

1 Introduction to the World of Vision

Supplementary content at <http://bit.ly/2TqTDt5>

Understanding how the brain works constitutes the greatest scientific challenge of our times, arguably the greatest challenge of all times. We have sent spaceships to peek outside of our solar system, and we study galaxies far away to build theories about the origin of the universe. We have built powerful accelerators to scrutinize the secrets of subatomic particles. We have uncovered the secrets to heredity hidden in the billions of base pairs in DNA. But we still have to figure out how the three pounds of brain tissue inside our skulls work to enable us to do physics, biology, music, literature, and politics.

The conversations and maneuvers of about a hundred billion neurons in our brains are responsible for our ability to interpret sensory information, to navigate, to communicate, to feel and to love, to make decisions and plans for the future, to learn. Understanding how neural circuits give rise to cognitive functions will transform our lives: it will help us alleviate the ubiquitous mental health conditions that afflict hundreds of millions, it will lead to building true artificial intelligence machines that are as smart as or smarter than we are, and it will open the doors to finally understanding who we are.

As a paradigmatic example of brain function, we will focus on one of the most exquisite pieces of neural machinery ever evolved: the visual system. In a small fraction of a second, we can get a glimpse of an image and capture a substantial amount of information. For example, we can take a look at Figure 1.1 and answer an infinite series of questions about it: *who is there, what is there, where is this place, what is the weather like, how many people are there, and what are they doing?* We can even make educated guesses about a potential narrative, including describing *the relationship between people in the picture, what happened before, or what will happen next*. At the heart of these questions is our capacity for visual cognition and intelligent inference based on visual inputs.

Our remarkable ability to interpret complex spatiotemporal input sequences, which we can loosely ascribe to part of “common sense,” does not require us to sit down and solve complex differential equations. In fact, a four-year-old can answer most of the questions outlined before quite accurately, younger kids can answer most of them, and many non-human animal species can also be trained to correctly describe many aspects of a visual scene. Furthermore, it takes only a few hundred milliseconds to deduce such profound information from an image. Even though we have computers that excel at



Figure 1.1 We can visually interpret an image at a glance. Who is there? What is there? Where is it? What are they doing? What will happen next? These are just examples among the infinite number of questions that we can answer after a few hundred milliseconds of exposure to a novel image.

tasks such as solving complex differential equations, computers still fall short of human performance at answering common-sense questions about an image.

1.1 Evolution of the Visual System

Vision is essential for most everyday tasks, including navigation, reading, and socialization. Reading this text involves identifying shape patterns. Walking home involves detecting pedestrians, cars, and routes. Vision is critical to recognize our friends and decipher their emotions. It is, therefore, not much of a strain to conceive that the expansion of the visual cortex has played a significant role in the evolution of mammals in general and primates in particular. It is likely that the evolution of enhanced algorithms for recognizing patterns based on visual inputs yielded an increase in adaptive value through improvements in navigation, discrimination of friend versus foe, differentiation between food and poison, and through the savoir faire of deciphering social interactions. In contrast to tactile and gustatory inputs and, to some extent, even auditory inputs, visual signals bring knowledge from vast and faraway areas. While olfactory signals can also diffuse through long distances, the speed of propagation and information content is lower than that of photons.

The ability of biological organisms to capture light is ancient. For example, many bacteria use light to perform photosynthesis, a precursor to a similar process that

captures energy in plants. What is particularly astounding about vision is the possibility of using light to capture *information* about the world. The selective advantage conveyed by visual processing is so impactful that it has led the zoologist Andrew Parker to propose the so-called light switch theory to explain the rapid expansion in number and diversity of life on Earth.

About five hundred million years ago, during the early Cambrian period, there was an extraordinary outburst in the number of different species. It is also at around the same time where fossil evidence suggests the emergence of the first species with eyes, the trilobites (Figure 1.2). Trilobites are extinct arthropods (distant relatives of insects and spiders) that conquered the world and expanded throughout approximately three hundred million years. The light switch theory posits that the emergence of eyes and the explosion in animal diversity is not a mere coincidence. Several investigators have argued that eyes emerged right before the Cambrian explosion. Eyes enabled some lucky early trilobite, or its great grandfather, to capture information from farther away, detecting the presence of prey or predator, endowing it with a selective advantage over other creatures without eyes, who had to rely on slower and coarser information for survival. Using these new toys, the eyes, an evolutionary arms race commenced between prey and predators to make inferences about the world around them and to hide from those scrutinizing and powerful new sensors. All of a sudden, body shapes, textures, and colors became fascinating, powerful, and dangerous. It seems likely that



Figure 1.2 Fossil record of a trilobite, circa 500 million years ago. Trilobites such as the one shown in this picture had compound eyes, probably not too different from those found in modern invertebrate species like flies. Trilobites proliferated and diversified throughout the world for about 300 million years. By Dwergenpaartje, CC BY-SA 3.0

body shapes and colors began to change to avoid detection through the initial versions of camouflage – in turn, leading to keener and better eyes to be more sensitive to motion and to subtle changes through the ability to better discriminate shapes. Let there be light, and let light be used to convey information.

1.2 The Future of Vision

Fast-forward several hundred million years, the fundamental role of vision in human evolution is hard to underestimate. Well before the advent of language as it is known today, vision played a critical role in communication, interpreting emotions and intentions, and facilitating social interactions. The ability to visually identify patterns in the position of the moon, the sun, and the stars led to understanding and predicting seasonal changes, which eventually gave rise to agriculture, transforming nomadic societies into sedentary ones, begetting the precursors of future towns. Art, symbols, and eventually the development of written language also fundamentally relied on visual pattern-recognition capabilities.

