In this chapter we continue our exploration of perception of the third dimension. We concentrate on two main topics. The first is the perception of depth—the three-dimensional shapes of objects. As was the case with distance, there are a variety of different cues to depth. We first explore four of these cues, along with their accompanying heuristics and their limitations. We then explore the way in which the visual system combines depth cues to provide a unified perception of the three-dimensional shapes of objects.

The second major topic concerns physiological studies of the tuning of single neurons in various parts of the visual cortex for variations in distance and depth. We review the neurophysiology of distance and depth together because there are few (if any) neurophysiological studies that make the distinction between distance and depth. Our review of distance and depth coding will bring us back to a question we have seen before: What form do we expect particular kinds of visual information to take within the brain? If we perceive the distances, sizes, and depths of the objects in a scene as a single unified perception, what does this imply about neural representations? Do we expect to find a separate representation of each distance and depth cue, followed by a single representation in which all of the depth or distance cues are combined? If not, what are the options?
24.1 Cues to depth

In the previous chapter, we made a distinction between distance and depth, reserving the term depth to refer to the three-dimensional shape of an object. Of course, it’s likely that distance and depth perception will depend on overlapping sets of cues. However, there are some cues that provide information about distance but not about three-dimensional shape, and vice versa. Let’s go back to the beginning, and discuss four cues to depth: shading, texture, motion, and disparity. In much of the current literature, three-dimensional shape is referred to simply as shape, and the use of the various cues is called shape-from-X: shape-from-shading, shape-from-texture, shape-from-motion, and shape-from-disparity. We will adopt this terminology. (There are other shape cues, including shape-from-contour and shape-from-highlights, but we will omit these for simplicity.)

24.1.1 Shape-from-shading

The first depth cue is shading – lighting and shadows. In the physical world, if there is only one source of illumination (say, the sun), then the surface of an object that faces toward the sun will have the highest illumination level, whereas the surface away from the sun will be in shadow, and there will be a pattern of shading from light to dark in between. The retinal illuminance in the image of the object will change regularly over the differently angled surfaces of the object, and the pattern of change will be different for objects of different three-dimensional shapes – think of a cylindrical silo vs. a square barn. These shading patterns are a potential cue to three-dimensional shape. An example of an artist using shading patterns to provide cues to depth is shown in Figure 24.1.

There are three things to notice about the shading cue. First, you may notice that shading is a
new cue in this chapter – it was not part of our list of distance cues in the prior chapter. In fact, shading is a cue to depth, but not a cue to distance. That is, three cylinders of different sizes will yield three shading patterns that are scaled in size but otherwise identical. Nothing in the shading pattern provides information about the distance or the size of the cylinder. Shape-from-shading is a pure shape cue – the shading pattern provides information about shape and shape alone.

Second, shading is a pictorial cue; and as with all pictorial cues, we should suspect that heuristics will play a major role in the use of shape-from-shading cues. The basic assumption that we introduced above – that the surface reflectance is constant over the object, is a heuristic. Usually the surface reflectance is constant; but if the paint on the silo were shaded cleverly from white to grey, the shape-from-shading mechanism would be defeated. The heuristics involved in using shading cues are doubtless complex, but in general they would be something like: a specific gradual shading pattern should be perceived as a cylinder; a different gradual pattern as a sphere; a sharp change in illuminance as a corner; and so on.

### 24.1.2 Shape-from-texture-compression

A second pictorial cue to three-dimensional shape is the cue of texture compression. We mentioned texture cues earlier – they are a major pictorial cue to distance. However, it turns out that there are several aspects to texture. The aspect that matters to distance perception is variations in the sizes (or average sizes) of the texture elements. But a different aspect of texture – the relative compression of texture elements – turns out to provide an important depth cue. Examples of texture compression patterns that yield perceptions of depth are shown in Figure 24.2.

Like shading, texture compression is a cue to depth but not to size or distance. Consider the figure with its portrayed surface covered with a texture of circles and ovals. In the left panel, the entire figure is covered with circles. In the center panel, the texture at the center is composed of circles, but as you approach the left and right edges, the circles give way to more and more...
Figure 24.3: A schematic illustration of how to generate the kinetic depth effect. The figure on the left represents a transparent, rotating cylinder with dots on its surface. The lengths of the arrows indicate the directions and speeds of the dots in the sequence of images generated by the cylinder as it rotates. The figure on the right represents the 2D image from these motions. Remarkably, subjects group the dots on the two surfaces separately, and perceive the temporal sequence of images as a transparent, rotating cylinder. [From Siegel and Andersen (1988, p. 260)].

24.1.3 Shape-from-motion: The kinetic depth effect

As we discussed in Chapter 20, particular kinds of temporal sequences in the retinal image are cues to the perception of physical motion. A third depth cue arises from motion sequences generated by moving objects. This cue is called shape-from-motion, and the resulting perception is called the kinetic depth effect (KDE). That is, a physical object of a particular three-dimensional shape, moving in front of you, creates a particular temporal pattern of two-dimensional images on your retina; and objects of different shapes will create different temporal sequences. These sequences are potential depth cues, and in fact they provide vivid impressions of objects of particular three-dimensional shapes, moving in depth. Figure 24.3, illustrates an example of how to construct a stimulus that is perceived as a cylinder rotating in depth.

