2 The Travels of a Photon

Natural Image Statistics and the Retina

Supplementary content at http://bit.ly/3aeW07Z

And there was light. Vision starts when photons reflected from objects in the world impinge on the retina. Although this may seem rather clear to us right now, it took humanity several centuries, if not more, to arrive at this conclusion. The compartmentalization of the study of optics as a branch of physics and visual perception as a branch of neuroscience is a recent development. Ideas about the nature of perception were interwoven with ideas about optics throughout antiquity and the middle ages. Giants of the caliber of Plato (~428–~348 BC) and Euclid (~300 BC) supported a *projection* theory according to which cones of light emanating from the eyes either reached the objects themselves or met halfway with other rays of light coming from the objects, giving rise to the sense of vision. The distinction between light and vision can be traced back to Aristotle (384–322 BC) but did not reach widespread acceptance until the investigations of properties of the eye by Johannes Kepler (1571–1630).

Light is transduced into electrical signals by photoreceptor cells, one of the astounding feats of evolution, rapidly allowing the organism to make inferences about distant objects and events in the environment. The function of the visual system is to rapidly extract information about what may be out there. Therefore, the structure of the environment plays a critical role in dictating the pattern of connections and physiological responses throughout the visual system and marks the beginning of our journey.

2.1 Natural Images Are Special

Let us consider a digital image of 100 × 100 pixels, and let us further restrict ourselves to a monochromatic world where each pixel can take 256 shades of gray (0 = black, 255 = white). Such small, colorless image patches constitute a far cry from the complexity of real visual input. Nevertheless, even under these constraints, there is a vast number of possible images. There are 256 one-pixel images, 256^2 two-pixel images, etc. All in all, there are $256^{10,000}$ possible 100 × 100-pixel images. This number is bigger than a one followed by 24,000 zeros: there are more of these image patches than the current estimate for the total number of stars in the universe.

Now take a digital camera, a rather old one with a sensor comprising only 100 × 100 pixels, turn the settings to gray images with eight bits ($2^8 = 256$), and go around shooting random pictures (Figure 2.1). If you shoot one picture per second, and if

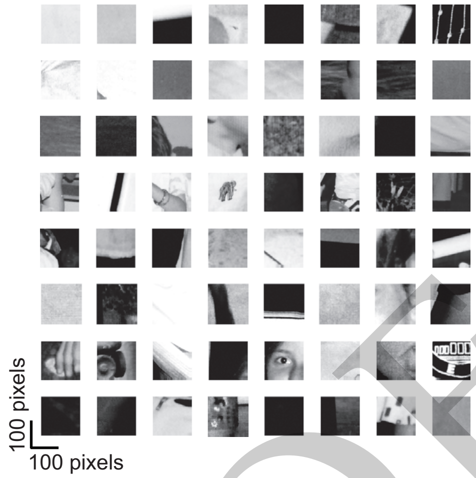


Figure 2.1 Natural images are special. Sixty-four example grayscale patches of 100 × 100 pixels extracted from photographs. Naturally occurring patches constitute a tiny subset of all possible random 100 × 100 image patches.

you spend an entire week collecting pictures without sleeping or pausing to eat, you will have accrued less than a million pictures, a very tiny fraction of all possible image patches. However, even with this tiny sample, you will start to notice rather curious regular patterns. The distribution of *natural* image patches that you collected tends to have peculiar properties that span an interesting subset of all possible image patches.

In principle, any of the $256^{10,000}$ grayscale patches could show up in the natural world. However, there are strong correlations and constraints in the way natural images look. A particularly striking pattern is that there tends to be a strong correlation between the grayscale intensities of any two adjacent pixels (Figure 2.2). In other words, grayscale intensities in natural images typically change smoothly and contain surfaces of approximately uniform intensity. Those surfaces are separated by edges that represent discontinuities, where such correlations between adjacent pixels break; these edges tend to be the exception rather than the rule. Edges play a significant role in vision (Chapter 5), yet they constitute a small fraction of the image.

One way of quantifying these spatial patterns is to compute the *autocorrelation function*. To simplify, consider an image in one dimension only. If *f(x)* denotes the grayscale intensity at position x, then the autocorrelation function A measures the average correlation in the pixel intensities as a function of the separation Δ between two points:

$$A(\Delta) = \int f(x) f(x - \Delta) dx, \tag{2.1}$$

where the integral goes over the entire image. This definition can be readily extended to images with more dimensions and colored images. The autocorrelation function of a natural image typically shows a peak at small pixel separations, followed by a gradual drop.

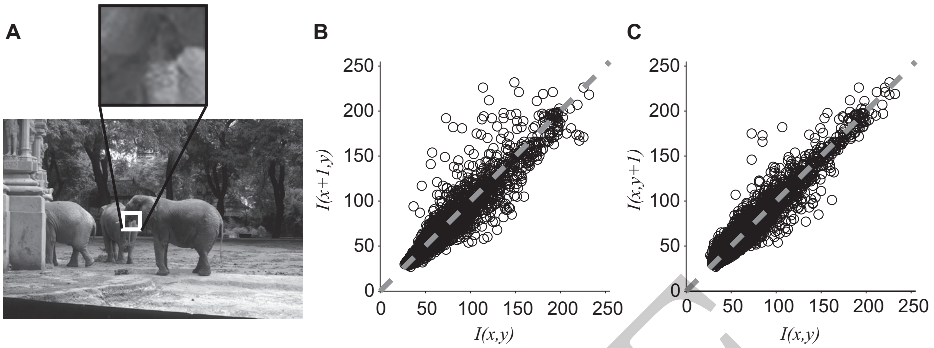


Figure 2.2 The world is rather smooth. For the small 100 × 100 pixel patch from the image in part **A** (white box, enlarged in the inset), the scatterplots show the grayscale intensity at position (*x*,*y*) versus the grayscale intensity at position (*x* + 1,*y*) (**B**, horizontally adjacent pixel) or position (*x*,*y* + 1) (**C**, vertically adjacent pixel). There is a strong correlation in the intensities of nearby pixels in natural images.

A related way of evaluating the spatial correlations in an image is to compute its power spectrum. Intuitively, one can convert correlations from the *pixel domain* into the *frequency domain*. Note that here when we say frequency, we are referring to *spatial* frequencies – that is, how fast things change in space. If there is much power at high frequencies, that implies substantial changes across small pixel distances, as one might observe when there is an edge. Conversely, much power at low frequencies implies more gradual changes and smoothness in the pixel domain. If *P* denotes power and *f* denotes the spatial frequency, natural images typically show that power decreases with *f* approximately as

$$P \sim 1/f^2. \tag{2.2}$$

There is significantly more power at low frequencies than at high frequencies in natural images. Such a function is called a power law. Power laws are pervasive throughout multiple natural phenomena: the sizes of craters on the moon, the frequency of word usage, the sizes of power outages, the number of criminal charges per convict, and the human judgments of stimulus intensities all follow power-law distributions. An important property of power laws is scale invariance. Specifically, if $P(f) = a \cdot 1/f^2$, where a is a constant, and if we multiply f by a scalar *c*, $f' = c\ f$, then $P(f') = a \cdot 1/(cf)^2 = a/c^2 \cdot 1/f^2 = a' \cdot 1/f^2$, with the new constant *a' = a/c*^{2}. If we change the scale of the image, its power spectrum will still have the same shape defined by the preceding equation.

