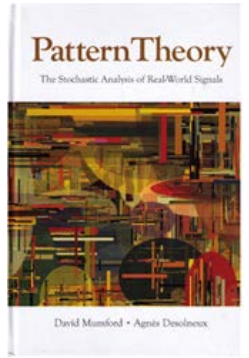


Fig. 36. In the "ball-in-a-box" experiments the motion of the shadow affects the perceived motion of the ball. The ball is perceived to rise from the ground if the shadow follows a horizontal trajectory in the image; but is perceived to move towards the back of the box if the shadow follows a diagonal trajectory. See `http://youtu.be/hdFCJepvJXU`. Left panel shows the first frame and the last frames for the two movies. Right panel. The explanation is that the observer resolves the ambiguities in the projection of a three-dimensional scene to perceive the 3D trajectory of the ball ([?].

Analysis by Synthesis: Mumford & Grenander

- Grenander (1960's) had proposed that vision could be formulated as pattern theory and proposed the idea of "analysis by synthesis". This is naturally expressed in Bayesian terms. (S. Geman was a student of Grenander).



- Mumford embraced Analysis by Synthesis and Pattern Theory.
- Analysis by Synthesis emphasizes pattern synthesis as well as pattern analysis. Bayesian inference requires you construct a prior probability model of whatever signals or situations you are modeling and you should always test your prior by sampling to see which features it models accurately and which it does not.

Mumford's Bold Hypothesis.

- Mumford (1991) boldly proposed a model for how a primate brain could perform analysis by synthesis using bottom-up and top-down processing.
- He proposed that each area of the cortex carries on its calculations with the active participation of a nucleus in the thalamus with which it is reciprocally and topographically connected. This nucleus plays the role of an 'active blackboard' on which the current best reconstruction of some aspect of the world is always displayed
- Each cortical area maintains and updates the organism's knowledge of a specific aspect of the world, ranging from low level raw data to high level abstract representations, and involving interpreting stimuli and generating actions.
- It draws on multiple sources of expertise, learned from experience, creating multiple, often conflicting, hypotheses which are integrated by the action of the thalamic neurons and then sent back to the standard input layer of the cortex.
- .

Mumford's bold hypothesis for the architecture of the neocortex

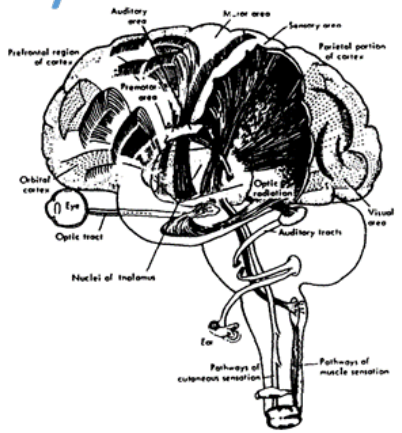


Fig. 2. The location of the thalamus within the cortex (from Luria 1969)

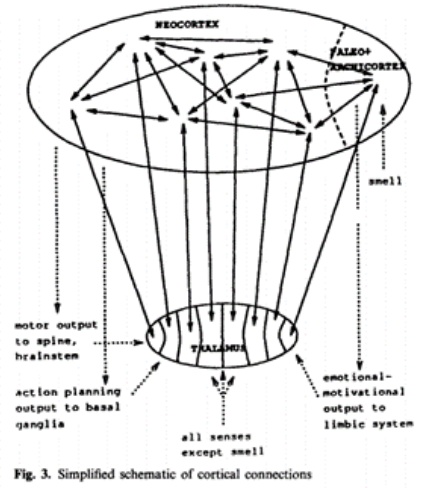


Fig. 3. Simplified schematic of cortical connections

- The higher areas of the neocortex attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view.
- The lower areas attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area.
- The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops.
- *Neuroscience experiments give increasing support for top-down models and maybe for analysis by synthesis.*

Vision as Bayesian Inference: Yuille & Kersten.

- A. Yuille & D. Kersten. Trends in Cognitive Science. 2006.