You can demonstrate the kinetic depth effect informally by rotating a three-dimensional object in such a way as to cast a temporal sequence of two-dimensional shadows. Almost any object will do, but an object that is not immediately recognizable from a single shadow will make a more convincing demonstration. For example, try bending a large paper clip into a novel (but fairly flat) three-dimensional shape, with a handle sticking out that you can hold to rotate it with. Place the wire just above the display plane of an overhead projector, with its handle pointed toward you. Its three-dimensional shape will not be easy to discern. Now rotate the bent wire by its handle, just above the display plane of the projector. Even though it is a novel object, the shadows of the bent wire will create a vivid perception of three-dimensional shape, and wires with different bending patterns can be recognized by the temporal sequences of their shadows. [These shadow patterns, of course, show that a heuristic is involved in perceiving shape-from-motion, because they
Figure 24.4: An example stereogram of shape-from-disparity. Disparity and texture compression cues give shape in depth. You should see two ellipsoids at different depths. Instructions on how to view stereograms are in Figure 23.6 of the previous chapter. [From Brenner and Landy (1999, p. 3836, Fig. 3)].

are actually just sequences of two-dimensional patterns arranged to mimic the temporal sequences of retinal images that would come from real objects. If you take an analytical perspective, you can also just rotate any irregular object, and notice the sequence of shapes it presents.]

You can probably also imagine that if you varied the distance from the projector to the screen, the sizes of all of the images would be scaled up, but the temporal sequence would remain the same. That is, like shading patterns and texture compression patterns, temporal sequences arising from object rotation yield cues to shape but not to distance or size.

24.1.4 Shape-from-binocular disparity

Finally, as a fourth cue to depth, we return to binocular disparity. Like two objects at different distances, any three-dimensional object produces slightly different images in your left and right eyes, and this binocular disparity provides a potential depth cue. A striking example of the use of disparity to produce the perception of three-dimensional shape is shown in Figure 24.4.

24.2 Each cue has its limitations

We now come to one final but important point about depth cues. We have said that three of these cues are purely depth cues: shape-from-shading, shape-from-texture-compression, and shape-from-motion. This is because the patterns they depend on scale up proportionally with distance. In other words, if we used these cues alone, we could perceive the 3-D shapes of objects, but not their sizes or distances. These cues would need to be augmented by absolute cues to distance (such as accommodation and convergence) before we could know the sizes and distances of objects.

It turns out that binocular disparity also has a weakness, but its weakness differs from the weakness displayed in common by the first three cues. We learned in the section on distance that the size of the retinal image produced by an object of a given size depends not only on the size of the object but also on the viewing distance. In consequence, the retinal image size is not a reliable cue to the physical size unless it is augmented by a separate estimate of distance. Similarly, it turns out that the binocular disparity produced by an object of a given depth depends not only on the
depth of the object, but also on the distance at which the object is viewed\(^1\). In consequence, the binocular disparity produced by an object is not by itself a reliable cue to the 3-D shape of the object. This is a critical flaw – even if the visual system calculates the disparity, the disparity by itself does not provide sufficient information to specify the depth of the object. This problem is sometimes referred to as the *stereo scaling problem*.

The ambiguity of disparity as a depth cue is nicely illustrated in a study by Elizabeth Johnston (1991). In her study, subjects were shown a random texture display with disparity as the only available cue to depth. The display depicted a half-cylinder standing out in front of the plane of fixation, like that shown in the stereogram of Figure 24.5A, and as shown schematically in Figure 24.5B. The subject could control the amount of disparity in the display. Johnston used a task she called the *apparently circular cylinder* task, shown in Figure 24.5C. In this task, the subject was asked to vary the perceived ratio of the cylinder’s half diameter, \( a \), to its depth, \( b \), until the cylinder appeared to have a circular cross-section. Most people report that the cylinder in Figure 24.5A appears flattened in depth, and would increase the disparity in the display to produce an apparently circular cylinder.

Using the apparently circular cylinder task, Johnston asked subjects to make judgments at three distances: 53.5, 107, and 214 cm. The experiment showed that the settings depended on viewing distance, and were often strongly non-veridical. Johnston’s data are shown in Figure 24.6. At the short viewing distance, a circular cylinder appeared elongated, and a decrease of disparity was required to make it perceptually cylindrical; whereas at the long viewing distance the opposite was true. Only at the intermediate viewing distance was depth perception approximately veridical.

The problem revealed is that without information about viewing distance the disparity cue cannot reveal a veridical depth value.

Moreover, the pattern of results suggests a novel and interesting heuristic. When it receives no information about absolute distance, the visual system apparently adopts a default distance value somewhere near 100 cm, and interprets the amount of disparity by this default distance to come up with a perception of depth. If the distance is less than the default distance, the physical depth is overestimated, whereas if the distance is greater than the default distance, the physical depth is underestimated. The suggested heuristic is: in the absence of a certain bit of information, make a particular pre-specified guess, and get on with the perceptual task. Here, the perceptual guess is that the distance of an object that provides no distance cues is about 100 cm. In DT’s mind, it is only a short jump to a very general and interesting depth and distance heuristic: the physical world, and the objects in it, are three-dimensional (!), and should be interpreted as such unless there is evidence to the contrary.

---

\(^1\)In fact, the horizontal binocular disparity produced by an object of a given depth decreases with the \textit{square} of the distance from the observer to the object. That is, an object \( A \) of a fixed depth produces four times the disparity at 10 cm than it produces at 20 cm; and a second object \( B \) of four times the depth produces the same disparity at 20 cm that \( A \) produces at 10 cm. Try to work out this geometry. Hint – a) with retinal images of a fixed size, disparity would decrease with distance (there is no binocular disparity created by the moon); and b) the sizes of the retinal image also decrease with distance, as always. Together these two effects cause binocular disparity to decrease with the square of the viewing distance.
Figure 24.5: The apparently circular cylinder task. A. A stereogram of a cylinder similar to that used by Johnston (1991) with only disparity cues. B. A schematic view of the stimulus, showing a circular cylinder, with its depth $b$, equal to its halfwidth $a$. C. Schematics of various flattened and elongated cylinders. Cylinder #3 is the circular cylinder. Most subjects judge the binocularly fused cylinder in A to be flattened, like cylinder #1 or #2, and would increase the amount of disparity to produce an apparently circular cylinder. [A From Cumming, Johnston, and Parker (1993, p. 828, Fig. 1B); B from Johnston (1991, p. 1353, Fig. 1); C from Johnston (1991, p. 1355, Fig. 3)].
CHAPTER 24. DEPTH AND THE PHYSIOLOGY OF DISTANCE AND DEPTH

