Lecture 1: Winter 2014. Stat 238

A.L. Yuille

Abstract

First (Introductory) Lecture: 6/Jan/2014.

1. Introduction to Vision

Note: each bullet point is intended to take about five minutes in a lecture.

1. *What is vision?* It is the task of extracting information about the external world from light rays imaged by a camera or an eye ("to know what is where by looking" Aristotle). Humans can perform a large number of "visual tasks". For example, in figure (1)(A)(from Kersten and Yuille 2014), humans can detect and localize the fox and its parts (e.g., ears, face, legs), determine its 3D shape, detect the other objects in the scene (the tree, grass, and twigs), estimate the position of the fox in the scene, reason about its activity, and estimate the time of year. The goal of computer vision is to build systems with similar, or even better abilities. Computer vision is part of a bigger enterprize – designing artificial intelligence systems. Many aspects of vision — "high-level vision" – are closely related to other forms of intelligence – reasoning, causal learning, and natural language processing.

2. *Why is vision challenging?* Vision is extremely difficult. This is due to the enormous complexity of images, their local ambiguity, and the large number of visual tasks. The space of images is practically infinite. A typical images is $1,024 \times 1,024$ pixels and each pixel can take values $1-255$ which gives the number of images to be $(1,024 \times 1,024)^{256}$ which is much bigger than the number of atoms in the known Universe. Moreover, images are locally ambiguous as can be observed by looking at the two circular apertures in figure (1)(B). Without context from the rest of the image, humans cannot determine what is inside the apertures. Finally, the number of visual tasks is enormous. For example, in figure (1), the scene consists of a ground plane (covered by grass at the front and twigs at the back) with a tree on the left and a fox emerging from behind it. But there an enormous number of objects (estimated at 20,000) which can occur in many different configurations in a variety of scene structures (though a ground plane is fairly common). The difficulty of performing vision was first appreciated when scientists started trying to build computer vision systems with similar abilities to humans. They rapidly discovered that it was extremely difficult compared to other "intelligent tasks" that seemed much harder (e.g., by 1995 there were automatic chess playing programs which could beat the world champion – but computer vision researchers were unable to detect faces in images).

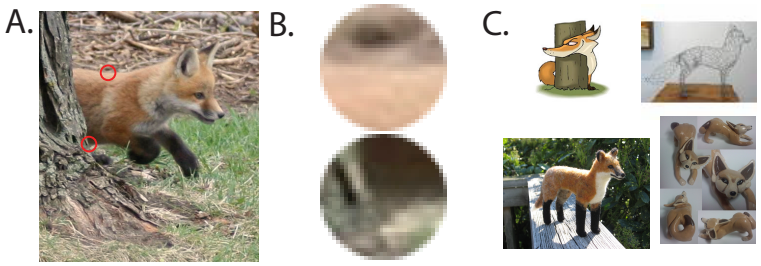

Figure 1. A: Humans can extract an enormous amount of information from a single image and perform a large range of "visual tasks". B: Images are locally ambiguous if the context is removed. C: Humans can easily recognize unusual images.