2.2 Efficient Coding by Allocating More Resources Where They Are Needed

One of the reasons why we are interested in characterizing the properties of natural images is the conjecture that the brain is especially well adapted to represent the real world.

This idea, known in the field as the *efficient coding principle*, posits that the visual system is specialized to represent the type of variations that occur in nature. If only a fraction of the $256^{10,000}$ possible image patches is present in any typical image, it may be smart to use most of the neurons to represent this fraction of image space that is occupied. Evolution places a constraint on brain sizes, and it is tempting to assume that brains are not filled with neurons that encode characteristics of images that would never show up in the natural world. Additionally, brains are costly from an energetic viewpoint, and therefore, it makes sense to allocate more resources where they are needed.

By understanding the structure and properties of natural images, it is possible to generate testable hypotheses about the preferences of neurons representing visual information. We will come back to this topic when we delve into the neural circuitry involved in processing visual information (later in this chapter and also in Section 6.12). Such specialization to represent the properties of natural images could arise as a consequence of evolution (*nature*) and as a consequence of learning via visual exposure to the world (*nurture*). The question of nature versus nurture appears repeatedly throughout the study of virtually all aspects of brain function. As in other domains of the nature-versus-nurture dilemma, it seems quite likely that both are true.

Certain aspects of the visual system are hardwired, yet visual experience plays a central role in shaping neuronal tuning properties. For example, the type of light-sensitive molecules in photoreceptors are hardwired; we cannot start to see colors outside the visible spectrum, no matter how much exposure we have to such frequencies. On the other hand, altering the statistics of the visual regime can lead to changes in how neurons respond to visual stimuli. We will come back to the question of what aspects of the neural circuitry are hardwired and which ones are plastic when we discuss the visual cortex (Section 6.12). As an initial guideline, a reasonable conjecture is that plasticity increases as we move up the visual system from the basic sensory elements to the cortical responses. According to this conjecture, the initial processing of visual information discussed in this chapter is mostly hardwired.

2.3 The Visual World Is Slow

The visual properties of nearby locations in the natural world are similar. In addition to those *spatial* correlations, there are also strong *temporal* correlations in the natural world. Extending the collection of natural world photographs in Section 2.1, imagine that you go back to the same locations, and now collect short videos of two-second duration while keeping the camera still. Because the camera is not allowed to move, the only changes across frames in the video will be dictated by the movement of objects in the natural world. If you use a camera that captures 30 frames per second, in most cases, adjacent frames in those videos will look remarkably similar. With some exceptions, objects in the world move rather slowly. Consider a cheetah, or a car, moving at a rather impressive speed of 50 miles per hour. Assuming that we have a camera capturing a distance of about 40 yards in 2,000 pixels, the cheetah will move approximately 30 pixels from one frame to the next. Most objects move at slower speeds. Therefore, the *temporal* power spectrum of the natural world also shows a peak at low temporal frequencies, with large changes typically occurring over tens to hundreds of milliseconds. The visual world is slow and mostly continuous.

Several computational models have taken advantage of the continuity of the visual input under natural viewing conditions to develop algorithms that can learn about objects and their transformations, a theme that we will revisit when discussing computational accounts of learning in the visual system (Chapter 8). Because movement is rather slow and continuous, we can assume that a sequence of images that reach the eyes typically contains the same object, thus automatically generating multiple slightly transformed examples of the same object. These multiple examples can be used to achieve the type of tolerance to transformations highlighted in Chapter 1. The notion of using temporal continuity as a constraint for learning is often referred to as the "slowness" principle.

2.4 We Continuously Move Our Eyes

The assumption that the camera is perfectly still in the previous section is not quite right when considering real brains. To begin with, we can move our heads, therefore changing the information impinging on the eyes. However, head movements are also rather sparse and relatively slow. Even with our heads perfectly still, it turns out that humans and other primates move their eyes all the time. The observation that the eyes are in almost continuous motion might seem somewhat counterintuitive. Unless you have reflected on eye movements or spent time scrutinizing another person's eye movements, introspection might suggest that the visual world around us does not change at all in the absence of external movements or head movements. However, it is dangerous to accept concepts derived from introspection without questioning our assumptions and testing them via experimental measurements.

Nowadays, it is relatively straightforward to measure eye positions quite precisely and rapidly in a laboratory, but this was not always the case, and physicists built ingenious contraptions to capture these rapid eye movements. Figure 2.3 shows an example of a sequence of eye movements during the presentation of a static image. The eyes typically stay more or less in one location, and then rapidly jump to another location, exploring the new location, before adventuring yet again into a new target. These rapid jumps are called visual *saccades* and typically take a few tens of milliseconds to execute from initial position to final position. The approximately constant positions in between saccades are called *fixations*.

During scene perception, subjects typically make saccades of approximately four degrees of visual angle. Degrees of visual angle are the most relevant and standard unit to measure sizes and positions in the visual field and capture the fact that there are many combinations of object sizes and distances to the eye that subtend the same angle (Figure 2.4). One degree of visual angle approximately corresponds to the size of your thumb at arm's length. Under natural scene perception, subjects tend to make saccades approximately every 250–300 milliseconds.

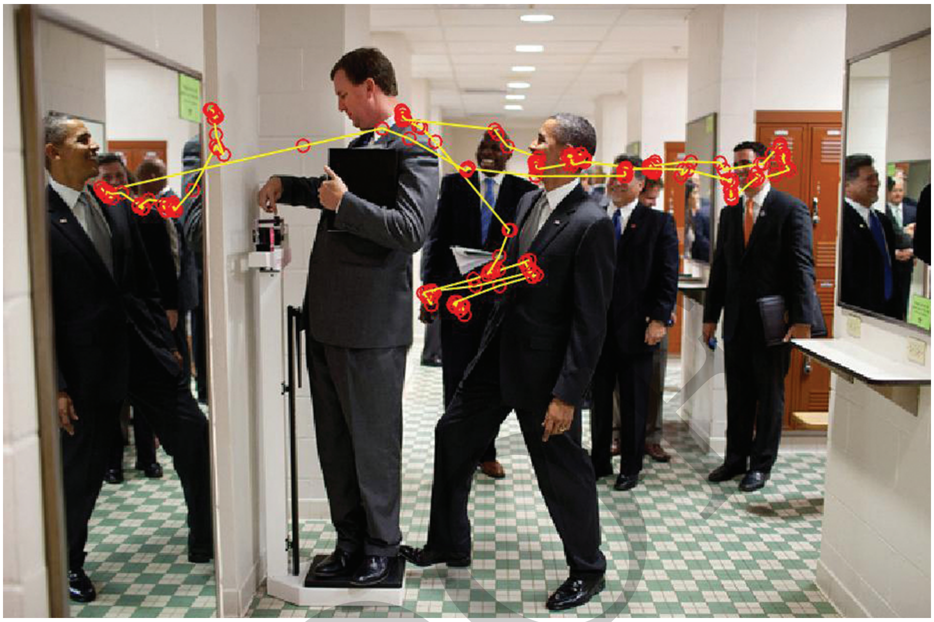


Figure 2.3 Humans frequently move their eyes. Pattern of fixations while a subject observed the image for 12 seconds. This figure shows the eye positions averaged every 33 milliseconds (red circles), and the yellow lines join consecutive eye positions. The whole display was approximately 20 × 30 degrees of visual angle.