Figure 24.6: Data from Johnston’s experiment (1991) for one subject. The matched depth of an apparently circular cylinder is shown as a function of the half-height of the cylinder. For physically circular cylinders, the depth equals the half height, so veridically perceived circular cylinders would fall along the dashed line. The curve parameter is the viewing distance in cm. Distance had an effect. Close cylinders were seen as elongated in depth, and disparity had to be decreased to make them perceptually circular. Distant cylinders were seen as flattened, and disparity had to be increased. At the intermediate distance of 107 cm, the cylinders were nearly veridically perceived. [From Johnston (1991, p. 1355, Fig. 4A)].

24.3 How do depth cues combine?

As discussed above for the case of distance, a long-standing question in psychophysics has concerned cue integration: how the visual system integrates several potential cues to yield a unified perception. In the case of distance perception, we did not explore their various combinations, but instead relied on Cutting and Vishton’s analysis. Here we address cue integration directly.

The different cues to depth have different weaknesses. Binocular disparity produces a depth cue, but one that cannot be used without the addition of a distance cue; and shading, texture compression, and the kinetic depth effect each allow estimation of the three-dimensional shape of an object, but not of its absolute distance nor its absolute size. Yet somehow we piece together veridical perceptions of the three-dimensional shapes of objects most of the time. How does the system work?

When an observer is viewing an ordinary object under ordinary viewing conditions, all four of these depth cues—shading, motion, disparity, and texture compression— if they were present, all indicate a consistent depth for a single object. And in general, the greater the number of consistent depth cues the better the perception of depth.

But in artificial laboratory situations, we can create stimuli that incorporate only, say, two of these depth cues. We can then perturb the degree of depth portrayed by one of the cues, and find the degree to which the other must be perturbed in the opposite direction to re-establish the
24.3. HOW DO DEPTH CUES COMBINE?

The results of Experiment 1 in Johnston et al. (1994) concerning the combination of motion and disparity cues. The two cues were tested separately and together and the results varied with viewing distance. At 50 cm, depth perception was nearly veridical under all conditions. But at 200 cm, the stereo cue presented alone yielded systematic underestimations of depth. The motion cue presented alone yielded quite veridical estimates, as did both cues together. [From Johnston et al. (1994, p. 2266, Fig. 5)].

There are actually two parts to this question. The first question is, are the different cues initially analyzed separately, or are they analyzed together with some joint optimization algorithm? Some theorists have argued that the visual system analyses all available depth information jointly, and comes out with a single best estimate of the depth of the object (a theoretical position called strong fusion)\(^2\). Other theorists have argued that the visual system has a set of separate modules for initial processing of the different depth cues separately, and combines them only later (a theoretical position called weak fusion). The second question is, if the different depth cues are analyzed separately – if weak fusion prevails – by what rules are the different depth estimates then combined to yield a single perceived depth? Do we use a linear weighted sum, or a more complicated combination rule? And if a linear weighted sum, what determines the weights?

A number of cue combinations have been investigated. For example, Johnston, Cumming and Landy (1994) followed up Johnston’s (1991) study of disparity that we described above, but they added a motion cue to the disparity cue. Using the apparently circular cylinder task, they studied the veridicality of disparity and motion cues separately and in combination. Their results are shown in Figure 24.7. As Johnston had shown before, the subject’s settings varied with viewing distance. At 50 cm, the disparity cue yielded a small over-estimation of depth, whereas motion gave reasonably veridical estimates of depth, as did both cues together. But at 200 cm, the disparity cue yielded large systematic under-estimations of depth, whereas motion alone and both motion

---

\(^2\)In the literature on binocular disparity, the term “fusion” is traditionally used to refer to the fact that the left and right eye images images “fuse” perceptually to yield a single perceived depth or distance. Here we are using the term differently, to refer to the combination of different cues to yield a single percept.
and stereo together yielded quite veridical depth estimates. This initial result, particularly at 200 cm, would seem to suggest that the motion cue simply dominates the disparity cue, but that’s too simple.

These authors did a second experiment, in which they made the two cues discrepant – disparity indicated one depth and motion indicated another. These results are shown in Figure 24.8. They found that the amount of disparity present influenced the amount of motion needed for the subject to perceive the apparently circular cylinder! The greater the disparity, the less motion-based depth was needed. That is, the notion that disparity is not used is incorrect. Moreover, the two cues traded off in a remarkably regular way, consistent with the hypothesis that they were combined as a weighted linear sum. But the weightings were not fixed, but varied with the viewing distance.

Finally, these authors found that if they weakened the motion cue by reducing the length of the motion sequence to two frames, its weight was diminished while the weight of the disparity cue was increased. These and other studies have shown that the weightings given to two cues can vary with circumstances. For example, if noise is added to the texture compression cue, it is weighted less heavily; and similar findings have been presented for other cues and cue combinations.

From findings such as these, Landy, Maloney, Johnston, and Young (1995) have proposed an intermediate theory of depth cue combination, which they call *modified weak fusion*. They build on the fact that each depth cue is imperfect – each is missing the specification of one or more parameters. As we said earlier, the disparity cue is missing a specification of absolute distance, and shading, texture compression, and motion are all missing a specification of absolute size. Landy and his colleagues suggest that the cue combination process acts in three stages. First, each cue is analyzed separately, as far as it can be, given its missing parameters. Second, the cues interact in a very limited way – each provides the missing parameters for the others. In their terms, the
cues promote each other – each cue process gleans what it can from the other cues to fill in its missing parameters. And third, information from the different cues is then combined in a weighted linear sum to yield an overall estimate of the depth of a particular object. Finally, the values of the weights vary depending on the reliability of each particular cue provided by the particular object in the particular situation.

