Chapter II. The travels of a photon: Natural image statistics and the retina

Let there be light. And there was light. Vision starts when photons reflected from objects in the world impinge on the retina. Light is transduced into electrical signals at the level of the photoreceptors, one of the astounding feats of evolution, rapidly allowing the organism to make inferences about distant objects and events. The structure of the environment plays a critical role in dictating the pattern of connections and responses throughout the visual system and marks the beginning of our journey.

2.1. Natural images are special

Let us consider a digital grayscale image of 100 x 100 pixels, let us further restrict ourselves to a gray world where each pixel can take 256 shades of gray. Such small colorless image patches constitute a far cry from the complexity of real visual input. Yet, even under these constraints, there is an extremely large number of possible images. There are 256 possible one-pixel images. There are 256 x 256 possible two-pixel images. All in all, there are $256^{10,000}$ possible 100 x 100 images. This is a pretty big number; there are more of these image patches than the current estimate for the total number of stars in the universe.

Now take a digital camera, a rather old one with a sensor comprising only 100 x 100 pixels, turn the settings to gray images, and go around shooting random pictures. If you are very fast and shoot one picture per second, and if you spend an entire week without sleeping or eating, just collecting pictures in the city, at the beach, in the forest, or at home, you will have accrued less than a million pictures, a very tiny fraction of a percent of all possible images. Yet, you will note rather interesting patterns. It turns out that the distribution of *natural* image patches that you collected in the world tends to have peculiar properties that span an interesting subset of all possible images.

In principle, any of the $256^{10,000}$ grayscale patches could show up in the natural world. However, there are strong correlations and constraints in the way natural images look. A particularly striking pattern is that there tends to be a strong correlation between the grayscale intensities of two adjacent pixels (**Figure 2.1**). In other words, grayscale intensities in natural images typically change in a smooth manner and contain surfaces of approximately uniform intensity. Those surfaces are separated by edges that represent discontinuities, where such correlations between adjacent pixels break, and which tend to be the exception rather than the rule. Overall, edges constitute a small fraction of the image.

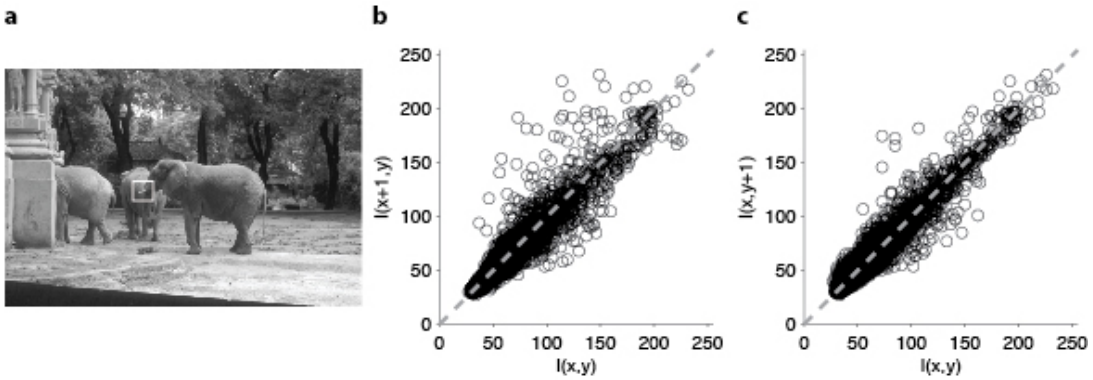


FIGURE 2.1. There is a strong correlation in the intensities of nearby pixels in natural images. For the small 100x100 pixel from the image in part **a** (white box), the scatter plots show the grayscale intensity in pixel with coordinates (x,y) versus the grayscale intensity in position (x+1,y) (**b**, horizontally adjacent pixel) or in position (x,y+1) (**c**, vertically adjacent pixel).

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

$$A(\Delta) = \int f(x) f(x-\Delta) dx$$

where the integral goes over the entire image. This definition can be readily extended to more dimensions and colored images. The autocorrelation function of a natural image typically shows a strong peak at small pixel separations followed by a gradual drop (for a review of the properties of natural images, see (Simoncelli and Olshausen, 2001)).