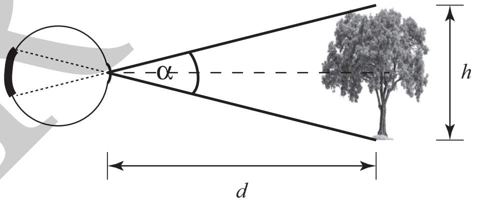


Figure 2.4 Sizes are measured in degrees of visual angle. The size of the tree is characterized by the angle α subtended in the eye. Different combinations of heights *h* and distances *d* give the same visual size in degrees of visual angle.

The intuition that our eyes are mostly still is simply wrong. Why is it that the world does not appear to be jumping from one fixation to the next several times per second? Watching a movie where the camera moves in a ballistic fashion three to four times a second can be quite irksome. The brain takes those retinal inputs that change a couple of times per second and creates the illusion of stability. Additionally, saccades are one of the fastest movements produced by the human body, reaching peak velocities of up to 900 degrees of visual angle per second. Considering a typical saccade spanning five inches in 20 milliseconds, this amounts to almost 15 miles per hour; peak velocities can be much greater than 100 miles per hour. During the few tens of milliseconds when the eyes are moving from one location to another, the sensory inputs change so fast that it is virtually impossible to see anything during a saccade. Every time we make a saccade, we are virtually blind to sensory inputs for a few tens of milliseconds. However, we are usually not aware of these saccades. Our brains have a saccade suppression mechanism so that we perceive a stable world. Even faster than saccades are blinks, which happen about 15 times a minute and typically last about 100–200 milliseconds. There is essentially no input to our eyes for more than 100 milliseconds, about 15 times a minute, and yet we are mostly oblivious to blinks unless we pay special attention to them. Saccadic suppression, blink suppression, and the stability of the visual world when the eyes are jumping from one place to another constitute persuasive examples that show that our subjective perception of the world is a construct. Perception constitutes an interpretation built by our brain based on the incoming sensory information, combined with expectations and with our general knowledge of the world. What we see is not a mere copy of what the eyes dictate.

The pattern of fixations depends on the image, temporal history, and current goals. The characteristics of the image influence eye movements: for example, high-contrast regions are more salient and tend to attract eye movements. The temporal history of previous fixations is also relevant: on average, subjects tend to avoid returning to a location they recently fixated on, a phenomenon known as *inhibition of return*. Current goals also play a critical role as well: if you are looking for your car in the parking lot, you will probably make more fixations on cars, and nearby objects of the same color as your car.

Zooming into Figure 2.3, in addition to the ballistic eye movements spanning several degrees of visual angle and occurring every 200–300 milliseconds (saccades), there are also many other smaller and faster eye movements. These eye movements are called *microsaccades* and typically span a fraction of a visual degree. Because these eye movements take place during the more or less stable fixations, they are referred to as fixational eye movements. Most saccades are involuntary (as noted before, we are typically not even aware that we are making saccades), but, of course, we can volitionally control our saccades. In contrast, microsaccades are involuntary. Together with other fixational eye movements, these small shifts in eye position may play a critical role in preventing adaptation. As we will see in Section 6.9, in the absence of any type of external movement, head movement, or eye movement, neurons quickly adapt to the inputs by reducing their activity. In fact, surprising experiments have shown that if the image on the retina is perfectly stabilized – through an apparatus that is capable of slightly moving the image to account for small eye movements – then the image quickly fades from perception. In other words, without constant eye movements, we would not be able to see anything except for transient changes due to moving objects or head movements.

2.5 The Retina Extracts Information from Light

The adventure of visual processing in the brain begins with the conversion of photons into electrical signals in the *retina* (diminutive form of the word *net*, in Latin). Due to its accessibility, the retina is the most studied part of the visual system. The conversion of

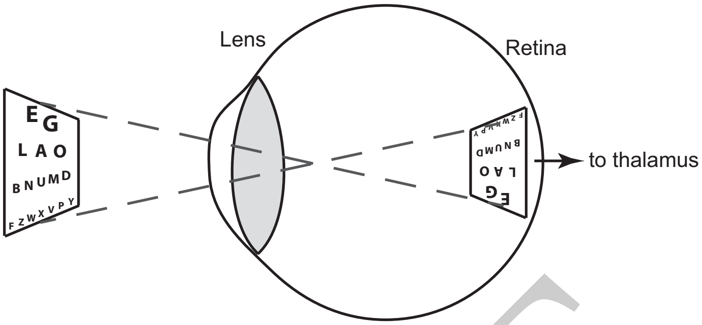


Figure 2.5 The eye lens inverts the image. As in many other types of lenses, the image is inverted when focused on the retina.

light into electrical signals, combined with the precise retinal circuitry, can well be considered one of the great achievements of evolution. The ability to convert light patterns into information and the structure of the eye made Charles Darwin (1809–1882) ponder whether such a feat could be achieved via natural selection. Elegant biochemical and electrophysiological work has characterized the signal transduction cascade responsible for capturing photons and using the photon's energy to trigger a set of chemical reactions that lead to voltage changes in photoreceptor cells.

Light information reaches the eye through a lens. When the light reaches the focal plane, the retina, the image is inverted (upside down and left/right, Figure 2.5). This basic fact of optics sometimes puzzles those who reflect about perception for the first time. Why don't we see everything upside down? This question has also tormented some of the brightest minds ever since the basic principles of optics were discovered. None other than the great Leonardo Da Vinci (1452–1519) erroneously assumed that we do not see upside down because of a second lens in the eye inverting the image again. Moreover, Johannes Kepler (1571–1630), who otherwise played a central role in advancing our thinking about visual perception, clearly described the inversion by the eye and left the problem of perception to be solved by natural philosophers (at the time, a mixture of what we would now call physicists and philosophers). Other philosophers assumed that newborn infants do see objects upside down and that this percept is eventually "corrected" by virtue of aligning visual inputs with the sense of touch. These philosophical ideas are another example of erroneous interpretations based on introspective models without an anchor on real experiments: there is no evidence that the sense of touch is needed to develop a visual system capable of interpreting what is up and down in the world.

