

BEWARE: These are preliminary notes. In the future, they will become part of a textbook on Visual Object Recognition.

Chapter II. Starting from the very beginning. Visuals input and natural image statistics. The retina and thalamus.

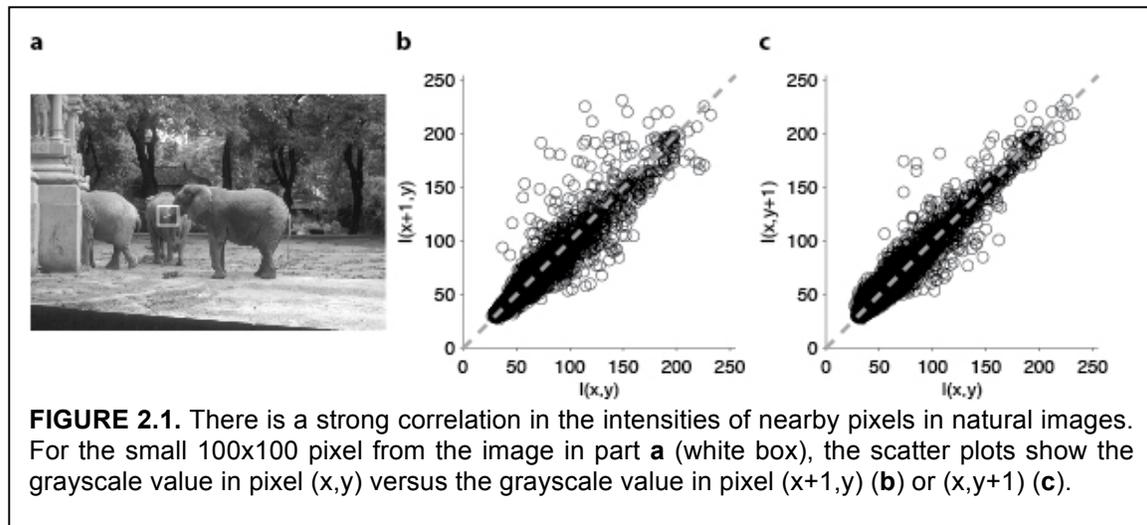
Let there be light. And there was light. Vision starts when photons reflected from objects in the world impinge on the retina. This light signal is transduced into electrical signals at the level of the photoreceptors, one of the astounding feats of evolution, allowing the organism to make inferences 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 image statistics

Let us consider a digital grayscale image of 100 x 100 pixels. This is a far cry from the complexity of real visual input. Yet, if each pixel can take 256 possible shades of gray, even for such a simple patch, there is a large number of possible images. There are 256 possible one-pixel images. There are 256x256 possible two-pixel images. All in all, there are $256^{10,000}$ possible 100x100 images. This is a pretty large number.

It turns out that the distribution of 100x100 pixel *natural* image patches that we encounter in the world includes only a small subset of this number. Before describing why this is so, let us reflect a minute on the meaning of “natural”. Imagine that we attach a digital camera to our forehead and go around a forest, a street or a beach, taking several pictures per second. We next extract all possible 100x100 pixel patches from those digital images. This gives a pragmatic definition of natural images and patches of natural images.

In principle, any of the $256^{10,000}$ patches could show up in the natural world. However, there are strong correlations and constraints in the way natural images look. First, there is 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 separated by edges that represent discontinuities. Overall, edges constitute a small fraction of the image. The



