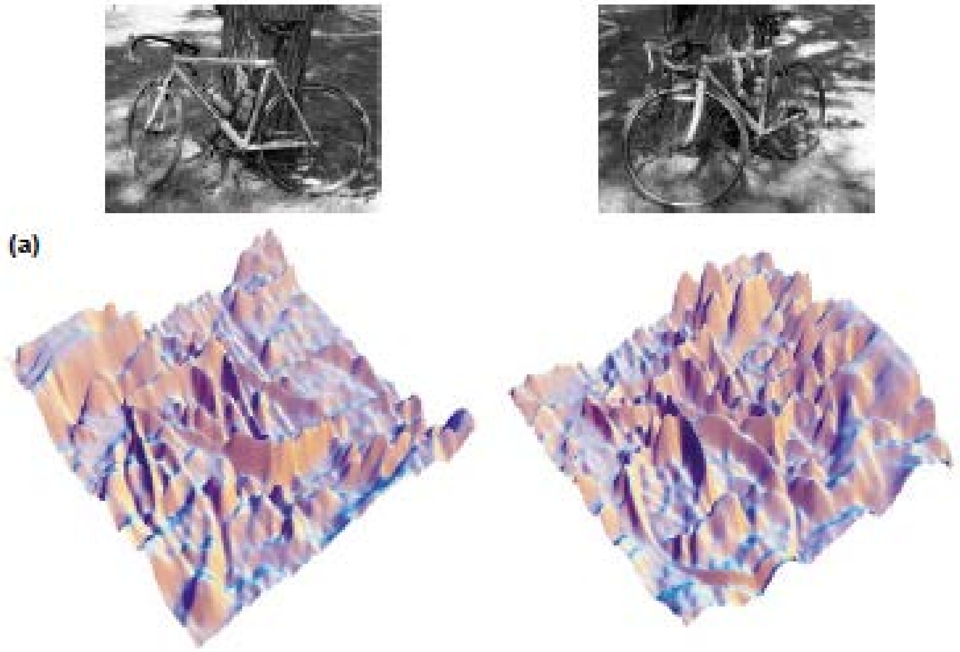
High-Level Vision: Beyond This Course

A.L. Yuille

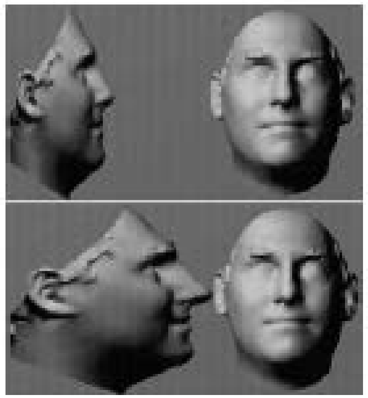
Why is Vision Hard?

- Complexity and Ambiguity of Images. Range of Vision Tasks.
- More 10x10 images -- 256^100 = 6.7 x 10 ^240 -- than the total number of images seen by all humans throughout history 3 x 10^21.
- (50 billion people, live 20 billion seconds, 30 image per second)

(a)



(b)



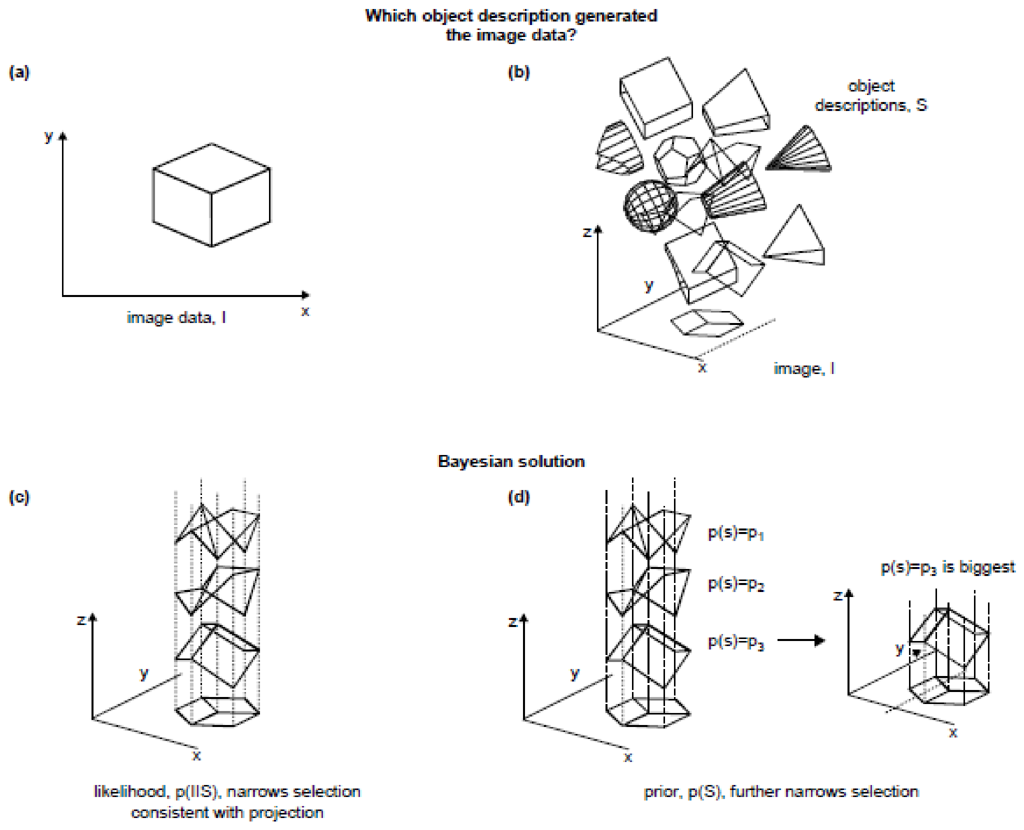
Bayes and Vision.

