Chapter III. The phenomenology of seeing

Supplementary contents at **http://bit.ly/38buAhB**

We want to understand the neural mechanisms responsible for visual cognition, and we want to instantiate these mechanisms into computational algorithms that resemble and perhaps even surpass human performance. In order to build such biologically inspired visually intelligent machines, we first need to define visual cognition capabilities at the behavioral level. What types of shapes can be recognized, and when and how? Under what conditions do people make mistakes during visual processing? How much experience and what type of experience with the world is required to learn to see? To answer these questions, we need to quantify human performance under well-controlled visual tasks. A discipline with the picturesque and attractive name of *Psychophysics* aims to rigorously characterize, quantify, and understand behavior during cognitive tasks.

III.1. What you get ain't what you see

As already introduced in **Chapter II**, 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 around us by constructing an interpretation that is consistent with our experiences. The notion that our brains make up stuff 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 predict where things are going, and to discern whether a friend is happy or not. It is extremely tempting to assume that our visual system actually captures a perfect literal rendering of the outside world.

Visual illusions constitute convincing examples of the dissociation between what is in the real world and what we end up perceiving. **Chapter II** presented several examples of the dissociation between inputs and percepts: the blind spot, the complete elimination of inputs during blinks, and the ultra-rapid input changes during saccadic eye movements. In all of these cases, our brains fill in the missing information.

Visual illusions are not the exception to the rule; they illustrate the fundamental principle that our perception is a construct, a confabulation, inspired by the visual inputs. There is substantial information in the world that we just do not see. For example, we cannot perceive with our eyes information in the ultraviolet portion of the light spectrum (but other animals like mice do). As another example, our visual acuity has a limit: there are small things like bacteria that we cannot see with our eyes.

There are things out there that we cannot see, and there are things that do not exist, but we do see. For example, when we watch a movie, the screen depicts a sequence of frames in rapid succession, typically presented at a rate of 30 frames per second. Our brains do not perceive this sequence of frames. Instead, the brain interprets the presence of objects that are moving on the screen. As another example, consider the triangle illustrated in ***Figure III-1***, known as the Kanizsa triangle. We perceive a white triangle in the center of the image, and we can trace each of the sides of said triangle. However, those edges are composed of *illusory contours*: in between the edge of one Pacman and the edge of the adjacent Pacman, there is no white edge. The triangle is purely in our brains.

INSERT *Figure III-1* AROUND HERE
Figure III-1: Our brains make up stuff*.* ***A****. The brain creates a white triangle from the incomplete information provided by the Pacman in the figure. The illusion is broken by closing the circles (****B****) or rotating the Pacman (****C****).*

III.2. Perception depends on adequately grouping parts of an image through specific rules

INSERT ***Figure III-2*** AROUND HERE
Figure III-2*.* ***Figure-ground segregation****. We tend to separate figure, here a person running, from the background, here uniform black.*

Our brains are confabulators, pretty useful confabulators that follow systematical rules to create our perceptual worlds. 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 elementary constraints about how patterns of light are integrated into perceptual sensations. These rules arose from attempts to understand the basic principles that lead to interpreting objects as wholes rather than the constituent isolated lines or elements that give rise to them. These grouping laws are usually summarized by pointing out that the forms are more than the mere sum of the parts.

■ *Figure-ground segregation.* We readily separate the figure from the background based on the relative contrast, size, color, and other properties. (***Figure III-2***). The famous artist M.C. Escher (1898–1972) capitalized on this aspect of cognition to render ambiguous images where the figure and background merge back and forth in different regions. Evolution probably discovered the importance of separating figure from ground when detecting a prey, leading to the phenomenon of camouflage whereby the figure blends into the background, making it difficult to spot.

■ *Closure*. We complete lines and extrapolate to complete known patterns or regular figures. We tend to put together different parts of the image to make a

single, recognizable, shape. For example, our brain creates a triangle in the middle of the Kanizsa image from incomplete information (***Figure III-1***).

INSERT ***Figure III-3*** *AROUND HERE*
Figure III-3*.* ***Grouping by similarity****. We tend to group objects that share common properties.* ***A****. We perceive horizontal lines composed of black squares interleaved with horizontal lines composed of white squares, grouping the items by their color.* ***B****. We perceive five distinct groups based on grouping shapes.*

■ *Similarity*. We tend to group similar objects together. Similarity can be defined by shape, color, size, brightness, and other properties (***Figure III-3***).

INSERT ***Figure III-4*** AROUND HERE
Figure III-4*.* ***Grouping by proximity****. We perceive this figure as vertical lines.*

■ *Proximity*. We tend to group objects based on their relative distances (***Figure III-4***). Proximity is a potent cue that can often trump some of the other grouping criteria.

■ S*ymmetry*. We tend to group symmetrical images.

INSERT ***Figure III-5*** AROUND HERE
Figure III-5*.* ***Grouping by continuity****. We tend to assume that the dark gray circles form a continuous line.*

■ *Continuity*. We tend to continue regular patterns (***Figure III-5***).

■ *Common fate*. Elements with the same moving direction tend to be grouped together. Movement is one of the strongest and most reliable cues for grouping and segmentation of an image, superseding the other criteria. Because of this, an animal that wants to camouflage with the background should stay very still.

III.3. The whole can be more than the sum of its parts

The Gestalt grouping rules dictate the organization of elements in an image into higher-order structures, new interpretable combinations of simple elements. A demonstration of the combination of elements beyond what can be discerned from the individual components is referred to as *holistic processing*. A particularly extensively studied form of holistic processing is the interpretation of faces.

Three main observations have been put forward to document the holism of face processing. First, the *inversion effect* describes how difficult it is to distinguish local changes in a face when it is turned upside down. An illusion known as the “Thatcher effect” illustrates this point: distorted images of Britain’s prime minister can be easily distinguished from the original when they are right side up

but not when they are upside down. The second observation suggesting holistic processing is the *composite face illusion*: putting together the upper part of a given face A and the bottom part of another face B, creates a novel face that appears to be perceptually distinct everywhere from the two original ones. The 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. The observation that the whole can be more than the sum of its parts is not restricted to faces; expertise in other domains, including fingerprint identification or recognition of novel arbitrary shapes, also leads to similar holistic effects.

III.4. The visual system tolerates large image transformations

The observation that the interpretation of the whole object is not merely a list of components makes it challenging to build models of object recognition that are based on a checklist of specific object parts. Another serious challenge to this type of checklist model of recognition is that often several of the parts may not be visible or may be severely distorted. A hallmark of visual recognition is the ability to identify and categorize objects despite 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, and other variables. This tolerance to image transformations is critical to recognition, it constitutes one of the fundamental challenges in vision (**Chapter I**), and it is, therefore, one of the key goals for computational models (**Chapter VIII**). Visual recognition capabilities would be quite useless without the ability to abstract away image changes.