Another way of evaluating the spatial correlations in an image is to compute its power spectrum. Intuitively, one can convert those correlations from the pixel domain into the frequency domain. If there is a lot of power at high frequencies, that implies large changes across small pixel distances as one might observe when there is an edge. Conversely, a lot of power at low frequencies implies more gradual changes and smoothness in the pixel domain. If P denotes power and f denotes the spatial frequency, natural images typically show that power decreases with *f* approximately as

$$P \sim 1/f^2$$

There is significantly more power at low frequencies than at high frequencies. Such a function is called a power law. Power laws are pervasive throughout multiple natural phenomena and have interesting properties. One important property of power laws is scale invariance. If we change the scale of the image, its power spectrum will still have the same shape defined by the equation above.

2.2. Efficient coding by allocating more resources where they are needed

One of the reasons why we are interested in characterizing the properties of natural images is the conjecture that the brain is especially well adapted to represent the real world. This idea, known in the field as the *efficient coding principle*, posits that the visual system is particularly good at representing the type of variations that occur in Nature. If only a fraction of the $256^{10,000}$ possible image patches are present in any typical image, it may be smart to use most of the neurons to represent the fraction of this space that is occupied. Brain sizes are constrained by evolution and it is tempting to assume that they are not filled with neurons that encode images that would never show up. Additionally, brains are extremely expensive from an energetic viewpoint [REFERENCE], and it makes sense to allocate more resources where they are needed.

By understanding the structure and properties of natural images, it is possible to generate testable hypothesis about the preferences of neurons representing visual information (Barlow, 1972; Olshausen and Field, 1996; Simoncelli and Olshausen, 2001; Smith and Lewicki, 2006), a topic that we will come back to once we delve into the neural circuitry involved in processing visual information.

Such specialization to represent the properties of natural images could arise as a consequence of evolution (Nature) or as a consequence of learning via visual exposure to the real world (Nurture). As in other domains of the Nature versus Nurture dilemma, it seems quite likely that both are true. Certain aspects of the visual system are hard-wired, yet visual experience plays a central role in shaping neuronal tuning properties.

2.3. The visual world is slow

The visual properties of nearby locations in the natural world are correlated. In addition to those spatial correlations, there are also strong temporal constraints in the natural world. Expanding on the collection of natural world photographs, imagine that you go back to the same locations and now collect short videos while keeping the camera still. Because the camera is not allowed to move, the only changes across frames will be dictated by the movement of objects in the natural world. Assuming that you use a camera that captures about 30 frames per second, in most cases, adjacent frames in those videos will look extremely similar. With some exceptions, objects in the world move rather slowly. Consider a cheetah, or a car, moving at a rather impressive speed of 50 miles per hour. Assuming that we have a camera capturing about 40 yards in 2000 pixels, the cheetah will move approximately 30 pixels from one frame to the next. Most objects move at slower speeds. Therefore, the *temporal* autocorrelation of the natural world also shows a peak at short temporal scales spanning tens to hundreds of milliseconds.

Several computational models have taken advantage of the continuity of the input under natural viewing conditions in order to develop algorithms that can

learn about objects and their transformations (Foldiak, 1991; Stringer et al., 2006; Wiskott and Sejnowski, 2002), a theme that we will revisit when discussing computational accounts of learning in the visual system. The notion of using temporal continuity as a constraint for learning is often referred to as the "slowness" principle.

2.4. We are continuously moving our eyes

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

Nowadays, it is relatively straightforward to measure eye movements in a laboratory. Figure 2.2 shows an example of a sequence of eye movements during presentation of a static image. The eyes typically stay in one location, then rapidly jump to another location, exploring that location briefly, before adventuring again into a new location. The rapid jumps are denominated visual *saccades* and typically take a few tens of milliseconds to execute from initial position to final position [add reference here]. The positions in between saccades are called fixations.

The pattern of fixations depends on the image, temporal history and goals. The characteristics of the image influence eye movements: for example, high contrast regions are more salient and tend to attract eye movements. The temporal history of previous fixations is also relevant: on average, subjects tend to avoid returning to a location they recently fixated on, a phenomenon known as *inhibition of return*.

