Chapter IV. Psychophysical studies of visual object recognition

We want to understand the neural mechanisms responsible for visual object recognition and we want to instantiate these mechanisms into computational algorithms that resemble and perhaps eventually surpass human performance. In order to untangle the mechanisms orchestrating visual recognition and build adequate computational models, we need to define visual recognition capabilities at the behavioral level. What shapes can humans recognize and when and how? Under what conditions do humans make mistakes? How fast can humans recognize complex objects? How much experience and what type of experience with the world is required to learn to recognize objects?

We can learn about visual object recognition by carefully quantifying human performance under a variety of well-controlled visual tasks. A discipline with the peculiar and attractive name of "Psychophysics" aims to rigorously characterize, quantify and understand behavior during cognitive tasks.

4.1. What you get ain't what you see

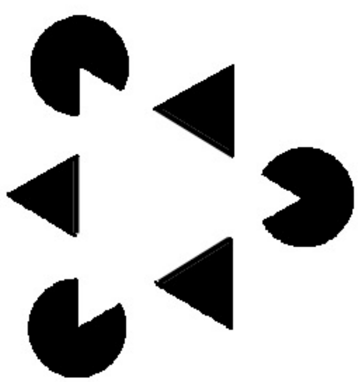

Figure 4.1: The Kanizsa triangle. The mind creates a white triangle from the incomplete information provided by the pacmen or other shapes in the figure.

It is clear that what we end up perceiving is a significantly transformed version of the pattern of photons impinging on the retina. Our brains filter and process visual inputs to understand the physical world by constructing an interpretation that is consistent with our experiences. This observation may seem counterintuitive at first: our perception is a sufficiently reasonable representation of the outside world to allow us to navigate, to grasp objects, to interpret where things are going and whether a friend is happy or not. It is extremely tempting to assume that our visual system is actually capturing a perfect rendering of the outside world.

Visual illusions constitute strong examples of the dissociation between what is in the real world and what we end up perceiving. A simple example of the dissociation between inputs and percepts is given by the blind spot. If you close one eye, there is a part of the visual field that is not mapped onto retinal ganglion cells, the spot where these cells leave the retina to form the optic nerve. It is possible to distinguish this blind spot by closing one eye, fixating on a given spot and slowly moving a finger from the center to the

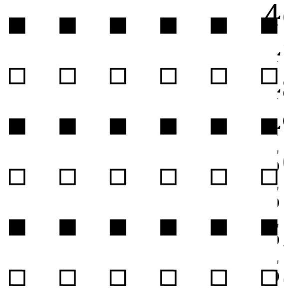

Figure 4.2. **Law of similarity**. We tend to group objects that share some properties.

periphery until part of it disappears from view (but not in its entirety which would imply that you moved your finger completely outside of your visual field).

Visual illusions are not the exception, rather they illustrate the fundamental principle that our perception is a construct, a confabulation, inspired by the visual inputs. There is a lot of information in the world that we just do not see. As a simple example, we do not perceive with our eyes information in the ultraviolet portion of the light spectrum (but other animals do). Another simple example is when we are watching a movie. A movie is nothing more than a sequence of frames, typically presented at a rate of 30 frames per second or more. Our brains do not perceive this rate and instead we interpret objects as moving on the screen.

In addition to not being able to perceive a lot of what's happening in the real world, our brains invent a lot of information that does not exist. Consider for example, the Kanizsa triangle illustrated in Figure 4.1. We perceive a large white triangle in the center of the image and we can trace each of the sides of said triangle. Yet, those edges are composed of illusory contours: in between the edge of a pacman and the adjacent small black triangle, there is no white edge.

4.2. Gestalt laws of grouping

One of the early and founding attempts at establishing basic principles of visual perception originated from the German philosophers and experimental psychologists in the late nineteenth century. The so-called *Gestalt* laws (in German "gestalt" means shape) provide basic constraints about how patterns of light are integrated into perceptual sensations (Reagan, 2000). These rules arose from attempts to understand the basic perceptual principles that lead to interpreting objects as wholes rather than the constituent isolated lines or elements that give rise to them.