Further studies have also begun to identify situations in which cues combine in different ways. For example, Hillis, Ernst, Banks, and Landy (2002) found that within a modality (2 vision cues), there was “strong fusion” in the sense that an observer is unable to report the individual cues. In contrast, they found that across modality (a vision and a haptic cue) there was “weak fusion” in the sense that observers can report the individual cues.

The theory of cue integration has been strongly influenced by ideas from Bayesian decision theory. It provides a probabilistic framework to describe how to combine multiple sources of partially reliable information. For a nice introduction see Mamassian, Landy, and Maloney (2002). Bayesian decision theory provides a rationale for models with weighted linear sums of the various cues and for how cues should be weighted in accord with their reliability. In addition, it provides a way to formalize some kinds of heuristic arguments in terms of conditional probabilities. For example, the heuristic of using a particular default distance can be reformulated into what are the prior probabilities of a distance given certain viewing condition. Moreover, multiple such heuristics can be sensibly combined by Bayesian inferences (Mamassian and Landy, 2001). It has also led to new ways to test models of how such heuristics and cues combine (Maloney and Mamassian, 2009).

24.4 Depth constancy

Consider next human depth constancy: the veridical perception of three-dimensional depth (shape) of an object despite changes in distance between the subject and the object. What do the experiments of Johnston and colleagues say about depth constancy? To be specific, the manipulations of these experiments included the distance to the object along with the nature of the depth cues. The results in Figures 24.6 show the matched depth of an apparently circular cylinder as a function of distance. A veridical judgment would be for the matched depth to be equal to the half-height as shown by the dashed diagonal line. The observed data are roughly veridical for the 1 m distance but deviate substantially for 0.5 and 2.0 m. Thus, in this experiment, there is some degree of depth constancy but does not generally hold over distances from 0.5 to 2.0 m.

Turning to the literature, there are diverse results for the quality of depth constancy and the related issue of shape constancy considered in a following chapter. At one extreme are investigators arguing there is little if any depth constancy (Todd and Norman, 2003). At the other extreme are investigators arguing that depth constancy is nearly perfect for full cue conditions (Frisby, Buckley, and Duke, 1996). While this debate is not settled, we will focus on a nearly full-cue experiment that provides one of the best candidates for showing relatively good depth constancy.

24.4.1 A nearly full-cue depth constancy experiment

Consider the depth constancy experiment conducted by Durkin, Proffitt, Olson, and Reinke (1995, Experiment 3). In this experiment, subjects viewed “real” physical cones in a brightly lit and structured environment. The shape of the cones was defined by their length relative to the size of the base. As illustrated in Figure 24.9, a subject viewed a single cone with the tip pointed toward
CHAPTER 24. DEPTH AND THE PHYSIOLOGY OF DISTANCE AND DEPTH

Figure 24.9: An overhead view of the apparatus used in the depth constancy experiment of Durgin and colleagues. Physical cones were presented at variable distances from the subject. The cones pointed toward the subject and varied in length. A subject adjusted the aspect ratio of a triangular image in a nearby video monitor to indicate the shape of the cone.

The subject's task was to view the physical cone in one display and adjust a schematic cone in a second display to match in its depth-height ratio. This second display is illustrated in the figure by the rectangle labeled “video monitor” near the subject. Specifically, a subject viewed an adjustable 2D triangular image on a video monitor. They used a mouse to match the shape of this image to the perceived shape of the cone. The subjects were specifically instructed to “pay attention to the shape rather than to the absolute size of the cone”. In this fashion, they obtained a judgment of the apparent depth of the cone relative to its base.

The results of this experiment are shown in Panel A of Figure 24.10. The graph shows the judged relative depth as a function of the physical depth of the physical cone. The curve parameter is the distance to the cone. For all distances, the judged depth grows in rough proportion to the physical depth. There is no reliable effect of viewing distance. The linear fit of these functions reveals a slope of 1.3 which is reliably greater than 1. Thus, these results are consistent with good but not completely veridical depth constancy. Any deviations with distance were on the order of 10%, which is better than the 100% deviations shown by Johnston (1991) with only one or two depth cues.

This experiment also included a control condition with monocular viewing. In this condition, the subjects wore an eye patch over one eye. Otherwise, the stimuli and procedure were the same. The results are shown in Panel B of Figure 24.10. Now there is essentially no effect of the cone's depth on the judgment of depth from either distance. Thus, for this control there is no constancy
for depth.

24.4.2 Interpretation

On the face of it, this experiment demonstrates quite good depth constancy over distances from 1 to 3 m. Furthermore, the lack of constancy in the monocular viewing condition indicates that there were no unintended cues in the procedure. In short, for this range of distances and nearly full cue conditions, one can achieve good depth constancy.

There are two issues that temper these conclusions. First, there are only a few studies that show this degree of constancy even under full cue conditions. The concern is that these few studies may have somehow inadvertently introduced cues that allowed the veridical result (Todd and Norman, 2003). This debate can only be resolved by considering both replications and more diverse generalizations of these results. Think about how these studies generalize to other stimuli and other ways to make the judgments. And, following Holway and Boring (1941), how could you maintain a constant proximal stimulus for depth over changes in distance?

