1 Principles and Consequences of the Initial Visual Encoding

Brian Wandell and David Brainard

Only infrequently is it possible to subject the manifold phenomena of life to simple and strict forms of mathematical treatment without forcing the data and encountering contradiction, probably never without a certain abandonment of the immense multiplicity of details to which those phenomena owe their aesthetic attractiveness. Nevertheless, however, it has often proved to be possible and useful to establish, for wide fields of biological processes and organic arrangements, comparatively simple mathematical formulas which, though they are probably not applicable with absolute accuracy, nevertheless simulate to a certain approximation a large number of phenomena. Such representations not only offer preliminary orientation in a field that at first seems completely incomprehensible, but they also often direct research into a correct course, in as much as first an insight into those fundamental formulations is sought, and then the deviations from their strict validity, which become apparent here and there, are made the subject of special investigations. Among the fields of physiology which have permitted the establishment of such guiding formulas the theory of visual sensations and of color mixture assumes a particularly distinguished position. (von Kries, 1902)

1.1 Introduction

Vision research has many purposes. Medical investigators aim to diagnose and repair visual disorders ranging from optical focus to retinal dysfunction to cortical lesions. Psychologists aim to identify and quantify the systematic rules of perception, including models of visual sensitivity, image quality, and the laws that predict percepts such as brightness, color, motion, size, and depth. Systems neuroscientists seek to relate visual experience and performance to the neural signals in the visual pathways, and computational investigators seek principles and models of perceptual and neural processes. Image systems engineers ask how to design sensors and processing to provide effective artificial vision systems.

Vision science draws upon findings from many fields, including biology, computer science, electrical engineering, neuroscience, psychology, and physics. Clear communication among people trained in different disciplines is not always straightforward. One of the ways that vision science has flourished is by using the language of mathematics to communicate core ideas. Vision science uses many types of mathematics; here we describe methods that have been used for many decades. These are certain linear methods, descriptions of noise distributions, and Bayesian inference. Many other linear methods (e.g., principal components, Fourier and Gabor bases, and independent components analysis) and nonlinear methods (e.g., linear–nonlinear cascades, normalization, information theory, and neural networks) can be found throughout the vision science literature. For this chapter, we focus on a few core mathematical methods and the complementary role of computation.

Physics – the field that quantifies the input to the visual system – provides mathematical representations of the light signal and definitions of physical units. The field of physiological optics quantifies the optical and biological properties of the lens. These properties are summarized as a mathematical transformation that maps the physical stimulus to the image focused on the retina, generally referred to as the retinal image. At each retinal location the image is characterized as the spectral irradiance (power per unit area as a function of wavelength). Retinal anatomy and electrophysiology identify the properties of the rod and cone photoreceptors, enabling us to calculate the photopigment excitations from the retinal image using linear algebraic methods.

Perhaps the most famous use of mathematics in vision science is at the intersection of physics and psychology: the laws of color matching formalize the relationship between the physics of light and certain aspects of color appearance. The mathematical principles of color matching are also deeply connected to Thomas Young's biological insight that there are only three types of cone photopigment (Young, 1802). This insight implies a low-dimensional biological encoding of the high-dimensional spectral light. The linear algebraic techniques used to describe the laws of color matching were developed by the mathematician Hermann Grassmann. Indeed, he developed vector spaces in part for this purpose (Grassmann, 1853). The mathematics he introduced remains central to color imaging technologies and throughout science and engineering.

While acknowledging the importance of mathematical foundations, it is also important to recognize that there is much to be gained by building computational methods that account for specific system properties. The added value of computations is clear in many different fields, not just vision science. The laws of gravity are simple, but predicting the tides at a particular location on earth is not done via analytic application of Newton's formulas. Similarly, that color vision is three-dimensional is a profound principle, yet precise stimulus control requires accounting for many factors, such as variations of the inert pigments across the retinal surface (CIE, 2007; Whitehead, Mares, & Danis, 2006) and the wavelength-dependent blur of chromatic aberration (Marimont & Wandell, 1994). The mathematical principles guide, but we need detailed computations to predict precisely how color matches vary from central to peripheral vision.

We hope this chapter helps the reader value principles expressed by equations and computations embodied in software. Establishing the principles first provides a foundation for implementing accurate computations. Historically, our knowledge about vision has been built up by developing principles, testing them against experiments, and combining them with computation; this remains a useful and important approach. Indeed, we believe the goal of vision science includes not only producing models that account for performance and enable engineering advances, but also leveraging those models to extract new principles that help us think about how visual circuits work.

There are competing views: some would argue that large data sets combined with analyses using machine learning provide the best way forward to understanding, and recent years have seen impressive engineering advances achieved with this approach (D. D. Cox & Dean, 2014). We are certainly interested in the performance of such models as a point of departure, but here we emphasize principles and data-guided computational implementations of these principles.

