Vision as Bayesian Inference. Supplement to Lecture 1

Alan Yuille
Department of Statistics, UCLA
Los Angeles, CA 90095
yuille@stat.ucla.edu

Abstract

NOTE: NOT FOR DISTRIBUTION.

1. Introduction

What is vision? Image I is generated by the state of the world W. You need to estimate W from I.

Vision is very very hard. Humans are "vision machines". The large size of our cortex distinguishes us from other animals (except close relatives like monkeys). Roughly half the cortex is involved in vision (maybe sixty percent for monkeys). It is possible that the size of the cortex developed because of the evolutionary need to see and interpret the world. So intelligence might be a parasite of vision. Most animals have poor vision – e.g., cats and cobras can only see movement – and often rely on other senses (e.g., smell). Animals may also only use vision, and other senses, for basic tasks such as detecting predators/prey and for navigation. The ability to interpret the entire visual scene may be unique to humans and seem to develop comparatively late (claims that 18 year old adults are still learning scene perception).

Why is vision difficult? *The world W is complex and ambiguous*, see figure (1) – the world consists of many objects (20,000 based on estimates made by counting the names of objects in dictionaries) and plenty of "stuff" (texture, vegetation, etc.). *The mapping from the world W to the image I is very complex* – the image is generated by light rays bouncing off objects and reaching the eye/camera. Images are like an *encoding of the world*, but it is an encoding that is not designed for communication (unlike speech, or morse code, or telephones), see figure (2). Decoding an image I to determine the state of the world W that caused it is extremely difficult. The difficulty was first appreciated when AI workers started trying to design computer vision systems (originally thinking it would only take a summer).

Another way to understand the complexity of images is by looking at how many images there can be. A typical images is $1,024 \times 1,024$ pixels. Each pixel can take values $1 - 255$. This gives a number of images to be $(1,024 \times 1,024)^{256}$ which is enormous (much much bigger than the number of atoms in the Universe). If you only consider 10×10 images, you find there are more of them than can have been seen by all humans over evolutionary history (allowing 40 year for life, 30 images a second, and other plausible assumptions). So the space of images is really enormous.

If vision is so hard then how is it possible to see? Several people (Gibson, Marr) have proposed "ecological constraints" and "natural constraints" which mean that the nature of the world provides constraints which reduce the ambiguities of images (e.g., most surfaces are smooth, most objects move rigidly). More recently, due to the growing availability of datasets (some with groundtruth) it has become possible to determine statistical regularities and statistical constraints (as will be described in this course). In short, there must be a lot of structure and regularity in images I, the world W, and their relationship which can be exploited in order to make vision possible.

But how can we learn these structures/regularities? How can infants learn them and develop the sophisticated human visual system? How can researchers learn them and build working computer vision systems? It seems incredibly difficult, if not impossible, to learn a full vision system in one go. Instead it seems that the only hope is proceed incrementally learning the simpler parts of vision first and then proceeding to the more complex parts. The study of infants (e.g., Kellman) suggests that visual abilities develop in an orchestrated manner with certain abilities developing within particular time periods – and, in particular, infants are able to perform motion tasks before they can deal with static images. Hence vision may be developed as a series of modules where easily learnable modules may be used as prequisites to train more complex modules.