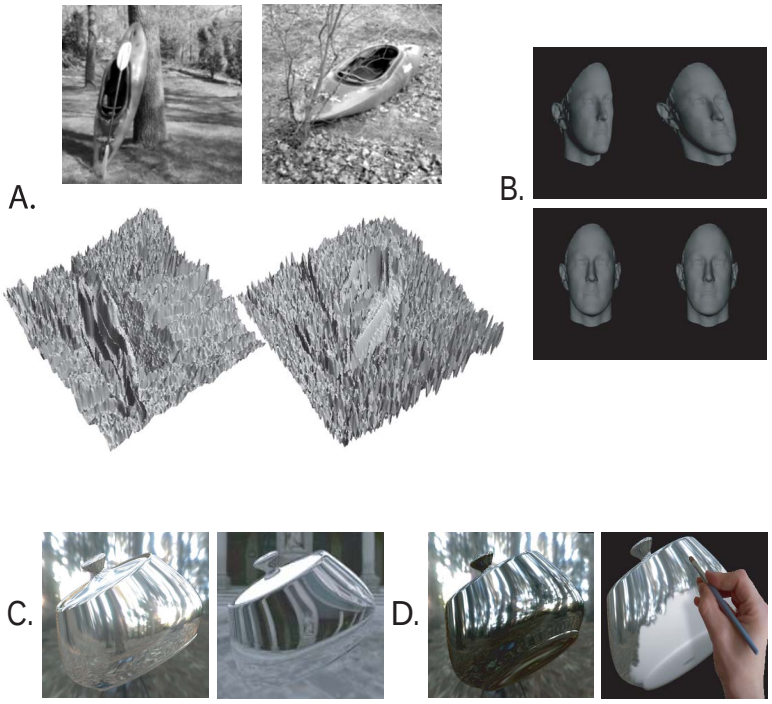

Figure 2. The Complexity and Ambiguity of Images. (A) The two images are of the same object (Dan Kersten's canoe) but the intensity profiles below (plots of the image intensity as a function of position) are very different. It would be very hard to look at these images (represented as plots) and determine 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.

3. *Computer vision as "inverse computer graphics"*. Computer graphics models how to generate images from a specification of the visual scene. The image depends on the illumination (light sources), the shapes and material properties/textures of objects, and their configuration in the viewed scene. This can be complex as shown in figure (2) where: (A) the images of the canoe can appear differently due to their position in the image, (B) two objects of different shapes can generate the identical image, and (C) the image of s shiny (i.e. highly specular) object is extremely dependent on the lighting. The computer graphics metaphor suggests that computer vision can be thought of an inverse problem, where the goal is to estimate the scene configuration which has generated the image. But inverse problems are much harder than "forward problems" (like computer graphics) and the difficulty is compounded by the enormous number of possible scene configurations. A naive "analysis by synthesis" strategy is to estimate the scene configuration by searching over the set of possible configurations (objects, illumination, scene layout), synthesizing the image (by computer graphics), and stopping when the synthesized image equals the observed image. But this is impractical because of the enormous number of scene configurations. We will return to this complexity issue later.

4. *Inverse Problems and Bayes*. Bayesian probability theory gives an idealized way to solve inverse problems like vision. Let $P(I|S)$ be the likelihood function, which is the probability of generating the image I from the scene configuration S. $P(S)$ is the prior on the world. Then $P(S|I) = P(I|S)P(S)/P(I)$ (Bayes Theorem) gives the posterior probability of the state S of the world conditioned on the image I. Estimating the best state $\hat{S} = \arg\max P(S|I)$ combines evidence from the likelihood function $P(I|S)$ with the prior $P(S)$. To see why a prior is necessary, figure (3), gives an example where the image is consistent with many possible interpretations but the prior suggests that one is most likely. In this example, S refers to the object and its three-dimensional configuration. Hence the choice is between a cube at a standard configuration or unusual, and unlikely, objects (expect in an art gallery) seen from unlikely viewpoints (unlikely, because they make the image consistent with the image of a cube, but a small change of viewpoints would make them look different). Apply Bayes to the full vision problem is extremely challenging because of the complexity issues (both of modeling $P(I|S)$, $P(S)$, and of estimating $\hat{S}$).

5. *Why does vision seem to be "easy"?* Vision appears easy and effortless to humans because we are "vision machines".

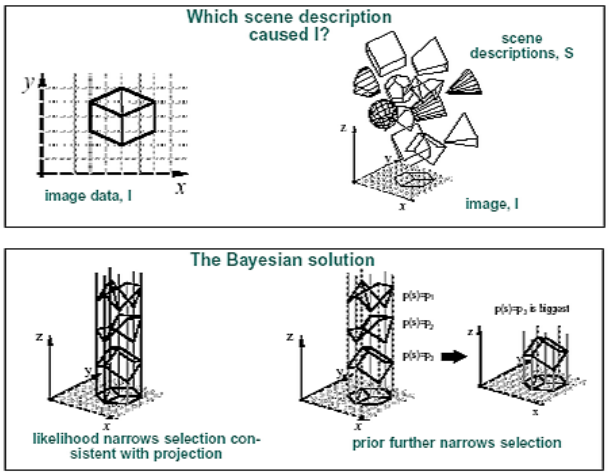

Figure 3. Sinha figure. The likelihood term $P(I|S)$ constrains the interpretation of I to scenes/objects which are consistent with it. But there remain many possibilities. The prior $P(S)$ is needed to give a unique interpretation by biasing towards those S which are most likely.