To further illustrate the critical role of tolerance to image transformations in visual recognition, consider a straightforward algorithm that we will refer to as "the rote memorization machine" (**Figure I-4**). This algorithm receives inputs from a digital camera and perfectly remembers every single pixel. It can remember the Van Gogh sunflowers, it can remember a selfie taken two weeks ago on Monday at 2:30 pm, it can remember precisely what your car looked like three years ago on a Saturday at 5:01 pm. While such extraordinary pixel-based memory might seem quite remarkable at first, it turns out that this would constitute a 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 different illumination, a different angle, and with different amounts of dust than in any of the memorized photographs. The problem with the rote memorization machine is beautifully illustrated in a short story by Argentinian fiction writer Jorge Luis Borges, titled "Funes the memorious." The story relates the misadventures of a character called Funes, who acquires infinite memory due to a brain accident. Funes' initial enthusiasm with his extraordinary memory soon fades when he cannot achieve visual invariance as manifested, for example, by failing to understand that a dog at 3 pm is the same dog at 3:01 pm when seen from a slightly different angle. Borges concludes: "To think is to forget differences, generalize, make abstractions."

INSERT ***Figure III-6*** AROUND HERE

***Figure III-6**. **Tolerance in visual recognition**. The lighthouse can be readily recognized despite large changes in the appearance of the image.*

Our visual system can abstract away many image transformations to recognize objects (***Figure III-6***), demonstrating a degree of robustness to changes in several image properties, including the following ones:

- Tolerance to scale changes, i.e., recognizing an object at different sizes. In vision, object sizes are typically measured in degrees of visual angle (***Figure II-4***). Now, consider again the sketch of a person running in ***Figure III-2***. If you are holding the page approximately at arms-length, the person will subtend approximately two degrees of visual angle. Moving the page closer and closer will lead to a multiple-fold increase in its size, mostly without affecting recognition. There are limits to recognition imposed by visual acuity (if the page is moved too far away), and there are also limits to visual recognition at the other end, if the image becomes too large (if the page touches your nose). However, there is a broad range of scales over which we can recognize objects.
- Position with respect to fixation, i.e., recognizing an object placed at different distances from the fixation point. For example, fixate on a given point, say your right thumb. Make sure not to move your eyes or your thumb. Then move the running man in ***Figure III-2*** to different positions. You can still recognize the image at different locations with respect to the fixation point. As discussed in **Chapter II**, acuity decreases sharply as we move away from the fixation point. Therefore, if you keep moving the page away from fixation (and then you stop, because motion is easily detected in the periphery), eventually, the image of the running man will become unrecognizable. However, there is a wide range of positions where recognition still works.
- 2D rotation, i.e., recognizing an object that is rotated in the same plane (***Figure III-6G***). You can recognize the running man even if you rotate the page, or if you tilt your head. Recognition performance is not completely invariant to 2D rotation, as mentioned above in the case of the Thatcher illusion.
- 3D rotation, i.e., recognizing an object from different viewpoints. Recognition shows some degree of tolerance to 3D rotation of an object, but it is not quite completely invariant to viewpoint changes. Rotation in the three-dimensional world is a particularly challenging transformation because the types of features revealed about the object can depend quite strongly on the viewpoint. In particular, some objects are much easier to recognize from certain canonical viewpoints rather than from other viewing angles.
- Color. In many cases, objects can be readily recognized in a photograph, whether it is in color, sepia, or grayscale (***Figure III-6E***). Color can certainly add valuable information and can enhance recognition, yet recognition abilities are quite robust to color changes.
- Illumination. In most cases, objects can be readily identified regardless of whether they are illuminated from the left, right, top, or bottom.
- New transformations. To some extent, we can also identify objects under novel transformations that we have not experienced before. Perhaps we have never

seen a lighthouse depicted as in ***Figure III-6F*** *or* ***K****.* The ability to extrapolate to such new conditions is particularly remarkable and a formidable challenge for computational models of visual recognition.

These are but a few of the myriad transformations an object can go through, with minimal impact on recognition, many other examples are illustrated in ***Figure III-6***. The visual system can also tolerate many types of non-rigid transformations, such as recognizing faces even with changes in expression, aging, make-up, or shaving. The examples in ***Figure III-6*** all depend on identifying the lighthouse based on its sharp contrast edges, but objects can be readily identified even without such edges. For example, motion cues can be used to define an object's shape.

INSERT ***Figure III-7*** AROUND HERE
Figure III-7*.* ***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.*

An intriguing example of tolerance is given by the capability to recognize caricatures and line drawings (***Figure III-7***). At the pixel level, these images bear little resemblance to the actual objects, and yet, we can recognize them quite efficiently, sometimes even better than the real images. It is likely that the ability to interpret line drawings like the ones in ***Figure III-7*** depends on specifically learning to identify symbols and certain conventions about how to sketch those objects more than on visual shape similarity with the objects represented by those drawings. In the case of face caricatures, artists capture essential recognizable features of the person, as opposed to the symbols and conventions in other simple line drawings, therefore highlighting a strong degree of invariance for image transformations.

In all of these cases, recognition is robust to image changes, but it is not perfectly invariant to those changes. It is possible to break recognition by changing the image. Thus, although many investigators refer to *invariant* visual recognition, a better term is probably *transformation-tolerant* visual recognition to emphasize that we do not expect complete invariance to any amount of image change.

III.5. Pattern completion: inferring the whole from visible parts

A particularly challenging form of tolerance that is rather ubiquitous during natural vision is the recognition of occluded objects. Looking at the objects around us, oftentimes, we only have direct access to partial information due to poor illumination or because another object is in front. Deciphering what an object is when only parts of it are visible requires extrapolating to complete patterns. A crude example of occlusion is shown in ***Figure III-6A***. It is easy to identify the lighthouse even though less than half of its pixels are visible. The visual system has a remarkable ability to make inferences from incomplete information. This ability is not exclusive to vision, but rather it is apparent in many other modalities,

including understanding speech corrupted by noise, or even in higher domains of cognition such as imagining a story from a few words printed on a page or deciphering social interactions from sparse information.

INSERT ***Figure III-8*** AROUND HERE
Figure III-8. Pattern completion. A*. It is possible to recognize the rotated B letters despite partial information. B. It is easier to recognize the objects when an explicit occluder is present (**A**) compared to the same object parts when the occluder is absent (**B**).*

Vision is an ill-posed problem because the solution is not unique. In general, there could be infinite interpretations of the world that are consistent with a given retinal image. The infinity of solutions is easy to appreciate in the case of occlusion. There are infinitely many ways to complete contours from partial information. For example, in ***Figure III-9A****,* the lighthouse might have a large hole in it, or there could well be an elephant hidden behind the black box. However, this is not how we would usually interpret the image. Despite these infinite possible solutions, the visual system typically lands on a single interpretation of the image, which is, in most cases, the correct one. Investigators refer to *amodal* completion when there is an explicit occluder (e.g., ***Figure III-8A***) and *modal* completion when illusory contours are created to complete the object without an occluder (e.g., ***Figure III-1A***). The presence of an occluder leads to inferring depth between the occluder shape and the occluded object. Such inferences about depth help create a surface-based representation of the scene. The occluder helps interpret the occluded object, as demonstrated in the famous illusion by Bregman with rotated B letters (compare ***Figure III-8A*** versus ***B***).

The visual system can work with tiny amounts of information. It is possible to occlude up to 80% of the pixels of an object with only a small deterioration in recognition performance. Recognition depends on which specific object features are occluded. Certain parts of an object are more diagnostic than others. One approach to investigating which object parts are diagnostic is to present objects through *bubbles* randomly positioned in the image, controlling which parts of the object are visible and which ones are not. Averaging performance over multiple recognition experiments, it is possible to estimate which object features lead to enhanced recognition and which object features provide less useful information. Instead of presenting an image through an occluder, or revealing features through bubbles, another approach to studying pattern completion is to reduce an image by cropping or blurring until it becomes unrecognizable. Using this approach, investigators have described *minimal images* that can be readily recognized but which are rendered unrecognizable upon further reduction in size.