The evolution of the visual system is only poorly understood and remains an interesting topic for further investigation. The future of the visual system will be equally fascinating. While speculating about the biological changes in vision in animals over evolutionary time scales is rather challenging, it is easier to imagine what might be accomplished in the near future over shorter time scales, via machines with suitable cameras and computational algorithms for image processing. We will come back to the future of vision in Chapter 9; as a teaser, let us briefly consider machines that can achieve, and perhaps surpass, human-level capabilities in visual tasks. Such machines may combine high-speed and high-resolution video sensors that convey information to computers implementing sophisticated simulations that approximate the functions of the visual brain in real time.

Machines may soon excel in face-recognition tasks to a level where an ATM will greet you by your name without the need for a password, where you may not need a key to enter your home or car, where your face may become your credit card and your passport. Self-driving vehicles propelled by machine vision algorithms have escaped the science fiction pages and entered our streets. Computers may also be able to analyze images intelligently to search the web by image or video content (as opposed to keywords and text descriptors). Doctors may rely more and more on artificial vision systems to analyze X-rays, MRIs, and other images, to a point where image-based diagnosis becomes the domain of computer science entirely. Future generations may be intrigued by the notion that we once let fallible humans make diagnostic decisions. The classification of distant galaxies, or the discovery of different plant and animal species, might be led by machine vision systems rather than astronomers or biologists.

Adventuring further into the domain of science fiction, one could conceive of brain-machine interfaces that might be implanted in the human brain to augment visual capabilities for people with visual impairment or blind people. While we are at it, why not also use such interfaces to augment visual function in normally sighted people

to endow humans with the capability to see in 360 degrees, to detect infrared or ultraviolet wavelengths, to see through opaque objects such as walls, or even directly witness remote events?

When debates arose about the possibility that computers could one day play competitive chess against humans, most people were skeptical. Simple computers today can beat even sophisticated chess aficionados, and good computers can beat world champions. Recently, computers have also excelled in the ancient and complex game of Go. Despite the obvious fact that most people can recognize objects much better than they can play chess or Go, visual cognition is actually more challenging than these games from a computational perspective. However, we may not be too far from building accurate computational approximations to visual systems, where we will be able to trust computers' eyes as much as, or even more than, our own eyes. Instead of "seeing is believing," the future motto may become "computing is believing."

1.3 Why Is Vision Difficult?

The notion that seeing is computationally more complicated than playing Go may be counterintuitive. After all, a two-year-old child can open her eyes and rapidly recognize and interpret her environment to navigate the room and grab her favorite teddy bear, which may be half-covered behind other toys. She does not know how to play Go. She certainly has not gone through the millions of hours of training via reinforcement learning that neural network machines had to go through to play Go. She has had approximately ten thousand hours of visual experience. These ten thousand hours are mostly unsupervised; there were adults nearby most of the time, but, by and large, those adults were not providing continuous information about object labels or continuous reward and punishment signals (there certainly were labels and rewards, but they probably constituted a small fraction of her visual learning).

Why is it so difficult for computers to perform pattern-recognition tasks that appear to be so simple to us? The primate visual system excels at recognizing patterns even when those patterns change radically from one instantiation to another. Consider the simple line schematics in Figure 1.3. It is straightforward to recognize those handwritten symbols even though, at the pixel level, they show considerable variation within each row. These drawings only have a few traces. The problem is far more complicated with real scenes and objects. Imagine the myriad of possible variations of pictures taken at Piazza San Marco in Venice (Figure 1.1) and how the visual system can interpret them with ease. Any object can cast an *infinite* number of projections onto the eyes. These variations include changes in scale, position, viewpoint, and illumination, among other transformations. In a seemingly effortless fashion, our visual systems can map all of those images onto a particular object.

Identifying specific objects is but one of the important functions that the visual system must solve. The visual system can estimate distances to objects, predict where objects are heading, infer the identity of objects that are heavily occluded or camouflaged, determine which objects are in front of which other objects, and make educated

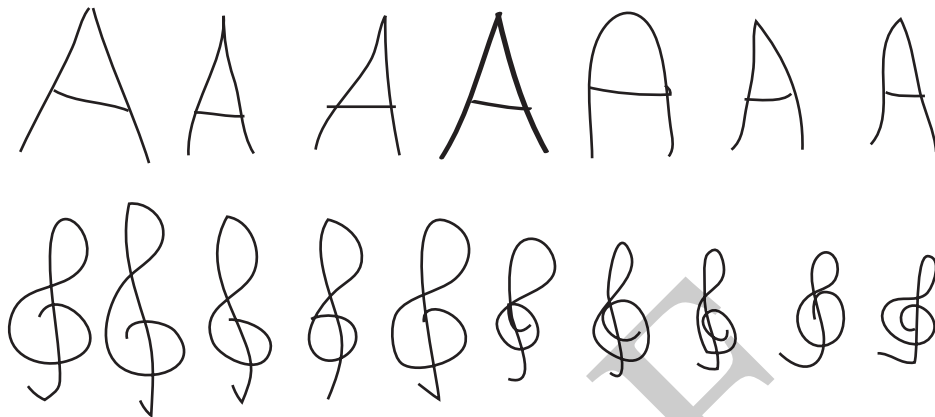


Figure 1.3 Any object can cast an infinite number of projections onto the eyes. Even though we can easily recognize these patterns, there is considerable variability among different renderings of each shape at the pixel level.

guesses as to the composition and weight of objects. The visual system can even infer intentions in the case of living agents. In all of these tasks, vision is an ill-posed problem, in the sense that multiple possible solutions are consistent with a given pattern of inputs onto the eyes.