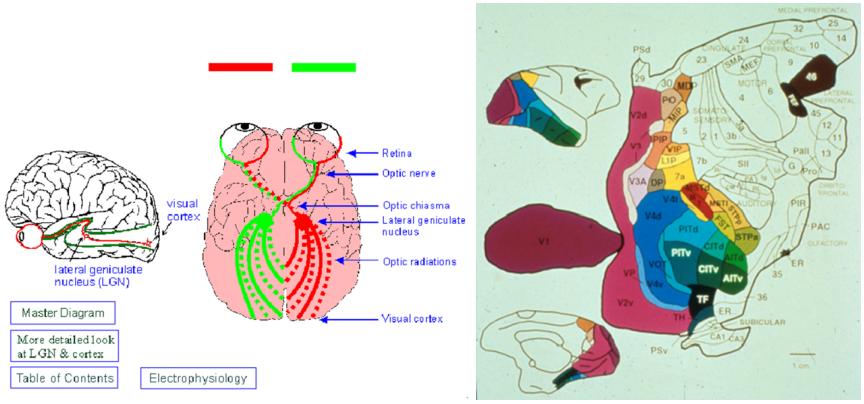

Figure 4. Left Panel: the eye captures images at the retina and transmits them to the early visual cortex (at back of head) via the laternal geniculate nucleus (LGN). Right Panel: the cortex is a sheet of neurons (hundreds of millions) which can be laid out on a flat sheet. Different regions of the cortex can be found (by anatomical and functional methods). These show that roughly forty percent of neurons in the cortex, those areas shown in color, are involved in visual processing.

As shown in figure (4), roughly forty percent of neurons in our cortex are involved in visual processing (the cortex is where intelligence is performed). Much of this processing is performed unconsciously so we are unaware of it (Helmholts described vision as "unconscious inference") unlike, for example, the conscious effort we may need to solve a chess problem or do mathematics (even though these involve much fewer neurons that doing vision). Humans can probably extract much more information from images than most animals can. Although an animal like a chicken can "see", it devotes far fewer neurons to doing vision and hence presumably cannot perform the large range of visual tasks that humans can (otherwise our brains are really badly designed). Indeed you can speculate that humans ability to obtain visual information, combined with our exceptional ability at long distance athletics (at which we outperform almost all animals), is the reason why humans are the dominant species on this planet.

6. *But human vision is fallible.* Vision has been described as controlled hallucination. There are many visual illusions which show that humans can fail to perceive the structure of visual scenes correctly, see figure (**??**). These illusions typically occur when the visual information is impoverished – and it is argued that human biases are due to using strategies will give the correct answers on natural images which contain richer cues. Moreover, humans can move and observe more images which can correct for some of these mistakes. But, when doing computer vision research, it is useful to know that human vision is not always correct – also what you perceive in an image is a function of processing by a large part of your brain and it does not correspond directly to the image that enters your eyes. Humans

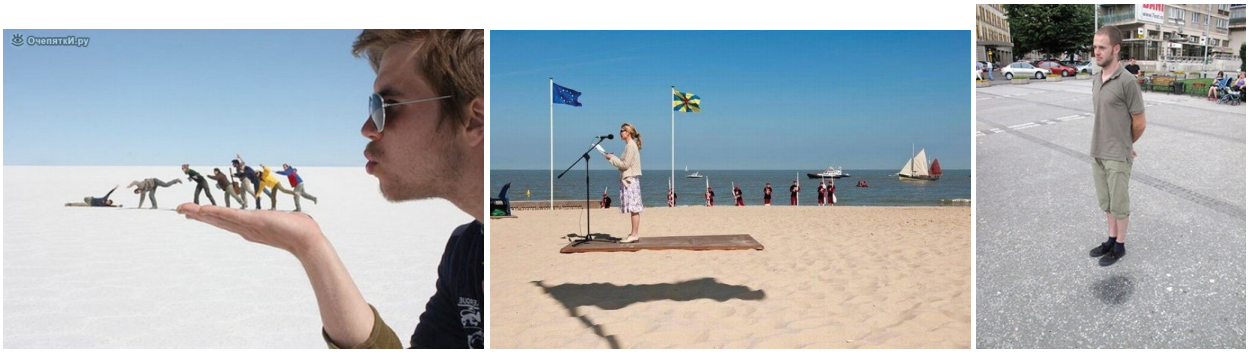

Figure 5. These images illustrate how human perception can fail. Left Panel (from the web): an illusion caused by the accidental alignment of the people in the background and the hand of the man in the foreground. Center Panel (Kersten): the woman appears to be on a flying carpet due to the accidental alignment of the shadow (which looks like it is caused by the beach towel, but instead is caused by a flag out of view). Right Panel: a man seems to be levitating, due to the wet spot which is mistaken for his shadow.

