Lecture 1: Introduction to visual recognition. Why is vision difficult?

Visual object recognition is essential for most everyday tasks. It is therefore not much of a strain to conceive that the expansion of visual cortex has played a significant role in the evolution of mammals and primates in particular. The evolution of enhanced algorithms for recognizing patterns based on visual input is likely to have yielded a significant increase in adaptive value through improvement in navigation, recognition of danger and food as well as social interactions. In contrast to tactile and to some extent even auditory inputs, visual signals provide information from far away and larger areas. While olfactory signals can also propagate long distances, the speed of propagation is significantly lower.

The primate visual system excels at recognizing patterns. Even when those patterns change from one instantiation to another. Consider the simple line schemas in Figure 1.1. It straightforward to recognize those handwritten symbols in spite of the fact that at the pixel level, they show considerable variation within each row. These drawings have only a few traces. The problem is far more complicated with real scenes and objects. Imagine the enormous variation that the visual system has be able to cope with to recognize a tiger camouflaged in the dense jungle.

Figure 1.1: The same pattern can look quite differently...

The history and evolution of the visual system is only poorly understood and an interesting topic for further investigation. The future of the visual system is arguably equally fascinating. It is easier to speculate on the technological advances that will become feasible as we understand more about the neural
circuitry involved in visual recognition. One may imagine that in the not-too-distant future, we may be able to build high-speed high-resolution video sensors that convey information to computers implementing sophisticated simulations of the visual cortex in real time. So-called robot vision applications may reach (or even surpass) human performance level in recognition tasks. Computers may excel in face recognition tasks to a level where an ATM machine will greet you by your name without the need of a password. Computers may also be able to analyze images intelligently to be able to search the web by image content (as opposed to image name). Doctors may rely more and more on artificial vision systems to screen and analyze clinical images. Cars may be equipped with automatic systems to avoid collision with other cars or to recognize pedestrians. Robots may even be able to navigate a complex cluttered terrain.

When debates arose about the possibility that computers could one day play competitive chess against humans, most people were skeptic. Yet, computers today can surpass even sophisticated chess aficionados. In spite of the obvious fact that most people can recognize objects much better than they can play chess, I will argue that Visual Object Recognition is actually more difficult than chess from a computational perspective. However, we may not be too far from accurate approximations where we will be able to trust “computers’ eyes” as much as we trust ours.

1.1 **Four key features of visual object recognition**

To be able to explain how the visual system tackles the identification of complex patterns, we need to account for four key aspects of visual recognition: selectivity, robustness, 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 primate 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. One trivial and useless way of implementing *Selectivity* in a computational algorithm is to memorize all the pixels in the image. Upon encountering the exact same pixels, the computer will be able to “recognize” the image and the computer will be very selective because it will not respond to any other possible image. The problem with this implementation is that it lacks *Robustness*.

*Robustness* refers to the ability of recognizing an object in spite of 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, colors, etc. We can even recognize objects where the image undergoes non-rigid transformations such as the one a face goes through upon smiling. Combining *Selectivity* and *Robustness* is arguably the key challenge in developing computer vision algorithms.
Given the combinatorial explosion of the number of images that map onto the same object, one would imagine that visual recognition is a very hard task that requires many years of learning at school. Of course, this is far from the case. By the time a first grader it starting to learn the basics of addition and subtraction, he is already quite proficient at visual recognition. In spite of the number of possible images cast by a given object, recognizing objects is very fast. In a stream of objects presented at a rate of 10 Hz (100 milliseconds per image), each image can be readily recognized (Potter and Levy, 1969). Furthermore, both scalp as well as invasive recordings from the human brain reveal signals that can discriminate among complex objects as early as ~150 milliseconds after stimulus onset (Liu et al., 2009; Thorpe et al., 1996). 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 ms. Many important visual signals arise or develop well after 150 ms. Moreover, recognition performance does improve with longer presentation times (e.g. (Serre et al., 2007)). However, a basic understanding of an image or the main objects within the image can be accomplished in ~150 ms. We denote this regime as “rapid visual recognition”.

