Chapter 20

Models for visual words

In most of the models in this book, the observed data are treated as continuous. Hence, for generative models the data likelihood is usually based on the normal distribution. In this chapter, we explore generative models that treat the observed data as discrete. The data likelihoods are now based on the categorical distribution; they describe the probability of observing the different possible values of the discrete variable.

As a motivating example for the models in this chapter, consider the problem of *scene classification* (Figure 20.1). We are given example training images of different scene categories (e.g., office, coastline, forest, mountain) and we are asked to learn a model that can classify new examples. Studying the scenes in Figure 20.1 demonstrates how challenging a problem this is. Different images of the same scene may have very little in common with one another, yet we must somehow learn to identify them as the same. In this chapter, we will also discuss object recognition, which has many of the same characteristics; the appearance of an object such as a tree, bicycle, or chair can vary dramatically from one image to another, and we must somehow capture this variation.

The key to modeling these complex scenes is to encode the image as a collection of *visual words*, and use the frequencies with which these words occur as the substrate for further calculations. We start this chapter by describing this transformation.

20.1 Images as collections of visual words

To encode an image in terms of visual words, we need first to establish a *dictionary*. This is computed from a large set of training images that are unlabeled, but known to contain examples of all of the scenes or objects that will ultimately be classified. To compute the dictionary, we take the following steps:

1. For every one of the I training images, select a set of J_i spatial locations. One possibility is to identify interest points (Section 13.2) in the image. Alternately, the image can be sampled in a regular grid.
2. Compute a descriptor at each spatial location in each image that characterizes the surrounding region with a low dimensional vector. For example, we might compute the SIFT descriptor (Section 13.3.2).
3. Cluster all of these descriptor vectors into K groups using a method such as the K-means algorithm (Section 13.4.4).
4. The means of the K clusters are used as the K prototype vectors in the dictionary.