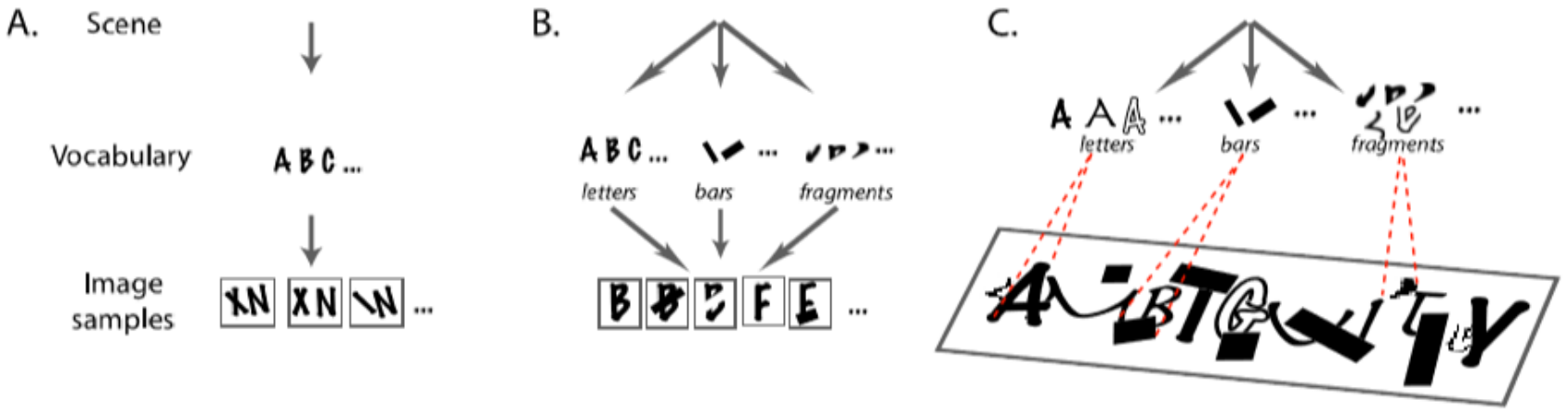


Figure 1: Left Panel (A). A simple vocabulary for generating the image. There is little, or no, ambiguity in interpreting images. At worst, the letter X may be confused with a slanted I partially occluding a vertical I. Centre Panel (B). A richer vocabulary. A given cause, such as a particular letter, 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.