We do not see objects upside down because perception constitutes our brain's construction of the outside world based on the pattern of activity from neurons in the retina. Since the day we are born, our brains learn that a specific pattern of activation in the retina is the way things are in the world. The brain does not know about what is right side up; it is all electrical signals. It is even possible to teach the brain to adapt to images with different rules, for example, by wearing glasses that invert the image. It is not easy to adapt to such glasses, and it takes dedication, but people can learn to ride a bicycle wearing glasses containing lenses that shift the world upside down or glasses that shift the image left and right. After adapting to these new rules, taking the glasses off becomes quite confusing, and subjects need to learn again to interpret the visual world without the inversions. Upon taking these nasty glasses off, relearning to adjust to the natural world is much faster than the initial brain training with the reversed world.

The network of neurons in the retina is a particularly beautiful structure that has mesmerized neuroscientists for more than a century. The history of retinal studies is intimately connected to the history of neuroscience and commences with the drawings of the famous Santiago Ramón y Cajal (1852–1934). Santiago Ramón y Cajal, considered to be the father of neuroscience, had a skillful hand for drawing and wanted to become an artist. However, his parents had other plans; Ramón y Cajal ended up following their advice and becoming a medical doctor. After obtaining his medical degree, he studied the techniques to stain neural tissue from the great Camillo Golgi (1843–1926), with whom he would engage in a ferocious scientific dispute about the fundamental structure of brain tissue, and with whom he shared the Nobel Prize in 1906.

The retina soon became a persistent passion for Ramón y Cajal. The retina is located at the back of the eyes; in humans, it has a thickness of approximately 250 μm and encompasses the surface area of about half a sphere of one-inch diameter. The retina is part of the central nervous system: it originates from the same embryonic structures that give rise to the rest of the brain, and it has a blood barrier similar to the one in the rest of the brain.

The schematic diagram of the retina in Figure 2.6 illustrates the stereotypical connectivity composed of three main cellular layers (photoreceptors, bipolar cells, and ganglion cells), interconnected through two additional intermediate layers (horizontal cells and amacrine cells). In vertebrate animals, light has to traverse through all the other cell types to get to the photoreceptors, shown at the top in Figure 2.6. Photoreceptors come in two main varieties: rods and cones. There are about 10^8 rods; these cells are very sensitive to light, and they are specialized for capturing photons under low-light conditions. Night vision depends on rods. Because the cones have different spectral sensitivities that enable interpretation of colors, and because cones are much less sensitive than rods to low illumination, we barely see colors at night. Rods are so sensitive that they can capture and transmit a single photon, which constitutes about 10^{-19} joules of energy in the visible portion of the spectrum. Meticulous experiments suggest that sometimes humans can detect single photons above chance.

In addition to the rods, there are about 10^7 cones specialized for vision under bright-light conditions. Most people have three types of cones: long-wavelength sensitive peaking at ~560 nanometers, medium-wavelength sensitive peaking at ~530 nanometers, and short-wavelength sensitive peaking at ~420 nanometers. Color vision relies on the activity of cones. Some humans show variations of color blindness – in most cases, due to deficiencies or even absence of one of these types of cones; in rare cases, there is an absence of more than one type of cone. Even with only two types of cones, people can still see different hues. For example, if people are missing the short-wavelength cones, they can still distinguish light of 400 nanometers versus 500 nanometers wavelength because of the differential responses triggered in the long- and

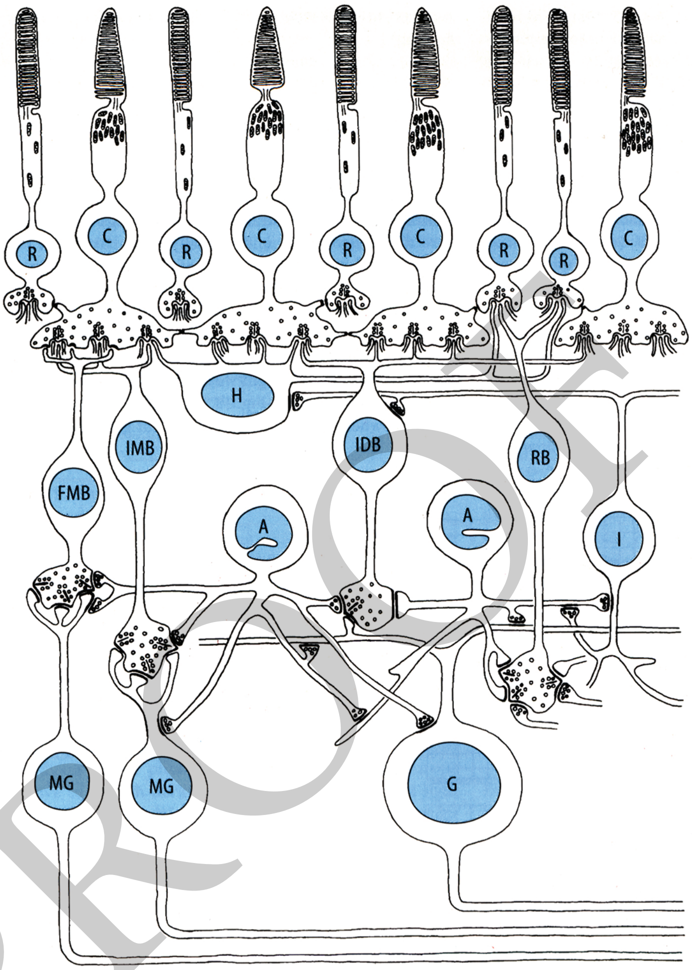


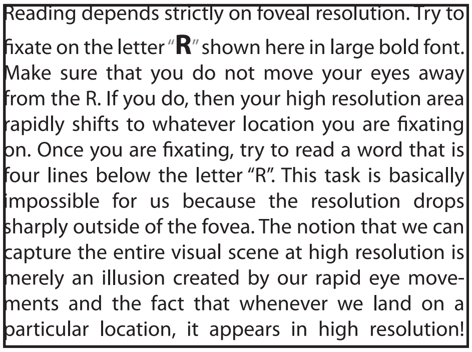
Figure 2.6 Schematic diagram of the cell types and connectivity in the primate retina. In this diagram, light comes from the bottom and goes through all the layers to reach the photoreceptors. R = rod photoreceptors; C = cone photoreceptors; FMB = flat midget bipolar cells; IMB = invaginating midget bipolar cells; H = horizontal cells; IDB invaginating diffuse bipolar cells; RB = rod bipolar cells; I = inner plexiform cell; A = amacrine cells; G = ganglion cells; MG = midget ganglion cells. Reproduced with permission from Dowling 2012

medium-wavelength sensitive cones. Color *blindness* is, therefore, a misnomer and should be reserved only for people who see in grayscale – that is, people who are only sensitive to intensity without any color sensation. A condition known as *achromatopsia* – caused by damage in the brain, not in the eye – can lead to complete color blindness, as related brilliantly by the famous British neurologist and author Oliver Sacks (1933–2015) in one of his books (Section 4.8).