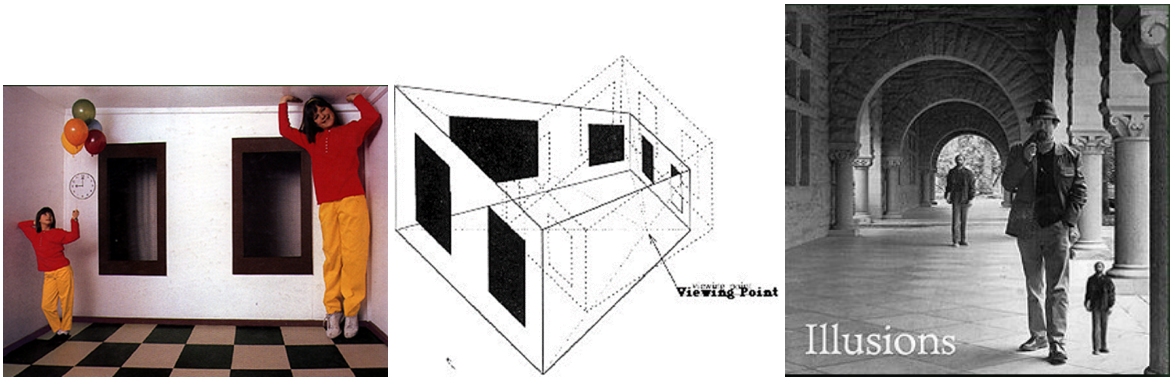

Figure 6. Ames Room: Left Panel: the two girls in this image are the sane size. The girl on the right appears to be larger only because the room has an unusual structure (center panel). *The girl on the left will appear to grow in size if she moves to the girl on the right*. Center Panel: the Ames room is constructed so that an observer at the viewpoint perceives it as a normal room (with the walls perpendicular to the floor, ceiling, and to each other). The observer uses this "Manhattan world" assumption to interpret the image, and hence makes errors whenestimating the size of the two girls. Right Panel: a related phenomena where humans estimate a three-dimensional structure of the scene using perspective cues and judge the size of people relative to their estimated three-dimensional potion. hence the man on the bottom right looks much smaller than the man in the middle left, even though their size in the image is identical.

are particularly likely to make mistakes if they only see subparts of the image, as shown in figure (1)(B).

Image sequences and sampling the world. An image is merely a sample of the set of light rays in space. If we fix a camera, or our eyes, we will receive a sequence of images over time that will give us more information about those parts of the scene which are changing (e.g., we can use structure from motion to estimate the three-dimensional geometry of those objects which are moving). We will gain much more information if we move the cameras (or eyes) because this yields additional cues for estimating the three-dimensional structure of the scene and estimate foreground/background relationships. Finally, humans and robots can gain even more information if we allow them to actively explore scenes in a purposeful manner (e.g., a nightwatchman can explore a dark building using a torch). In particular, the use of image sequences enables algorithms to resolve ambiguities which can occur in static images. For example, many of the perceptual mistakes made in figure (5) will be corrected if we move and get more images. But some ambiguities remain even when there is motion. For example, a person moving in the Ames room still appears to change size when they move, see figure (6) (and an inverted face mask still looks like a real face, even when you move). (*Say something about visual attention.*)

7. *How to design computer vision systems that can perform visual tasks given these enormous challenges of complexity?* (And how can biological visual systems perform them?). The most standard strategy is to divide vision up into a large number of topics/components which can addressed independently (e.g., "edge detection", "object detection", "depth estimation", and so on). Most computer vision conferences are organized in terms of these topics. This is a reasonable

