Parameterizing Exemplars of Categories

David Mumford
Department of Mathematics
Harvard University

This volume contains two papers dealing with computational approaches to face recognition, and it seems worthwhile to make explicit one of the key differences between them.

Turk and Pentland's article extends a method due to Kohonen for learning categories and reconstructing exemplars of these categories from degraded input, using linear methods. This approach is based on the idea of describing exemplars by individual feature vectors in some vector space V, and the hope that all exemplars of a single category will have feature vectors in or near a high codimensional subspace W of V; in particular, the distribution of these feature vectors might be approximately multidimensional gaussian, with only a few large eigenvalues, and W would be the span of the corresponding eigenvectors.

Yuille's article extends ideas of Elschlager and Fischler on characterizing categories of images by a configuration of feature points. This approach is based on the idea of describing exemplars by vectors from one feature point to another, or some geometrically derived quantities. This idea is extended by having a measure of fit of some kind of template for each feature at each location, and doing a gradient descent over ways to overlay the composite template with suitable proportions over the original image.

I want to make explicit what seems to me a key difference between these approaches, which contrasts the method of representation used by the two approaches. To illustrate the difference, it is convenient to take a toy example in which we can work everything out, instead of trying to imagine what happens right away with such complex stimuli as faces. Let us imagine a category of widgets, which are simply black lines with one white dot on them. Faces are often imagined as relatively bland areas of skin on which the eyes, nose, and mouth are placed in differing configurations (e.g., the eyes closer or further apart), and what we seek to abstract with widgets is the placement of such strongly salient readily located parts as the pupils of the eyes within the face. We make the example one-dimensional to simplify the math.

We imagine the image of the widget normalized to fixed position and scale (to eliminate the background), and sampled on a fixed grid, say with N sample points. The widget then produces a feature vector of N values. Depending on where the white dot is, the feature vector will be $N - 1$ zeroes and 1 value of some positive quantity, call it one. Or if we allow interpixel locations, we may instead have two adjacent pixel values of t and $1 - t$, some t between 0 and 1. The mean of this "cluster" of possible feature vectors is simply

$$(1/N, 1/N, \ldots, 1/N)$$

which may be thought of as a very blurry view of a generic widget. The "shape" of the cluster is either N points forming a basis of V, or, if interpixel dots are allowed, a polygonal curve snaking around V, spanning it. The eigenvalues of this cluster are all equal or nearly equal and the only possible widget subspace W is all of V. This leads to N coordinates for a widget. The alternative, of course, is to measure the widget by the location of the white dot, using 1 coordinate! In other words, if the spatial location of key features is one of the main variables among exemplars of a category, the eigenspace approach leads to an inefficient representation, in which the number of coordinates goes up like the number of possible locations of the feature (up to the desired accuracy of the representation). On the other hand, the strength of the eigenspace approach is that many examples of cluster variability do lead to approximately gaussian distribution in feature space, and then this approach does isolate the independent dimensions of variability very well.

This tension between these two methods of describing a signal also occurs in the theory of wavelets, a recent attempt to use linear methods to achieve the most compact representations of various classes of signals. The idea of wavelets is to start with a basic wave function $f(x)$ and represent all signals as superpositions of translates and scalings of f, specifically via the functions:

$$f_{i,j}(x) = f(2^i x + j)$$

This approach has problems if the signal has discontinuities which are not located on the integral grid, or its refinements by powers of 2. Mallat and others instead have proposed a radically new approach[1] in which many more functions than needed for a basis are used to represent the signal, e.g., the whole two-dimensional family:

$$f_{a,b}(x) = f(ax + b)$$

Then, in the expansion of each signal, only those functions $f_{a,b}$ with large coefficients are retained. This represents a merging of the two schemes for representing signals, or exemplars. To summarize, I feel that to efficiently describe faces, it will be very useful to make a thorough empirical study of what the "shape" of the cluster of face images really is, and what is the most efficient way to encode the variation present.

Note

1. cf. Mallat & Zhong, *Complete signal representation with multiscale edges,* preprint, Courant Institute, NYU, 1990.