People missing one type of cone have specific confusion points – that is, certain combinations of wavelengths that they cannot distinguish. To be able to demonstrate these colors that they cannot differentiate, it is critical to equalize light intensity. Under natural conditions, colors are often correlated with different intensities, and, therefore, people with cone deficiencies may use those intensity cues to circumvent their reduced resolution in the color spectrum. The *Ishihara* test is a common way of assessing color deficiencies, and there are plenty of such tests available online. Many people are surprised when they take these tests and find out that they cannot distinguish certain color combinations. Color vision deficiency is actually quite common in males (about one in 12!), with a much lower prevalence in women (about one in 200). A politically incorrect joke states that women know hundreds of colors and men only know five. This joke is not entirely wrong for some men (though strictly speaking, even with only two cones, it is possible to distinguish lots of different colors).

Rods and cones are not uniformly distributed throughout the retina. In particular, there is a part of the retina, called the *fovea*, which is specialized for high acuity. This ~300 μm region contains no rods and a high density of cones, with an astonishing 17,500 cones. This high density leads to a fine sampling of the visual field, thereby providing subjects with higher resolution at the point of fixation. For example, our ability to read depends strictly on the fovea: try fixating on the letter "R" on the second line in Figure 2.7. Next, try to read a word that is five words away and two lines below the "R," *without moving your eyes*. Cellular density and the degree of convergence from cones to downstream neurons decreases with eccentricity – that is, with distance from the fovea. In addition, the optics of the eye lens has enhanced contrast modulation transfer at the fovea. Because of the optics of the eye and the nonuniform sampling, we only see in high resolution in the fovea (Figure 2.8B). Therefore, saccadic eye movements bring the center of fixation into sharp focus to obtain detailed information. People with *macular degeneration* show progressively more damage in the foveal area, leading to a deterioration of the quality of high-resolution information, eventually perceiving noise or a blurry version of the image (Figure 2.8C).

Even though locations that are far from the fovea have coarser sampling, we have the illusion of perceiving approximately equal resolution throughout the visual field. Eye movements are partly responsible for this illusion: every time we move our eyes, we



Reading depends strictly on foveal resolution. Try to fixate on the letter "**R**" shown here in large bold font. Make sure that you do not move your eyes away from the R. If you do, then your high resolution area rapidly shifts to whatever location you are fixating on. Once you are fixating, try to read a word that is four lines below the letter "R". This task is basically impossible for us because the resolution drops sharply outside of the fovea. The notion that we can capture the entire visual scene at high resolution is merely an illusion created by our rapid eye movements and the fact that whenever we land on a particular location, it appears in high resolution!

Figure 2.7 We can only read in the foveal region. Fixate on the large bolded R on the second line and try to read words on another line without moving your eyes.

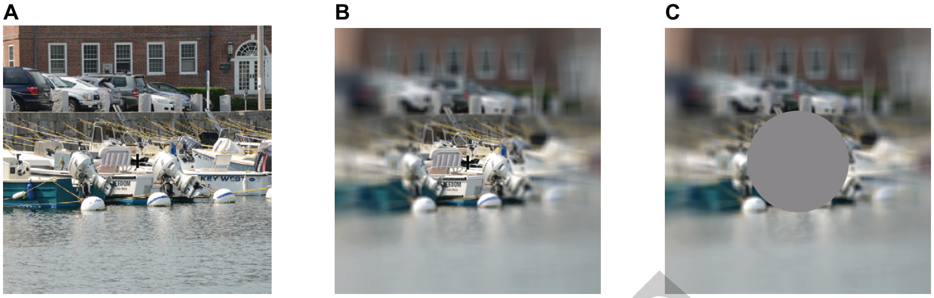


Figure 2.8 Only the area around fixation is seen in high resolution. (**A**) Original photograph. If you were at this place, fixating on the location indicated by the + sign, you would have the illusion that the entire field is full of details. (**B**) However, the image conveyed to the brain by the retina is closer to the one in **B**, with high resolution at the fixation location and increasingly more blurring toward the periphery. Our perception seems to be closer to **A** than to **B**, because we constantly move our eyes, sampling new locations at high resolution. (**C**) People with macular degeneration see noise or a blurry image in the center, in addition to the regular blurriness of the periphery.

fixate on a new location, which appears in high resolution. We naturally assume that the whole visual field has the same resolution. Additionally, there is probably information stored about previous fixations. When we move our eyes to a new location, the old fixation location now appears in the periphery, with lower resolution. However, the low-resolution version could be combined with a version stored in working memory based on the previous high-resolution fixation.

There is a region in the back of each eye that contains no photoreceptors. This region is where the axons of the retinal output cells, the retinal ganglion cells (RGCs), exit the eye. People cannot detect light that is focused on precisely this region, which is thus denominated the *blind spot*. The easiest way to detect the blind spot is to close one eye, fixate on a given distant spot, and slowly move your index finger from the center to the periphery until part of the finger disappears from view (but not in its entirety, which would imply that you moved your finger entirely outside of your visual field). There are many demos online to help detect the blind spot. Legend has it that King Charles II of England was fascinated with the blind spot and used to entertain himself by placing the head of a prisoner in his blind spot to imagine him headless before the actual decapitation.

Under normal circumstances, we are not aware of the blind spot – i.e., we have the subjective feeling that we can see the entire field in front of us (even with one eye closed). Given that we do not normally perceive the blind spot, one may surmise that it is actually rather small. However, you can fit the projection of nine full moons in the sky into the blind spot. How is it possible to be so utterly oblivious to such a large and empty region of the visual field? We are generally not aware of the blind spot because the brain fills in and compensates for the lack of receptors in the blind spot. This filling-in process emphasizes again the notion that our visual percepts are not a literal reflection of reality but rather a reconstruction concocted by our brains. We will return to the notion of vision as a subjective construction when we discuss visual illusions (Section 3.1) and visual consciousness (Chapter 10).

Information from the photoreceptors is conveyed to a second cellular layer consisting of horizontal cells, bipolar cells and amacrine cells, and finally to retinal ganglion cells (RGCs). The human retina contains approximately 6.4 million cones, about 110 million rods, and about one million retinal ganglion cells. Thus, on average, there is a convergence of about 100 photoreceptors to one ganglion cell, but these numbers vary depending on the location in the retina. As noted before, convergence is minimal in the fovea and more extensive in the distant periphery. In the fovea, one cone is upstream of one RGC, and in the periphery, there are about 15 cones per RGC and hundreds of rods per RGC.