This chapter begins by describing the representation of the visual stimulus, and how light rays in the scene pass through the optics of the eye and arrive at the retina. Next, we explain how the retinal photoreceptors (a) transform the retinal spectral irradiance into photoreceptor excitations, and (b) spatially sample the retinal image. Each of these steps can be expressed by a crisp mathematical formulation. To describe the real system with quantitative precision, we implemented software that models specific features of the scene, optics, and retina (ISETBio; Cottaris *et al.*, 2019, 2020; https://github.com/isetbio/isetbio/wiki), and we illustrate the use of these models in several examples.

The frontiers of vision science use mathematics to understand visual percepts, which provide a useful basis for thought and action. The information provided by light-driven photopigment excitations is used to create these percepts, but knowledge of the excitations alone falls far short of describing visual perception. The brain makes inferences about the external world from the retinal encoding of light, and throughout the history of vision science many investigators have suggested that the role of neural computation is to implement the principles that underlie these inferences. This point was emphasized as early as Helmholtz, who wrote:

> The general rule determining the ideas of vision that are formed whenever an impression is made on the eye, is that such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism. (Helmholtz, 1866; English translation Helmholtz, 1896)

Within psychology this idea is called unconscious inference, a phrase that emphasizes that we are not aware of the neural processes that produce our conscious experience, an idea that was important to Helmholtz. Perhaps more important in this context is the principle that the percepts represent critical properties of external objects in the field of view, such as depth, reflectance, shape, and motion.

The mathematics of perceptual inference can take many forms, and in common scientific practice the mathematics of inference depend on what is known about the input signal. If the scene properties are not uniquely determined by the sensory measurements, such as when only three spectral classes of cones sample the spectral irradiance of the retinal image, probabilistic reasoning about the likely state of the world is inevitable. In vision science, linear methods combined with the mathematical tools of probabilistic inference are commonly used to understand how the brain interprets the mosaic of photoreceptor excitations to see objects, depth, and color. In the final part of this chapter we close the loop between sensory measurements and perceptual inference by introducing the mathematics of such inferences, focusing on two specific examples relevant to the study of the initial visual encoding. The principles we introduce, however, apply generally.

1.2 Scene to Retinal Image

1.2.1 Light Field

Light is the most important visual stimulus.[1] The word light means the electromagnetic radiation that is visible to the human eye.[2] The mathematical representation of light has been developed over many centuries through a series of famous experiments, and these experiments provide several different ways to think about light. Many properties of how light is encoded by the eye can be understood by treating light as comprising rays of many different wavelengths.

In a passage in his 1509 notebook (Da Vinci, 1970), Leonardo da Vinci noted that an illuminated scene is filled with rays that travel in all directions.[3] As evidence, he described a pinhole camera (camera obscura) made by placing a small hole in a wall of a windowless room (Figure 1.1). The wall is adjacent to a brightly illuminated piazza; an image of the piazza (inverted) appears on a wall within the room. Leonardo noted that an image is formed wherever the pinhole is placed, and he concluded that the rays needed to form an image must be present at all of these

Figure 1.1 *Light field geometry. The complete set of rays in the environment is the light field. The rays that arrive at the imaging system, in this figure a large pinhole camera, are the incident light field. If the imaging system includes a lens, rather than just a pinhole, the incident light field is described by the positions and angles of the rays at the lens aperture. Figure reproduced from Ayscough (1755).*

1 Mechanical force on the retina (pressure phosphenes) and injecting current into the retina or brain (electrical phosphenes) can also cause a visual sensation.
2 www.merriam-webster.com/dictionary/light
3 From the section prove how all objects, placed in one position, are all everywhere.

positions. Leonardo compared the space-filling light rays to the traveling waves that arise after dropping a rock in a pond.

The Russian physicist, Andrey Gershun, provided a mathematical representation of the geometry of these rays, which he called the light field (Gershun, 1939). The mathematical representation of the light field quantifies the properties of the light rays at each position in space [Equation (1.1)]. Each ray travels from a location (x, y, z) in a direction (α, β) and has a wavelength and polarization (λ, ρ). To know these parameters and the intensity of every ray is to know the light field at a given moment in time:

$$LF(x, y, z, \alpha, \beta, \lambda, \rho). \tag{1.1}$$

The light field representation does not capture some phenomena of electromagnetic radiation such as interference (waves) or the Poisson character of light (photon) absorption by the photoreceptors. Even so, the light field representation provides an excellent model to describe the ways in which light interacts with surfaces, and the geometric description of the light field is important in the mathematics of computer graphics, a technology that is important for illumination engineering, photography, and cinema (Pharr, Jakob, & Humphreys, 2016; Wald *et al.*, 2003, 2006).

