Chapter 6

Visual Perception

This chapter continues where Chapter 5 left off by transitioning from the physiology of human vision to perception. If we were computers, then this transition might seem like going from low-level hardware to higher-level software and algorithms. How do our brains interpret the world around us so effectively in spite of our limited biological hardware? To understand how we may be fooled by visual stimuli presented by a display, you must first understand how we perceive or interpret the real world under normal circumstances. It is not always clear what we will perceive. We have already seen several optical illusions. VR itself can be considered as a grand optical illusion. Under what conditions will it succeed or fail?

Section 6.1 covers perception of the distance of objects from our eyes, which is also related to the perception of object scale. Section 6.2 explains how we perceive motion. An important part of this is the illusion of motion that we perceive from videos, which are merely a sequence of pictures. Section 6.3 covers the perception of color, which may help explain why displays use only three colors (red, green, and blue) to simulate the entire spectral power distribution of light (recall from Section 4.1). Finally, Section 6.4 presents a statistically based model of how information, called cues, are combined from multiple sources to produce a perceptual experience.

6.1 Perception of Depth

This section explains how humans judge the distance from their eyes to objects in the real world using vision. The perceived distance could be metric, which means that an estimate of the absolute distance could be given. For example, a house may appear to be about 100 meters away. Alternatively, the distance information could be ordinal, which means that the relative arrangement of visible objects can be inferred. For example, one house appears to be closer than another one because it is partially blocking the view of the further one.

Monocular vs. stereo cue A piece of information that is derived from sensor stimulation and is relevant for perception is called a sensory cue or simply a cue. In this section, we consider only depth cues, which provide information that contributes toward depth perception. If the depth cue is derived from the photoreceptors or movements of a single eye, then it is called a monocular depth cue. If both eyes are required, then it is a stereo depth cue. There are many more monocular depth cues than stereo, which explains why we are able to infer so much depth information from a single photograph. Figure 6.1 shows an example. The illusions in Figure ?? show that even simple line drawings are enough to

Figure 6.1: This painting uses a monocular depth cue called a texture graduate to enhance depth perception. The bricks become smaller and thinner as the depth increase. (Gustave Caillebotte. Paris Street, Rainy Day, 1877. Art Institute of Chicago.)
6.1. PERCEPTION OF DEPTH

(a) (b)

Figure 6.2: Even simple line drawings provide significant depth cues. (a) The Ponzo illusion: The upper yellow bar appears to be longer, but both are the same length. (b) The Müller-Lyer illusion: The lower horizontal segment appears to be shorter than the one above, but they are the same length.

provide strong cues. In an interesting connection to engineering, the cues used by humans also work in computer vision algorithms to extract depth information from images.

6.1.1 Monocular depth cues

Retinal image size Many cues result from the geometric distortions caused by perspective projection; recall the “3D” appearance of Figure 1.19(c). For a familiar object, such as a human, coin, or basketball, we often judge its distance by how “large” it appears to be. Recalling the perspective projection math from Section 3.4, the size of the image on the retina is proportional to \(1/z\), in which \(z\) is the distance from the eye (or the common convergence point for all projection lines). See Figure 6.3. The same thing happens when taking a picture with a camera: A picture of a basketball would occupy larger part of the image, covering more pixels, as it becomes closer to the camera. Two important factors exist. First, the viewer must be familiar with the object to the point of comfortably knowing its true size. Second, it must be appear naturally so that it does not conflict with other depth cues. If there is significant uncertainty about the size of an object, then knowledge of its distance should contribute to estimating its size. This falls under size perception, which is closely coupled to depth perception. Cues for each influence the other, in a way discussed in Section 6.4.

One controversial theory is that our perceived visual angle differs from the actual visual angle. The visual angle is proportional to the retinal image size. This theory is used to explain the illusion that the moon appears to be larger when it is near the horizon. For another example, see Figure 6.4.

Figure 6.3: The retinal image size of a familiar object is a strong monocular depth cue. The closer object projects onto a larger number of photoreceptors, which cover a larger portion of the retina.

Figure 6.4: For the Ebbinghaus illusion, the inner disc appears larger when surrounded by smaller discs. The inner disc is the same size in either case. This may be evidence of discrepancy between the true visual angle (or retinal image size) and the perceived visual angle.
6.1. PERCEPTION OF DEPTH

Figure 6.5: Height in visual field. (a) Trees closer to the horizon appear to be further away, even though all yield the same retinal image size. (b) Incorrect placement of people in the visual field illustrates size constancy scaling, which is closely coupled with depth cues.