Figure 2.6 shows a simplified schematic of the connectivity in the retina from photoreceptors to horizontal and bipolar cells, then onto amacrine cells and ganglion cells. Molecular and anatomical markers have helped define different types of horizontal and bipolar cells and even more types of amacrine cells and ganglion cells, each of which is involved in specific computations to capture different aspects of the incoming images. Furthermore, serial electron microscopy is beginning to elucidate the *retinal connectome* – that is, the precise pattern of synaptic connections in the retina. In the not-too-distant future, it is conceivable that we may have access to a rather complete anatomical map of the retina.

2.6 It Takes Time for Information to Reach the Optic Nerve

At first glance, vision may seem to be instantaneous. We open our eyes, and the world emerges rapidly in all its glory. However, there is no such thing as instantaneous signal propagation. It takes time for the cascade of processes that converts incoming photons into the spiking activity of retinal ganglion cells. The latency of retinal ganglion cell responses to a stimulus flash depends on multiple factors – including the previous history of visual stimulation, the intensity of the stimulus flash, its size, and its color, among others.

The axons from the retinal ganglion cells that convey information to the rest of the brain are collectively known as the optic nerve. On average, it takes 30–50 milliseconds from the onset of a stimulus flash for spikes to emerge from the optic nerve and propagate down to the rest of the brain. This latency is further combined with the computational time required to interpret the information in the brain, to be elaborated upon in Section 5.12. Because of these delays, what we see reflects what transpired in the world in the recent past. The delays are sufficiently short to trick our perception and allow us to get a rapid assessment of what happens in the world.

2.7 Visual Neurons Respond to a Specific Region within the Visual Field

Like most neurons throughout the brain, retinal ganglion cells (RGCs) convey information by emitting action potentials, also known as spikes. Cells before RGCs in the

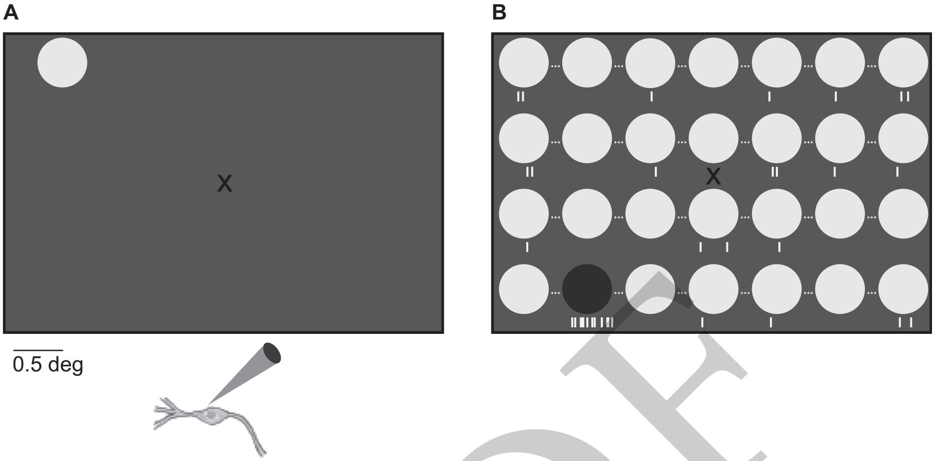


Figure 2.9 Neurons have localized receptive fields. (**A**) A light stimulus (white circle) is flashed in a circumscribed location while recording the activity of a neuron in a fixating animal ("X" denotes the fixation location). (**B**) The procedure is repeated in multiple different locations. The small vertical ticks denote neuronal activity. The location of maximum activity (black circle) denotes the neuron's receptive field. The stimulus size is also changed to map the boundaries of the receptive field. The neuron also shows a low spontaneous rate at other locations.

retina constitute the exception to this rule and communicate using graded voltage signals without emitting spikes. To understand how RGCs represent visual information, we need to examine how different inputs map onto spiking responses. The functional properties of RGCs have been extensively examined by electrophysiological recordings that go back to the prominent work of Haldan Hartline (1903–1983), Horace Barlow (1921–2020), and Stephen Kuffler (1913–1980). RGCs (as well as most neurons in the visual cortex) respond most strongly to a circumscribed region of the visual field called the *receptive field* (Figure 2.9). The receptive field can be mapped by flashing a stimulus at different locations and different sizes to locate the areas that trigger neuronal activation. Neurons tend to also fire spontaneously so that there are small neuronal responses even when the retina is in complete darkness or when the stimulus is very far from the receptive field. In other words, neuronal firing rates are not necessarily zero in the absence of visual stimulation inside the receptive field. It should be emphasized that the location of the receptive field is always specified with respect to the fixation point, not with respect to a fixed location in space. If subjects move their eyes, the location of the receptive field in the environment changes, but the position with respect to the fixation point does not.

These receptive fields tile the entire visual field. Without moving your eyes, any location in the visual field where you can see anything implies that there is an RGC with a receptive field that encompasses that location. The receptive fields of RGCs are topographically organized – that is, nearby RGCs in the retina represent nearby locations in the visual field. This topography is preserved in the projections from RGCs onto the thalamus, and from there onto the cortex as well. The nonuniform distribution of neurons from the fovea to the periphery means that there is a consistent eccentricity dependence in the size of the receptive fields. In the fovea, there is a one-to-one mapping between cones and RGCs. Receptive fields near the fovea are smallest, and receptive field sizes grow approximately linearly with eccentricity. The large receptive fields in the periphery are one of the main reasons why we have less resolution outside of the fovea.

The RGC schematically illustrated in Figure 2.9 increases its firing rate with increased luminance inside the receptive field. This type of cell is referred to as an *on-center* cell. There are also other RGCs, *off-center* cells, which increase their firing rate when there is a decrease in luminance in the center of their receptive fields.

RGC activity does not directly reflect the pattern of photons arriving at the retina due to the distortions introduced by the eye lens, due to the temporal delays and intermediate processing introduced by the previous cellular layers, and due to the eccentricity-dependent variations in convergence from photoreceptors onto RGCs. However, it is still possible to make an educated guess about incoming visual stimuli by examining RGC responses. We do not have the tools to record the activity of every RGC. Current technologies only allow simultaneously registering the activity of a few hundred RGCs. Even with such a small population, it is possible to reconstruct a rather accurate version of the light patterns reaching the retina.

2.8 The Difference-of-Gaussians Operator Extracts Salient Information and Discards Uniform Surfaces

Even when the center of an on-center cell is bombarded with a high-luminance flash, its response will be modulated by what is outside of the receptive field center. In particular, for most RGCs, a perfectly uniform high-luminance white wall will *not* trigger high activation. Consider the following experiment: a small uniform white circle is shown in the center of the receptive field, and the neuron fires above baseline levels (Figure 2.9). Next, the circle is slightly enlarged, and the neuron shows a higher firing rate. If we keep increasing the size of the circle, at some point, the firing rate reaches its peak value. Making the circle any larger leads to a *reduction* in firing rate; this phenomenon is known as *surround inhibition*. Surround inhibition is observed not only for RGCs; it is also prevalent throughout the entire visual system. On-center neurons are particularly interested in spatial changes – i.e., increased luminance within the receptive field combined with decreased luminance outside the receptive field. The converse is true for off-center neurons.