III.6. Visual recognition is very fast

To recap, what we perceive is a subjective construct created by our brains following a series of phenomenological rules to group elements in the image. Our brains make inferences to arrive at a unique solution for an ill-posed problem,

giving rise to a representation that allows us to interpret a scene and identify objects and their interactions. Given the complexity of this process, one might imagine that it would take an enormous amount of computational time to see anything. On the contrary, vision *seems* almost instantaneous.

The German physicist and physician Hermann von Helmholtz (1821 – 1894) demonstrated that conduction of signals in nerve tissue had a finite and measurable speed, which was a rather revolutionary concept at the time. As we discussed in **Chapter II**, there is no such thing as instantaneous vision: even the conversion of incoming light signals into the output of the retinal ganglion cells takes time, on the order of 40 milliseconds. Subsequent processing of the image by the rest of the brain also takes additional time. What is quite remarkable is that all the processing of sensory inputs, tolerance to transformations, and inferences from incomplete information, can be accomplished in a small fraction of a second. This speed is quite critical: vision would be far less useful if it took many seconds to arrive at an answer (**Chapter I**).

Reaction time measurements have been used to study the mechanisms of perception since the very beginnings of psychophysics. Measuring reaction times provided investigators with an objective measurement as opposed to introspective evaluations. For example, these measurements allowed psychophysicists to quantify the notion of a trade-off between speed and accuracy, evident throughout visual and other tasks and forming the basis of models of decision making.

One of the original studies to document the speed of vision consisted of showing images in a rapid sequence (known in the field as rapid serial visual presentation tasks). Subjects could interpret each of the individual images even when objects were presented at rates of 8 per second. Nowadays, it is relatively easy to present stimuli on a screen for short periods spanning tens of milliseconds or even shorter timescales. In earlier days, investigators had to go through ingenious maneuvers to ensure that stimuli were presented only briefly. A device invented in 1859 to accomplish rapid exposure to light signals, called a *tachistoscope*, uses a projector and a shutter similar to the ones in single-lens reflex photo cameras. This device was subsequently used during World War II to train pilots to rapidly discriminate silhouettes of aircraft. Complex objects can be recognized when presented tachistoscopically for < 50 milliseconds, even in the absence of any prior expectation or other knowledge.

Reaction times measured in response to visual stimuli take much longer than 50 milliseconds. Emitting any type of response (pressing a button, uttering a verbal response, or moving the eyes) requires several steps beyond visual processing, including decision making and the neural steps to prepare and execute the behavior. In an attempt to constrain the amount of time required for visual recognition, Simon Thorpe and colleagues recorded evoked response potentials from scalp electroencephalographic (EEG) signals while subjects performed a go/no-go animal categorization task. Subjects were shown a photograph that either

contained an animal or not and were instructed to press a key whenever they detected an animal. What exactly these EEG signals measure remains unclear. However, it is possible to measure minute voltages, on the order of a few microvolts at the scalp level, and detect changes that are evoked by the presentation of visual stimuli. The investigators found that EEG revealed a signal at about 150 milliseconds after stimulus onset that was different between trials when an animal was shown versus those trials when no animal was present. It is not known whether this EEG measurement constitutes a visual signal, a decision signal, a motor signal, or some combination of all of these types of processes. Regardless of the exact interpretation of these measurements, the results impose an upper bound for this specific recognition task; the investigators argued that visual discrimination of animals versus non-animals embedded in natural scenes should happen *before* 150 milliseconds. Similar behavioral and physiological reports have been observed in macaque monkeys. Consistent with this temporal bound, in another study, subjects had to make a saccade as soon as possible to one of two alternative locations to discriminate the presence of a face versus non-face stimulus. Saccades are appealing to measure behavioral reaction times because they are faster than pressing buttons or verbally producing a response. It took subjects, on average, 140 milliseconds from stimulus onset to initiate an eye movement in this task. These observations place a strong constraint on the computational mechanisms that underlie visual processing (**Chapter VIII**).

INSERT ***Figure III-10*** AROUND HERE

Figure III-10. ***Spatiotemporal pattern completion: subjects can integrate asynchronously presented object information***. *Subjects were presented with different parts of an object asynchronously (in this example, a camel). The middle part of the diagram shows the sequence of steps in the experiment. Subjects fixated for 500 ms, and then observed a sequence of frames in which the object fragments were separated by a stimulus onset asynchrony (SOA). Subjects performed a 5-alternative forced-choice categorization task. Subjects could integrate information up to asynchronies of about 30 ms.*

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 occlusion and object movement). By disrupting such synchronous access to the parts of an object, it is possible to probe the speed of temporal integration in vision. In a behavioral experiment to quantify the speed of integration, investigators presented different parts of an object asynchronously (***Figure III-10***), like breaking Humpty Dumpty and trying to put the pieces back together again. In between the presentation of object parts, subjects were presented with noise for a given amount of time known as the stimulus onset asynchrony (SOA). The researchers conjectured that if there were a long interval between the presentation of different objects parts (long SOA), subjects would be unable to interpret what the object was. Conversely, if the parts were presented in close temporal proximity, the brain would be able to integrate the parts back to a unified perception of the

object. The results showed that subjects could integrate information up to asynchronies of about 30 ms.

Another striking example of rapid temporal integration is the phenomenon known as *anorthoscopic perception*, defined as the interpretation of a whole object in cases where only a part of it is seen at a given time. In classical experiments, an image is shown through a slit. 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 perception of motion from snapshots in this and related experiments eventually inspired the development of movies, where a sequence of slightly displaced frames presented at a sufficient rate are integrated by the brain to give rise to a continuous visual experience.

The power of temporal integration is also nicely illustrated in experiments where an actor wearing black attire is in a completely dark room with only a few sources of light placed along his body. With just a handful of light points, it is possible to infer the actor's motion patterns. Related studies have shown that it is possible to dynamically group and segment information purely based on temporal integration.

Not all visual tasks are so fast. Finding a needle in a haystack is famously challenging. Searching for Waldo can be somewhat infuriating and takes several seconds or more during which the observer will typically move his/her eyes multiple times, sequentially scrutinizing different parts of the image. Even without making eye movements, certain visual tasks require more time. One example task that requires more time, even in the absence of saccades, is the pattern completion problem described in the previous section. Experiments where subjects have limited time to process an image show that completion of heavily occluded objects requires more time than recognition of the fully visible object counterparts. The simplest of these experiments are time-forced tasks, where identification of heavily occluded objects typically lags recognition of fully visible objects by 50 to 150 milliseconds.

Another situation where subjects have limited computational time is *priming* experiments. Priming refers to a form of temporal contextual modulation whereby an image A is preceded in time by another image P called the prime. The perception of A depends on P, P is said to prime perception of A. For example, the presentation of the prime P might influence how well or how fast subjects recognize the stimulus A. Priming is not restricted to the visual domain. For example, consider the planets in the solar system: Mercury, Venus, Earth, Mars, Jupiter. Now try to complete the following word: M _ _ N. It is quite likely that you thought about "moon," although the word "mean" would be at least as good an answer. In fact, according to Google's Ngrams, the word "mean" is three times for frequent in the English language compared to the word "moon." Therefore, it should be more

likely for people to think of “mean” rather than “moon”; the previous sentence listing several planets primed the reader to think about the moon.