1.4 Four Key Features of Visual Recognition

In order to explain how the visual system tackles the identification of patterns, we need to account for four key features of visual recognition: *selectivity*, *tolerance*, *speed*, and *capacity*.

Selectivity involves the ability to discriminate among shapes that are very similar at the pixel level. Examples of the exquisite selectivity of the visual system include face identification and reading. In both cases, the visual system can distinguish between inputs that are very close if we compare them side by side at the pixel level. A trivial and useless way of implementing *Selectivity* in a computational algorithm is to memorize all the pixels in the image (Figure 1.4A). Upon encountering the same pixels, the computer would be able to “recognize” the image. The computer would be very selective because it would not respond to any other possible image. The problem with this implementation is that it lacks *tolerance*.

Tolerance refers to the ability to recognize an object despite multiple transformations of the object’s image. For example, we can recognize objects even if they are presented in a different position, scale, viewpoint, contrast, illumination, or color. We can even recognize objects where the image undergoes nonrigid transformations, such as the changes a face goes through upon smiling. A straightforward but useless way of implementing tolerance is to build a model that will output a flat response no matter the input. While the model would show tolerance to image transformations, it would not

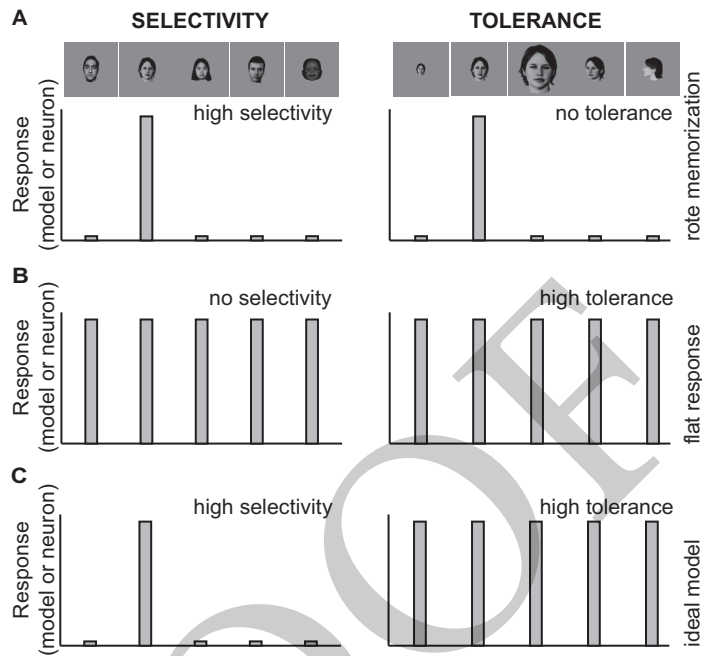


Figure 1.4 A naïve (and not very useful) approach to model visual recognition. Two simple models that are easy to implement, easy to understand, and not very useful. A rote memorization model (A) can have exquisite selectivity but does not generalize. In contrast, a flat response model (B) can generalize but lacks any selectivity. (C) An ideal model should combine selectivity and tolerance.

show any selectivity to different shapes (Figure 1.4B). Combining *selectivity* and *tolerance* (Figure 1.4C) is arguably the key challenge in developing computer vision algorithms for recognition tasks. To consider a real-world example, a self-driving car needs to selectively distinguish pedestrians from many other types of objects, no matter how tall those pedestrians are, what they are wearing, what they are doing, or what they are holding.

Given the combinatorial explosion in the number of images that map onto the same “object,” one could imagine that visual recognition requires many years of learning at school. Of course, this is far from the case. Well before a first grader starts to learn the basics of addition and subtraction (rather trivial problems for computers), she is already quite proficient at visual recognition, a task that she can accomplish in a glimpse. Objects can be readily recognized in a stream of other objects presented at a rate of 100 milliseconds per image. Subjects can make an eye movement to indicate the presence of an object in a two-alternative forced-choice task about 200 milliseconds after showing the visual stimulus. Furthermore, both scalp and invasive recordings from the human brain reveal signals that can discriminate among complex objects as early as ~150 milliseconds after stimulus onset. The *speed* of visual recognition constrains the number of computational steps that any theory of recognition can use to

account for recognition performance. To be sure, vision does not stop at 150 milliseconds. Many aspects of visual cognition emerge over hundreds of milliseconds, and recognition performance under challenging tasks improves with longer presentation times. However, a basic understanding of an image or the main objects within the image can be accomplished in ~150 milliseconds. We denote this regime as *rapid visual recognition*.

One way of making progress toward combining selectivity, tolerance, and speed has been to focus on object-specific or category-specific algorithms. An example of this approach would be the development of algorithms for detecting cars in natural scenes by taking advantage of the idiosyncrasies of cars and the scenes in which they typically appear. Another example would be face recognition. Some of these category- and context-specific heuristics are useful, and the brain may learn to take advantage of them. For example, if most of the image is blue, suggesting that the image background may represent the sky, then the prior probabilities for seeing a car would be low (cars typically do not fly), and the prior probabilities for seeing a bird would be high (birds are often seen against a blue sky). We will discuss the regularities in the visual world and the statistics of natural images in Chapter 2. Despite these correlations, in the more general scenario, the visual recognition machinery is capable of combining selectivity, tolerance, and speed for an enormous range of objects and images. For example, the Chinese language has more than three thousand characters. Estimations of the *capacity* of the human visual recognition system vary substantially across studies. Several studies cite numbers that are considerably more than ten thousand items.