During scene perception, subjects typically make saccades of approximately 4 degrees of visual angle. Degrees of visual angle is the most relevant and common unit to measure sizes and positions in the visual field. One degree of visual angle approximately corresponds to the size of your thumb at arm's length. Under natural scene perception circumstances, subjects tend to make saccades approximately every 300 ms (Rayner, 1998).

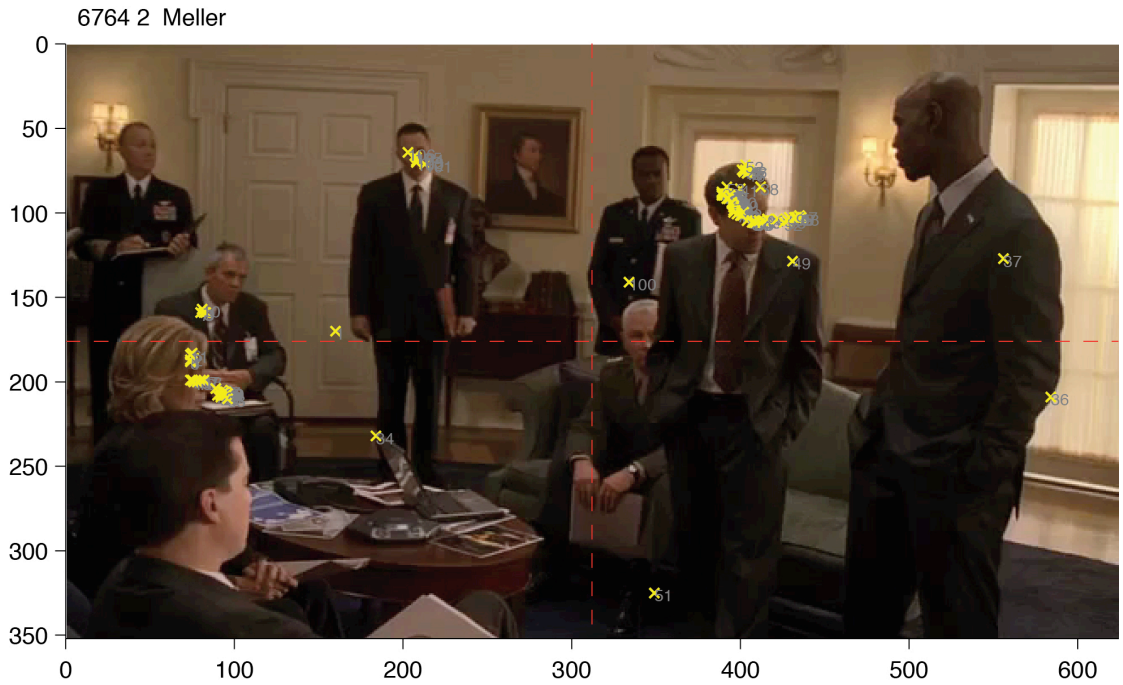
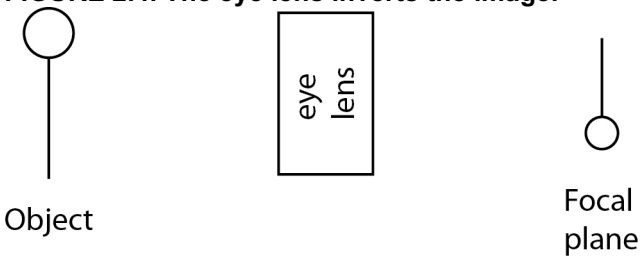


FIGURE 2.2. Pattern of fixations (yellow x) while a subject observes an image. Here we show the average eye positions averaged every 33 ms. The units are given in terms of pixels (100 pixels correspond to ~2 degrees of visual angle under these viewing conditions).

2.5. The retina

The adventure of visual processing in the brain begins with the conversion of photons into electrical signals in the retina (diminutive form of the word *net*, in Latin). The net of neurons in the retina is a particularly beautiful structure that has mesmerized Neuroscientists for decades. Due to its accessibility, the retina is the most studied part of the visual system. The retina is located at the back of the eye and has a thickness of approximately 500 μm. From a developmental point of view, the retina is part of the central nervous system. The retina encompasses an area of about 5x5 cm. A schematic diagram of the retina is shown in **Figure 2.3**, illustrating the stereotypical connectivity composed of three main cellular layers.