- *Law of closure*. We complete lines and extrapolate to complete known patterns or regular figures. An example of this is given by the famous Kanizsa triangle. Our mind creates a triangle in the middle of the image from incomplete information (**Figure 4.1**).
- *Law of similarity*. We tend to group similar objects together. Similarity could be defined by shape, color, size or brightness (**Figure 4.2**)
- *Law of proximity*. We tend to group objects based on their distance (**Figure 4.3**).
- Law of *symmetry*. We tend to group symmetrical images.

- *Law of continuity*. We tend to continue regular patterns (**Figure 4.4**).
- *Law of common fate*. Elements with the same moving direction tend to be grouped.

These laws are usually summarized by pointing out that the forms (Gestalten) are more than the mere sum of the component parts.

4.3. Holistic processing of faces

An interesting example of the processing and interpretation of a whole image beyond what can be discerned from the individual components is the f holistic processing of faces. Three main observations have been put forward to document the holism of face processing. First is the inversion effect (Yin, 1969; Valentine, 1988), which describes how difficult it can be to distinguish local changes in a face when it is turned upside down (this is also called the “Thatcher effect” alluding to the images of Britain’s prime minister originally used to demonstrate the perceptual illusion). The second observation is the composite face illusion: by putting together the upper part of face 1 and the bottom part of face 2, one can create a novel face that appears to be perceptually distinct from the two original ones (Young et al., 1987). A third argument for holistic processing is the parts and wholes effect: changing a local aspect of a face distorts the overall perception of the entire face (Tanaka and Farah, 1993).

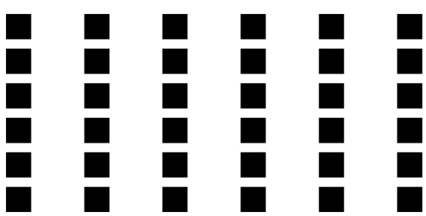

Figure 4.3. **Law of proximity**. We tend to see this figure as vertical lines.

4.4. Tolerance to object transformations

A hallmark of visual recognition is our ability to identify and categorize objects in spite of large transformations in the image. An object can cast an *infinite* number of projections onto the retina due to changes in position, scale, rotation, illumination, color, etc. This invariance to image transformations is critical to recognition. Our visual recognition capabilities would be quite useless without the ability to abstract away those changes.

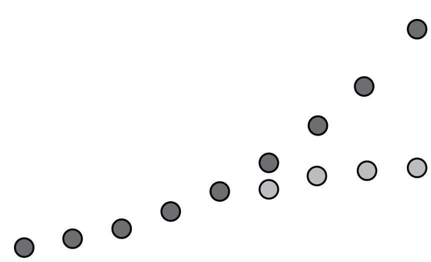

Figure 4.4. Law of continuity. We tend to assume that the dark gray circles form a line.

To further illustrate the critical role of tolerance to object transformations in visual recogntiion, consider a very simple algorithm that we will refer to as "the rote memorization machine". This algorithm receives inputs from a digital camera and remembers every single pixel. It can remember the Van Gogh sunflowers, it can remember a picture with your face taken two weeks ago on Monday at 2:30pm, it can remember exactly what your car looked like three years ago on a Saturday at 5:01pm. While such extraordinary memory might seem quite remarkable at first, it turns out that this would constitute a rather brittle approach to recognition. This algorithm would not be able to recognize your car in the parking lot today, because you may see it under a different illumination, a different angle, and with different amounts of dust than in any of the memorized photographs. This problem is beautifully illustrated in a short story by Argentinian fiction writer Jorge Luis Borges in "Funes the memorious", relating the story of a character who has infinite memory due to a brain accident. Borges concludes: "To think is to forget differences, generalize, make abstractions".