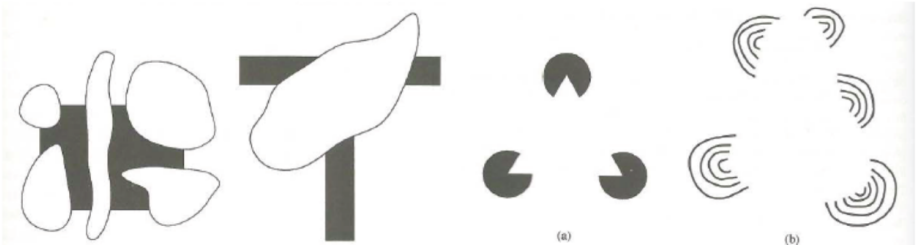

Figure 7. Intermediate level vision includes grouping regions together. These examples from Kanisza suggest that grouping includes knowledge of surfaces. Foreground structures (white blobs) explain why parts of the objects are not seen

organizational strategy but it treats these topics as being independent and ignores the dependencies between them (e.g., edge detection and object detection are dependent problems). Ultimately a "big picture" framework is needed which can put all these components together in a unified manner. We will sketch a "big picture" framework below which is based on hierarchies. More fundamentally, we can think of images in terms of Pattern Theory (e.g., Mumford and Desolnuex 2010). This proposes that images, and other data, should be studied by determining the elementary set of "visual patterns" which occur and how these elementary patterns can be combined to form more complex patterns. This requires systematically understanding and modeling the types of "visual patterns" than can happen in images. This requires developing and testing vision theories on datasets, and visual tasks, which are sufficiently large to be representative of the complexity of the real world (only in the last ten years has this started being practical due to technological developments). These datasets allows us to *learn* models which capture the structure of real world scenes (the ability to learn models from data has lead to major advances in vision over the past 15 years). There are, however, problems associated with some datasets. They can be unrealistic and unrepresentative of the real world (leading to models which over-fit the data and do not generalize to images outside the dataset). They can also cause many researchers to concentrate on the specific task(s) evaluated by the dataset. And, too often, they can lead to research which simply combines techniques in an complicated way to yield slight improvements in performance without adding any insight. *Say more about learning?*.

8. *The Hierarchical Structure of Vision*. It is natural to organize vision hierarchically into low-,intermediate-, and high-level vision. *Low-level vision* performs tasks such as finding salient structures like edges, grouping image pixels which have similar intensity properties, and estimating the motion of images. It uses no knowledge about objects, surfaces, or other scene structures. Or, perhaps more accurately, it performs processing that is "generic" and hence is common to many different scene structures. Note that this assumes the decoupling between different visual tasks discussed above, there are certainly examples of human perception where motion estimation is influenced by the object being viewed rather than by the local structure of the image. Low-level vision typically involves local processing of images which, as discussed above, is often ambiguous. Hence low-level vision should output a set of hypotheses about the image which can be validated or rejected by intermediate- and high-level vision. *Intermediate-level* vision performs a range of tasks including estimating depth, surface shape, and occlusion (where one surface partially hides another surface). There are many ways to estimate depth (sometimes called modules) – e.g., binocular stereo, shape from shading, shape from texture, structure from motion, depth from de-focus, depth from haze, and so on. Intermediate-level vision exploits knowledge about properties of the natural world – e.g., surfaces tend to be smooth, objects tend to move rigidly. *High-level vision* includes object recognition, scene understanding, action recognition, and many others – see figure (8). It exploits detailed knowledge about objects and scenes and so has considerably more knowledge about the world than the lower levels.

9. *Hierarchical Models: Representation, Inference, and Learning* A full hierarchical model consists of a series of hierarchical representations corresponding to low-, intermediate-, and high-level vision respectively. These representations can be computed by processes which include feedforward (from low to high) and feedback (from high to low). For example, to detect a face you first perform the low-level task of edge detection, then group the edges together to find coherent contours (intermediate-level vision), and then recognize that these contours include the boundary of a face and the boundaries of its salient features (e.g., eyes, mouth, nose). But this bottom-up approach does not work on most images (at least not in the simple form described here). The problem is that low-level vision is highly ambiguous – if you look at a small part of an image through an aperture it is very hard to detect edges, people seem to need bigger spatial context and high-level knowledge (e.g., recognizing that the object is a face and hence inferring where its edges