FIGURE 2.4. The eye lens inverts the image.



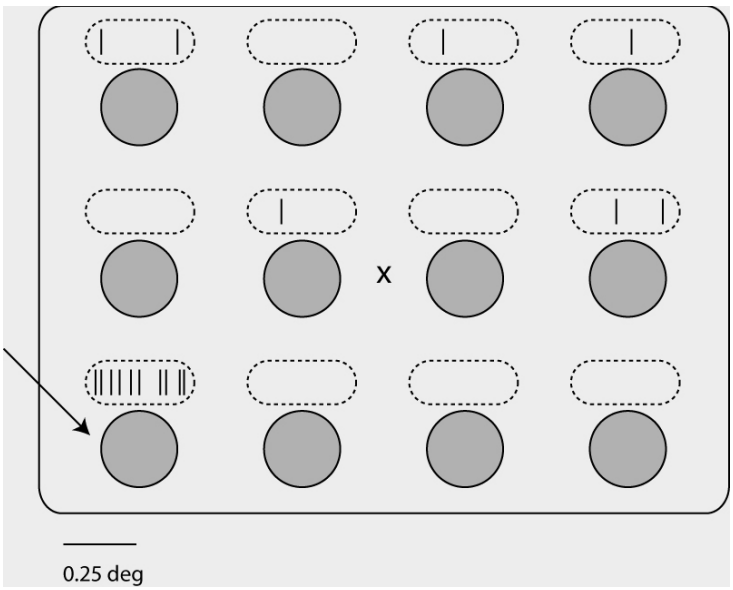


FIGURE 2.5. Mapping receptive fields. Neurons throughout the visual system typically respond to stimuli only when presented within a certain location in the visual field. Here the "x" stands for the fixation point, the circles indicate different stimulus locations and each vertical line denotes an action potential. The neuron fires vigorously when a stimulus is presented in the lower left corner (arrow) but not elsewhere.

FIGURE 2.3.
Schematic diagram of the cell types and connectivity in the primate retina. R = rod photoreceptors; C = cone photoreceptors; FMB = flat midget bipolar cells; IMB = invaginating midget bipolar cells; H = horizontal cells; IDB invaginating diffuse bipolar cells; RB = rod bipolar cells; I = interplexiform cell; A = amacrine cells; G = ganglion cells; MG = midget ganglion cells. Reproduced from Dowling (2007), Scholarpedia, 2(12):3487.

Photoreceptors come in two main varieties: rods and cones. There are about 10^8 rods; these cells are particularly specialized for capturing photons under low-light conditions. Night vision depends on rods. There are about 10^6 cones specialized for vision under bright light conditions. There are three types of cones depending on their wavelength sensitivity. Color vision relies on the activity of cones. There is extensive biochemical work characterizing the signal transduction cascades responsible for converting light into electrical signals by photoreceptors (Yau, 1994).

There is a special part of the retina, called the fovea, that is specialized for high acuity. This ~500 μm region of the retina contains a high density of cones (and no rods) and provides a finer sampling of the visual field, thereby providing subjects with higher resolution at the point of fixation (~1.7 degrees). For example, our ability to read depends on the fovea (try fixating on a word without moving your eyes and reading five words away).

There is a part of the visual field projection in each eye, denominated the blind spot, which does not map onto to photoreceptors. The easiest way to detect the blind spot is to close one eye and slowly move a small object in the opposite hemifield until the object disappears. Under normal circumstances, we are not aware of the blind spot, i.e., we have the subjective feeling that we can see the entire field in front of us (even with one eye closed). This is because the brain fills

in and compensates for the lack of receptors in the blind spot. This fill-in process introduces the notion that our visual perception is a constructive process whereby our brains build an interpretation of the outside world. We will return to the notion of vision as a subjective construction when we discuss visual consciousness.

Similarly, the eye lens inverts the image (upside down and left/right, **Figure 2.4**). This basic fact of Optics sometimes puzzles those who reflect about perception for the first time. Why don't we see everything upside down? Because visual perception (as well as other modalities) constitutes our brain's construction of the outside world based on the pattern of activity from neurons in the retina. Our brains learn that a certain pattern of activation is right side up. In fact, it is possible to teach the brain to adapt to different images with different rules, for example, by wearing glasses that invert the image (Stratton, 1896).