The second issue is to not overstate the role of binocular cues in depth judgments. This experiment used stimuli that minimized pictorial cues to depth. Under these conditions, monocular viewing allowed no depth constancy. This says nothing about depth constancy with good pictorial cues for depth. We consider an related example of shape constancy with somewhat better pictorial cues in the next chapter.
24.4.3 Relation between cue combination and constancy experiments

So far, this chapter has presented two quite different kinds of experiments. In the first part, we described experiments identifying different depth cues and investigating their combination. These studies depended on well-controlled but necessarily artificial stimuli. They revealed a variety of depth cues and the rules for how they are combined. In the second part, we described experiments using more naturalistic stimuli to investigate the degree of depth constancy. These experiments showed quite good depth constancy for distances from 1 to 3 m. These studies represent two distinct and useful approaches to the issues of depth. One emphasizes well-controlled experiments at the cost of artificial stimuli, while the other emphasizes less-controlled but more naturalistic stimuli. A future more complete treatment will need to encompass both of these situations.

24.5 Physiology of distance and depth

We now come back to one of the fundamental themes of this book – the art and logic of forging links between perceptual and physiological events. Suppose you are a neurophysiologist, and you are interested in studying the encoding of distance, size, and depth. You want to attack the question of what parts of the visual brain carry out the analysis of distance and depth cues, and by what sequence of information transformations. What hints can we glean from combining logic with system properties, to mount an intelligent attack on this question? Again, this is the point at which the neuroscientist places her bets on the experiments she thinks will best reveal the fundamental processing features of the visual system.

Perceptual studies tell us that distance and depth information comes in many different forms – cues based on the analysis of pictures, the analysis of motion, and so on. Given the idea of parallel processing, we tend to think of some of these different cues – motion and form, for example – as being analyzed separately in early visual processing. On this line of thinking, one line of attack would be to look for neurons in early visual processing that are tuned to each of the different distance cues separately, and then for other neurons later in the visual system that respond to various combinations of cues or even provide a single unified representation of the distances of objects. Similarly, following on modified weak fusion theory, perhaps we should look for three stages of processing. We might look for an early, separate processing of the different depth cues – shading, texture, motion patterns, disparity; a second stage at which each of these cues is promoted; and a third and final stage at which a combined representation of the three-dimensional shapes of objects emerges.

The fact that there are so many distance and depth cues suggests that the neural analyses underlying distance and depth perception will be complex. Moreover, the depths and distance of objects are important both to our perception of where objects are and to our recognition of what they are, so taking the Ungerleider and Mishkin (1982) characterization of visual streams, we might expect to see representations of depth and distance in both the dorsal and the ventral stream. Taking the Milner and Goodale (1995) action vs. perception approach, we might expect that both distance and depth should be represented in the dorsal stream, as we need both to guide our motor responses (think of reaching for an object and picking it up). We might also expect that depth will be represented in the ventral stream, since the three-dimensional shape of an object is critical to object recognition.

Curiously, there has been very little physiological work motivated by these lines of thinking.
Instead, most physiologists have concentrated almost exclusively on looking for neurons that are tuned for a single depth cue – binocular disparity – and almost exclusively in the dorsal stream\(^3\).

In this section we begin with studies of disparity tuning in individual neurons in visual cortex. We will see that disparity tuning emerges first at early cortical levels – V1 and V2. Moving to the dorsal stream, we will find that MT neurons, which we previously learned were tuned to the direction and speed of motion, are also tuned to disparity. We will then move to a few studies of the responses of dorsal stream neurons to depth cues other than disparity. And finally, we will review some work on responses to depth cues in neurons in the ventral stream.

### 24.6 Depth cues in V1 and V2

In the earliest published study of disparity tuning in cortical neurons, Horace Barlow, Colin Blakemore, and Jack Pettigrew (1967) recorded from V1 neurons in the anesthetized cat. The orientation tuning and the binocularity of most cortical cells had been established by Hubel and Wiesel a few years earlier. Building on this work, Barlow and colleagues used moving black or white bars, or black/white edges of various orientations, in each eye separately, to locate the receptive fields of individual cortical cells in each eye. They then placed the optimal stimulus in both the left and right eyes, and varied the horizontal displacement between the two stimuli, thus varying the binocular disparity of the combined stimulus. They found that many cat V1 neurons were tuned for binocular disparity, and even more importantly, that different V1 neurons responded optimally to different disparities.

Rather than plotting the disparity tuning of their cells per se, Barlow and colleagues presented their data in terms of distance coding. For purposes of illustration, they “solved” the stereo scaling problem by assuming a fixed viewing distance of 50 cm. They then back calculated from each neuron’s optimal disparity to find the distance (relative to 50 cm) that the stimulus would have had to be from the cat, in order to elicit the best response from the neuron.

The results are shown in Figure 24.11. These data showed for the first time that lines and edges at different relative distances from the cat will excite different neurons in the cat’s visual cortex. In other words, *activity in this ensemble of neurons could represent the presence of lines and edges at different relative distances.* These data thus provide us with our first glimpse of a possible neural representation of distance and/or depth in the mammalian cortex.

A decade later, Gian Poggio and Burkhart Fischer (1977) undertook a study of disparity tuning in V1 and V2 neurons in awake, behaving macaque monkeys. The monkeys were trained to fixate a fixation point while the researchers isolated a single neuron, and located its receptive field in each eye. Moving bright and dark bars were used as stimuli. Like Barlow’s study, Poggio and Fischer then varied the relative locations of the stimuli within the two receptive fields – the binocular disparity – and plotted the neuron’s disparity tuning curve. They found that more than 80% of the neurons in both V1 and V2 were tuned for binocular disparity. Moreover, both simple cells and complex cells were disparity tuned in about equal numbers.