- History of Bayes and Vision dates to the early 1980's and before. (Ulf Grenander's pattern theory, 1960's).
- Vision as an inverse inference problem.
- Decode images by inverting image formation.
- As argued by Gibson and Marr, this requires knowledge about the world Natural constraints (Marr), Ecological constraints (Gibson).

- Bayesian formulations are natural. Constraints are priors and can be learnt from examples.

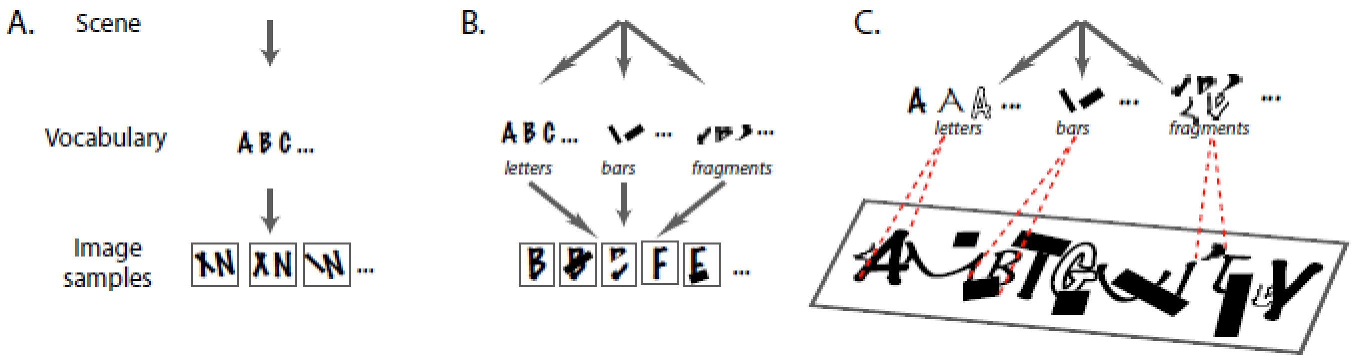
Bayes for Vision

- Courtesy of Pavan Sinha (MIT)
- The likelihood is not enough.



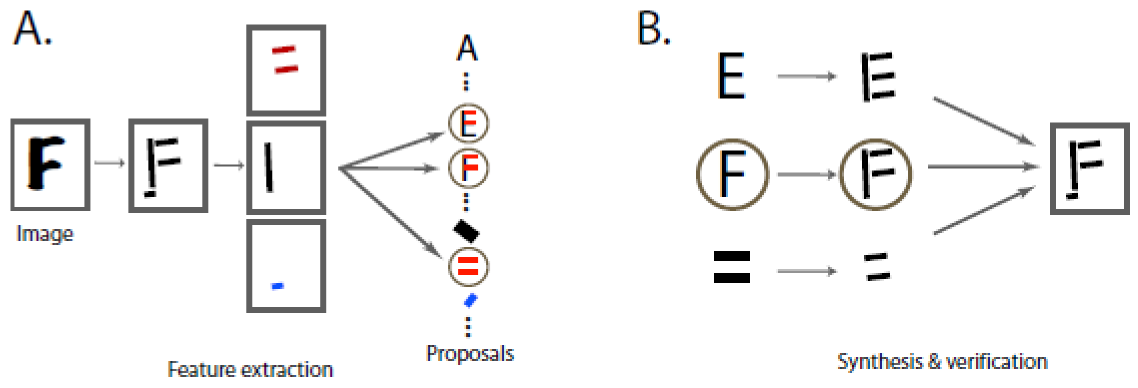
Models for Generating Images:

- Grammars (Grenander, Fu, Mjolsness, Biederman).
- Simple to Complex Grammars: Easy to hard Inference