In sum, a theory of visual recognition must be able to account for the high selectivity, tolerance, speed, and capacity of the visual system. Despite the apparent immediacy of seeing, combining these four key features is by no means a simple task.

1.5 The Travels and Adventures of a Photon

The challenge of solving the ill-posed problem of selecting among infinite possible interpretations of a scene in a transformation-tolerant manner within 150 milliseconds of processing seems daunting. How does the brain accomplish this feat? We start by providing a global overview of the transformations of visual information in the brain.

Light arrives at the retina after being reflected by objects in the environment. The patterns of light impinging on our eyes are far from random, and the natural image statistics of those patterns play an important role in the development and evolution of the visual system (Chapter 2). In the retina, light is transduced into an electrical signal by specialized photoreceptor cells. Information is processed in the retina through a cascade of computations before it passes on to a structure called the thalamus and, from there, on to the cortical sheet. The cortex directs the sequence of visual computation steps, converting photons into percepts. Several visual recognition models treat the retina as analogous to the pixel-by-pixel representation in a digital camera. A digital

camera is an oversimplified description of the computational power in the retina, yet it has permeated into the general jargon as introduced by manufacturers who boast of a “retina display” for monitors.

It is not unusual for commercially available monitors these days to display several million pixels. Commercially available digital cameras also boast tens of millions of pixels. The number of pixels in such devices approximates or even surpasses the number of primary sensors in some biological retinas; for example, the human retina contains ~6.4 million so-called cone sensors and ~110 million so-called rod sensors (we will discuss those sensors in Chapter 2). Despite these technological feats, electronic cameras still lag behind biological eyes in essential properties such as luminance adaptation, motion detection, focusing, energy efficiency, and speed.

The output of the retina is conveyed to multiple areas, including the superior colliculus, the suprachiasmatic nucleus, and the thalamus. The superficial layers of the superior colliculus can be thought of as an ancient visual brain. Indeed, for many species that do not have a cortex, the superior colliculus (referred to as optic tectum in these species) is where the main visual elaborations of the input take place. The suprachiasmatic nucleus plays a central role in regulating the circadian rhythm. Humans have an internal daily clock that runs slightly longer than the usual 24-hour day, and light inputs via the suprachiasmatic nucleus help modulate and adjust this cycle.

The main visual pathway carries information from the retina to a part of the thalamus called the lateral geniculate nucleus (LGN). The LGN projects to the primary visual cortex, located in the back of our brains. Without the primary visual cortex, humans are mostly blind, highlighting the critical importance of the pathway conveying information into the cortex for most visual functions. Investigators refer to the processing steps in the retina, LGN, and primary visual cortex as “early vision” (Chapter 5). The primary visual cortex is only the first stage in the processing of visual information in the cortex. Researchers have discovered tens of areas responsible for different aspects of vision (the actual number is still a matter of debate and depends on what is meant by “area”). An influential way of depicting these multiple areas and their interconnections is the diagram proposed by Felleman and Van Essen, shown in Figure 1.5. To the untrained eye, this diagram appears to depict a bewildering complexity, not unlike the circuit diagrams typically employed by electrical engineers. We will delve into this diagram in more detail in Chapters 5 and 6 and discuss the areas and connections that play a crucial role in visual cognition.

Despite the apparent complexity of the neural circuitry in Figure 1.5, this scheme is an oversimplification of the actual wiring diagram. First, each of the boxes in this diagram contains millions of neurons. There are many different types of neurons. The arrangement of neurons within each box can be described in terms of six main layers of the cortex (some of which have different sublayers) and the topographical arrangement of neurons within and across layers. Second, we are still far from characterizing *all* the connections in the visual system. One of the exciting advances of the last decade is the development of techniques to scrutinize the connectivity of neural circuitry at high resolution and in a high-throughput manner.

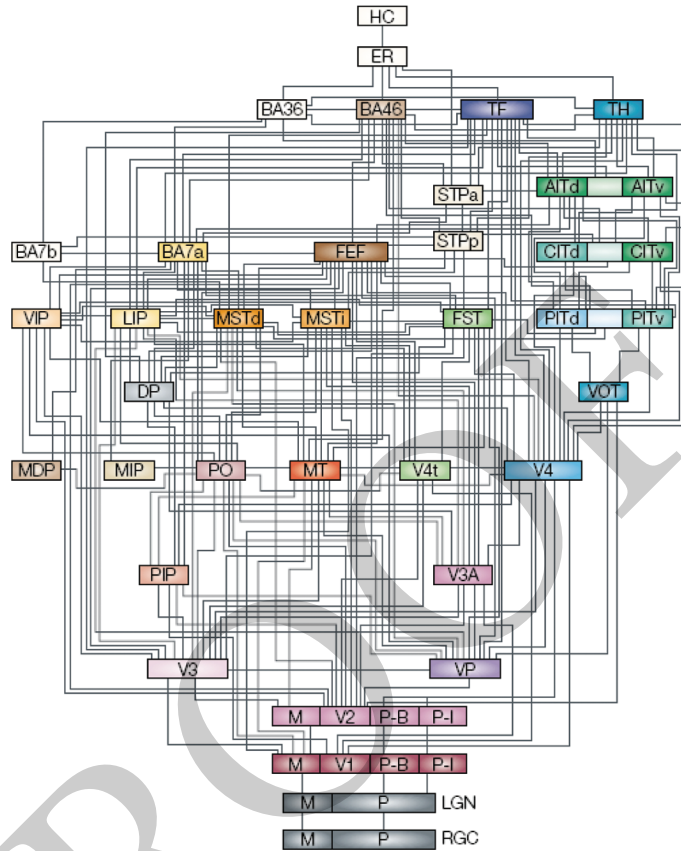


Figure 1.5 The adventures of a photon. Schematic diagram of the connectivity in the primate visual system. Adapted from Felleman and Van Essen 1991

For a small animal like a one-millimeter worm with the fancy name of *Caenorhabditis elegans*, we have known for a few decades the detailed connectivity pattern of each one of its 302 neurons, thanks to the work of Sydney Brenner (1927–2019). However, the cortex is an entirely different beast, with a neuronal density of tens of thousands of neurons per square millimeter. Heroic efforts in the burgeoning field of “connectomics” are now providing the first glimpses of which neurons are friends with which other neurons in the cortex. Major surprises in neuroanatomy will likely come from the use of novel tools that take advantage of the high specificity of molecular biology.

