Generative Hierarchical Models for Image Analysis

Stuart Geman
Division of Applied Mathematics
Brown University

A probabilistic grammar for the groupings and labeling of parts and objects, when taken together with pose and part-dependent appearance models, constitutes a generative scene model and a Bayesian framework for image analysis. To the extent that the generative model generates features, as opposed to pixel intensities, the “inverse” or “posterior distribution” on interpretations given images is based on incomplete information; feature vectors are generally insufficient to recover the original intensities. I will argue for fully generative scene models, meaning models that in principle generate actual digital pictures. I will outline an approach to the construction of fully generative models through an extension of context-sensitive grammars and a re-formulation of the popular template models for image fragments.

The grammars that I will propose are context-sensitive by virtue of their content-dependent composition rules. An ’L’ is more than a horizontal and vertical stroke found in close proximity. At the least, there are probabilistic constraints on the relative poses of the pieces. And in fact there are many expectations about the common style and color of the two strokes. In general, the likelihood of a composition is dependent on the details of the instantiations of the parts, as in a desk and matching chair or two girls that look like twins. Most compositions are content-dependent. I will propose an approach to constructing probabilistic context-sensitive hierarchical models from a series of content-dependent composition rules.

Mostly I will focus on the problem of constructing pixel-level appearance models. I will propose an approach based on image-fragment templates, as introduced by Ullman and others. However, rather than using a correlation between a template and a given image patch as an extracted feature, I will define a conditional data model on pixel intensities under which the correlation is assumed to be a sufficient statistic. This produces a tractable forward model, and furthermore a fully specified likelihood function that can be used to learn templates from image data. A training set of eyes, for example, yields an ensemble of left and right eyes, of familiar and natural character, but not actually coming from any particular individuals in the training set.

In so far as these templates fully specify a generative model of image patches, they can be used to build simple detection/recognition systems that operate locally, directly on image patches, and are independent of any context or hierarchical modeling. This offers a simple test for the effectiveness of the data model. In fact the result can be of some utility by itself. My colleagues and I have experimented with ethnic classification of East Asian versus Indian faces. Templates were trained on labeled data from each of the two classes. The resulting maximum-likelihood classification of test data, using only the eye regions of faces, is 97% accurate.