**Height in the visual field** Figure 6.5 illustrates another important cue, which is the height of the object in the visual field. The Ponzo illusion exploits this cue. Suppose that we can see over a long distance without obstructions. Due to perspective projection, the horizon is a line that divides the view in half. The upper half is perceived as the sky, and the lower half is the ground. The distance of objects from the horizon line corresponds directly to their distance due to perspective projection: The closer to the horizon, the further the perceived distance. For familiar objects, such as people, our brains performance size constancy scaling by assuming that the distance, rather than the size, of the person is changing if they come closer. Size constancy falls of the general heading of subjective constancy, which appears through many aspects of perception, including shape, size, and color.

**Accommodation** Recall from Section 4.4 that the human eye lens can change its optical power through the process of accommodation. For young adults, the amount of change is around 10D (diopters), but it decreases to less than 1D for adults over 50 years old. The ciliary muscles control the lens and their tension level is reported to the brain through efference copies of the motor control signal. This is the first depth cue that does not depend on signals generated by the photoreceptors.

Figure 6.6: Motion parallax: As the perspective changes laterally, closer objects have larger image displacements than further objects. (Figure from Wikipedia.)

**Motion parallax** Up until now, the depth cues have not exploited motions. If you have ever looked out the side window of a fast-moving vehicle, you might have noticed that the nearby objects race by much faster than further objects. The relative difference in speeds is called parallax and is an important depth cue; see Figure 6.6. Even just providing multiple images from varying viewpoints within a short amount of time provides strong depth information. Imagine trying to simulate a stereo rig of cameras my snapping one photo and quickly moving the camera sideways to snap another. If the rest of the world is stationary, then the result is roughly equivalent to having two side-by-side cameras. Pigeons frequently bob their heads back and forth to obtain stronger depth information than is provided by their pair of eyes. Finally, closely related to motion parallax is optical flow, which is a characterization of the rates at which features move across the retina. This will be revisited in Sections 6.2 and 8.5. In a VR system, to preserve most depth cues based on motion, it is important to track head position, in addition to orientation; see Section 9.2.

**Other monocular cues** Figure 6.7 shows several other monocular cues. As shown in Figure 6.7(a), shadows that are cast by a light source encountering an object provide an important cue. Figure 6.7(b) shows a simple drawing that provides an ordinal depth cue by indicating which objects are in front of others. This cue is called interposition. Figure 6.7(c) illustrates the image blur cue, where levels are depth are inferred from the varying sharpness of focus. Figure 6.7(d) shows an atmospheric cue in air humidity causes far away scenery to have lower contrast, thereby appearing to be further away.
6.1. PERCEPTION OF DEPTH

6.1.1 Monocular depth cues

(a) Shadows resolve ambiguous depth in the ball and shadow illusion. (b) The interposition of objects provides an ordinal depth cue. (c) Due to image blur, one gnome appears to be much closer than the others. (d) This scene provides an atmospheric cue: Some scenery is perceived to be further away because it has lower contrast.

6.1.2 Stereo depth cues

As you may expect, focusing both eyes on the same object enhances depth perception. Humans perceive a single focused image over a surface in space called the horopter; see Figure 6.8. Recall the vergence motions from Section 5.3. Similar to the accommodation cue case, motor control of the eye muscles for vergence motions provides information to the brain about the amount of convergence, thereby providing a direct estimate of distance. Each eye provides a different viewpoint, which results in different images on the retina. This phenomenon is called binocular disparity. Recall from (3.49) in Section 3.5 that the viewpoint is shifted to the right or left to provide a lateral offset for each of the eyes. The transform essentially shifts the virtual world to either side. The same shift would happen for a stereo rig of side-by-side cameras in the real world. However, the binocular disparity for humans is different because the eyes can rotate to converge, in
addition to having a lateral offset. Thus, when fixating on an object, the retinal images between the left and right eyes may vary only slightly, but this nevertheless provides a powerful cue used by the brain. Furthermore, when converging on an object at one depth, we perceive double images of objects at other depths (although we usually pay no attention to it). This double-image effect is called diplopia. You can perceive it by placing your finger about 20cm in front of your face and converging on it. While fixating on your finger, you should perceive double images on other objects around the periphery. You can also stare into the distance while keeping your finger in the same place. You will then see a double image of your finger. If you additionally roll your head back and forth, you should appear as if the left and right versions of your finger are moving up and down with respect to each other. These correspond to dramatic differences in the retinal image, but we are usually not aware of them because our attention is on the single perceived image.

