CHAPTER 9

Segmentation by Clustering

A crucial problem in mid-level vision involves coming up with image representations that are simultaneously compact and expressive. These representations must summarize information available from the first stages of visual processing, and pass them on. Summaries are necessary because early vision produces vast quantities of information. The richness of the available representation tends to overwhelm what is significant. Useful summaries could be computed from pixels or from groups of pixels—for example, by constructing groups of pixels that all have the same color or texture. They could also be computed from local pattern elements—for example, by collecting together edge points that seem to lie on a line or on a circle, or close to some complex geometric structure. The core idea is collecting together pixels or pattern elements into summary representations that emphasize important, interesting, or distinctive properties.

Obtaining such representation is known variously as *segmentation*, *grouping*, *perceptual organization*, or *fitting*. We use the term *segmentation* for a wide range of activities because, although techniques may differ, the motivation for all these activities is the same: obtain a compact representation of what is helpful in the image. It's hard to see that there could be a comprehensive theory of segmentation, not least because what is interesting and what is not depends on the application. There is certainly no comprehensive theory of segmentation at time of writing, and the term is used in different ways in different quarters.

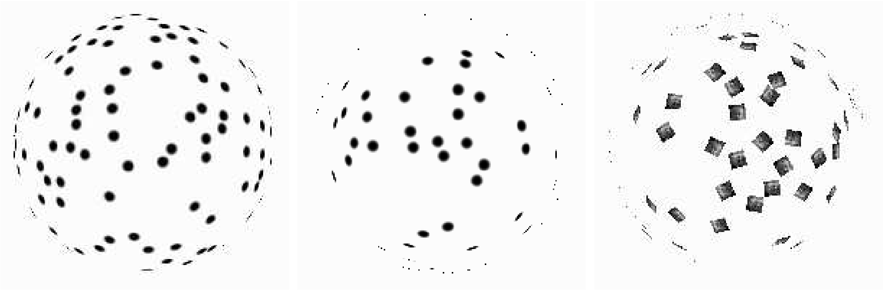

FIGURE 9.1: As these images suggest, an important component of vision involves organizing image information into meaningful assemblies. The human vision system seems to do so rather well. In each of these three images, blobs are organized together to form textured surfaces that appear to bulge out of the page (you may feel that they are hemispheres). The blobs appear to be assembled "because they form surfaces," hardly a satisfactory explanation and one that begs difficult computational questions. Notice that saying that they are assembled because together they form the same texture also begs questions (how do we know?). In the case of the surface on the **left**, it might be quite difficult to write programs that can recognize a single coherent texture. This process of organization can be applied to many different kinds of input.

The details of what the summary representation should be depend on the task, but there are a number of quite general desirable features. First, there should be relatively few (that is, not more than later algorithms can cope with) components in the representation computed for typical pictures. Second, these components should be suggestive. It should be pretty obvious from these components whether the objects we are looking for are present, again for typical pictures.

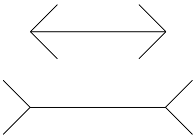

FIGURE 9.2: The famous Müller-Lyer illusion; the horizontal lines are in fact the same length, although that belonging to the lower figure looks longer. Clearly, this effect arises from some property of the relationships that form the whole (the *gestaltqualität*), rather than from properties of each separate segment.

There are two important threads in segmentation, which aren't wholly different. In the first, our summary is assembled purely locally, by clustering methods that focus on local relations between items. Here we are trying to assemble items that look like one another. This approach allows us, for example, to assemble together clumps of pixels that look similar; such clumps are commonly called *regions.* Generally, this approach uses clustering methods, and is the focus of this chapter. In the second approach, we assemble together items based on global relations—for example, all items that lie on a straight line. Figure 9.1 shows a collection of small groups of pixels. When one looks at this figure, these groups of pixels appear to belong together, most likely because taken together they suggest the presence of a surface. In this approach, we are interested in methods that can collect together tokens or pixels of groups of pixels that, when taken together, suggest the presence of a structure of some form. This approach emphasizes methods that can identify parametric models in pools of data; we describe such methods in Chapter 10.