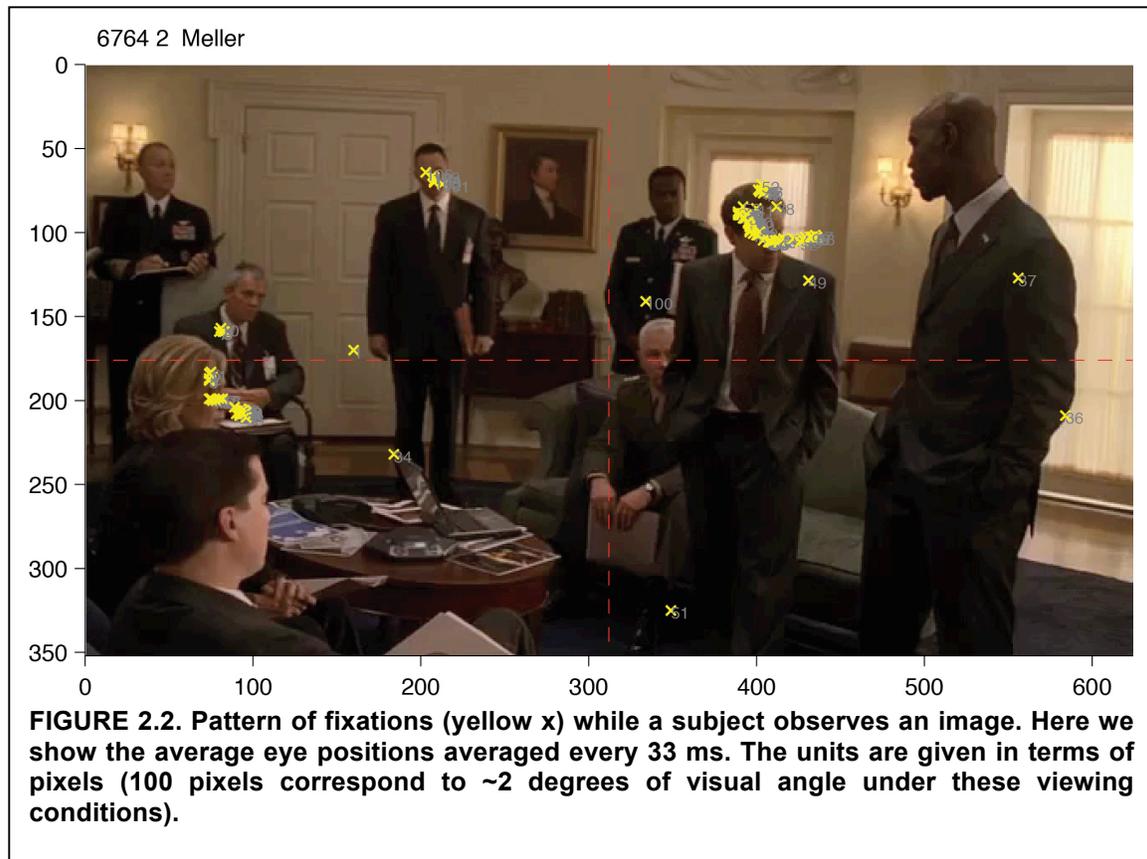
autocorrelation function¹ of a natural image typically shows a strong peak at small pixel separations followed by a gradual drop (for a review of the properties of natural images, see (Simoncelli and Olshausen, 2001)).

Another well-characterized property of natural images is the power spectrum. Typically, natural images show a power law that is approximately defined by a $1/f^2$ power spectrum where f is the spatial frequency. There is significantly more power at low frequencies than at high frequencies and the decay goes approximately as f^2 . Power laws are pervasive throughout multiple natural phenomena and have interesting properties such as scale invariance.

One of the reasons why we are interested in characterizing the properties of natural images is the conjecture that the brain (and the visual system in particular) has adapted to represent specifically 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. This idea is known in the field as the *efficient coding principle*. 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).

The eyes stay in one location, then jump to another location, and so on. It is interesting to note that these eye movements typically go unnoticed from an introspective viewpoint. Yet, humans (and other primates) are constantly moving their eyes, making several saccades per second. The pattern of fixations is dictated by the characteristics of the image (e.g. high contrast regions are more salient), by the history of previous fixations (e.g. on average, subjects tend to

¹ For a single valued function $f(x)$ defined over a domain D , the autocorrelation function is: $R(\lambda) = \int_D f(x)f(x + \lambda)dx$



avoid returning to a location they recently fixated on) and by behavioral goals (e.g. fixating on red objects while looking for a red car).

During scene perception, subjects typically make ~4 degrees² saccades every 260 to 330 ms (Rayner, 1998). 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 transformations³ (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.

2.2. 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

² One degree of visual angle is approximately equal to the size of your thumb when you extend your arm.

³ The notion of using temporal continuity as a constraint for learning is often referred to as the “slowness” principle.