6.1.3 Implications for VR

Incorrect scale perception A virtual world may be filled with objects that are not familiar to us in the real world. In many cases, they might resemble familiar objects, but their precise scale might be difficult to determine. Consider the Tuscany demo world from Oculus VR, shown in Figure 6.9. The virtual villa is designed to be inhabited with humans, but it is difficult to judge the relative sizes and distances of objects because there are not enough familiar objects. Further complicating the problem is that the user’s height in VR might not match his height in the virtual world. Is the user too short, or is the world too big? A common and confusing occurrence is that the user might be sitting down in the real world, but standing in the virtual world. An additional complication occurs if the interpupillary distance (recall from Section 4.4) is not match with the real world. For example, if the user’s pupils are 64mm apart in the real world but only 50mm apart in the virtual world, then the virtual world will seem much larger, which dramatically affects depth perception. Likewise, if the pupils are very far apart, the user could either feel enormous or the virtual world might seem small. Imagine simulating a Godzilla experience, where the user is 200 meters tall and the entire city appears to be a model. It is fine to experiment with such scale and depth distortions in VR, but it is important to understand their implications on the user experience.

Mismatched In the real world, all of the depth cues work together in harmony. We are sometimes fooled by optical illusions that are designed to intentioning cause inconsistencies among cues. Sometimes a simple drawing is sufficient. Figure 6.10 shows an elaborate illusion that requires building a distorted room in the real world. It is perfectly designed so that when viewed under perspective projection from one location, it appears to be a rectangular box. Once our brains accept this, we unexpectedly perceive the size of people changing as they walk across the room! This is because all of the cues based on perspective appear to be functioning
6.1. PERCEPTION OF DEPTH

Section 6.4 may help you to understand how multiple cues are resolved, even in the case of inconsistencies.

In a VR system, it is easy to cause mismatches and in many cases they are unavoidable. Recall from Section 5.4 that vergence-accommodation mismatch occurs in VR headsets. Another source of mismatch may occur from imperfect head tracking. If there is significant latency, then the visual stimuli will not appear in the expected place at the expected time. Furthermore, many tracking systems track the head orientation only. This makes it impossible to use motion parallax as a depth cue if the user moves from side to side without any rotation. To preserve most depth cues based on motion, it is important to track head position, in addition to orientation; see Section 9.2. Optical distortions may cause even more mismatch.

Monocular cues are powerful! A common misunderstanding among the general public is that depth perception enabled by stereo cues alone. We are bombarded with marketing of “3D” movies and stereo displays. The most common instance today is the use of circularly polarized 3D glasses in movie theaters so that each eye receives a different image when looking at the screen. VR is no exception to this common misunderstanding. CAVE systems provided 3D glasses with an active shutter inside so that alternating left and right frames can be presented to the eyes. Note that this cuts the frame rate in half. Now that we have comfortable headsets, presenting separate visual stimuli to each eye is much simpler. One drawback is that the rendering effort (the subject of Chapter 7) is doubled, although this can be improved through some context-specific tricks.

As you have seen in this section, there are many more monocular depth cues than stereo cues. Therefore, it is wrong to assume that the world is perceived as “3D” only if there are stereo images. This is particularly valuable for leveraging captured data from the real world. Recall from Section 1.1 that the virtual world may be synthetic or captured. It is generally more costly to create synthetic worlds, but it is then simple to generate stereo viewpoints (other than paying the double rendering cost). On the other hand, capturing panoramic, monoscopic images and movies is fast and inexpensive (examples were shown in Figure 1.7). There are already smartphone apps that stitch pictures together to make a panoramic photo and direct capture of panoramic video is likely to be a standard future on smartphones within a few years [?]. By recognizing that this content is sufficiently “3D” due to the wide field of view and monocular depth cues, it becomes a powerful way to create VR experiences. There are already hundreds of millions of images in Google Street View, shown in Figure 6.11, which can be easily viewed using Google Cardboard or other headsets. They provide a highly immersive experience with substantial depth perception, even though there is no stereo. There is even strong evidence that stereo displays cause significant fatigue and discomfort, especially for objects as a close depth [?]. Therefore, one should think very carefully about the use of stereo. In many cases, it might be more