9.1 HUMAN VISION: GROUPING AND GESTALT

A key feature of the human vision system is that context affects how things are perceived (e.g., see the illusion of Figure 9.2). This observation led the Gestalt school of psychologists to reject the study of responses to stimuli and to emphasize grouping as the key to understanding visual perception. To them, grouping meant the tendency of the visual system to assemble some components of a picture together and to perceive them together (this supplies a rather rough meaning to the word context used above). Grouping, for example, is what causes the Müller-Lyer illusion of Figure 9.2: the vision system assembles the components of the two arrows, and the horizontal lines look different from one another because they are peceived as components of a whole, rather than as lines. Furthermore, many grouping effects can't be disrupted by cognitive input; for example, you can't make the lines in Figure 9.2 look equal in length by deciding not to group the arrows.

A common experience of segmentation is the way that an image can resolve itself into a *figure*— typically, the significant, important object—and a *ground*—

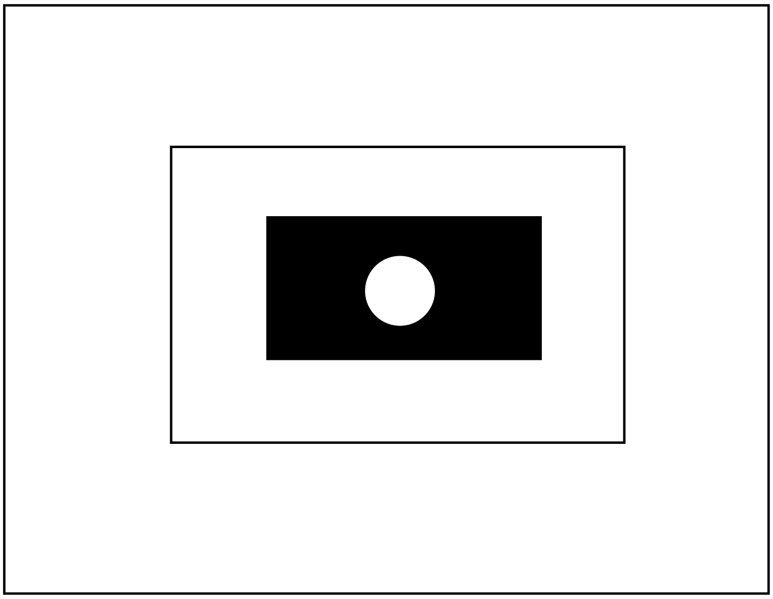

FIGURE 9.3: One view of segmentation is that it determines which component of the image forms the figure and which the ground. The figure illustrates one form of ambiguity that results from this view. The white circle can be seen as figure on the black rectangular ground, or as ground where the figure is a black rectangle with a circular hole in it and the ground is then a white square.

the background on which the figure lies. However, as Figure 9.3 illustrates, what is figure and what is ground can be profoundly ambiguous, meaning that a richer theory is required.

The Gestalt school used the notion of a *gestalt*—a whole or a group—and of its *gestaltqualität*—the set of internal relationships that makes it a whole (e.g., Figure 9.2) as central components in their ideas. Their work was characterized by attempts to write down a series of rules by which image elements would be associated together and interpreted as a group. There were also attempts to construct algorithms, which are of purely historical interest (see Gordon (1997) for an introductory account that places their work in a broad context).

The Gestalt psychologists identified a series of factors, which they felt predisposed a set of elements to be grouped. These factors are important because it is quite clear that the human vision system uses them in some way. Furthermore, it is reasonable to expect that they represent a set of preferences about when tokens belong together that lead to a useful intermediate representation.

There are a variety of factors, some of which postdate the main Gestalt movement:

- **Proximity:** Tokens that are nearby tend to be grouped.
- **Similarity:** Similar tokens tend to be grouped together.
- **Common fate:** Tokens that have coherent motion tend to be grouped to-