\(^3\)It is an interesting historical question to ask, when the multiplicity of distance and depth cues has been appreciated for so long, why have physiologists concentrated so exclusively on binocular disparity, rather than exploring the many other cues to depth and distance? Early comments in the physiological literature even sometimes suggest that binocular disparity is the *only* depth or distance cue. DT’s speculation is that disparity is easy to comprehend from the geometry of having two eyes, especially if the disparity scaling problem is set aside. Perhaps it is harder to be convinced of the value of cues that depend more obviously on (fallible) heuristics.
Figure 24.11: A code for relative distance in cat V1 cells. In this figure, the cat is assumed to be fixated at 50 cm. The Vieth-Muller circle is the locus of all point that yield zero disparity for the 50 cm fixation distance. Each of the numbers at the right of the figure corresponds to a different cortical neuron. The numbers are plotted at the distances that would have led to optimal firing for that neuron. [From Barlow et al. (1967, p. 339, Fig. 6)].
Figure 24.12: Six prototypical neurons with different types of disparity coding. In each panel the abscissa shows the amount of disparity between the stimuli in the two eyes. By convention, negative disparity values (-) represent crossed disparities (created by near objects in normal viewing) and positive disparity values represent uncrossed disparities (created by far objects in normal viewing). The ordinate shows the firing rate of the neuron. The two traces in each panel show responses to two opposite directions of motion. The horizontal dashed and dotted lines show the responses to left and right eye stimuli alone. The top row of panels shows three tuned excitatory neurons, tuned for negative (“tuned excitatory near”), zero (“tuned excitatory zero”), and positive (“tuned excitatory far”) disparities, respectively. The bottom center panel shows a tuned inhibitory neuron. More broadly tuned neurons responsive to negative disparities (“near” cells) and to positive disparities (“far” cells) are shown at the bottom left and right respectively. [From Poggio (1991)].
Some of the patterns of disparity tuning seen in V1 and V2 in this and later papers are shown in Figure 24.12. Some V1 and V2 neurons are narrowly tuned for binocular disparity. The most common type of neuron is the tuned excitatory neuron, which responds best over a narrow range of disparities, and is inhibited by disparities outside this range. Different tuned excitatory neurons respond best to negative (crossed) disparities, to disparities near zero, or to positive (uncrossed) disparities. Similarly, tuned inhibitory neurons are inhibited across a narrow range of near-zero disparities. Other neurons are much more broadly tuned, and are either excited over a broad range of near disparities and inhibited over a broad range of far disparities (near cells) or vice versa (far cells). Other authors have suggested that these cell types are not mutually exclusive, and that there is probably a continuous variation of disparity tuning among V1 and V2 cells. Thus, at least some analysis of binocular disparity apparently occurs at the earliest levels of processing in visual cortex.

In later studies, Poggio and his colleagues also tested V1 and V2 neurons with random dot stereograms. Interestingly, simple cells were generally unresponsive to random dot stereograms. A few complex cells did respond, and some of these were tuned to the amount and direction of disparity in the random dot stereograms (reviewed in Poggio, 1991). Thus, as Julesz would have wanted us to believe from the psychophysical data, at least a few neurons in early cortical processing can signal the amount of binocular disparity even in the absence of other cues to the shape of the stimulus. We will need to wait to see, however, whether these relatively infrequently encountered early cortical neurons form a major basis of the analysis of random dot stereograms.

24.7 Depth cues in MT

In Chapter 21 we learned that neurons in cortical area MT, located at the beginning of the dorsal stream, are tuned for the direction and speed of motion, and we concluded that MT plays a major role in the processing of motion signals. However, it has been known since the early 1980s that MT neurons are also tuned for binocular disparity.

Greg DeAngelis, William Newsome and their colleagues have carried out a series of studies of the disparity tuning of MT neurons (DeAngelis, Cumming, and Newsome, 2000). The logic of the work is similar to that we have already seen for studies of motion from the Newsome laboratory (see Chapter 21), and combines single unit recording, behavioral testing, and microstimulation.

To study disparity tuning in area MT, DeAngelis and colleagues used sparse, moving random dot patterns similar to those used in their earlier studies of motion, but adding binocular disparity as a major independent variable. In the disparity studies, the authors chose to record both the activity of a single neuron, and the background activity of multiple units in the vicinity of that single neuron. The locations and sizes of receptive fields of the individual neurons and multi-unit recording sites were characterized, along with the preferred direction of motion and the preferred speed of the neurons. In most cases, the responses of the single neuron and the multiunit response had similar tuning on all of these dimensions.

The authors then varied the amount of binocular disparity between left and right eye inputs.

---

4In this and later figures, the abscissa shows the disparity between the stimuli presented to the left and right eyes. By convention, crossed disparity – which arises from objects closer than the fixation point in natural viewing – is labelled with a (-) sign, and neurons that respond best to crossed disparity are often labelled near neurons. Uncrossed disparity – which arises from objects farther away than the fixation point in natural viewing – is labelled with a (+) sign, and neurons that respond best to uncrossed disparity are often labelled far neurons.
within the receptive field of the neuron. Figure 24.13 shows six examples of disparity tuned neurons in area MT. In these experiments, the binocular disparity was varied in small steps from \textit{near} (negative) to \textit{far} (positive). Both single-unit and multi-unit responses are shown. As in the case of the V1 and V2 neurons recorded by Poggio and Fischer (1977), MT neurons show clear disparity tuning, but vary in their sharpness of tuning. In each case, both the single-unit activity and the multi-unit activity varied together with the amount of binocular disparity, suggesting a patchy or columnar organization, with cells with similar disparity tuning curves grouped together anatomically.

### 24.7.1 Microstimulation and behavioral studies

A patchy or columnar structure, of course, is very advantageous experimentally because it makes possible the use of microstimulation. To have an effect one must be able to stimulate a group of similarly selective neurons and not neurons that are selective in contradictory ways. Since neurons of the same disparity tuning occurred in clusters, DeAngelis and colleagues had reason to proceed with microstimulation. In these studies they introduce a new stimulus variable: the percent binocular correlation of the dots.