6.2 Perception of Motion

We rely on our vision to perceive motion for many crucial activities. One use to separate a moving figure from a stationary background. For example, a camouflaged animal in the forest might only become noticeable when moving. This is clearly useful whether humans are the hunter or the hunted. Motion also helps us to assess the 3D structure of an object. Imagine assessing the value of a piece of fruit in the market by rotating it around. Another use is to visually guide actions, such as walking down the street or hammering a nail. VR systems have the tall order of replicating these uses in a virtual world in spite of our limited technology. Just as important as the perception of motion is the perception of non-motion, which we called perception of stationarity in Section 2.3. For example, if we apply the VOR by turning our heads, do the virtual world objects move correctly on the display so that they appear to be stationary? Slight errors in time or image position might inadvertently trigger the perception of motion.
6.2.1 Detection mechanisms

Reichardt detector Figure 6.12 shows a neural circuitry model, called a Reichardt detector, which respond to directional motion in the human vision system. Neurons in the ganglion layer and LGN detect simple features in different spots in the retinal image. At higher levels, motion detection neurons exist that respond when the feature moves from one spot on the retina to another nearby. The motion detection neuron activates for a feature speed that depends on the difference in path lengths from its input neurons. It is also sensitive to a particular direction of motion based on the relative locations of the receptive fields of the input neurons. Due to the simplicity of the motion detector, it can be easily fooled. Figure 6.12 shows a feature moving from right to left. Suppose that a train of features moves from left to right. Based on the speed of the train and the spacing between the features, the detector may inadvertently fire, causing motion to be perceived in the opposite direction. This is the basis of the wagon-wheel effect, for which a wheel with spokes or a propeller may appear to be rotating in the opposite direction, depending on the speed. The process can be further disrupted by causing eye vibrations from humming [7]. This simulates stroboscopic conditions, which discussed in Section 6.2.2.

Aperture problem.

Many detectors come together to perceive rigid body motion.

Waterfall illusion?

Optical flow. Vection (at least mention)?

Distinguishing object motion from observer motion Figure ?? shows two cases that produce the same images across the retina over time. In Figure ??(a), the eye is fixed while the object moves by. In Figure ??(b), the situation is reversed: The object is fixed, but the eye moves. The brain uses several cues. Saccadic suppression, which was mentioned in Section 5.3, suppresses vision during movements, which may suppress motion detectors in the second case. Another cue is provided by proprioception, which is the body’s ability to estimate its own motions due to motor comments. This includes the use of eye muscles in the second case. Finally, information is provided by large-scale motion. It if appears that the entire scene is moving, then the brain assumes the most likely interpretation, which is that the user must be moving. This is why the haunted swing illusion, shown in Figure 2.20, is so effective. This is an example of vection, which is problematic for VR: Vision strongly indicates motion while the other senses disagree. This is a leading cause of VR sickness, discussed in Sections 8.5 and 10.1.

6.2.2 Stroboscopic apparent motion

Nearly everyone on Earth has seen a motion picture, whether through a TV, smartphone, or movie screen. The motions we see are an illusion because a sequence of still pictures is being flashed onto the screen. This phenomenon is called stroboscopic apparent motion; it was discovered and refined across the 19th century. The zoetrope, shown in Figure 6.13 was developed around 1834. It consists of a rotating drum with slits that allow each frame to be visible for an instant while the drum rotates. In Section 1.3, Figure 1.20 showed the Horse in Motion
### 6.2. PERCEPTION OF MOTION

<table>
<thead>
<tr>
<th>FPS Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>72</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>1000</td>
</tr>
<tr>
<td>5000</td>
</tr>
</tbody>
</table>

Figure 6.14: Various frame rates and comments on the corresponding stroboscopic apparent motion. Units are in Frames Per Second (FPS).

film from 1878.

Why does this illusion of motion work? An early theory, which has largely been refuted in recent years, is called persistence of vision. The theory states that images persist in the vision system during the intervals in between frames, thereby causing them to be perceived as continuous. One piece of evidence against this theory is that images persist for up to 100ms [7], which implies that the 10 FPS (Frames Per Second) is the slowest speed that stroboscopic apparent motion would work; however, it is perceived down to 2 FPS [8]. Another piece of evidence against the persistence of vision is the existence of stroboscopic apparent motions that cannot be accounted for by it. The phi phenomenon and beta movement are examples of motion perceived in a sequence of blinking lights, rather than flashing frames (see Figure fig:). The most likely reason that stroboscopic apparent motion works is that it triggers the neural motion detection circuitry illustrated in Figure 6.12 [7, 8].