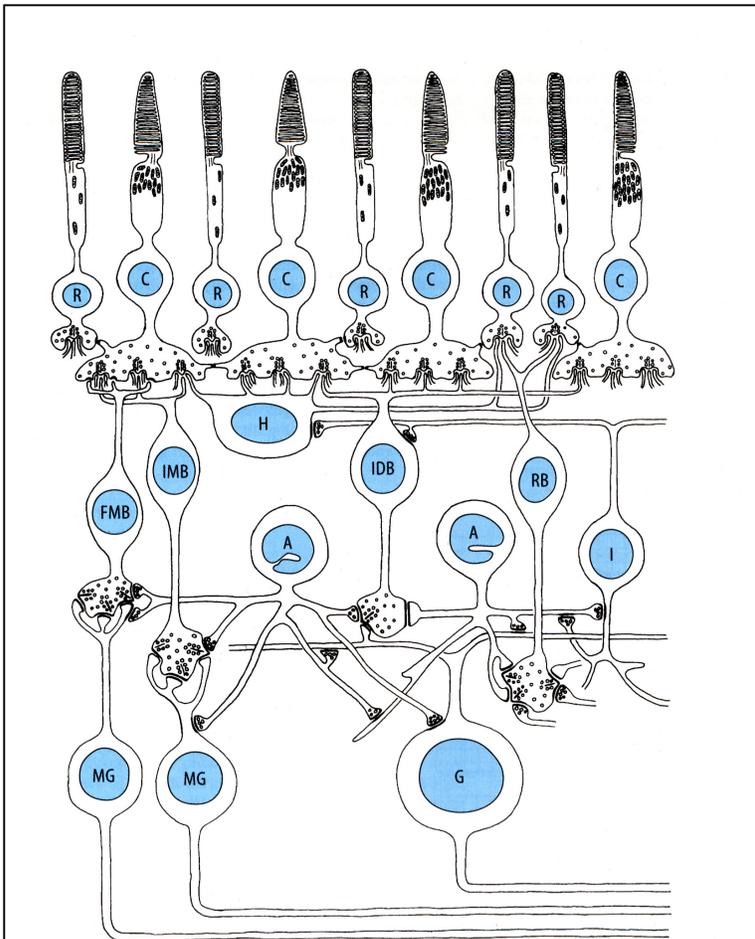


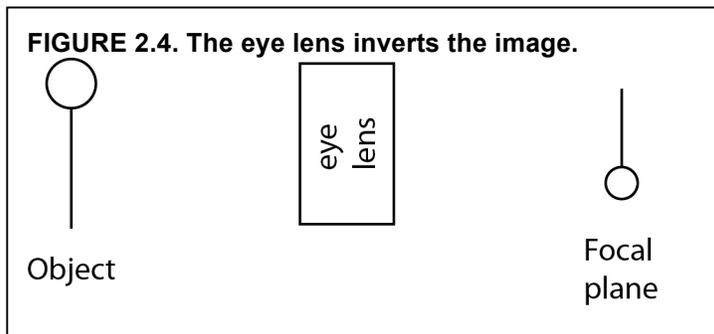
FIGURE 2.3. Schematic diagram of the cell types and connectivity in the primate retina. R = rod photoreceptors; C = cone photoreceptors; FMB = flat midget bipolar cells; IMB = invaginating midget bipolar cells; H = horizontal cells; IDB = invaginating diffuse bipolar cells; RB = rod bipolar cells; I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

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.