Similar experiments can be done in the visual domain by showing a picture as a prime instead of a list of words. By changing the amount of exposure to the prime, we can assess whether the prime image was recognized or not by evaluating its influence on subsequent perception. When the prime P is a heavily occluded object, the magnitude of the priming effect depends on the time interval between P and A. If this interval is less than 50 milliseconds, the priming effect vanishes, suggesting that 50 milliseconds was not enough to complete the pattern and therefore to have any impact on subsequent recognition.

Finally, another common tool to limit processing time in the psychophysicist’s arsenal is *backward masking.* In backward masking experiments, a stimulus A is closely followed by a noise pattern B. If the interval between A and B is very short, typically less than 20 milliseconds, the initial stimulus A is essentially invisible. With longer intervals, subjects can still see the initial stimulus A, but recognition is impaired. When A is a heavily occluded object, and a noise pattern B is introduced about 50 milliseconds after A, it becomes challenging to complete the pattern in A. Investigators argue that the noise pattern interrupts the computations required for pattern completion. If the interval between A and the noise pattern B is longer than approximately 100 milliseconds, the effect of backward masking disappears. These different types of experiments show converging evidence that putting together the parts to infer the whole, during a single fixation, requires additional computational steps manifested through longer reaction times.

III.7. Spatial context matters

In addition to temporal integration, visual recognition also exploits the possibility of integrating spatial information. Essential aspects of recognition are missed if we take vision out of context.

INSERT ***Figure III-11*** AROUND HERE
Figure III-11. Context matters*. The dark circle in the center appears to be larger on the right than on the left, but they are actually the same size*.

Several visual illusions demonstrate strong contextual effects in visual recognition. In a simple yet elegant demonstration, the perceived size of a circle can be strongly influenced by the size of the neighboring stimuli (***Figure III-11***). Another 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. These strong contextual dependencies show that the visual system spatially integrates information, and the perception of local features can also depend on the surround and even on global image properties.

INSERT ***Figure III-12*** AROUND HERE
Figure III-12. Context matters in the real world too*. What is the object in the white box? Warning: do not peek into the next figure before trying to answer this question!*

Such contextual effects are not restricted to visual illusions and psychophysics demos like the one in ***Figure III-11***. Everyday vision capitalizes on contextual information. Consider ***Figure III-12*** (and do *not* peek into ***Figure III-13*** yet): what is the object in the white box? It is typically hard to answer this question with any degree of certainty. If you are not sure, take a guess. Write down your top 5 wild guesses. Now, turn your attention to ***Figure III-13***. What is the object in the white box? Recognizing the same object in ***Figure III-13*** 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. One could imagine that the observer may examine multiple different parts of the image before fixating on the white box to deduce what the object is. However, in laboratory experiments where we can precisely monitor eye gaze, subjects show a notable and rapid improvement in recognition performance even when they are only fixating on the white box and the image disappears before subjects can move their eyes. These contextual effects are fast, depend on the amount of context, and can be at least partly triggered by presenting even simpler and blurred version of the background information. These effects also emphasize that perception constitutes an interpretation of the sensory inputs in the light of temporal and spatial context.

INSERT ***Figure III-13*** AS FAR AS POSSIBLE FROM ***Figure III-12*** WHILE APPROXIMATELY IN THIS SECTION
Figure III-13. **Context matters in the real world too**. What is the object in the white box?

III.8. The value of experience

INSERT *Figure III-14* AROUND HERE
Figure III-14. **Color aftereffect**. Fixate on the center x without moving your eyes and count slowly to 30. Then move your eyes to a white surface. What do you see?

Our percepts are influenced by previous visual experience at multiple temporal scales. The phenomena that we have described so far, including the ability to discriminate animals from non-animals, to detect faces, the integration of spatially discontinuous object fragments, span temporal scales of tens to a few hundred milliseconds. We also considered two examples of temporal integration that also span tens to hundreds of milliseconds, priming, and backward masking.

Several visual illusions and phenomena show the powerful effects of temporal context on longer time scales spanning several seconds. One example is *visual adaptation*. A famous example of visual adaptation is the waterfall effect: after staring at a waterfall for about 30 seconds, shifting the gaze to other static objects, those objects appear to be moving upward. The visual system is adapted to downward movement, and things that are not moving appear to be moving

upwards, in the opposite direction to what we are adapted to. Adaptation is not restricted to motion. Similar after-effects can be observed after adapting to colors, textures, or objects like faces. For example, fixate on the x in the center in ***Figure III-14*** for about 30 seconds, then move your eyes to a white surface. You will experience an after effect: the white surface will appear to show blobs of color approximately at the same positions in the retina as the circles in ***Figure III-14*** but with complementary colors.

The role of experience in perception extends well beyond the scale of seconds and minutes. Even lifelong expertise can play a dramatic effect on how we perceive the visual world. For example, the interpretation of an image can strongly depend on whether one has seen that particular image before or not. In the first exposure to the so-called Dalmatian dog illusion, observers think that the image consists of a smudge of black and white spots. However, after recognizing the dog, subjects can immediately interpret the scene and spot the dog again the next time. Several similar images created by Craig Mooney are commonly used to assess the role of experience in perceptual grouping.

We could say that the naïve observer cannot interpret the Dalmatian dog image, but he/she can learn to understand the image. In this case, the learning process is quite fast: a brief explanation, or briefly tracing the contours of the dog immediately reveals the image's content. Interestingly, once the dog is recognized, the viewer can also interpret other parts of the image as well.

There are many other situations where images may seem unintelligible to the novice observer. You may have seen clinical images such as X-rays or magnetic resonance images. In many cases, those images may reveal nothing beyond strange grayscale surfaces and textures to the untrained brain (as a side comment, note that *the brain is trained* to interpret images, not the *eyes*, it does not make any sense to speak of an untrained eye!). However, an experienced clinician can rapidly interpret the image to come up with a diagnosis. Similarly, if you do not read Chinese, Chinese text may look like a collection of picturesque hieroglyphs.

Another aspect of how our experience with the world influences our perceptions is the interpretation of 3D structure from 2D images. Many visual illusions are based on intriguing 3D interpretations. For example, street artists create striking illusions that convey a stunning 3D scene when a 2D image painted on the street is seen from the right angle. Even when we know that these are illusions, they are so powerful that our brains, laden with years of experience, cannot help but send their top-down cognitive influences to enforce a robust perceptual experience.

Another example of how our pre-conceived experience-dependent notions with the 3D world influence what we see is the *hollow-face illusion*. A 3D face mask rotates in such a way that in certain angles, it appears convex, protruding towards

the viewer, whereas in other angles, it should be concave and appear hollow. However, the concave version is still perceived as a convex face by the viewer. There is a robust top-down bias to interpret the face as convex, probably because we rarely, if ever, encounter concave hollow versions of a face.