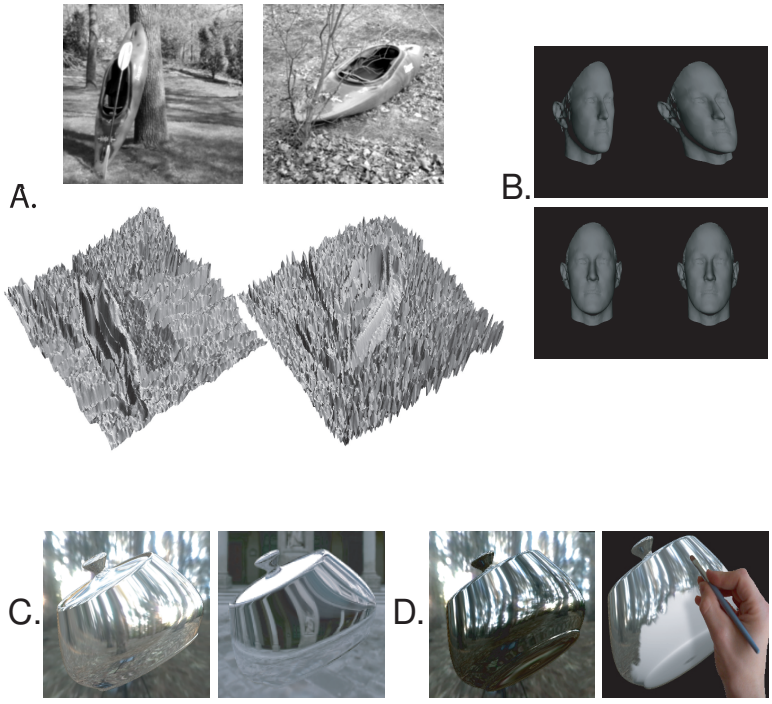

Figure 1. The Complexity and Ambiguity of Images. (A) The two images are of the same object (Dan Kersten's canoe) but the intensity profiles below (plots of the image intensity as a function of position) are very different. It would be very hard to look at these images (represented as plots) and determine that there is a canoe in each. (B) The face appears identical in the two bottom images, but the top two images show that one face is normal and the other is very distorted (more on this in the Lighting chapter). (C) Images of certain obkects (particularly of specular one – like polished metal) depend very strongly on the illumination conditions.

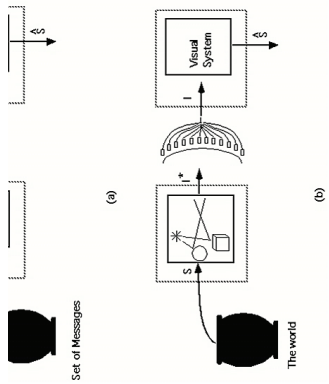

Figure 2. Information Decoding. In standard decoding (e.g., telephone, morse code) the information is encoded and decoded by a system that knows the encoder. The encoding is designed to make transmitting and decoding signals efficient. In vision, the encoding is performed by the physics of the world and hence the decoding is considerably harder.

But why is this incremental/modular strategy possible? Why is it possible to vision incrementally? (You cannot learn to ride a bicycle or parachuting incrementally). Stuart Geman suggests this is because the structure of the world, and of images, is compositional (i.e., is build out of elementary parts). He provocatively suggests that if the world is not compositional then God exists! What he means that if the world/images are compositional (i.e. can be built up from elementary parts) then it is possible to imagine that an infant could learn these models. But if the world is non-compositional, then it seems impossible to understand how an infant could learnt at all (the idea of compositionality will be clarified as the course develops). Less religiously, Chomsky in the 1950's proposed that language grammars were innate (i.e., specified in the genes) because of the apparent difficulty of learning them (it is known that children develop grammars even if they are not formally taught – e.g., children of slaves). But this raises the question of how genes could "learn the grammars" in the first place. Others (e.g.,

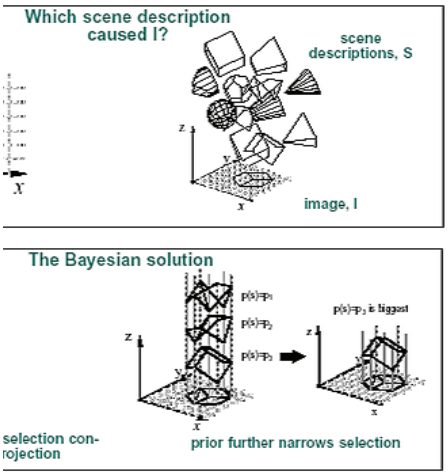

Figure 3. Sinha figure. The likelihood term $P(I|W)$ constrains the interpretation of I to scenes/objects which are consistent with it. But there remain many possibilities. The prior $P(W)$ is needed to give a unique interpretation by biasing towards those W which are most likely.