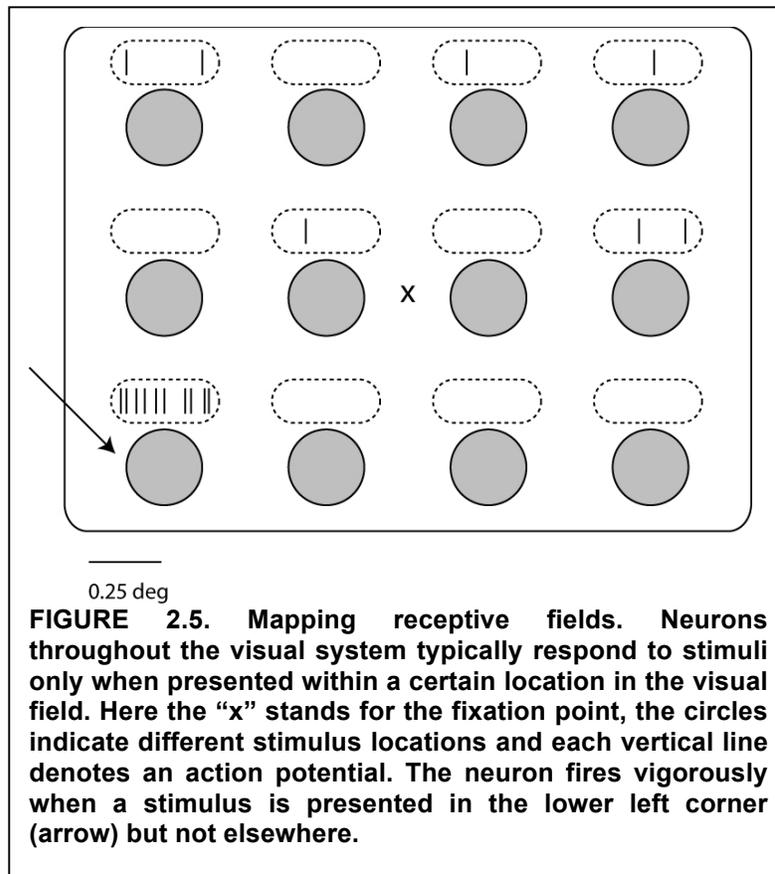
Photoreceptors come in two main varieties: rods and cones. There are about 10^8 rods; these cells are particularly specialized for capturing photons under low-light conditions. Night vision depends on rods. There are about 10^6 cones specialized for vision under bright light conditions. There are three 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 $\sim 500 \mu\text{m}$



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 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.

2.3. 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 with 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.

2.4. The lateral geniculate nucleus (LGN)

The retina projects to a part of the thalamus called the lateral geniculate nucleus (LGN)⁴. 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

⁴ 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, C.G. (1994). How inferior temporal cortex became a visual area. *Cerebral cortex* 5, 455-469. 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.

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 input⁵. 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 from 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.

2.5. 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**):

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

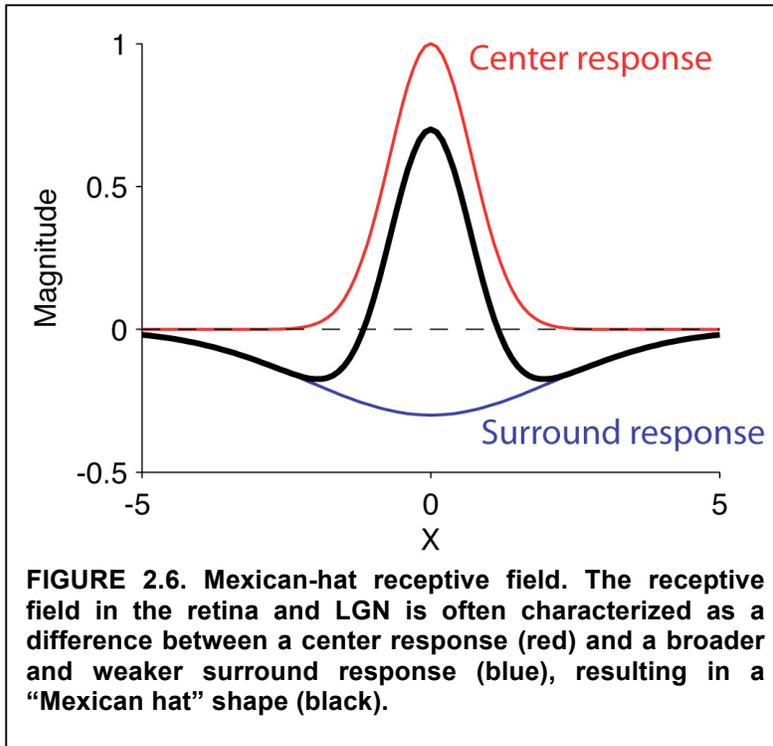
The first term indicates the influence of the center and is characterized by the width σ_{cen} . The second term indicates the influence of the surround and is characterized by the width σ_{sur} and the scaling factor B. The difference between these two terms yields a “Mexican-hat” structure with a peak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

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

⁵ Ipsilateral input means that the right LGN receives input from the right eye.

where



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

2.6. 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 Ophthalmology and Visual Science* 35, 9-32.