Day 1: Summary.

Lecture 1. Vision is very difficult because images are extremely ambiguous and complex. The set of all possible images is enormous. All humans (50 billion), living 60 years, seeing 30 images a second have only seem a tiny fraction of all possible 10-10 images. Vision seems easy for humans only because we use a very big part of our brain to do vision (half the cortex). In the 1970s, people thought that vision was easy (a summer project) and the chess was much harder – but by 1995 there were computers which could beat the world chess champion, but computer vision could still not find faces in images. Vision is a decoding problem which inverts the process which generates the image – light is reflected off objects and is received by an eye or camera. Bayesian probability theory gives a way to address inverse problems. Generative models and discriminative models. Probability distributions over structured representations – a theory for all intelligence, not just vision. Realize that most animals probably do not understand entire images (unlike humans) but instead they may only use vision to detect prey, predators, and to navigate. They are often only able to detect motion and may not understand static images. See `Vision\_and\_the\_Brain.pdf’.

Lecture 2:

An image is a rectangular array (lattice) of pixels. Each pixel takes values 0 to 255. Look at an image with intensity values given by numbers. Can you say what it is? If you can, you probably segment it into two regions – by making a boundary at places where the intensity changes rapidly and grouping together (into the same region) neighboring pixels which have similar intensity values. In short, you assume that images consist of regions, with slowly varying intensities, separated by sharp edges where the intensity changes rapidly. Is this true? To understand more, we have to introduce methods which allow us to measure properties of images. Filters (smooth/derivative) are techniques for processing images (tutorial on these). We compute histograms – non-parametric marginal probability distributions of the response of derivative filters. These have a characteristic form (similar for almost all images) – most pixels have small derivatives, but some have large derivatives (non-Gaussian). Regions where intensity changes slowly, separated by rapid jumps (edges). This is too simple, but a good start.

Lecture 3:

Statistical edge detection. Standard step edge model (Canny). But is this realistic? To understand better, get people to label the edges -- give the groundtruth. Learn the filter responses on and off groundtruth edges, to get the conditional distribution of the response (conditioned on and off edges). This shows that edges often have high derivatives but non-edges can also have high derivatives, so edge detection is not easy. Do Bayesian inference to estimate the probability that a pixel is an edge. – requires also learning the prior probability of a pixel being an edge (Bayes Decision Theory). We can learn the probability of multiple filter responses conditioned on and off edges – this gives better performance. More recent statistical methods (Berkeley, Kokkinos) get performance very close to humans. Regression models using filter responses. (Need to discuss “learning” – why it is limited by the amount of training data – Vapnik). See `EdgeDetectionExamples.pdf’.

Lecture 4: Piecewise smoothness (weak smoothness, weak membrane). The histograms of derivatives suggest that images are piecewise smooth (as a first approximation). Introduce the Total Variation (TV) model (Osher et al). This minimizes an energy function – data term is quadratic penalty, prior term is the modulus of the gradient. The input is an image – the output is a smoothed version which preserves edges (mostly). It can be used to denoise images, or to do edge detection by thresholding the gradient of the output. How to find the minimum of the energy? This energy is convex and bounded below, so it only has one minimum. So steepest descent algorithms will converge to this minimum. There are several variants of steepest descent (split Bregman, total variations, level sets). The TV norm encourages sparsity (mathematical argument shows bias to exactly zero). By contrast, if we use a quadratic prior term – we will smooth out edges and also allow many small values. We can use this energy to make a Gibbs distribution – quadratic prior is a Gaussian. It is known that Gaussian are non-robust. The TV norm was state of the art for image denoising until a few years ago. See `EdgeDetectionExamples.pdf’. And ExtrasDay1.

Lecture 5: Manhattan World. Images are formed by perspective projection (simple pinhole camera model) of a three-dimensional scene. Parallel straight lines in the image (e.g., the sides of a railway line) are projected to straight lines in the image that converge to a vanishing point. Many scenes have a Manhattan structure where edges mostly occur in the x,y, and z directions. Humans seem to use this knowledge to estimate the three-dimensional structure of the scene – and to interpret properties of images (e.g., size of people) in terms of this three-dimensional structure (and not in terms of their size in the image). This can cause visual illusions is the scene does not have a Manhattan structure – such as the Ames room, where people appear to change size as they move in the room. But how do people, or a computer vision system, determine: (i) is the image of a Manhattan scene?, and (ii) if so, what is the orientation of the eye/camera relative to the three dimensional x,y,z coordinate system of the scene? We describe a simple generative model: the input E is the intensity gradient at each pixel in the image. This image is generated by a process that depends on the (unknown) orientation Psi and on hidden variable m (one for each pixel) which specifies if the pixel is the image of an edge in the x,y, or z directions, of an edge in a different direction, or is not the image of an edge. The probability of the magnitude of the gradient is given by the conditional distributions (specified in Lecture 2) and is a function of the hidden variable m. The probability of the gradient direction depends on the hidden variable m – whether the pixel is an x-edge, a y-edge, a z-edge, a random edge, or not an edge – and, if it is an x,y,z edge then it also depends on the camera orientation. We put a prior on the hidden variable which is independent between pixels (a bad approximation to the real world, because x,y,z edges are typically continuous). We assume that each pixel is generated independently. We use Bayes to compute the posterior probability of Psi summing out the hidden variable m. We can find the most probable value of Psi by exhaustive search. Results are good on images with Manhattan structure (i.e. we superimpose our prediction of the x,y,z axes on the images, and show that they align to the edges in the image). To check that an image is Manhattan we can do model selection – we design a null model for generative the intensity gradient at every image pixel. This null model is a special case of the Manhattan model where no pixel is aligned to the x,y,z axis. For an input image, we compute its intensity gradients E. We calculate the probability of generating them using the null model and the Manhattan model. If the probability is higher with the Manhattan model – then we decide that the image is Manhattan – otherwise we decide it is not-Manhattan. Note, this model is much too simple (i.e. ignores the spatial continuity of x,y,z, edges) and is not state of the art. But the term Manhattan world is popular for describing images with an x,y,z structure. See `ManhattanWorld.pdf’.