Our visual system is able to abstract away many of those image transformations to recognize objects. The visual system shows a significant degree of robustness to changes in many image properties, including the following:

- Scale (e.g. you can recognize an object at different sizes). You can easily demonstrate the strong degree of tolerance for object transformations. For example, take a piece of text with 12pt font size, hold it at arm's length and focus on any given letter, say "A". The A will subtend a fraction of one degree of visual angle (approximately the size of your thumb at arm's length).
- Position with respect to fixation (e.g. we can recognize an object placed at different distances to the fixation point)
- 2D rotation (e.g. we can recognize an object after turning our head sideways or rotating the object within the plane)
- 3D rotation (e.g. we can recognize an object from different viewpoints)
- Color (e.g. we can recognize the objects in a photograph whether it's in color, sepia, grayscale)
- Illumination (e.g. consider illuminating an object from the left, right, top or bottom)
- Cues (e.g. an object's shape can be determined by edges, by motion cues, by completion without sharp edges)
- Clutter (e.g. we can recognize objects despite the presence of other objects in the image)
- Occlusion (e.g. we can recognize objects from partial information)
- Other non-rigid transformations (e.g. we can recognize faces even with changes in expression, aging, even from the line drawing sketches in **Figure 4.5**!)

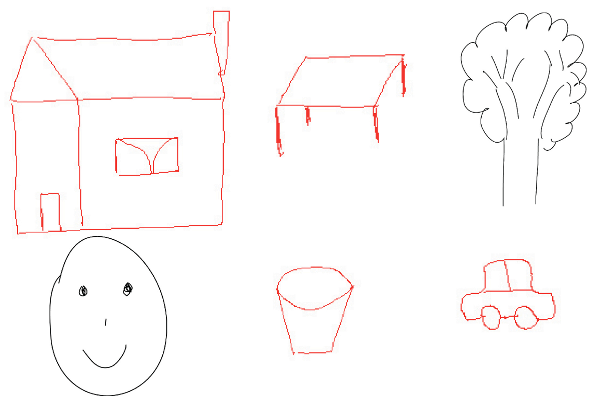

Figure 4.5. Recognition of line drawings. We can identify the objects in these line drawings despite the extreme simplicity in the traces and the minimal degree of resemblance to the actual objects.

A particularly intriguing example of tolerance is given by the capability to recognize caricatures and line drawings. At the pixel level, these images seem to bear little resemblance to the actual objects and yet, we can recognize them quite efficiently, sometimes even better than the real images!

4.5. Speed of visual recognition

Visual recognition *seems* almost instantaneous. Several investigators have shown that we can recognize complex objects in a small fraction of a second.

One of the original studies by Mary Potter consisted of showing a sequence of images in a rapid sequence (RSVP, rapid serial visual presentation) and showing that subjects could detect the individual images even when presented at rates of 8 per second (Potter and Levy, 1969). Complex objects can be recognized when presented tachistoscopically for < 50 ms without a mask, even in the absence of any prior expectation or other knowledge (Vernon, 1954).

Part of the delays in reaction time measurements are associated with the behavioral response. In an attempt to constrain the amount of time required for visual recognition, Thorpe and colleagues recorded evoked response potentials from scalp electroencephalographic (EEG) signals while subjects performed a go/no-go animal categorization task (Thorpe et al., 1996). They found that frontal cortex electrodes showed a differential signal at about 150 ms; they argued that visual discrimination of animals versus non-animals in complex scenes should happen before that time. Kirchner *et al* used eye movements to elicit rapid responses and showed that subjects could make a saccade to discriminate the presence of a face or non-face stimulus in slightly more than 100 ms (Kirchner and Thorpe, 2006). These observations place a strong constraint into the mechanisms that underlie visual recognition.

Such speed in object recognition also suggests that the mechanisms that integrate information in time must occur rather rapidly. Under normal viewing conditions, all parts of an object reach the eye more or less simultaneously (in the absence of occlusions and object movement). By disrupting such synchronous access, one can probe the speed of temporal integration in vision. In a behavioral experiment to quantify the speed of integration, Jed Singer presented different parts of an object asynchronously (Figure 4.6), like breaking