The beauty of the retinal circuitry, combined with its accessibility for experimental examination and manipulations make it an attractive area of intense research. Photoreceptors connect to bipolar and horizontal cells, which in turn communicate with amacrine and ganglion cells. There is a large number of different types of amacrine cells and there is ongoing work trying to characterize the function of these different types of cells and their role in information processing. Similarly, there is variety in the type of ganglion cells and how these cells respond to different light input patterns. Whereas rods, cones, bipolar and horizontal cells are non-spiking neurons, ganglion cells do fire action potentials and carry the output of retinal computations.

2.6. Receptive fields

The functional properties of ganglion cells have been extensively examined by electrophysiological recordings that go back to the prominent work of Kuffler (Kuffler, 1953). Retinal neurons (as well as most neurons examined in visual cortex so far) respond most strongly to a circumscribed region of the visual field called the receptive field (**Figure 2.5**). Two main types of ganglion cell responses are often described depending on the region of the visual field that activates the neurons. "On-center" cells are activated whith light input in the center of the receptive field and they are inhibited by the presence of light input in the borders of the receptive field. The opposite holds for "off-center" ganglion cells. Some ganglion cells are also strongly activated by the direction of motion of a bar within the receptive field. In addition to these spatial properties, most neurons respond with a strong transient upon stimulus onset and the response rate decays over time. Although it seems that vision happens very fast, information is not propagated instantaneously; it takes several tens of milliseconds to elicit a response at the level of retinal ganglion cells in the retina.

2.7. The lateral geniculate nucleus (LGN)

The retina projects to a part of the thalamus called the lateral geniculate nucleus (LGN). The retina also projects to the superior colliculus, the pretectum, accessory optic system, pregeniculate and the suprachiasmatic nucleus among other regions. Primates can recognize objects after lesions to the superior colliculus but not after lesions to V1 (see {Gross, 1994 #90} for a historical overview). To a good first approximation, the key connectivity involved in visual object recognition involves the pathway traveling to the LGN and to cortex.

Throughout the visual system, as we will discuss later, there are massive backprojections. One of the few exceptions to this claim is the connection from the retina to the LGN. There are no connections from the LGN back to the retina. The thalamus has been often succinctly (and somewhat unfairly) called the "gateway to cortex". This nomenclature advocates the idea that the thalamus is a relay area involved in controlling the on-off of the visual information conveyed to the cortex. This is likely to be only an oversimplification and the picture will change dramatically as we understand more about the neuronal circuits and computations in the LGN.

Six distinct layers can be distinguished in the LGN. Layers 2, 3 and 5 receive ipsilateral input. Ipsilateral input means that the right LGN receives input from the right eye. Layers 1, 4 and 6 receive contralateral input. Therefore, the input from the right and left visual hemifields is kept separate at the level of the input to the LGN. Layers 1 and 2 are called magnocellular layers and receive input from M-type ganglion cells. Layers 3-6 are called parvocellular layers and receive input form P-type ganglion cells. There are about 1.5 million cells in the LGN.

While we often think of the LGN predominantly in terms of the input from retinal ganglion cells, there is a large number of back-projections, predominantly from primary visual cortex, to the LGN (Douglas and Martin, 2004). To understand the function of the circuitry, in addition to the number of inputs, we need to know the corresponding weights or synaptic influence for the different type of projections. Our understanding of the different types of receptive fields in the LGN is guided by the retinal ganglion cell input.

2.8. Quantitative description of center-surround receptive fields

The receptive fields for LGN cells are slightly larger than the ones in the retina. The responses of LGN cells are typically described a difference of Gaussians operator (**Figure 2.6**):

$$D(x,y) = \pm\left(\frac{1}{2\pi\sigma_{cen}^2} \exp\left[-\frac{x^2+y^2}{2\sigma_{cen}^2} \right] - \frac{B}{2\pi\sigma_{sur}^2} \exp\left[-\frac{x^2+y^2}{2\sigma_{sur}^2} \right] \right) \qquad \textbf{Equation 2.1}$$

The first term indicates the influence of the center and is characterized by the width σ_{cen}. The second term indicates the influence of the surround and is characterized by the width σ_{sur} and the scaling factor B. The difference between these two terms yields a "Mexican-hat" structure with a peak in the center and an inhibitory dip in the surround.