Vision as Bayesian Inference: Yuille & Kersten

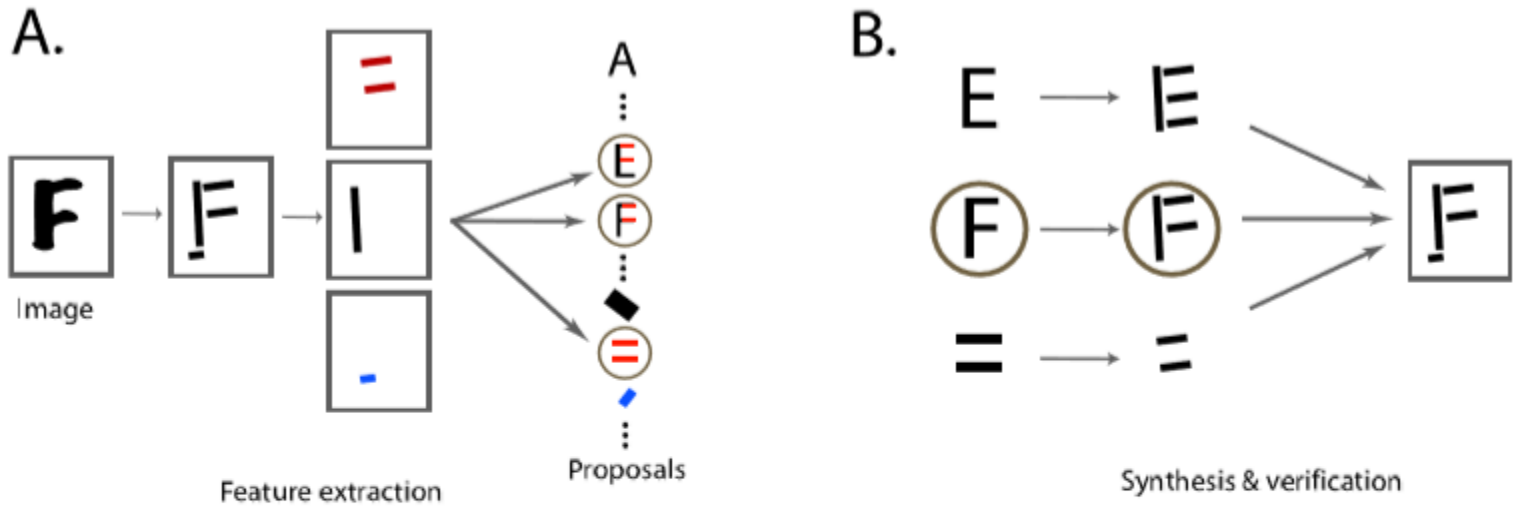


Figure 2: Analysis by synthesis. A. Low-level processing (left panel) can extract edge features, such as bars, and use conjunctions of these features to make bottom-up proposals to access the higher-level models of objects. B. The high-level objects access the image top-down to validate or reject the 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.

Vision as Bayesian Inference: Yuille & Kersten

- Adding more realism building on work by Z. Tu & S.C. Zhu 2002, Z. Tu et al. 2006.

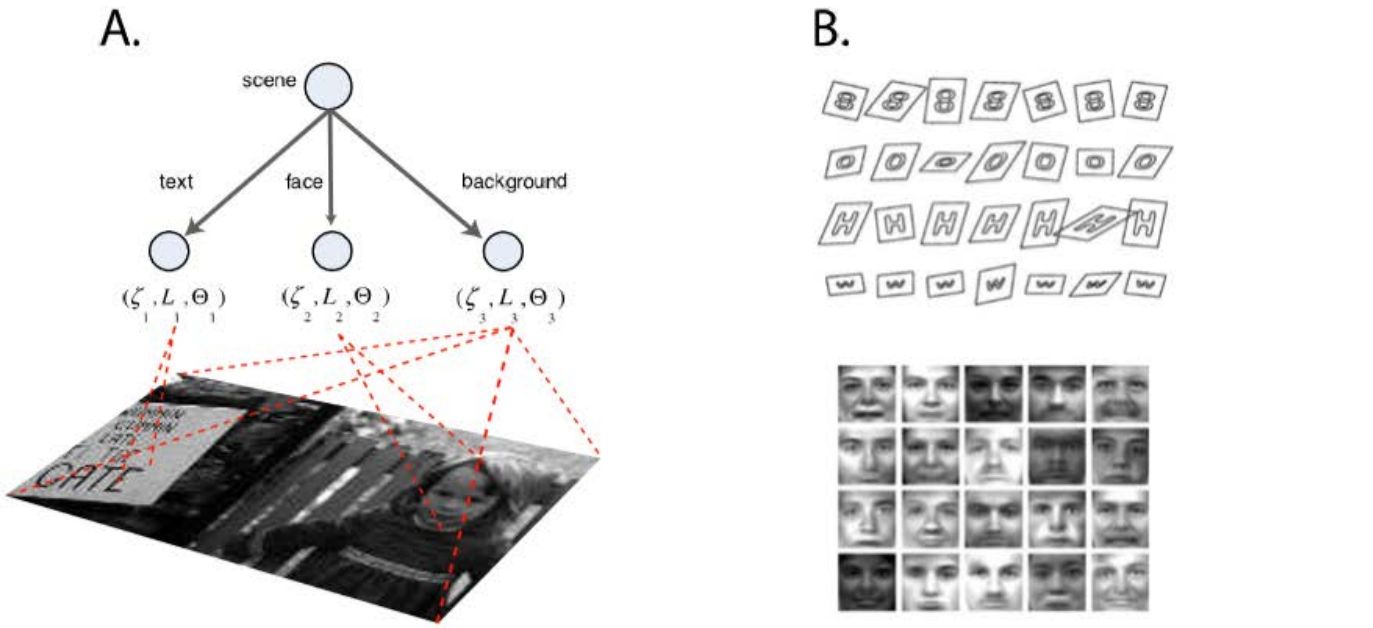


Figure 3: A. The image is generated (left panel) by a probabilistic context free grammar shown by a two layer graph with nodes with properties (ζ, l, θ) corresponding to regions L_i in the image. B. The right panel shows samples from the face model and the letter model – i.e. from $p(I_{R(L)}|\zeta,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

Reviving Marr's Dream

- Marr's dream lead to much technical progress. Many current CV models are based on these computational models.
- But recent work – e.g., Deep Nets – has stressed learning the posterior distribution directly. P(x|y), instead of P(y|x) and P(x).
- There has been reasonable interaction between BV and CV, at least at the behavioral level. Knill and Richards. 1996.
- There has been some neuroscience, for example by T.S. Lee's group at CMU. But most neuroscientists showed little interest in computational models.
- But overall Marr's dream has not been achieved Why not?