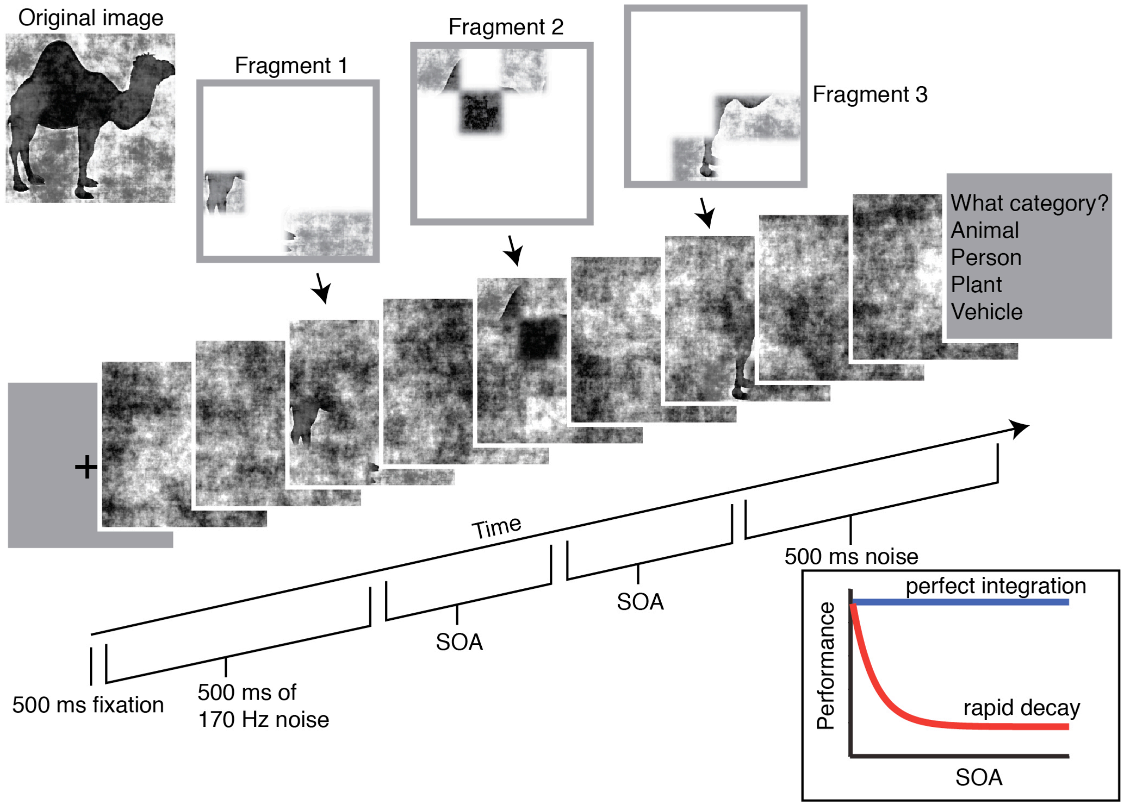

Figure 4.6. **Subjects can integrate asynchronously presented object information**. Subjects were presented with different parts of an object in an asynchronous fashion (in this example, a camel). The middle part of the diagram shows the sequence of steps in the experiment. Subjects fixate and they observe a sequence of frames in which the object fragments are separated by a stimulus onset asynchrony (SOA). Subjects perform a 5-alternative forced choice categorization task. Subjects could integrate information up to asynchronies of about 30 ms.

Humpty Dumpty and trying to put the pieces back together again. He reasoned that if there was a long interval between the presentation of different parts, subjects would be unable to interpret what the object was, but if the parts were presented very close in time, the brain would easily be able to integrate them back to a unified perception of the object.

Another striking example of temporal integration is the phenomenon known as *anorthoscopic perception*, defined as perception of a whole object in cases where only a part of which is seen at a given time, perhaps one of the very earliest attempts at cinema. In classical experiments, an image is seen through a slit and the image moves rapidly allowing the viewer to catch only a small part of the whole at any given time. The brain integrates all the snapshots and puts them together to create a perception of a whole object moving. The power of temporal integration is emphasized in cases where an actor is placed in a completely dark room wearing black with only a few sources of information placed along his body.

With just a handful of points it is possible to infer the actor's motion patterns(Johansson, 1973). Related studies have shown that it is possible to dynamically group and segment information purely based on temporal integratoin(Anstis, 1970; Kellman and Cohen, 1984).