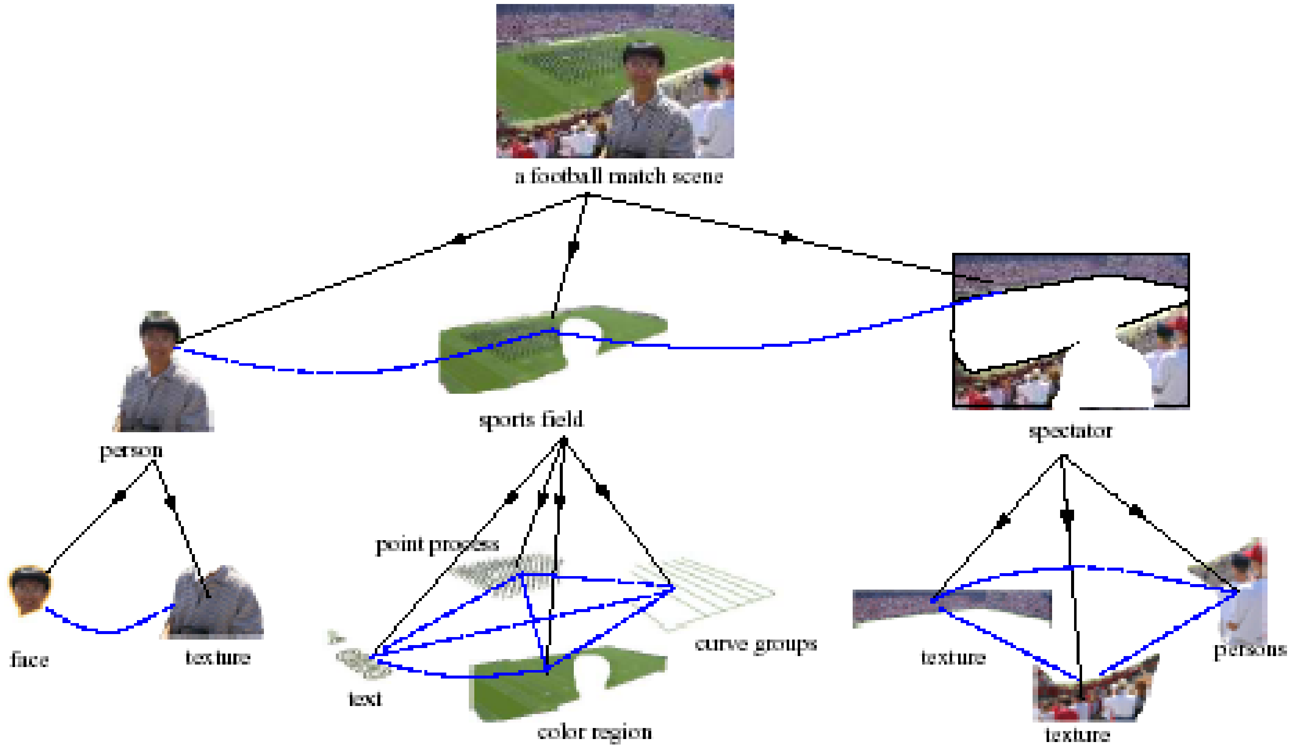
Analysis by Synthesis

- Analyze an image by inverting image formation.
- Proposals and Verification



Can we do this for Real Images?

- Image Parsing:
- Learn probabilistic models of the visual patterns that can appear in images.
- Interpret/understand an image by decomposing it into its constituent parts.



Vision Goals and Tasks

- Vision is often formalized as low, middle, and high-level.
- This seems to map onto different parts of the visual cortex (V1, V2,…, IT). (Previous Talks).
- High level vision relates very naturally to other aspects of cognition – reasoning, language.

Some Vision Goals (SC Zhu et al)

- Understanding objects, scenes, and events. Reasoning about functions and roles of objects, goals and intentions of agents, predicting the outcomes of events.

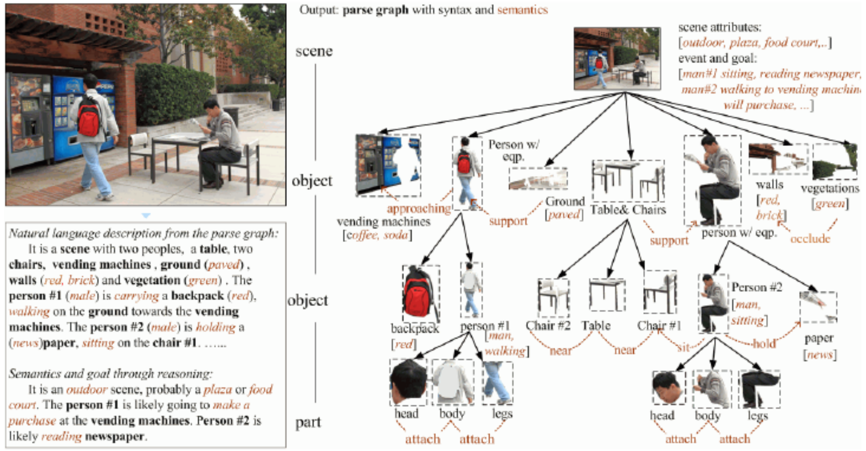


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

Converting Parse Graphs to Language

- Illustration: Perona and Fei-Fei Li.

Image shown to subjects	40ms	80ms	107ms	500ms
	"Possibly outdoor scene, maybe a farm. I could not tell for sure."	"There seem to be two people in the center of the scene."	" People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight."	"Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets..."

Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].

Reasoning about Objects in 3D Space

- Understanding the 3D scene structure enables reasoning.

Figure 9. Placing objects in a consistent geometric frame, such as children playing soccer, allows reasoning about objects in 3D space. Results from Koller's group ICCV09 [37]