Faces have always been a particularly fascinating domain of study for psychologists. Understanding and identifying faces is prone to the same experience-dependent effects as other visual stimuli. For example, psychologists have characterized the "other-race" effect whereby it is harder for people to identify faces from races that they do not have experience with. For example, imagine someone born in Asia who has not had contact with the western world either in person, or in movies, or via any other format. Western faces would all look similar to that person. The converse is also true: western people who have not been exposed to many Asian faces may find it difficult to discriminate among them. As another example switching away from faces, a shepherd who has spent years tending to sheep may be quite good at identifying individual sheep, whereas they may all look similar to the naïve observer.

III.9. People are approximately the same wherever you go, with notable exceptions

In the previous sections, and most of the psychophysics literature, we imagine a generic adult individual as a prototypical subject to discuss properties of human vision. To a good first-order approximation, the basic observations described so far hold regardless of the person's gender, skin color, religion, cultural background, even age, except for the first few years of life. People see the world in approximately the same way wherever we go.

There are exceptions to this rule. One exception discussed in the last section is due to the role of experience. Doctors may see structure when examining an X-ray image, and a shepherd can identify individual sheep. Other obvious exceptions include cases where the hardware is different or malfunctions. For example, as discussed in **Chapter II**, many males only have two types of cones in the retina. Several other distinct eye conditions have been described, including amblyopia (reduced vision in one eye) and nystagmus (repetitive, uncontrolled eye movements). Many people require corrective glasses to fix problems in accommodation by the eyes' lens. Albinism also leads to vision challenges under bright lighting conditions. As we will discuss in **Chapter IV**, there are also cortical lesions that lead to abnormal visual perception.

Age matters too. As people age, accommodation by the eye lens might change, some people develop cataracts, others suffer macular degeneration (***Figure II-8***). Infants and very young children also see the world differently, not only because of their expertise with the world but also because their visual system is not fully mature. Humans are not born with their fully developed visual system. The visual acuity of a newborn is approximately between 20/200 and 20/400, which means that what they see 20 feet away is comparable to what adults see at 200 or 400

feet. In the US, a person with a visual acuity of 20/200 or less is considered to be legally blind.

Once we take out all these factors, let us consider two people of approximately the same age, with approximately the same visual experience, without any visual deficit. How different are their perceptions of the world? Recently, there has been increased interest in understanding individual differences in visual perception among normally sighted individuals. Although the general principles outlined in this chapter apply generally, there remains an interesting amount of variation in perception. An example of such variations has been recently brought to the forefront during the rather passionate discussion about the color of a dress (**Chapter I**, ***Figure I-7***). There was an approximately bimodal distribution of the color names used by people to describe the dress.

Additionally, there have also been studies documenting variability in other visual domains. For example, there is considerable variability in the abilities to recognize faces, with some people being particularly good and others particularly bad. Moving into higher psychological territory, beauty is in the brain of the beholder: there is considerable variation in visual aesthetic preferences.

III.10. Animals excel at vision too

In the next chapters, we will delve into the brain to enquire about the neural computations responsible for visual perception. It is easier to investigate the insides of non-human animals' (henceforth animals) brains rather than the human brain. Therefore, most of the discussion in the next three chapters will focus on animals' brains. The converse is true about behavior: it is easier to study visual behavior in humans than in other animals. This chapter has focused on human visual behavior. Before we scrutinize brain circuits, it is important to ask whether animals share the amazing properties of vision described so far.

Almost every existing animal species has capitalized on the advantages of visual processing, from flies to fish to birds to rodents and primates. Nocturnal animals like bats, coyotes, or mice, have a well-developed visual system. Many subterranean species like moles still have vision. A recent study of the so-called "blind" mole, presumed to be blind because the eyes are permanently closed under the skin during its entire life, has shown that they have rods, cones, and retinal ganglion cells that project to the rest of the brain. The investigators even showed that these moles have light directed behavior! There are a few animal species that are entirely blind, including some types of spiders, fish, and flatworms. However, blindness is the exception in the animal kingdom.

Diversity rules in Biology: there is an extraordinary repertoire of variations in the visual system. We cannot do justice here to the flamboyant arsenal of visual capabilities displayed in the animal kingdom. Animals have adapted to their niche and survival needs by evolving specialized uses for visual processing. We will only

mention a few examples of similarities and differences between vision in animals and humans.

Some properties of animal vision are distinct from human visual capabilities. Humans are limited to the visible part of the spectrum (defined as visible by humans!), whereas other species can sense ultraviolet light (e.g., mice, dogs, many types of birds) and also infrared light (e.g., many types of snakes). While (most) humans have three types of cones (**Chapter II**), the number of cones varies widely across the animal kingdom. Some species have only one type of cone (e.g., various bats and the common raccoon), other animals have two types of cones (e.g., cats and dogs), and there are even species with sixteen (the mantis shrimp) or up to thirty types of cones (some species of dragonflies). Cuttlefish can also sense light polarization, which humans cannot.

Even the number and position of eyes show wide variation. Spiders have between eight and twelve eyes, five-arm starfish have five eyes, and horseshoe crabs have ten eyes. The position of the eyes dictates what regions of the visual field are accessible to the animal. Snails have eyes in their tentacles; starfish have their eyes located in each of their arms. Even for species with two eyes, the position of the eyes plays a critical role in vision and shows variability. Approximately forward-facing eyes imply that the central parts of the visual field are accessible to both eyes, enabling the capability of estimating depth from stereopsis (the small difference in sampling between the two eyes). On the other hand, more laterally facing eyes provide a wider field of view. In an extreme example of laterally positioned eyes, rabbits have a blind spot at the center of the visual field. Humans have approximately 120 degrees of binocular field and a total visual field of approximately 180 degrees. In contrast, other animals with two eyes positioned so that they face the sides can have more than 300 degrees of visual field (e.g., cows, goats, horses). The two eyes in humans are essentially yoked together so that their positions are strongly correlated (except in certain conditions like amblyopia). In contrast, other species like the chameleon can move each eye independently, and they can therefore focus on two completely different locations in the visual field.

The resolution of the visual system also shows enormous diversity from animals like the starfish that represent the entire visual scene with approximately 200 pixels all the way to species like preying birds that surpass human acuity. Nocturnal predators have higher sensitivity than humans in low light conditions than humans (e.g., owls, tigers, lions, jaguars, leopards, geckos).

The ability to detect movement is perhaps one of the few universal properties of visual systems, probably a testament to the importance of responding to moving predators and preys, as well as to other imminent looming danger. Many species are specialized to rapidly detect motion changes. For example, wing movements triggered by visual stimuli can be evoked in dragonflies in about 30 ms after

stimulus onset, faster than the time it takes for information to get out of the human retina.

Thus, the human visual system, as amazing as it is, is certainly not unique. There exist multiple species that display "better" vision in terms of the ability to detect ecologically relevant features, where what is ecologically relevant depends on the species, of course. Our sense of vision largely dictates how we perceive the world around us. Without the aid of other tools, we are confined to an interpretation of the world based on our senses, and we are often arrogant or unimaginative enough to think that the world *is* precisely as we see it. The short list of visual system properties outlined above emphasizes that our view of the world is but one limited representation, that we can see things that others cannot, and vice versa. We are missing much exciting visual action in the world.