Finally, even if we did know the connectivity of every single neuron in the visual cortex, this knowledge would not immediately reveal the computational functions (but knowing the connectivity would still be immensely helpful). In contrast to electrical circuits where we understand each element and the overall function can be appreciated by careful inspection of the wiring diagram, many neurobiological factors make the map from structure to function a nontrivial one.

1.6 Tampering with the Visual System

One way of finding out how something works is by taking it apart, removing different parts, and reevaluating its function. For example, if we remove the speakers from a car, the car will still function pretty well, but we will not be able to listen to music. If we take out the battery, the car will not start. Removing parts is an important way of studying the visual system as well. For this purpose, investigators typically consider the behavioral deficits that are apparent when parts of the brain are lesioned through studies in nonhuman animals.

In addition to the work in animals, there are various unfortunate circumstances where humans suffer from brain lesions that can also provide insightful clues as to the function of different parts of the visual pathway (as well as other aspects of cognition). Indeed, the fundamental role of the primary visual cortex in vision was discovered through the study of lesions. Ascending through the visual system beyond the primary visual cortex, lesions may yield specific behavioral deficits. For example, subjects who suffer from a rare but well known condition called *prosopagnosia* typically show a significant impairment in the ability to recognize faces (Chapter 4).

One of the challenges in interpreting the consequences of lesions in the human brain is that these lesions often encompass large brain areas and are not restricted to neuroanatomically- or neurophysiologically-defined loci. Several more controlled studies have been performed in animal models – including rodents, cats, and monkeys – to examine the behavioral deficits that arise after lesioning specific parts of the visual cortex. Are the lesion effects specific to one sensory modality, or are they multimodal? How selective are the visual impairments? Can learning effects be dissociated from representation effects? What is the neuroanatomical code? We will come back to these questions in Chapter 4.

Another important path to study brain function is to examine the consequences of externally activating specific brain circuits. One of the prominent ways to do so is by injecting currents via electrical stimulation. Coarse methods of electrically stimulating parts of the cortex often disrupt processing and mimic the effects of a circumscribed lesion. One advantage of electrical stimulation is that the effects can be rapidly reversed, and it is possible, therefore, to study the same animal performing the same task under the influence of electrical stimulation in a specific circuit or not. Intriguingly, in some cases, more refined forms of electrical stimulation can lead not to disrupted processing but instead to enhanced processing of specific types of information. For example, there is a part of the brain referred to as the MT (middle temporal cortex), which receives inputs from the primary visual cortex and is located near the center of the diagram in Figure 1.5. Neurons in this area play an important role in the ability to discriminate the direction of moving objects. Injecting localized electrical currents into area MT in macaque monkeys can bias the animal's perception of how things are moving in their visual world. In other words, it is possible to directly create visual motion thoughts by tickling subpopulations of neurons in area MT (Chapter 4). Combined with careful behavioral measurements, electrical stimulation can provide a glimpse at how influencing activity in a given cluster of neurons can affect perception.

There is also a long history of electrical stimulation studies in humans in subjects with epilepsy. Neurosurgeons need to decide on the possibility of resecting the epileptogenic tissue to treat seizures. Before the resection procedure, neurosurgeons use electrical stimulation to examine the function of the tissue that may undergo resection. The famous American-Canadian neurosurgeon Wilder Penfield (1891–1976) was among the pioneers in using this technique to map brain function. One of his famous discoveries is the “homunculus” map of the sensorimotor world: there is a topographical arrangement of regions where electrical stimulation leads to subjects moving or reporting tactile sensations in the toes, legs, fingers, torso, tongue, and face. Similarly, subjects report seeing localized flashes of light upon electrically stimulating the primary visual cortex.

How specific are the effects of electrical stimulation? Under what conditions is neuronal firing causally related to perception? How many neurons and what types of neurons are activated during electrical stimulation? How do stimulation effects depend on the timing, duration, and intensity of electrical stimulation? We will come back to these questions in Chapter 4.

1.7 Functions of Circuits in the Visual Cortex

The gold standard to examine function in brain circuits is to implant a microelectrode (or multiple microelectrodes) into the area of interest (Figure 1.6). A microelectrode is a thin piece of metal, typically with a diameter of about 50 μm , that can record voltage changes, often in the extracellular milieu. This technique was introduced by Edgar Adrian (1889–1977) in the early 1920s to examine the activity of single nerve fibers. These recordings required clever use of the electronics available at the time to be able to amplify the small voltage differences that characterize electrical communication within neurons. These extracellular recordings (as opposed to the much more challenging intracellular recordings) allow investigators to monitor the activity of one or a few neurons in the near vicinity of the electrode ($\sim 200 \mu\text{m}$) at neuronal resolution and sub-millisecond temporal resolution.