Reviving Marr's Dream

- Why not? Because Image are very complicated (see below).
- CV researchers had to work with the complexity of real images. BV researchers did not.
- And BV and CV researchers had different agendas. CV researchers want to design an entire vision system.
- BV researchers often settle for understanding components. Neuroscience is a "cottage industry". A. Movshon.

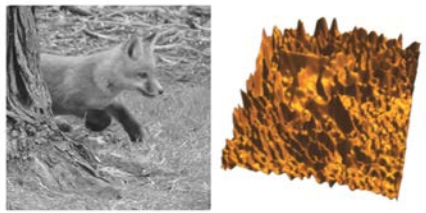


Fig. 2. Why is vision hard? The raw input to the Fox image (left panel) is the intensity values plotted as a function of spatial position (right panel). These intensity patterns vary depending on the pose of the fox, the lighting conditions, and other factors. The human visual system must decode this raw input, which is extremely difficult.

The Split

- Stimuli – in the 1980's CV and BV researchers both worked mostly with simple synthetic stimuli. Real world stimuli were too difficult.
- But CV researchers had to leave their comfort zone of synthetic stimuli because their algorithms had to work on real world stimuli. This started a long slow process where CV developed increasingly complicated mathematical and computational techniques.
- But BV researchers had no need to leave their comfort zone. Their research required controlled stimuli – almost impossible with real world images. As a side effect, BV researchers never needed to learn the mathematical and computational tools that CV researchers were developing.
- BV findings on synthetic stimuli could inspire CV algorithms. But CV researchers found that models that work on "toy stimuli" rarely worked on real stimuli. BV findings were increasingly considered to be irrelevant.

Why it is time to revive Marr's Dream.

- What has changed?
- ***Computer Graphics (CG) advances.*** It is possible to generate realistic visual stimuli which can be used as controlled stimuli (experimental design) for BV experiments. Behavioral, Electrophysiological, fMRI, These stimuli are also being gradually accepted by the CV community.
- ***Neuroscience Techniques.*** Great progress in methods for recording from the brain. Optogenetics. Mapping mice neural circuits.
- ***Mturk:*** The ability to do "big data" experiments using large numbers of experimental subjects.
- ***Machine Learning Methods***. These enable BV researchers to make predictions when they have enough "big data". They also enable the design of CV theories that serve as "data-driven ideal observers".
- CV can make predictions on realistic stimuli which are not embarrassingly bad and hence can be used as models of human and primate vision.

Reviving Marr's Dream: Conclusion.

- BV systems are much better than CV systems (expect for a few very special cases). BV systems can perform many more visual tasks, they require much less supervision, they are adaptive and flexible.
- BV can challenge CV to perform at human level.
- But directly studying BV without studying CV is problematic. CV researchers have an immense range of mathematical and computational techniques. They know what the really hard vision problems are (even if they do not know how to solve them).
- From another perspective: we best understand the brain by trying to reverse engineer it.

Some Reference Papers.

- S. Ullman. The Interpretation of Visual Motion. 1977.
- D. Marr. Vision. 1982.
- C. Koch, J. Marroquin, A.L. Yuille. Analog "Neuronal" Networks. PNAS. 1986
- A.L. Yuille and N.M. Grzywacz. A Computational Theory for the Perception of Coherent Visual Motion. Nature. 1988.
- D. Mumford. On the Computational Architecture of the Neocortex. 1991.
- D.K. Knill and W. Richards. Perception as Bayesian Inference. 1996.
- T.S. Lee. Computations in the Early Visual Cortex. 2003.
- A.L. Yuille and D.K. Kersten. Vision as Bayesian Inference. TICS. 2006.
- A.L. Yuille and D.K. Kersten. Early Vision. In From Neuron to Cognition via Computational Neuroscience. Eds. M. Arbib, J.J. Bonaiuto. 2016.