What about the perceptual properties described earlier in this chapter? What do the Gestalt rules look like for other species? Can animals also perform pattern completion? Deciphering what animals perceive is not an easy task and requires well-designed experiments and careful training. Monkeys, particularly rhesus macaque monkeys, constitute one of the main species of interest to study the visual system. Their eyes are quite similar to the human ones, and it is possible to train them to perform sophisticated visual tasks. Chimpanzees and bonobos have a visual system that is even more similar to the human one, but they have been less explored, particularly in terms of their brain properties.

Monkeys can be trained to perform multiple visual tasks, including discriminating the presence or absence of visual stimulation, reporting the direction of a moving stimulus, and detecting whether two stimuli are the same or not. Monkeys have been trained to discriminate complex objects, including faces as well as numeric symbols. They can trace lines and contours. They can even learn that the symbol 7 corresponds to 7 items on the screen, and is larger than the number 3. Monkeys can also learn to play simple video games.

How well can monkeys and other animals extrapolate to novel stimuli that they have not been trained on? For example, to what extent are their recognition abilities tolerant of the type of image transformations described earlier in this chapter (***Figure III-6***)? We can define multiple levels of increasingly more complex sophistication and abstraction in the ability to perform visual discrimination tasks: (1) discrimination, as in evaluating the presence or absence of a light source; (2) rote categorization, as in the ability to memorize a few exemplars within a class of objects and distinguish those exemplars from a few exemplars in a different class; (3) open-ended categorization, extending the previous ability to situations where there is an extensive and perhaps continuous number of exemplars within a category; (4) concepts, where animals can draw inferences across different exemplars; (5) abstract relations, dealing with relationships between exemplars as well as relations between concepts. Macaque monkeys do seem to be capable of a relatively sophisticated level of abstraction, including transformation-tolerant

visual categorization. After training with a set of visual object categories, their performance and pattern of mistakes resemble those of humans when tested under the same conditions. However, some tasks call into question how abstract monkeys' internal representations of the visual world are. For example, monkeys excelling at a visual discrimination task in the upper left visual hemifield may have to be re-trained extensively to perform the same task in the bottom right visual hemifield, whereas humans would rapidly transfer their learning across stimulus locations. Yet, this type of lack of extrapolation may not strictly reflect visual differences between species, but perhaps it is more related to task instructions and the communication between researchers and monkeys,

Over the last decade, there has also been increased interest in using rodents, particularly mice and rats, to investigate visual function. There are multiple exciting advantages and opportunities when considering the rodent visual system, including the number of individuals that can be examined, and the availability of an extraordinary repertoire of molecular tools. The type of visual discrimination tasks that rodents have been trained on is limited compared to the behavioral repertoire of macaque monkeys. However, rats do seem to be able to perform basic comparisons between visual shapes, even with some degree of extrapolation to novel renderings of the objects in terms of size, rotation, and illumination.

III.11. Summary

- *Psychophysics* is an exciting field that deals with quantifying behavior, including reaction time metrics, performance metrics, and eye movements.

- Brains make up stuff. Subjective perception is a construct that is constrained by sensory information in light of previous experience. Visual illusions illustrate the dissociation between sensory inputs and perception.

- The Gestalt rules of perception describe how we typically group image parts to construct objects. Such rules include closure, proximity, similarity, figure/ground separation, continuity, and common fate.

- Visual recognition performance shows tolerance to large transformations of an image.

- It is possible to make inferences from partial information, for example, during recognition of occluded objects.

- Visual recognition is fast. Many visual recognition questions can be answered in approximately 150 ms.

- Subjects can integrate information presented asynchronously but only over a few tens of milliseconds.

- Contextual information can help recognize objects.

- Humans are generally consistent with each other in their visual recognition abilities and visual perception. Yet, there is inter-individual variability, particularly when it comes to tasks requiring extensive prior experience.

- Animals excel at vision too, and it is essential to study animals in order to elucidate the mechanisms of vision.

3.1. Further reading

See **http://bit.ly/38buAhB** for more references.

- Cooper EE, Biederman I, Hummel JE. 1992. Metric invariance in object recognition: a review and further evidence. Can J Psychol 46: 191-214
- Eagleman DM. 2001. Visual illusions and neurobiology. Nat Rev Neurosci 2: 920-6
- Herrnstein RJ. 1990. Levels of stimulus control: a functional approach. Cognition 37: 133-66
- Nakayama K, He Z, Shimojo S. 1995. Visual surface representation: a critical link between lower-level and higher-level vision In Visual cognition, ed. S Kosslyn, D Osherson. Cambridge: The MIT press
- Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. Nature 381: 520-22