One way of making progress towards combining selectivity, robustness and speed has been to focus on object-specific or category-specific approaches. 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. Some of these specific heuristics may be extremely useful and the brain may learn to take advantage of them (e.g. if large parts of the image are blue, suggesting that the image background may represent the sky, then the priors for seeing a car would be low). However, in the more general scenario, the visual recognition machinery is capable of combining selectivity, robustness and speed for an enormous range of objects and images. For example, the Chinese language has over 2,000 characters. Estimations of the capacity of the human visual recognition system vary substantially across studies. Several studies cite numbers that are well over 10,000 items (e.g. (Biederman, 1987; Shepard, 1987; Standing, 1973)).

In sum, a theory of visual recognition needs to be able to account for the high selectivity, robustness, speed and capacity of the primate visual system. In spite of the apparent simplicity of “seeing”, combining these four key features is by no means a simple task.

1.2 The travels of a photon

We start by providing a global overview of the transformations of light information to the cortical signals that support visual recognition. For more information, see (Felleman and Van Essen, 1991; Maunsell, 1995; Wandell, 1995). Light arrives at the retina after reflection by objects. 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 is submitted to cortex. Several object recognition models treat the retina as analogous to the pixel-by-pixel representation in a digital camera. This is highly inaccurate. The retina is capable of performing multiple and complex computations on the input image. The output of the retina is conveyed to multiple areas including the superior colliculus and the suprachiasmatic nucleus. The pathway that carries information to cortex goes from the retina to a part of the thalamus called the lateral geniculate nucleus (LGN). The LGN projects to primary visual cortex, located in the back of our brains. Primary visual cortex is often referred to as V1. The fundamental role of primary visual cortex in visual processing and some of the basic properties of V1 were discovered through the study of the effects of bullet wounds during the First World War. Our Lecture 2 will discuss the processing of information in the retina, LGN and V1, coarsely, labeled “early vision” by many researchers.

Primary visual cortex is only the first stage in the processing of visual information in cortex. Researchers have discovered tens of areas responsible for different aspects of vision (the actual number is still a matter of debate and relies on what we mean by “area”). An influential way of depicting these multiple areas and their interconnections is the diagram proposed by Felleman and Van Essen and illustrated in Figure 1 (Felleman and Van Essen, 1991). To the untrained eye, this diagram appears to show a bewildering complexity, not unlike the type of circuit diagrams typically employed by electrical engineers. In subsequent Lectures, we will get into this diagram in more detail and discuss some of the areas and connections that play a key role in visual recognition. At this point, I would like to note that this scheme is an oversimplification of the actual wiring diagram. First, each of the boxes in this diagram contains millions of neurons. There are different types of neurons. The arrangement of neurons can be described in terms of the six main layers of cortex (some of which have different sublayers) and the topographical arrangement of neurons within and across layers. Second, we are still very far from characterizing all the connections in the visual system. It is likely that major surprises in neuroanatomy will come from the usage of novel tools that take advantage of the high specificity of molecular biology tools. Even if we did know the connectivity of every single neuron, this would not immediately reveal the functions or computations (but it would be immensely helpful). In contrast to electrical circuits where we understand each element and the overall function can be appreciated from the wiring diagram, many neurobiological factors make the map from structure to function a non-trivial one.
1.3 Lesion studies

One way of finding out how something works is by taking it apart, removing parts of it and re-evaluating function. This 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 in either macaque monkey studies or through natural lesions in humans.

As an example, we already mentioned above that studying the behavioral effects of bullet wounds during World War one provided important information about the architecture and function of V1. In this case, subjects typically revealed that there was a part of the visual field where they were essentially blind (scotoma). Ascending through the visual hierarchy, lesions may yield more specific behavioral deficits. For example, subjects who suffer from a rare but well-known condition called prosopagnosia typically show a significant impairment in recognizing faces.