This form of spatial context-dependent response pattern is known as center-surround receptive fields and is typically modeled as a difference of two Gaussian curves (Figure 2.10). Considering an on-center cell, and assuming that the center of the receptive field is at location $x = 0$, $y = 0$, the neuronal activity in response to illumination at a new position x, y will be driven by an excitatory component proportional to $\left(1/2\pi\sigma_{cen}^2\right)e^{-\left(x^2+y^2\right)/2\sigma_{cen}^2}$, where σ_{cen} reflects the spatial extent of the excitatory driving force (dashed line in Figure 2.10). This excitation is counterbalanced by a surround inhibitory component given by $\left(1/2\pi\sigma_{sur}^2\right)e^{-\left(x^2+y^2\right)/2\sigma_{sur}^2}$, where σ_{sur} reflects the spatial extent of the inhibitory driving force (dotted line in Figure 2.10). The difference-of-Gaussians operator is used to describe the receptive field structure of RGCs:

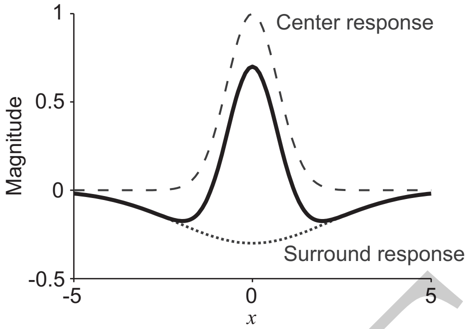


Figure 2.10 Mexican-hat receptive field. The receptive field in retinal ganglion cells is often characterized as a difference between a center response (dashed line) and a broader and weaker surround response (dotted line), resulting in a "Mexican-hat" shape (solid line).

$$D(x,y) = \pm\left(\frac{1}{2\pi\sigma_{cen}^2} e^{-\frac{x^2+y^2}{2\sigma_{cen}^2}} - \frac{B}{2\pi\sigma_{sur}^2} e^{-\frac{x^2+y^2}{2\sigma_{sur}^2}}\right), \qquad (2.3)$$

where the scaling factor $B < 1$ controls the relative strength of excitation and inhibition, where $\sigma_{sur} > \sigma_{cen}$, and where the $\pm$ corresponds to on-center and off-center cells, respectively. The difference between the two terms yields a "Mexican-hat" structure with a peak in the center and an inhibitory dip in the surround. Biology is full of surprises and exceptions. The responses of some RGCs cannot be accounted for by Equation (2.3).

2.9 Visual Neurons Show Transient Responses

In the same way that a large spatially uniform stimulus does not elicit strong activation because neurons are tuned to detect spatial changes, temporal changes are critical too. A constant stimulus generally does not lead to sustained neuronal responses. Some RGCs respond at the onset of the stimulus, others respond at the offset, and others respond at the onset and offset. In all these cases, the responses tend to rapidly adapt when the stimulus remains constant and in the absence of any other external changes (in the absence of eye or head movements). Some neurons maintain a constant response above baseline during the duration of the stimulus after the initial transient. In contrast, the firing rate in other neurons decreases to baseline levels after the initial transient. RGCs are, therefore, sensitive not only to spatial context but also to temporal context.

The incorporation of contextual information allows neurons to efficiently encode spatial changes and temporal changes without spending abundant and energetically expensive spikes to indicate that the stimulus is constant in space or time. The regularities in the structure of the visual stimulus described in Sections 2.1 and 2.2 are thus reflected in the firing properties of RGCs.

Equation (2.3) can be expanded to provide a quantitative description of the *dynamic* responses of retinal ganglion cells when presented with a stimulus that begains at $t = 0$ and stays constant:

$$D(x,y,t) = \pm\left(\frac{D_{cen}(t)}{2\pi\sigma_{cen}^2}e^{-\frac{x^2+y^2}{2\sigma_{cen}^2}} - \frac{BD_{sur}(t)}{2\pi\sigma_{sur}^2}e^{-\frac{x^2+y^2}{2\sigma_{sur}^2}}\right), \tag{2.4}$$

where $\mathrm{D}_{cen}(t) = \alpha_{cen}^2 te^{-\alpha_{cen}t} - \beta_{cen}^2 te^{-\beta_{cen}t}$ describes the dynamics of the center excitatory function and $\mathrm{D}_{sur}(t) = \alpha_{sur}^2 te^{-\alpha_{sur}t} - \beta_{sur}^2 te^{-\beta_{sur}t}$ describes the dynamics of the surround inhibitory function.

Equation (2.4) describes the internal dynamics of an RGC upon presentation of a stimulus that remains constant. In addition to these types of responses, some RGCs are also strongly activated by a stimulus that moves within the receptive field. One such type of cell is the *on–off direction-selective RGC*, which shows enhanced responses when a stimulus within the receptive field is moving in a specific direction. Such direction-selective responses are also modulated by the surrounding context: neurons respond most vigorously when there is a *difference* in the direction of motion between the receptive field and the surround. An entire visual field moving in the same direction constitutes a weak stimulus for this type of neuron. This contextual subtraction helps the neurons distinguish the movement of external objects from self-motion. In addition, locations of depth boundaries also lead to motion discontinuities during self-motion relative to a static scene. Motion-sensitive RGCs tend to have large dendritic arbors and are particularly abundant in the periphery. Because of this, detecting an object in the periphery is easier when it moves, an observation that you can readily test by fixating on any given letter here, extending your hand in the periphery, and comparing your perception of the hand when it is static versus when it is moving.

The conduction velocities of RGCs have been used to separate between *magnocellular cells* (M-type RGCs) and *parvocellular* cells (P-type RGCs, also called midget cells). M-type cells have large dendritic arbors, have fast conduction velocity, respond to low-contrast stimuli, show transient responses, and have little sensitivity for colors. In contrast, P-type cells show small dendritic arbors, have color sensitivity, and tend to exhibit more sustained responses and low conduction velocities.

There continues to be exciting research geared toward elucidating all the different types of functional and structural specializations of RGCs; current estimates suggest that there are at least tens of distinct ganglion cell types, depending on how exactly a "type" is defined. Except for the fovea, different ganglion cell types are approximately distributed throughout so that the same external stimulus features can be captured throughout the visual field.

2.10 On to the Rest of the Brain

The principal destination of the output of retinal ganglion cells is a part of the thalamus called the lateral geniculate nucleus (LGN). The retina also projects to the suprachiasmatic nucleus and the superior colliculus, among many of other regions (anatomical studies have mapped more than 40 brain regions that receive inputs from the retina). The suprachiasmatic nucleus plays a vital role in regulating circadian rhythms, while the superior colliculus constitutes the main visual processing center for many species before the expansion of the cerebral cortex. Primates can recognize objects after lesions to the superior colliculus, but not after lesions to visual cortical regions. Therefore, the key pathway for visual perception involves the one going from RGCs to the LGN to the cortex.

