Perception 1

Some early work in gesture recognition used principal component analysis on static images and looked at the values of the PCA projections over time to determine what gesture the user was performing. Example gestures included waving hello, good-bye, opening and closing the fist, etc. What role would dynamic time warping play in recognizing these gestures? What benefits could be achieved by using hidden Markov model framework (similar to modern speech systems) instead? Many modern techniques for analysis of video are based on particle filter algorithms. How does a particle filter compare to an HMM? Give an example of a scenario for gesture analysis where a particle filter might be helpful.

Perception 2

In computer vision, many algorithms depend on a coarse-to-fine approach. One of the first real-time target tracking results in computer vision was based on a fast hardware implementation of course-to-fine matching. An early demo video showed an example of tracking a truck as it drives down a road, using video obtained from a moving helicopter. Explain how coarse-to-fine template matching can be used to locate an object of interest in a single video frame. Explain how the method could be applied to track a simple target through a video sequence. Be sure to include specifics on how the coarse-to-fine technique would be implemented. How could you modify your method to deal with occlusions of the target object (for example when the truck passes in and out of tree cover)? Sketch out how you might design a system which could analyze such aerial video footage and extract the path of the truck, a model of the tree cover that it drives under, and the geometry of the roads that it drives on. You should assume that the truck would be tracked over a long period of time, such that the road could branch, etc.

Perception 3

Suppose you have instrumented a room with multiple synchronized and calibrated cameras covering the walls and ceiling. With this camera array you can capture images of an object at the center of the room simultaneously from different, known viewpoints. We are interested in constructing 3-D models of the object from this set of images. Compare and contrast the following two approaches to shape reconstruction: 1) Perform stereo reconstruction separately on pairs of images captured by adjacent cameras. Fit a surface model to the resulting 3-D point cloud. 2) Apply the space carving method of Kutulakos and Seitz (see “A Theory of Shape by Space Carving”) to directly construct a voxel model. Pick any standard stereo algorithm for use in method (1). Your comparison should discuss the assumptions about the object made by each method, the effect of image resolution and image noise on the result, and the computational cost. Give an example of an object for which (1) would produce better results than (2) and vice-versa.

Perception 4

Recently Pietro Perona’s group at Caltech has popularized an approach to object recognition based on the constellation model, which describes how to represent an object for recognition purposes. What are the strengths and weaknesses of this representation? Give an example of an object for which the constellation model might work well and one for which it should work poorly. What assumptions are they making in their learning method for object modeling (the process of fitting constellation models to training data)? What is the computational complexity of analyzing an input image using the constellation model? Felzenszwalb and Huttenlocher recently improved on this by allowing only star topologies of the network of relations between object parts. Describe why this is more efficient and relate their algorithm to belief propagation in junction trees.