One of the challenges in interpreting lesions in the human brain and localizing visual functions based on these studies is that these lesions often
encompass large brain area and are not restricted to neuroanatomically- and 
neurophysiologically-defined areas. 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 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? The 
lesion and neurological studies are discussed in Lecture 3.

1.4 Function of circuits in visual cortex

The gold standard to record the function in brain circuits is to implant a 
microelectrode (or multiple microelectrodes) into the area of interest. These 
extracellular recordings allow the investigators to monitor the activity of one or a 
few neurons in the near vicinity of the electrode (~200 µm) at neuronal resolution 
and sub-millisecond temporal resolution.

Recording the activity of neurons has defined the receptive field structure 
of neurons in the retina, LGN and primary visual cortex. The receptive field, 
loosely speaking, is defined as the area within the visual field where a neuronal 
response can be elicited. The size of these receptive fields typically increases 
monotonically from cells in the retina all the way to cells in the macaque inferior 
temporal cortex. In a classical neurophysiology experiment, Hubel and Wiesel 
inserted a thin microwire to isolate single neuron responses in the primary visual 
cortex of a monkey. After presenting different visual stimuli, they discovered that 
the neuron fired vigorously when a bar of a certain orientation was presented 
within the neuron’s receptive field. The response was significantly less strong 
when the bar showed a different orientation. This orientation preference 
constitutes a hallmark of a large fraction of the neurons in V1.

Recording from other parts of visual cortex, investigators have 
characterized neurons that show enhanced responses to stimuli moving in 
specific directions, neurons that show preference for complex shapes such as 
fractal patterns or faces, neurons that are particularly sensitive to color contrasts. 
Lecture 4 will focus on the neurophysiological responses beyond primary visual 
cortex. How does selectivity to complex shapes arise and what are the 
computational transformations that can take the simpler receptive field structure 
at the level of the retina to the complex object preferences in inferior temporal 
cortex? As noted above, it is important to combine selectivity with robustness to 
object transformations. How robust are the visual responses to object 
transformations? How fast do neurons along the visual cortex respond to new 
imput stimuli? What is the neural code, that is, what aspects of neuronal responses 
better reflect the input stimuli? What are the mechanisms to combine selectivity 
and invariance?
1.5 Moving beyond correlations

Neurophysiological recordings provide a correlation between the activity of neurons (or groups of neurons) and the visual stimulus presented to the subject. Neurophysiological recordings can also provide a correlation with the subject’s behavioral response (e.g. image recognized or not recognized). Yet, as often stated, correlations do not imply causation.

In addition to the lesion studies briefly mentioned above, an important tool to move beyond correlations is to try to bias the subject’s behavioral performance through the use of electrical stimulation. With the same electrodes used to record neural responses, it is possible to inject current and combine this with careful psychophysical measurements. In a classical study, the group of Newsome recorded the activity of neurons in an area called MT, located within the dorsal part of the macaque visual cortex. As observed previously, these neurons show strong motion direction preferences. The investigators trained the monkey to report the direction of motion of the stimulus. Once the monkeys were proficient in this task, they started introducing trials where they would perform electrical stimulation. Remarkably, they observed that electrical stimulation could bias the monkey’s performance by about 10 to 20% in the preferred direction of the recorded neurons (Salzman et al., 1990).

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 the epilepsy. Before the resection procedure, they use electrical stimulation to examine the function of the tissue that may undergo resection. Penfield was one of the pioneers in using this technique to map neural function and described the effects of stimulating many locations and in many subjects. Anecdotal reports provide a fascinating account of the potential behavioral output of stimulating cortex. For example, in one of many cases, a subject reported that it felt like “… being in a dance hall, like standing in the doorway, in a gymnasium…”

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? Is visual awareness better modeled by a threshold mechanism or by gradual transitions? Lecture 5 is devoted to the effects of electrical stimulation in the macaque and human brains.