Hinton) have questioned whether genes can convey enough information to specify grammars (at least for vision). Recent computational work, however, (Klein and Manning) has shown that grammars can, in fact, be learnt in an unsupervised manner (partially supervised – at least, the system must know something about the form of the grammar if not the parameter values). It seem more plausible that the brain has the ability to learn complex patterns (for vision, speech, language). But this only seems possible if the world is compositional. These ideas of compositionality and modularity will be developed during the course.

2. How to formally approach vision?

How to formulate vision taking into account the (unknown) structures/regularities and compositionality/modularity?

This course argues that vision should be formulated in terms of probability distributions formulated on structured representations (e.g., graphs and grammars). In the last few years, there has been a revolution in the scope of probabilistic modeling arising from combining ideas/concepts from Computer Science, Engineering, Mathematics, and Statistics. This revolution has been lead by the machine learning, neural networks, cognitive modeling, and artificial intelligence communities inspired by the need to design artificial systems capable of performing "intelligent tasks" (and also be researchers attempting to understand the complexities of human intelligence).

One main idea here is the formulation of vision as Bayesian inference. The idea is that we have probability distributions $P(I|W)$ for the image generation process and $P(W)$ for the state of the world. Then vision reduces to interpreting $W^* = \arg\max P(W|I)$, where $P(W|I) = P(I|W)P(W)/P(I)$. In vision, these ideas date back to Grenander's concept of Pattern Theory Grenander (1950's-1990's) was ahead of his time (slowness of computers, lack of effective inference algorithms, lack of knowledge about the complexities of vision).

The Bayesian approach can be illustrated by an example from Sinha, see figure (3). Suppose you have an image of a cube. The likelihood function $P(I|W)$ rules out all interpretations W which are not consistent with the image – so W could be a cube or another structure (abstract work of art) that can project to the image of a cube, but it cannot be a sphere, or a face, or a giraffe. But, after using the likelihood, there remain many possible interpretations. The prior is needed to narrow them down to a cube – because a cube is more likely (unless you are in an artist's studio).

Grenander's idea of inference was summarized as "analysis by synthesis". Taken literally, it corresponded to performing stochastic sampling from $P(I|W)P(W)$ until you generated an image that agreed with the input image. This is an example of top-down processing which is opposite to bottom-up processing (more standard in computer vision). Mumford speculated that pattern theory – analysis by synthesis – was able to explain the feedback connections observed in the brain (the feedforward connections correspond to bottom-up).

Specifying a problem by $P(I|W)P(W)$ is called *generative* since it enables us to obtain stochastic samples of images I corresponding to world state W. It can be contrasted to *discriminative* methods, used in machine learning, which (broadly

speaking) attempt to model the posterior distribution $P(W|I)$ directly. Recent work in machine learning develops discriminative models $P(W|I)$ which are also defined on complex structures (graphs and grammars). Hence the boundary between generative and discriminative models is increasingly getting blurred (as we will see in the course).

This leads to a research program which is to develop generative and discriminative probability models which can capture the complexities and ambiguities of images. We will first start with simple probability distributions defined on simple structures (e.g., simple graphs) and then proceed to increasingly complex structures. In all cases, we will have to deal with three major questions: (I) Are the models rich enough to *represent* the complexities of the visual tasks. (II) Do we have inference algorithms to estimate $W^* = \arg\max_W P(I|W)P(W)$ efficiently? (III) Do we have learning algorithms so that we can learn these probability distributions with supervision, or partial supervision? We have to balance the computational aspects of the models (inference and learning) with their ability to represent the visual tasks.

The course describes how to develop this program. The ability to do so is a consequence of the compositionality/modularity of vision. We are able to solve vision tasks without having to solve the entire vision problem – i.e., we do not need to know the full distributions $P(I|W)P(W)$ in order to make progress.