The stimuli and the response paradigms used in these studies are shown in Figure 24.14. The dots outside the receptive field of the neuron under study were all set to zero disparity, and appeared as a plane of twinkling dots to a human observer. Within the receptive field, a chosen fraction (say, 50\%) of the dots had a common disparity, while the other dots had randomly selected disparities. This fraction is the percent binocular correlation of the dot display. The dots with the common disparity were presented with either a positive or a negative disparity from one trial to the next, so that to a human observer the stimulus appeared as a region in which some of the dots were displaced either in front of or behind the surrounding plane of dots. The percentage of binocularly correlated dots was varied from one trial to the next, making the task harder or easier. In the psychophysical experiments, the monkey’s task was to report whether the dots with the common disparity appeared nearer or farther than the plane of fixation.

The results of the study are shown in Figure 24.15. Both of the data sets shown were obtained from multi-unit activity. In both cases the multi-unit activity was tuned for binocular disparity; and in both cases microstimulation increased the proportion of trials on which the monkey reported depth in the direction to which the local neurons were tuned. Thus, in all likelihood MT neurons are on the causal chain that controls our perceptual access to binocular disparity cues.

### 24.7.2 MT neurons tuned for tilt

In Chapter 23 we introduced the concept of \textit{velocity gradients} or \textit{motion flow fields}: the idea that differential motion between an observer and objects in the physical world gives rise to complex but highly specific patterns of motion in the retinal images; and that analysis of these patterns could provide information about the motions, shapes and locations of objects. The next question is, are there neurons that are specifically tuned to variations in velocity gradients?

When a textured surface is \textit{tilted} in depth and moved laterally in front of the observer, the texture elements at the near edge of the plane move relatively rapidly across the retina, whereas those at the far edge move relatively slowly, and there is a velocity gradient in between. For example, imagine a plane tilted around a horizontal axis, so that the top is farther away from you and the bottom is nearer. As the plane moves, the texture elements at the bottom will move more
Figure 24.13: Six patterns of disparity tuning in MT neurons. Each panel shows the responses of a single neuron (SU, open circles, right side axis) and the simultaneous multi-unit activity (MU, closed circles, left side axis) recorded at a particular site in MT. The single-unit and multi-unit activity correlate well, suggesting that neurons in local regions have similar disparity tuning. Panel F shows a recording site at which neurons showed much enhanced firing rates to binocular stimulation, but were not appreciably tuned for binocular disparity. The dashed and solid horizontal lines show spontaneous firing rates for single unit and multiunit responses, respectively. The isolated symbols marked L and R show responses to left or right eye stimulation alone. [From DeAngelis and Newsome (1999, p. 1402, Fig. 4)].
Figure 24.14: The stimuli used for microstimulation studies of disparity tuning. A. The stimulus field. A field of dots was placed over the receptive field of the cell. Filled and open dots show the patterns presented to the left and right eyes respectively. A variable percentage of the dots (e.g. 50%) had a common binocular disparity (percent binocular correlation), whereas the other dots within the receptive field had random disparities, and the dots outside the receptive field had zero disparity. The dots labelled “near target” and “far target” were the target lights to which the monkey was trained to shift fixation, in order to indicate his perception of whether the correlated dots were perceived to be in front of (“near”) vs. behind the plane of fixation (“far”). B. Sequence of events for each trial. The monkey first fixated the fixation point. The field of dots then appeared for 1 sec. In the microstimulation experiments, microstimulation was also applied during the same 1 sec. interval. After the field of dots went off, the monkey refixated to either the “far target” or the “near target” light to indicate his psychophysical response. C. A bird’s-eye schematic of the visual stimulus, with 50% binocular correlation. In the region of the receptive field of the cell, half of the dots – the six dots in the oval – share a common disparity, in this case indicating a “far” target. The other six dots within the receptive field have random disparities, shown as dots at various distances. The background dots all have zero disparity. The additional horizontal line shows the plane in which the correlated dots would appear on “near” trials. [From DeAngelis et al. (2000, p. 310, Fig. 21.4)].
Figure 24.15: The effect of microstimulation on the monkey’s reports of perceived depth. A. Disparity tuning of a multiunit site in area MT. The site is maximally responsive to far (+) disparities. The arrows show the two disparities chosen for behavioral testing. B. Results of the behavioral tests. The abscissa shows the percent binocular correlation in the test field. The ordinate shows the “proportion preferred decisions”; that is, the proportion of trials on which the monkey chose the preferred disparity (far, or +) of the neurons at the site under test. Without microstimulation (open circles), the monkey’s psychometric function is a steadily rising function of binocular correlations. With microstimulation (closed circles), the monkey’s psychometric function is shifted leftward, indicating that he saw the stimulus as “far” on an increased number of trials. C and D. A similar experiment at a site tuned to “near”. (From DeAngelis et al. (2000, p. 311, Fig. 21.5)).
Figure 24.16: Tilt tuning in MT neurons. The black patches with arrows show the patterns of motion of the stimuli that are consistent with tilted planes moving laterally in front of the observer. The lengths of the arrows represent the different speeds of motion. The histograms near the center show the responses of a single neuron to the different motion flow patterns. The graph in the center shows the response in a polar plot: the direction out from the center represents the tilt of the motion flow field, and the distance out from the center shows the magnitude of the response. This neuron shows a maximum response to the flow field shown at the right ($0^\circ$), and a minimal response to the flow field shown at the left ($180^\circ$). [From Xiao et al. (1997, p. 958, Fig. 2)].