Figure III-1

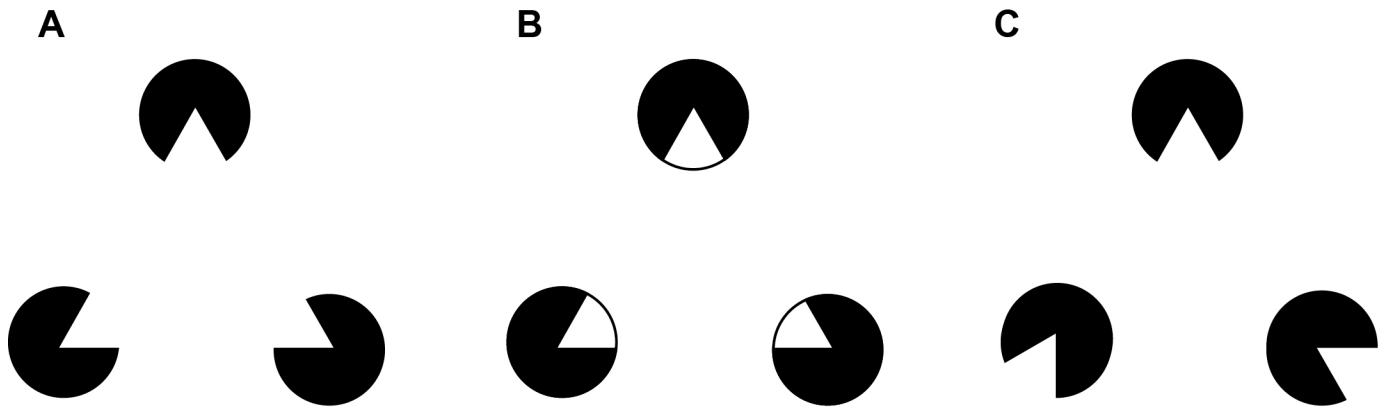


Figure III-2

Figure III-3

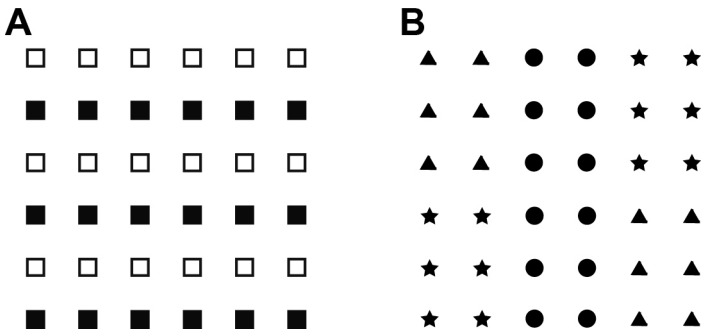


Figure III-4

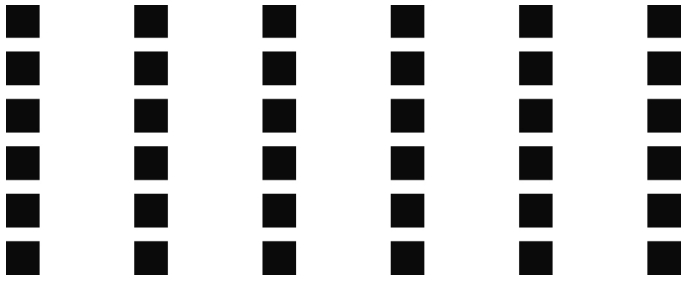


Figure III-5

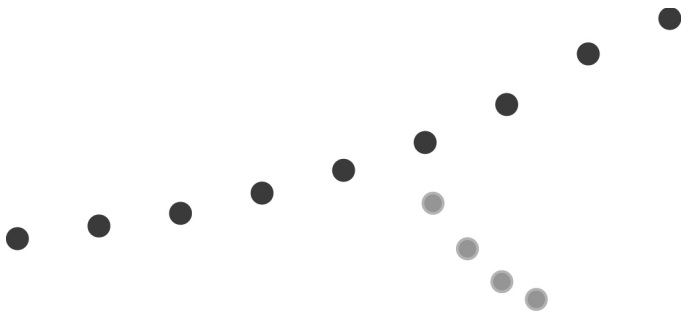


Figure III-6

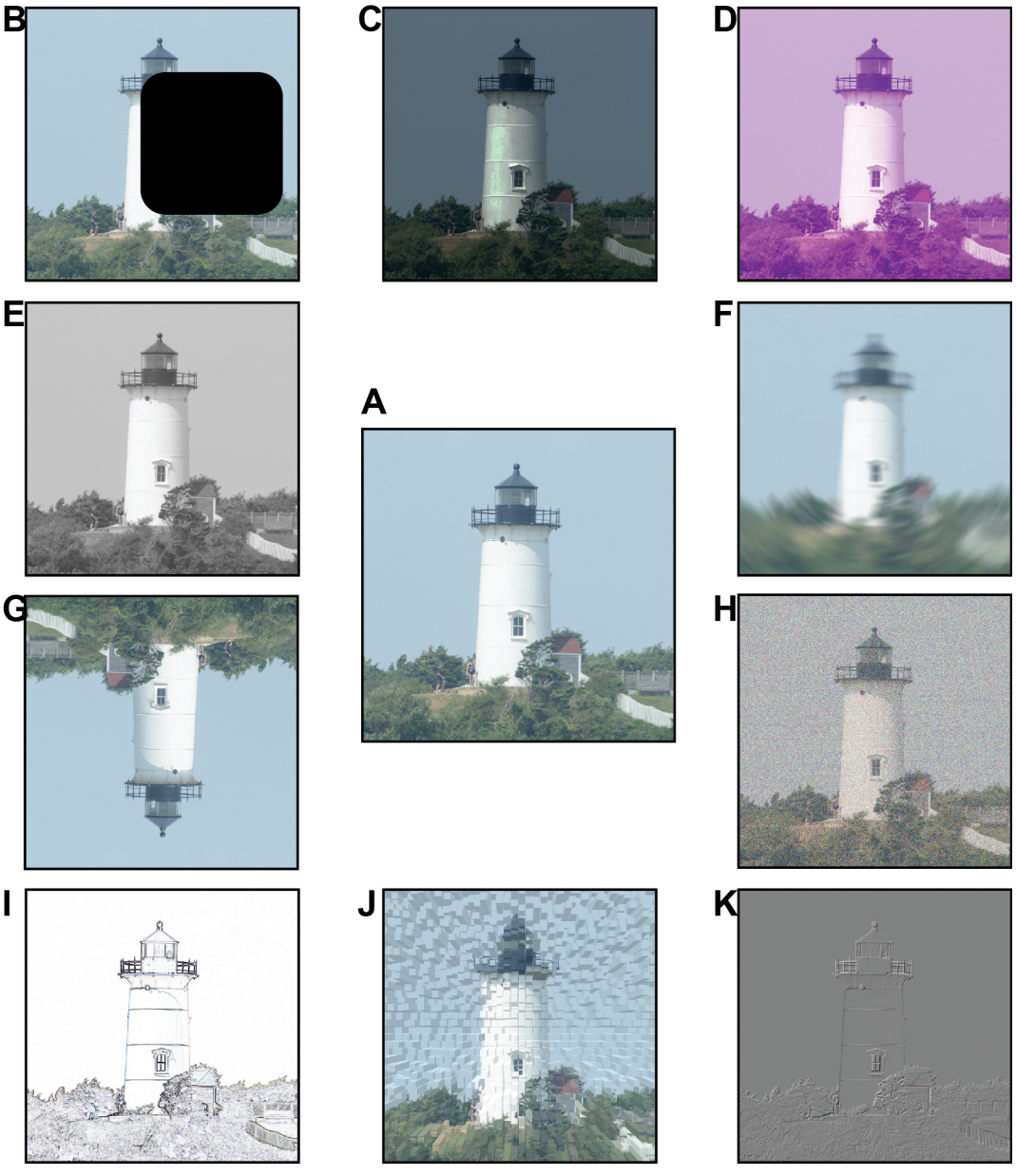
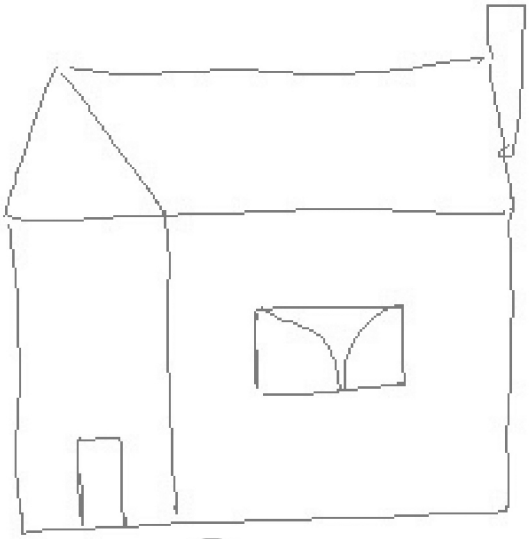
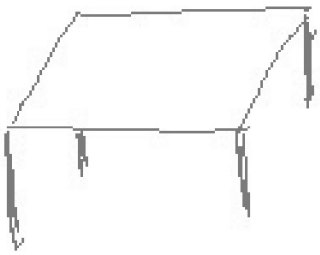
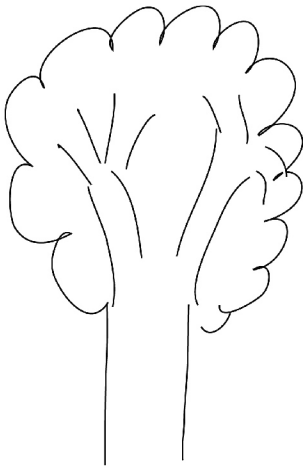
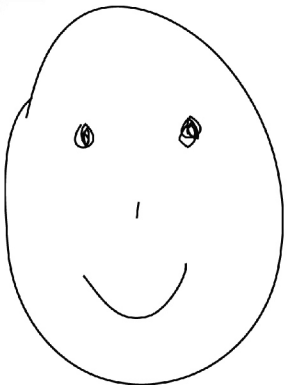