Many noninvasive techniques aim to examine what happens in the brain only in a very indirect fashion by measuring signals that have a weak correlation with the aggregate activity of millions of different cells. These techniques probably include an indirect assessment of neuronal activity but also of the myriad of other cells present in the brain. To make matters even worse, some noninvasive techniques average activity over many seconds, several thousand times slower than the actual interactions taking place in the brain. As an analogy, imagine a sociologist interested in what people in Paris think about climate change; she can interview many individuals, which is laborious but quite precise (equivalent to invasive single neuron recordings), or else she can average the total amount of sound produced in the whole city over an entire week, which is much easier but not very informative (equivalent to noninvasive measurements).

Recording the activity of neurons has defined what types of visual stimuli are most exciting in different brain areas. One of the earliest discoveries was the receptive field of

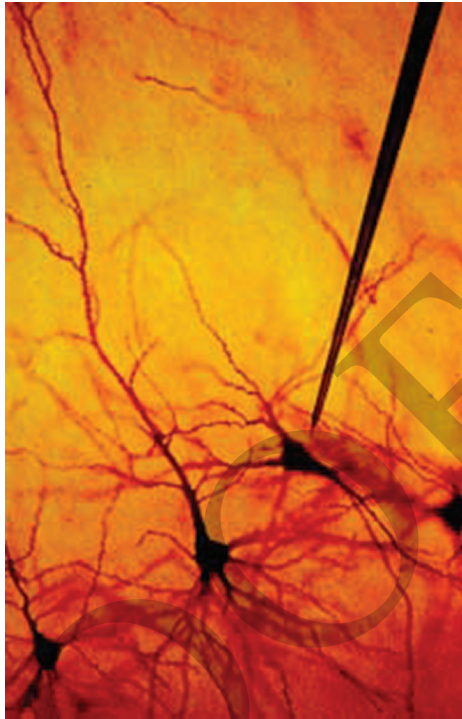


Figure 1.6 Listening to the activity of individual neurons with a microelectrode. Illustration of electrical recordings from microwire electrodes. Reproduced from Hubel 1995

neurons in the retina, LGN, and primary visual cortex. The receptive field is defined as the area within the visual field where a neuronal response can be elicited by visual stimulation (Figure 2.9, Chapter 2). Visual neurons are picky: they do not respond to changes in illumination at any part of the visual field. Each neuron is in charge of representing a circumscribed region of the visual space. Together, all the neurons in a given brain area form a map of the entire visual field – that is, a map of the accessible part of the visual field (e.g., humans do not have visual access to what is happening behind them). The size of these receptive fields typically increases from the retina all the way to those areas like the inferior temporal cortex situated near the top of the diagram in Figure 1.5.

Spatial and temporal changes in illumination within the receptive field of a neuron are necessary to activate visual neurons. However, not all light patterns are equal. Neurons are particularly excited in response to some visual stimuli, and they are oblivious to other stimuli. In a classical neurophysiology experiment, David Hubel (1926–2013) and Torsten Wiesel inserted a microelectrode to isolate single neuron responses in the primary visual cortex of a cat. After presenting different visual stimuli, they discovered that the neuron fired vigorously when a bar of a specific orientation was presented within the neuron's receptive field. The response was weaker when the bar showed a

different orientation. This orientation preference constitutes a hallmark of a large fraction of the neurons in the primary visual cortex (Chapter 5).

Hubel and Wiesel's discovery inspired generations of visual scientists to insert electrodes throughout the visual cortex to study the stimulus preferences in different brain areas. Recording from other parts of the visual cortex, investigators have characterized neurons that show enhanced responses to stimuli moving in specific directions, neurons that prefer complex shapes such as fractal patterns or faces, and neurons that are particularly sensitive to color contrast.

How does selectivity to complex shapes arise, and what are the computational transformations that can convert the simpler receptive field structure at the level of the retina into more complex shapes? How robust are the visual responses in the visual cortex to stimulus transformations such as the ones shown in Figure 1.3? How fast do neurons along the visual cortex respond to new stimuli? What is the neural code – that is, what aspects of neuronal responses better reflect the input stimuli? What are the biological circuits and mechanisms to combine selectivity and invariance? Chapters 5 and 6 delve into the examination of the neurophysiological responses in the visual cortex.

There is much more to vision than filtering and processing images for recognition. Visual processing is particularly relevant because it interfaces with cognition; it connects the outside world with memories, current goals, and internal models of the world. A full interpretation of an image such as Figure 1.1 and the ability to answer an infinite number of questions on the image rely on the bridge between vision and cognition, which we will discuss in Chapter 6.

1.8 Toward the Neural Correlates of Visual Consciousness

The complex cascade of interconnected processes along the visual system must give rise to our rich and subjective perception of the objects and scenes around us. We do not quite know how to directly assess subjective perception from the outside. How do we know that what one person calls red is the same as someone else's perception of red? Some time ago, there were wild discussions in the media about the color of a dress; the photograph became viral, so much so, that it is now known as *the dress* (Figure 1.7). Some people swear that the dress is blue and black. To me, this is as mysterious as if they told me that those people have thirty fingers in their right hand. Why would any honest human being try to convince me that this evidently white and gold dress is actually black and blue? And yet, some people see the dress as white and gold, and others perceive it as distinctly black and blue.

Perception is in the eye of the beholder. To be more precise, perception is in the *brain* of the beholder. If we only worked at the perceptual level without communicating, we would have never figured out that people can see the same dress in such drastically different ways. To indirectly access subjective perception, we need to study behavior. The dress emphasizes that we should not let our introspection guide the scientific agenda. Our intuitions are fallible, as we will discuss again and again.



Figure 1.7 The viral photograph of the dress.