1.6 Towards a theory of visual object recognition

Ultimately, a key goal is to be able to develop a theory of visual recognition that can explain the high levels of primate performance in rapid recognition tasks. A successful theory would be amenable for computational
implementation, in which case, one could directly compare the output of the computational model against behavioral performance measures (Serre et al., 2005). A complete theory would include the information from lesion studies, neurophysiological recordings, psychophysics, electrical stimulation studies, etc. Lecture 6 will discuss multiple approaches to building computational models and theories of visual recognition.

In the absence of a complete understanding of the wiring circuitry, only sparse knowledge about neurophysiological responses and other limitations, it may be worth wondering whether it is worth even thinking about theoretical efforts. My (biased) answer is that it is not only useful; it is essential to enhance progress. Computational models can integrate existing data across different laboratories, techniques and experimental conditions, explaining apparently disparate observations. Models can formalize knowledge and assumptions and provide a quantitative, systematic and rigorous path towards examining computations in visual cortex. A good model should be inspired by the empirical findings and should in turn be able to produce non-trivial (and hopefully experimentally-testable) predictions. These predictions can be empirically evaluated to validate, refute or expand the models.

How do we build and test computational models? How should we deal with the sparseness in knowledge and the large number of parameters often required in models? What are the approximations and abstractions that can be made? Too much simplification and we may miss the crucial aspects of the problem. Too little simplification and we may spend decades bogged down by non-essential details. Consider as a simple analogy, 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, the exact position where force is applied and so on. We should perhaps be thankful for the lack of computers in that time, such that there was no possibility of running simulations that included all these variables which are not essential 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 simple example) is not good either. Perhaps a central question in computational neuroscience is to achieve the right level of abstraction for each problem.

1.7 Towards the neural correlates of visual consciousness

Subjective feelings and percepts emerge from the activity of neuronal circuits in the brain. It seems that most scientists would agree with this statement. 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 constitutes an area of active research and passionate debates (Koch, 2005). Historically, many
neuroscientists avoided research in this field as a topic too complex or too far removed from what we understood to be worth a serious investment of time and effort. In recent years, however, this has begun to change: while still very far from a solution, systematic and rigorous approaches guided by neuroscience knowledge may one day unveil the answer to one of the greatest challenges of our times.

Due to several practical reasons, the underpinnings of subjective perception have been particularly (but not exclusively) studied in the domain of vision. There have been several heroic efforts to study the neuronal correlates of visual perception using animal models (e.g. (Leopold and Logothetis, 1999; Macknik, 2006) among many others). A prevalent experimental paradigm involves dissociating the visual input from perception. For example, in multistable percepts (such as the well-known Necker cube illusion) the same input can lead to two 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 may not be correlated with perception. In a perhaps 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 visual cortex. Neurophysiological, neuroanatomical and theoretical considerations suggest that subjective perception correlates with activity occurring after primary visual cortex (Koch, 2005; Leopold and Logothetis, 1999; Macknik, 2006)). Similarly investigators have suggested an upper bound on the circuits involved in subjective perception. There is some evidence suggesting constraints on how early in the processing pathway the representations must be, as well. Although lesions restricted to the hippocampus and frontal cortex (thought to underlie memory and association) yield severe cognitive impairments, these lesions seem to leave many aspects of visual perception largely intact. Thus, the neurophysiology and lesion studies seem to constrain the problem to the multiple stages involved in processing visual information along the ventral visual cortex. Ascending through the ventral visual cortex several neurophysiological studies suggest that there is an increase in the degree of correlation between neuronal activity and visual awareness (Koch, 2005; Leopold and Logothetis, 1999; Macknik, 2006).

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? Lecture 7 will provide some glimpses into what is known (and what is not known) about these fascinating questions.

References