**Frame rates** How many frames per second are appropriate for a motion picture? The answer depends on the intended use. Figure 6.14 shows a table of significant frame rates from 2 to 5000. Stroboscopic apparent motion begins at 2 FPS. Imagine watching a security video at this rate. It is easy to distinguish individual frames, but the motion of a person would also be perceived. Once 10 FPS is reached, the motion is obviously more smooth and we start to lose the ability to distinguish individual frames. Early silent films ranged from 16 to 24 FPS. The frame rates were often not constant and they were played at a faster rate than they were filmed. Once sound was added to film, incorrect speeds and fluctuations in the speed were no longer tolerated because both sound and video needed to be synchronized. This motivated a fixed rate of 24 FPS that is still used today by the movie industry. Personal video cameras remained at 16 or 18 FPS into the 1970s. The famous Zapruder film of the Kennedy assassination in 1963 was taken at 18.3 FPS. Although 24 FPS may be enough to perceive motions smoothly, a large part of cinematography is devoted to ensuring that motions are not so fast that jumps are visible due to the slow frame rate.

Such low frame rates unfortunately lead to perceptible flicker as the images rapidly flash on the screen with black in between. This motivated several workarounds. In the case of movie projectors, two-blade and three-blade shutters were invented so that they would show each frame two or three times, respectively. This enabled movies to be shown at 48 FPS and 72 FPS, thereby reducing discomfort from flickering. Analog television broadcasts in the 20th century were at 25 (PAL standard) or 30 FPS (NTSC standard), depending on the country. To double the frame rate and reduce perceived flicker, they used interlacing to draw half the image in one frame time, and then half in the other. Every other horizontal line is drawn in the first half, and the remaining lines are drawn in the second. This increased the frames rates on the television to 50 and 60 FPS. The game industry has used 60 FPS standard target for smooth game play.

As people started sitting close to giant CRT monitors in the early 1990s, the flicker problem became problematic again. Our perception of flicker is stronger at the periphery, particularly at about 30° from center [9]. Furthermore, even when flicker cannot be directly perceived, it may still contribute to fatigue or headaches. Therefore, frame rates were increased to even higher levels. A minimum acceptable ergonomic standard for large CRT monitors was 72 FPS, with 85 to 90 FPS being widely considered as sufficiently high to eliminate flicker problems. The problem has been carefully studied by psychologists under the heading of flicker fusion threshold; the precise rates at which flicker is perceptible or causes fatigue depends on many factors in addition to FPS, such as position on retina, age, color, and light intensity. This, the actual limits depends on the kind of display size, specifications, how it is used, and who is using it. Modern LCD and LED displays, used as televisions, computer screens, or smartphone screens have 60, 120, and even 240 FPS.

The story does not end there. If you connect an LED to a pulse generator (put a resistor in series), then flicker can be perceived at much higher rates. Go to a dark room and hold the LED in your hand. If you wave it around so fast that your eyes cannot tracking it, then the flicker becomes perceptible as a zipper pattern that appears. This happens because each time the LED pulses on, it is imaged in a different place on the retina. Without image stabilization, it appears as an array of pulses. The faster the motion, the further apart the images will appear. The higher the pulse rate (or FPS), the closer together the images will appear. Therefore, to see the zipper effect at very high speeds, you need to move the LED very quickly. It is possible to see the effect for a few thousand FPS.
6.3. PERCEPTION OF COLOR

6.2.3 Implications for VR

VOR, pursuit, and low persistence
- Slow-switching pixels in LCD
- 90 FPS for headsets (high FOV)

6.3 Perception of Color

Color spaces.
- Why does a display need only have R, G, B? Why is printing CMYK?
- Problem: Limited dynamic range of displays.

6.4 Combining Sources of Information

A Bayesian view, or statistical decision theory...
- Could be multiple cues per sense
- Could be multiple senses, as in vection
- Helps explain adaptation as more data weighs in

Further Reading

  - Muller-Lyer Illusion:
  - Motion detection circuitry: Barlow, Hill, 1963; Mikami, Newsome, Wurtz, 1986; Reichardt 1961
  - Phi phenomenon: Max Wertheimer, Experimental Studies on the Perception of Motion, 1912.
  - The Science of Illusions, J. Ninio.
  - LCD backlight scanning
Bibliography