A whole field with the charming name of *psychophysics* deals with careful quantification of behavior as a way of assessing perception (Chapter 3). We will examine where, when, and how rapidly subjects perceive different shapes to construct their own subjective interpretation of the world surrounding them. We will also discuss why brains can be easily deceived by visual illusions. Behavioral measurements will constitute the central constraint toward building a theory of visual processing.

Visual perception is certainly not in the toes and not even in the heart, as some of our ancestors believed. Most scientists would agree that subjective feelings and percepts emerge from the activity of neuronal circuits in the brain. Much less agreement can be reached as to the mechanisms responsible for subjective sensations. The “where,” “when,” and particularly “how” of the so-called neuronal correlates of consciousness constitute an area of active research and passionate debates. Historically, many neuroscientists avoided research in the field of consciousness as a topic too convoluted or too far removed from what we understood to be worth a serious investment of time and effort. In recent years, however, the tide has begun to change. While still very far from a solution, systematic and rigorous approaches guided by neuroscience may one day unveil the answer to one of the greatest mysteries of our times – namely, the physical basis for conscious perception.

Due to several practical reasons, the underpinnings of subjective perception have been mainly (but not exclusively) studied in the domain of vision. There have been heroic efforts to study the neuronal correlates of visual perception using animal models. A prevalent experimental paradigm involves dissociating the visual input from perception. For example, in multistable percepts (e.g., Figure 1.8), the same input can lead to

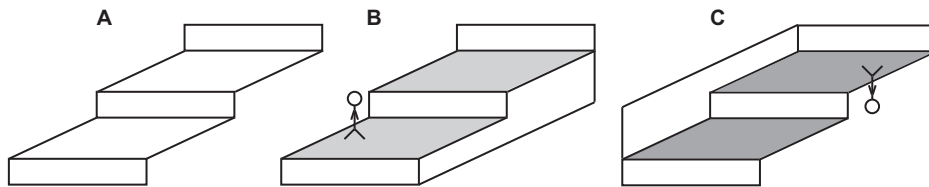


Figure 1.8 A bistable percept. (A) The image can be interpreted in two different ways. (B) In one version, the person is climbing up the stairs. (C) The other version involves an upside-down world.

two distinct percepts. Under these conditions, investigators ask which neuronal events correlate with the alternating subjective percepts.

It has become clear that the firing of neurons in many parts of the brain is *not* correlated with perception. In an arguably trivial example, activity in the retina is essential for seeing, but the perceptual experience does not arise until several synapses later when activity reaches higher stages within the visual cortex (Chapter 10). Neurophysiological, neuroanatomical, and theoretical considerations suggest that subjective perception correlates with activity occurring after the primary visual cortex. Similarly, investigators have suggested an upper bound in terms of where in the visual hierarchy the circuits involved in subjective perception could be. Although lesions restricted to the hippocampus and frontal cortex (thought to underlie memory and associations) yield severe cognitive impairments, these lesions leave visual perception largely intact. Thus, neurophysiology and lesion studies constrain the neural circuits involved in subjective visual perception to the multiple stages involved in processing visual information along the ventral cortex. Ascending through the ventral visual cortex, several neurophysiological studies show that there is an increase in the degree of correlation between neuronal activity and visual percepts.

How can visual consciousness be studied using scientific methods? Which brain areas, circuits, and mechanisms could be responsible for visual consciousness? What are the functions of visual consciousness? Which animals show consciousness? Can machines be conscious? Chapter 10 will provide initial glimpses into what is known (and what is not known) about these fascinating questions.

1.9 Toward a Theory of Visual Cognition

Richard Feynman (1918–1988), a Nobel-winning physicist from Caltech, famously claimed that understanding a device means that we should be able to build it. We aim to develop a theory of vision that can explain how humans and other animals perceive and interpret the world around them. In one of the seminal works on vision, David Marr (1945–1980) defined three levels of understanding, which we can loosely map onto (1) what is the function of the visual system?, (2) how does the visual system behave under different inputs and circumstances?, and (3) how does biological hardware instantiate those functions and behaviors?

A successful theory of vision should be amenable to computational implementation, in which case we can directly compare the output of the computational model against behavioral performance measures and neuronal recordings. A complete theory would include information from lesion studies, neurophysiological recordings, psychophysics, and electrical stimulation studies. Chapters 7 and 8 introduce state-of-the-art approaches to building computational models and theories of visual recognition.

In the absence of a complete understanding of the wiring circuitry, and with only sparse knowledge about neurophysiological responses, it is important to ponder whether it is worth even thinking about theoretical efforts. Not only is it useful to do so, but it is actually *essential* to develop theories and instantiate them through computational models to push the field forward. Computational models can integrate existing data across different laboratories, techniques, and experimental conditions, and help reconcile apparently disparate observations. Models can formalize knowledge and assumptions and provide a quantitative, systematic, and rigorous path toward examining computations in the visual cortex. A good model should be inspired by the empirical findings and should, in turn, produce nontrivial and experimentally testable predictions. These predictions can be empirically evaluated to validate, refute, or expand the models. Refuting models is not a bad thing. Showing that a model is wrong constitutes progress and helps us build better models.

How do we build and test computational models? How should we deal with the sparseness in knowledge and the vast number of parameters often required in models? What are the approximations and abstractions that can be made? If there is too much simplification, we may miss crucial aspects of the problem. With too little simplification, we may spend decades bogged down by nonessential details.