As we will discuss in Sections 5.17 and 6.11, there are massive *back projections* throughout the visual system (Figure 1.5). If area A projects to area B, then in most cases area B also projects back to area A. One of the few exceptions to this rule is the connection from the retina to the LGN. There are no connections from the LGN back to the retina. Therefore, the pathways from photoreceptors to RGCs to LGN can be thought of as mostly feedforward.

The thalamus has often been succinctly called the gateway to the cortex, modulating what type of sensory information reaches the cortex. The receptive fields of LGN cells also display the center-surround structure depicted in Figure 2.10 and can be approximated by Equations (2.3) and (2.4). Thalamic cells are often referred to – in a rather unfair fashion – as *relay cells*, advocating the idea that the thalamus merely copies and pastes the output of RGCs and conveys this output to the cortex.

One obvious distinction between RGC and LGN cells is the pattern of connectivity. While we often think of the LGN predominantly in terms of the input from RGCs, there is a large number of back-projections from diverse cortical areas, predominantly from the primary visual cortex, onto the LGN. Precisely how these feedback connections modulate the response to visual stimuli in the LGN is not well understood.

Like the vast majority of brain structures, there are two copies of the LGN, one in each hemisphere. The right LGN receives input from both eyes, but only from the left hemifield (mostly the part of the visual field to the left of the fixation point) while the converse holds for the left LGN. The right eye receives information from both hemifields and sends right hemifield information to the LGN in the left hemisphere and left hemifield information to the LGN in the right hemisphere.

Six layers can be distinguished in the LGN. Layers 2, 3, and 5 receive *ipsilateral* input (i.e., information from the eye on the same side). Layers 1, 4, and 6 receive *contralateral* input (i.e., information from the eye on the opposite side). A single point in space is therefore represented in six different maps at the level of the LGN. Information from the right and left eyes does not merge in the LGN. Layers 1 and 2 are called magnocellular layers and receive input from M-type RGCs. Layers 3–6 are called parvocellular layers and receive input from P-type RGCs. There are about 1.5 million cells in the human LGN. Thus, the overall density of LGN neurons allocated to different parts of the visual field is comparable to that in RGCs, whereas there is a large expansion in the number of neurons as we move on to the cortex.

Because the LGN, and the thalamus in general, is connected to multiple cortical areas, it sits in a rather unique position to integrate sensory inputs with different forms of processed information throughout the cortex. The description of the LGN as a relay structure is only a major oversimplification, and the picture of the LGN will change dramatically as we understand more about the neuronal circuits and computations in the LGN.

2.11 Digital Cameras versus the Eye

In Chapters 7–9, we will examine computational models of visual processing. By and large, state-of-the-art computer models start with the output of a regular digital camera that has captured a picture and represents it as a two-dimensional matrix of pixels, each one of which is coded in a three-dimensional color world (such as red, green, and blue intensities). However, the sophisticated series of computations by the retina is still not quite matched by even the best digital cameras out there.

The angle of view of a digital camera depends on the focal length of the lens. For a focal length of 17 mm (approximate distance from the optical center of the eye lens to the retina), the field of view is approximately 90 degrees, whereas the field of view for humans spans almost 180 degrees. The resolution of the human eye has been estimated to be on the order of 500 megapixels, still much more than some of the fanciest commercially available digital cameras out there.

Another difference is that digital cameras are approximately uniform in their sensitivity to light. In contrast, the retina allocates more resources than the best current cameras to process conditions with low illumination. If you have ever tried to take pictures at night, you probably have noticed that it is not easy to take digital pictures under low-light conditions. To circumvent these challenges, photographers may use contraptions such as tripods to stabilize the camera and leave the camera shutter open for many seconds, if not minutes or more. In contrast, the eye can convey accurate information and help us navigate in the forest even under starlight only. We would not want to have to wait for many seconds or minutes before we can see anything at night. One of the tricks to achieve this is that the retina can adapt to low-light conditions and change its gain to achieve higher sensitivity. The eye has to work under conditions of strong sunlight all the way to moonless nights, a difference of about nine orders of magnitude in light intensity. This adaptation takes time, as can be appreciated when going from a dark place out into the sunlight or vice versa.

In addition to this adaptation to the average illumination, the light intensity can vary over three log units within a scene. The retina can accommodate this because of adaptation mechanisms spanning different spatial and temporal scales. In contrast, taking digital pictures in a scene with such significant variations in illumination is tricky: either one part of the image is completely dark or another part of the image is completely overexposed.

Digital cameras typically lack many of the sophisticated motion detection and contextual correction mechanisms described in this chapter for RGCs. Images are rarely blurry for us, whereas digital cameras need to implement a lot of additional correction mechanisms to yield crisp images. Another striking difference is the way that we compensate for the spectral composition of the illuminant: we do not see those orangey photos that digital cameras give us. However, the most striking difference between biological vision and digital cameras is the presence of an exquisitely sophisticated computational device to process the output of RGCs, the cortex, which we begin to examine next.

2.12 Summary

- Natural images are special: they are spatially smooth and change slowly in time.
- The efficient coding hypothesis posits that neuronal resources are allocated optimally to represent the statistics of environmental inputs.
- Positions and sizes in the visual field are measured in degrees of visual angle. One degree corresponds approximately to the size of your thumb at arm's length.
- Humans and other primates make frequent eye movements denominated *saccades,* spanning multiple degrees of visual angle, and occurring three to four times a second.
- Two types of photoreceptors convert light into electrical signals for visual perception: *rods* and *cones*. Rods are primarily responsible for night vision and cones for color vision.
- Retinal ganglion cells communicate the output of the retina to the rest of the brain.
- Retinal ganglion cells respond to a localized region of the visual field denominated the *receptive field*.
- The center of focus is projected onto the *fovea*, an area populated by cones, with higher cellular density and smaller receptive field sizes, providing high resolution.
- On-center retinal ganglion cells are excited by light within their receptive field and inhibited by light in the surrounding region. Their responses can be described by a difference-of-Gaussians function.
- Information from retinal ganglion cells is conveyed to the lateral geniculate nucleus in the thalamus.
- As a coarse approximation, the eye can be considered to be a specialized digital camera, though eyes are capable of many sophisticated tricks that current digital cameras cannot perform.
- Perception is a construct, an interpretation made by the brain, inspired by sensory formation, but not a literal reflection of the outside world.

Further Reading

See http://bit.ly/3aeW07Z for more references.

- Barlow, H. (1972). Single units and sensation: a neuron doctrine for perception. *Perception* 1, 371–394.

- Helmstaedter, M.; Briggman, K. L.; Turaga, S. C.; Jain, V.; Seung, H. S.; and Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature* 500, 168–174.
- Kuffler, S. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* 16, 37–68.
- Simoncelli, E.; and Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience* 24, 193–216.
- Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.