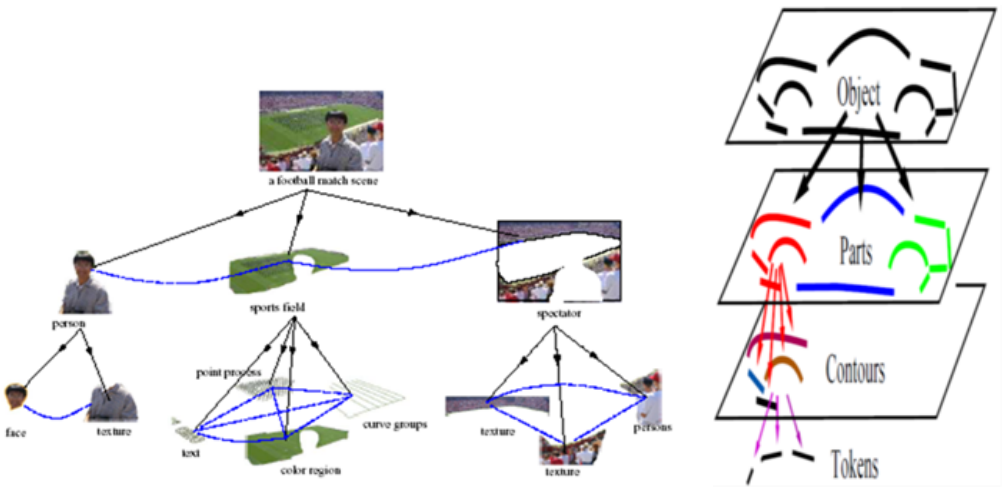

Figure 8. High-Level Vision. Left Panel: This includes parsing an image into its constituent patterns, such as faces, bands, spectators. Right Panel: It also includes parsing objects like cars into parts and subparts.

are likely to be). Also the perception of shape is also influenced by high-level knowledge – if you look at the inside of a face mask you will see it as a normal (convex) face even though the local depth cues indicate a concave object. Hence to get good results on some low-level tasks you need access to high-level knowledge, which precludes a simple bottom-up strategy. A more sophisticated alternative is to use bottom-up processing to propagate forward multiple hypotheses (e.g., about the positions of edges) but without committing to a single hypothesis (unless the bottom-up evidence is overwhelming). The high-levels can resolve the ambiguities of these hypotheses by exploiting higher level knowledge and propagate down to remove false hypotheses. Debates about the roles of bottom-up and top-down processing are common in biological vision.

10. *Computer vision techniques.* Computer vision is a dynamic interdisciplinary field which has absorbed techniques from many disciplines (e.g., CS, Engineering, Mathematics, Statistics). A list of "20 techniques that all computer vision researchers should know" has rapidly grown to over 80. Several of these techniques, however, are based on similar underlying principles. There are a core set of ideas that keep arising. The techniques include filtering, geometry, probabilities on graphs, inference algorithms, learning algorithms. This course will try to give examples of these core techniques.

11. *How to organize a course on Computer Vision?* Computer vision is extremely complex. There are three strategies: (I) Organize the course around big picture theories – e.g., focus on how to build a general purpose visual system that can perform all visual tasks. (II) Organize a course around specific visual tasks and applications – i.e. how to build computer vision systems that are of practical use? (III) Organize a course around techniques – e.g., the 20-plus, or maybe, 80-plus, techniques which everybody doing computer vision should know. There are limitations to each of these strategies (most vision course are either type II or type III). A common critique of big picture theories is that they rarely lead to practical real world systems. In the 198o's several big picture theories were developed but researchers found it hard to make progress using them and turned instead to narrower and more achievable projects. But focus on narrower projects can lead to tunnel vision and the break up of computer vision into an enormous bag of tricks/techniques. Hence this course attempts to cover the large range of techniques but within a big picture framework.

12. *Specific Visual Tasks and Applications*: In the last few years there have been several very impressive vision applications. These include Kinect, Face Detection and Recognition, Google Goggles, Build Rome in a Day, Automated Vehicles – and more examples keep arising (e.g., cosmetic surgery). All of these focus on specific tasks and work in specific domains, which greatly reduces the complexity.