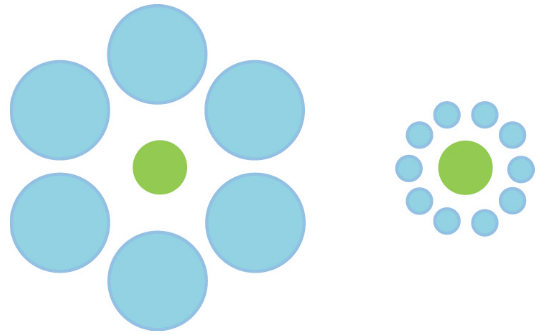

Figure 4.7. **Context matters**. The green circle on the right appears to be larger than the one on the left but they are the same size.

4.6. Beyond pixels – contextual effects

In addition to temporal integration, visual recognition also exploits the possibility of integrating spatial information. Several visual illusions demonstrate the existence of strong contextual effects in visual object recognition. For example, it is significantly more difficult to recognize faces when they are upside down (see "Holistic processing" above). In a simple yet elegant demonstration, the perceived size of a circle can be strongly influenced by the size of its neighbors (**Figure 4.7**). Another extremely simple example is the Muller-Lyer illusion: the perceived length of a line with arrows at the two ends depends on the directions of the two arrows. Several entertaining examples of contextual effects have been reported (e.g. (Sinha and Poggio, 1996; Eagleman, 2001)). These strong contextual dependences illustrate that the visual system spatially integrates information and the perception of local features may depend on the global surrounding properties.

Figure 4.8. **Context matters in the real world too**. What is the object in the white box?

Such contextual effects are not restricted to visual illusions and psychophysics demos like the one in Figure 4.7. Consider **Figure 4.8**. What is the object in the white box? It is typically very hard to answer this question with any degree of certainty. Now, turn your attention to **Figure 4.9**. What is the object in the white box? This is a much easier question! Even though the pixels inside the white box are identical in both figures, the surrounding contextual information dramatically changes the probability of correctly detecting the object. These contextual effects are very fast and can be triggered by presenting even simpler and blurred version of the background information(Wu et al., Submitted). These contextual effects also emphasize that perception constitutes an interpretation of the input in the light of context and experience.

4.5 The value of experience

Our percepts are also influenced by previous visual experience. This observation holds at multiple different temporal scales. At short time scales, several visual illusions show the powerful effects of visual adaptation. One such illusion is the waterfall effect: after staring at a waterfall for a minute or so, and then shifting the gaze to other static objects, those objects appear to be moving upward. At longer time scales, the interpretation of an image could depend on whether one has seen the image before. A typical example is the Dalmatian dog: for the first-time observer the image consists of a smudge of black and white spots. However, after recognizing the dog, people can immediately spot him the next time. Other similar examples are Mooney images (Mooney, 1957).

Figure 4.9. **Context matters in the real world too**. What is the object in the white box?

References

Anstis SM (1970) Phi movement as a subtraction process. Vision research.
Eagleman DM (2001) Visual illusions and neurobiology. Nat Rev Neurosci 2:920-926.
Johansson G (1973) Visual perception of biological motion and a model for its analysis. . Perception and Psychophysics 14:201-211.
Kellman PJ, Cohen MH (1984) Kinetic subjective contours. Perception and Psychophysics.
Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision research 46:1762-1776.
Mooney CM (1957) Age in the development of closure ability in children. Can J Psychol 11:219-226.
Potter M, Levy E (1969) Recognition memory for a rapid sequence of pictures. Journal of experimental psychology 81:10-15.
Reagan D (2000) Human perception of objects: Sinauer.
Sinha P, Poggio T (1996) I think I know that face. Nature 384:404.
Tanaka JW, Farah MJ (1993) Parts and wholes in face recognition. The Quarterly journal of experimental psychology A, Human experimental psychology 46:225-245.

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520-522.
Valentine T (1988) Upside-down faces: a review of the effect of inversion upon face recognition. Br J Psychol 79 (Pt 4):471-491.
Vernon M (1954) Visual perception. Cambridge: Harvard University Press.
Wu E, Wu K, Kreiman G (Submitted) Learning Scene Gist with Convolutional Neural Networks to Improve Object Recognition.
Yin R (1969) Looking at upside-down faces. Journal of experimental psychology 81:141-145.
Young AW, Hellawell D, Hay DC (1987) Configurational information in face perception. Perception 16:747-759.