Figure III-7









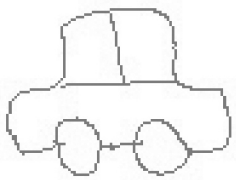


Figure III-8

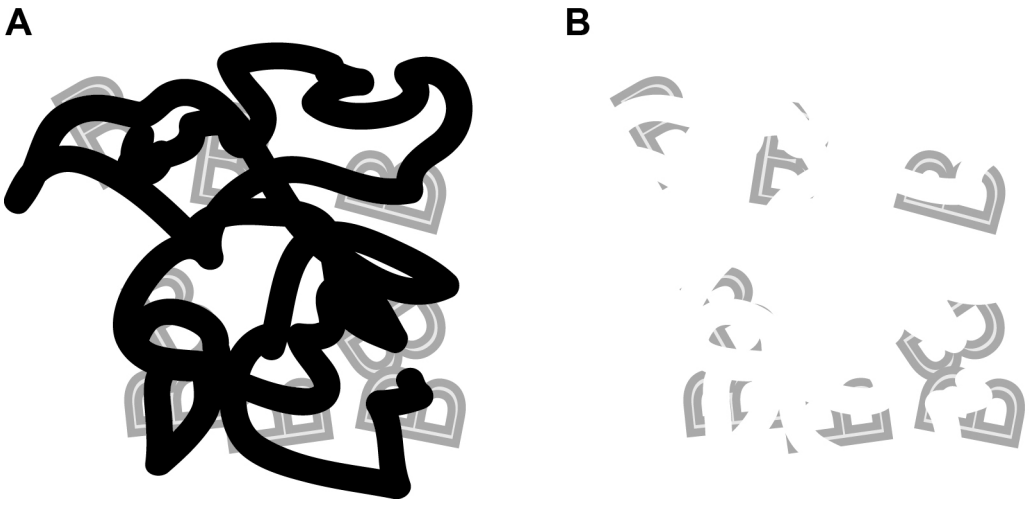


Figure III-9

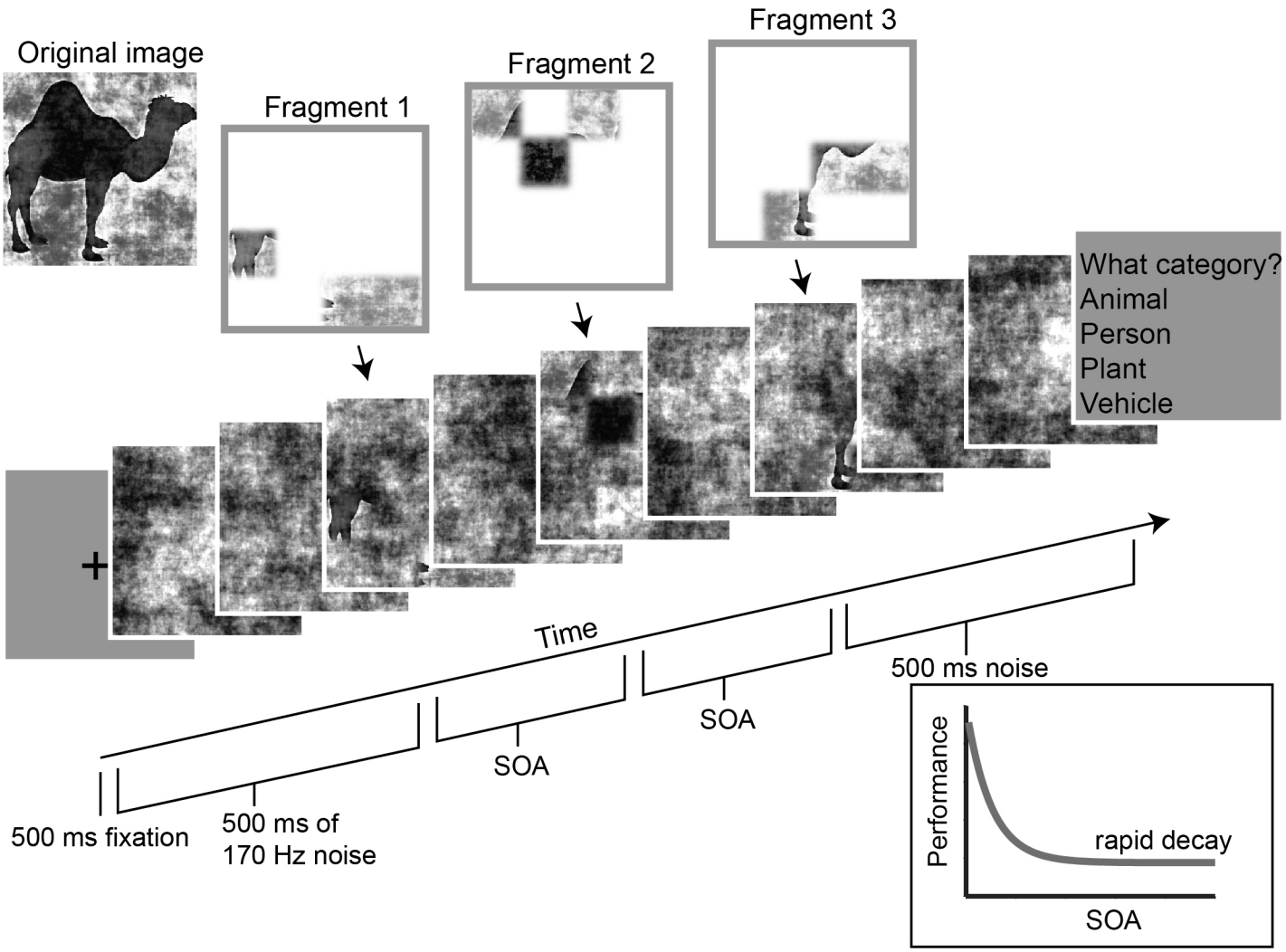

Original image
Fragment 1
Fragment 2
Fragment 3
What category?
Animal
Person
Plant
Vehicle
+
Time
500 ms fixation
500 ms of
170 Hz noise
SOA
SOA
500 ms noise
Performance
rapid decay
SOA

Figure III-10

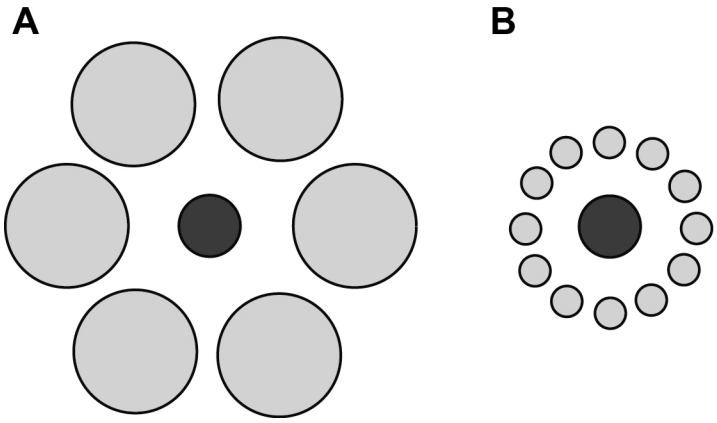


Figure III-11

Figure III-12

Figure III-13