This static description can be expanded to take into account the dynamical evolution of the receptive field structure:

$$D(x,y,t) = \pm\left(\frac{D_{cen}(t)}{2\pi\sigma_{cen}^2}\exp\left[-\frac{x^2+y^2}{2\sigma_{cen}^2}\right] - \frac{BD_{sur}(t)}{2\pi\sigma_{sur}^2}\exp\left[-\frac{x^2+y^2}{2\sigma_{sur}^2}\right]\right) \quad \textbf{Equation 2.2}$$

where

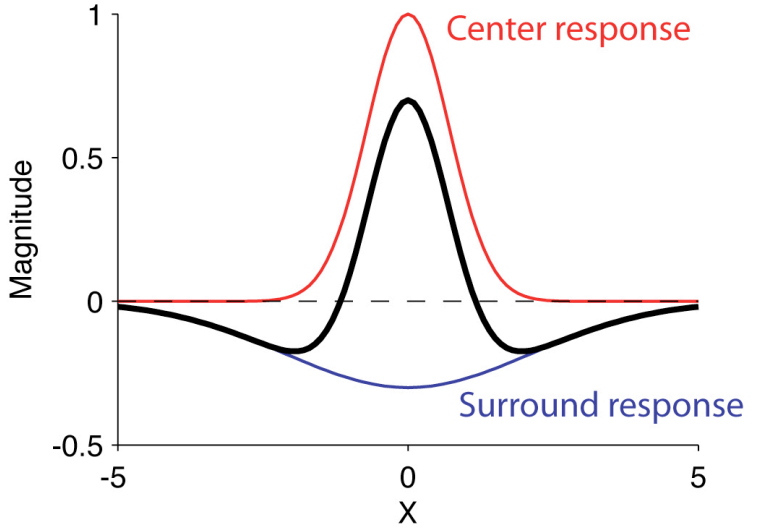


FIGURE 2.6. Mexican-hat receptive field. The receptive field in the retina and LGN is often characterized as a difference between a center response (red) and a broader and weaker surround response (blue), resulting in a "Mexican hat" shape (black).

$D_{cen}(t) = \alpha_{cen}^2 texp[-\alpha_{cen}t] - \beta_{cen}^2 texp[-\beta t]$ describes the dynamics of the center excitatory function and $D_{sur}(t) = \alpha_{sur}^2 texp[-\alpha_{sur}t] - \beta_{sur}^2 texp[-\beta_{sur}t]$ describes the dynamics of the surround inhibitory function (Dayan and Abbott, 2001; Wandell, 1995).

2.9. References

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

Dayan, P., and Abbott, L. (2001). Theoretical Neuroscience (Cambridge: MIT Press).

Douglas, R.J., and Martin, K.A. (2004). Neuronal circuits of the neocortex. Annu Rev Neurosci 27, 419-451.

Foldiak, P. (1991). Learning Invariance from Transformation Sequences. Neural Computation 3, 194-200.
Gross, C.G. (1994). How inferior temporal cortex became a visual area. Cerebral cortex 5, 455-469.
Kuffler, S. (1953). Discharge patterns and functional organization of mammalian retina. Journal of Neurophysiology 16, 37-68.
Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607-609.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychol Bull 124, 372-422.
Simoncelli, E., and Olshausen, B. (2001). Natural Image Statistics and Neural Representation. Annual Review of Neuroscience 24, 193-216.
Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. Nature 439, 978-982.
Stratton, G. (1896). Some preliminary experiments on vision without inversion of the retinal image. Psychological Review 3, 611-617.
Stringer, S.M., Perry, G., Rolls, E.T., and Proske, J.H. (2006). Learning invariant object recognition in the visual system with continuous transformations. Biol Cybern 94, 128-142.
Wandell, B.A. (1995). Foundations of vision (Sunderland: Sinauer Associates Inc.).
Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput 14, 715-770.
Yau, K. (1994). Phototransduction mechanism in retinal rods and cones. Investigative Opthalmology and Visual Science 35, 9-32.