As a simple analogy, consider physicists in the pre-Newton era discussing how to characterize the motion of an object when a force is applied. In principle, one of these scientists may think of many variables that might affect the object's motion – including the object's shape, its temperature, the time of the day, the object's material, the surface where it stands, and the exact position where force is applied. We should perhaps be thankful for the lack of computers in the time of Newton: there was no possibility of running complex machine learning simulations that included all these nonessential variables to understand the beauty of the linear relationship between force and acceleration. At the other extreme, oversimplification (e.g., ignoring the object's mass in this case) would render the model useless. A central goal in computational neuroscience is to achieve the right level of abstraction for each problem, the Goldilocks resolution that is neither unnecessarily detailed nor too simplified. Albert Einstein (1879–1955) referred to models that are as simple as possible, but not any simpler.

A particularly exciting practical corollary of building theories of vision is the possibility of teaching computers how to see (Chapters 8 and 9). We continuously use vision to solve a wide variety of everyday problems. If we can teach some of the tricks of the vision trade to computers, then machines can help us solve those tasks, and they can probably solve many of those tasks faster and better than we can. The last decade has witnessed a spectacular explosion in the availability of computer vision algorithms to solve many pattern-recognition tasks. From a phone that can recognize

faces to computers that can help doctors diagnose X-ray images, to cars that can detect pedestrians, to the classification of images of plants or galaxies, the list of exciting applications continues to proliferate.

Chapter 9 will provide an overview of the state-of-the-art of computer vision approaches to solve different problems in vision. Humans still outperform computers in many visual tasks, but the gap between humans and machines is closing rapidly. We trust machines to compute the square root of seven with as many decimals as we want, but we do not have yet the same level of rigor and efficacy in automatic pattern recognition. Many real-world applications may not require that type of precision, though. After all, humans make visual mistakes too. We may be content with an algorithm that makes fewer mistakes than humans for the same task. For instance, when automatically identifying faces in photographs, correctly labeling 99 percent of the faces might be pretty good. Blind people may be very excited to use devices to recognize where they are heading toward, even if their mobile device can only recognize a fraction of the buildings in a given location.

Alan Turing (1912–1954), a famous British mathematician who helped decipher the codes used by the Nazis to communicate, and who is considered to be one of the founding fathers of computer science, proposed a simple test to assess how smart a machine is. In the context of vision, imagine that we have two rooms with their doors closed. In one of the rooms, there is a human; in the other room, there is a machine that we want to test. We can pass any picture into the room, and we can ask any questions about the picture. The machine or the human return the answers in a typewritten piece of paper so that we cannot identify its voice or handwriting, and there is no other trick. Based on the questions and answers, we need to decide which room has the machine and which one has the human. If, for any picture and any question about the picture, we cannot identify which answers come from a machine and which answers come from the human, we say that the machine has passed the Turing test for vision.

It is tantalizing, exciting, and perhaps also a little bit scary, to think that well within our lifetimes, we may be able to build computers that pass at least some restricted forms of the Turing test for vision. Andrew Parker's theory proposed that animal life as we know it started with the light switch caused by the first eyes on Earth. We might be close to another momentous transformation, the machine visual switch. It is quite likely that life will change radically when machines can see the world the way we do. Perhaps a second Cambrian explosion is on the horizon – an explosion that might lead to the rapid appearance of new hybrid species with machine-augmented vision, where we may trust machine vision more than we trust our own eyes, and where machines can lead the way to discovery in the same way that our visual sense has guided us over the last millennia.

1.10 Summary

- The light switch theory posits that the appearance of eyes during the Cambrian explosion gave rise to the rapid growth in the number and diversity of animal species.

- A theory of visual recognition must account for four fundamental properties: selectivity, tolerance, speed, and large capacity.
- Brain lesions and electrical stimulation provide a window to causally intervene with vision and thus begin to uncover the functional architecture responsible for visual processing.
- Scrutinizing the activity of individual neurons in the visual system opens the door to elucidate the neural computations responsible for transforming pixels into percepts.
- Perception is in the brain of the beholder. Vision is a subjective construct.
- The search for the mechanisms of consciousness requires identifying neural correlates of subjective perception.
- Inspired and constrained by neurophysiological function, neuroanatomical circuits, and lesion studies, we can train computers to see and interpret the world the way humans do.

Further Reading

See <http://bit.ly/2TqTdt5> for more references.

- Hubel, D. (1979). The brain. *Scientific American* 241: 45–53.
- Marr, D. (1982). *Vision*. San Francisco: Freeman Publishers.
- Parker, A. (2004). *In the blink of an eye: how vision sparked the big bang of evolution*. New York: Basic Books.
- Poggio, T.; and Anselmi, F. (2016). *Visual cortex and deep networks*. Cambridge, MA: MIT Press.
- Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT Press.

- Ascending through the visual hierarchy, neurons show increased receptive field sizes, more complex tuning preferences, and longer latencies.
- Neurons in area V2 respond to illusory contours.
- Spatial context, temporal context, and task demands like attention can modulate neuronal responses along ventral visual cortex.

Further Reading

See <http://bit.ly/2TpAg3w> for more references.

- Carandini, M.; Demb, J. B.; Mante, V.; Tolhurst, D. J.; Dan, Y., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience* 25: 10577–10597.
- Hubel, D.H.; and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* 195: 215–243.
- Kremkow, J.; Jin, J.; Wang, Y.; and Alonso, J. M. (2016). Principles underlying sensory map topography in primary visual cortex. *Nature* 533: 52–57.
- Markov, N.T.; Ercsey-Ravasz, M. M.; Ribeiro Gomes, A. R.; Lamy, C.; Magrou, L., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex* 24: 17–36.
- Schmoleksy, M.; Wang, Y.; Hanes, D.; Thompson, K.; Leutgeb, S.; et al. (1998). Signal timing across the macaque visual system. *Journal of Neurophysiology* 79: 3272–3278.