Figure 24.16: Tilt tuning in MT neurons. The black patches with arrows show the patterns of motion of the stimuli that are consistent with tilted planes moving laterally in front of the observer. The lengths of the arrows represent the different speeds of motion. The histograms near the center show the responses of a single neuron to the different motion flow patterns. The graph in the center shows the response in a polar plot: the direction out from the center represents the tilt of the motion flow field, and the distance out from the center shows the magnitude of the response. This neuron shows a maximum response to the flow field shown at the right ($0^\circ$), and a minimal response to the flow field shown at the left ($180^\circ$). [From Xiao et al. (1997, p. 958, Fig. 2)].

Guy Orban and his colleagues (Xiao, Marcar, Raiguel, and Orban, 1997) asked whether or not MT neurons were tuned for such variations in tilt. To find out, they tested MT neurons with two-dimensional random dot patterns. They first determined the cell’s preferred direction and speed of motion and its preferred binocular disparity. Patterns with the preferred motion and the preferred disparity were then presented to the cell. These patterns also contained velocity gradients that were constructed as if the dots were placed on tilted planes. Xiao et al. found that some individual MT neurons were indeed tuned for the direction of tilt, with different neurons tuned to different tilts. An example of one such neuron is shown in Figure 24.16. This finding is interesting because it suggests that these neurons have properties useful for representing the orientations of tilted surfaces, and thus helping to represent the spatial layout of the visual world.

These experiments are important ones, because they open the door to the neurophysiological analysis of distance and depth cues other than binocular disparity, and they immediately raise many questions. It would be fascinating to know whether or not neurons like these respond to the
rotations of three-dimensional objects in depth, such that they could underlie our use of structure-from-motion cues. It would also be fascinating to know whether some individual MT neurons are tuned to disparity and others to motion, or whether the same neuron might even be tuned jointly to the same depth value coded by both motion and depth. If so, individual MT neurons might integrate two different shape-from-X cues – shape-from-disparity and shape-from-motion – to represent the depth of an object. Quantitative analysis of neurons like these could allow us to test the modified weak fusion theory, or lead to new ideas about the ways in which depth cues combine.

## 24.8 Depth cues in ventral stream neurons

The above review of the tuning of individual neurons to distance and depth cues has been confined to the early levels of visual processing and to the dorsal stream. But what of the ventral stream? We argued early in this chapter that, insofar is the ventral stream “does” pattern recognition, it too should have neurons tuned to the three-dimensional shapes of objects.

Orban and his colleagues (Janssen et al., 1999, 2000) used random dot patterns and complex shape contours to create pairs of stimuli, A and B, that differed in binocular disparity. These stereo pairs could be presented in two configurations – A to the left eye, B to the right, or vice versa. Human observers viewing these stimuli perceive surfaces that are wavy in three dimensions, with the pattern of waves reversing from concave to convex depending on which stimulus goes to which eye. An example of such a stimulus pair is shown in Figure 24.17. This is an example of shape-from-disparity.

Using monkeys, Janssen et al. presented these stimulus pairs in both configurations in the receptive fields of individual neurons in inferotemporal (IT) cortex – the classical terminus of the ventral stream. They found that more than 1/3 of neurons in IT cortex were tuned for such stimuli, responding preferentially to the stereo pair in one left eye-right eye configuration as opposed to the
other. Responses of a neuron tuned for these differences in binocular disparity is shown in Figure 24.17. These data provide the first evidence that some IT neurons are indeed capable of representing the three-dimensional structure of objects. Jansen et al. (2000) also showed that disparity tuned neurons are confined to one subdivision of IT cortex – area STS, a subpart of area TE that lies in the lower bank of the superior temporal sulcus. Such studies of 3D shape and orientation have also been pursued in V4 by Hinkle and Connor (2002) and Yamane, Carlson, Bowman, Wang, and Connor (2008).

Of course, these results raise many questions. Do these neurons respond to other depth cues? Do they respond to the same aspects of three-dimensional shape to which neurons in MT respond, or do MT neurons respond to only the tilts of planes while STS neurons respond to only the three-dimensional shapes of objects?

24.9 A larger question: The “role” of an anatomical structure

In Chapter 21 we learned that cortical area MT contains neurons tuned for the direction and speed of motion, and that microstimulation of area MT changes a monkey’s psychophysical reports of the perceived direction of stimulus motion. Based on these data, we argued that MT has the right characteristics to play a major role in coding the direction and speed of stimulus motion. But now we also know that many MT neurons are tuned for binocular disparity, and that microstimulation of MT changes a monkey’s reports of perceived distance (and/or depth). Given the new knowledge, and the same logic, we must also argue that MT is poised to play a major role in coding the distance (and/or depth) of visual stimuli. Thus, we must recharacterize our view of the “role” of area MT. Perhaps rather than coding motion per se, MT neurons respond to both shape-from-motion and shape-from-disparity, and are setting up a cue-invariant representation of the three-dimensional shapes of objects.

The moral of the story is that we are still at the beginning of our characterization of the properties of cortical neurons. Moreover, the data we do have on any area is largely dependent on the theories (or whims) of the individual investigators who have been drawn to study it, rather than on any systematic survey of the properties. We know, for example, that MT neurons are tuned for the tilts of planes, and that some TE neurons are tuned for shape-from-disparity; but we don’t know whether or not the reverse is true. Thus, we should not be too confident in our current characterizations of the “roles” of anatomically and physiologically defined structures in the visual brain. We can expect that our descriptions of neurons and structures at all levels of the visual system will be forced to change as visual scientists confront them with new questions, new paradigms and new stimuli.

For a yet more conservative view, Wandell (1995) suggests a different way of putting the question. Rather than speaking of the “role” of an anatomical structure, he suggests that we ask, what computations are accomplished within this structure. This way of thinking may help avoid overconfidence about the the role of a particular structure.