1.2.2 The Incident Light Field

An eye – or a camera – records a small subset of the light field, those rays arriving at the pupil or entrance aperture. We call these the incident light field. In Figure 1.1 the dashed and solid lines are the light field and the solid lines are the incident light field. The natural parameterization of the incident light differs from the general light field. We can represent the incident light field using only the position (u, v) and angle (α, β) of the rays at the entrance aperture of the imaging system:

$$ILF(u, v, \alpha, \beta, \lambda, \rho, t). \tag{1.2}$$

Equation (1.2) also represents time (t) explicitly, which allows it to describe effects of motion both in the scene and by the eye.

1.2.3 Spectral Irradiance and the Plenoptic Function

The eye and most cameras do not measure the full incident light field. Rather, the rays are focused to an image at the retina or sensor, and the photodetectors respond to the sum across all directions of the image rays. To be explicit about this, Adelson and Bergen (1991) introduced the term plenoptic function, a simplified version of the incident light field, that was chosen to guide thinking about the computations carried out in the human visual pathways [their Equation (2)]. First, they approximated the eye as a pinhole camera; with this approximation all rays have the same entrance position $\mathbf{p}$. Additionally, the retina/sensor surface defines the direction ($\mathbf{d}$) of the rays that pass through the pinhole. For the pinhole case, specifying two angles of a ray at the pinhole is equivalent to specifying the location where a ray will intersect the retina/sensor surface, (r_x, r_y). Finally, Adelson and Bergen ignored polarization as unimportant for human perception. With these restrictions, the plenoptic function for human vision is simply the retinal spectral irradiance, over time (t):

$$E(r_x, r_y, \lambda, t; \mathbf{p}, \mathbf{d}). \tag{1.3}$$

In Equation (1.3) we have explicitly reintroduced position and direction, but these are often implicit [as in the formulation of Equation (1.2) above]. Understanding the progression from light field to incident light field to retinal spectral irradiance is useful for understanding how the information available for visual processing relates to the complete set of potential information that could be sensed by a visual system.

Adelson and Bergen note that by placing the pinhole at many different positions and viewing directions, we can estimate the full light field from the set of spectral irradiances. It is possible to be more efficient and estimate the incident light field by using a lens, rather than a pinhole, inserting a microlens array over the photodetector array and placing multiple detectors behind each microlens. Both cameras and microscopes have employed this technology to support depth estimation (Adelson & Wang, 1992) and control focus and depth of field in post-processing (Ng *et al.*, 2005). Cameras that estimate the full incident light field are not currently in wide use (Wikipedia contributors, 2021); but, the widely used dual pixel autofocus technology obtains a coarse measure of the incident light field (Canon U.S.A., Inc., 2017; Mlinar, 2016). This is accomplished by inserting a microlens array over pairs of photodetectors. With this design rays from, say, the left and right sides of the lens are captured by adjacent detectors. This coarse estimate of the light field is useful for setting the lens focus and estimating depth.

1.2.4 The Initial Visual Encoding

Computational models of the early visual pathways define a series of transformations that characterize how the incident light field becomes a neural response. In this chapter, we introduce the mathematics used to characterize the initial visual encoding in the context of the first few of these transformations (Figure 1.2; see also Brainard & Stockman, 2010; Packer & Williams, 2003; Rodieck, 1998; Wandell, 1995). We focus on the encoding of the spectral radiance by the photoreceptors – subsequent neural processing operates on this visual encoding.

A visual scene's light field is generated by the properties and locations of the light sources and objects, and how the rays from the light sources are absorbed and reflected by the objects. Here we consider the special case of scenes presented on a flat display, so that in the idealized case where the display is the only object and there are no other light sources, the full light field is determined just by the spectral radiance emitted at each location of the display. Elsewhere, we consider

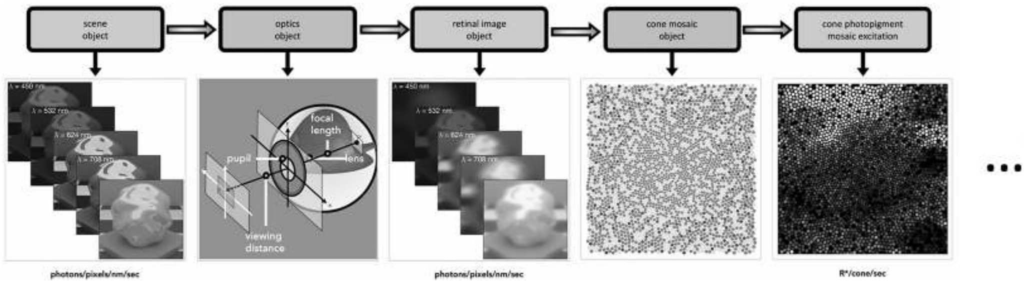

Figure 1.2 *The initial encoding of light by the visual system.* Scene*: An image on a display surface is characterized by the spectral radiance at each display location. Images of the display spectral radiance are shown at a few sample wavelengths, along with a rendering of the image.* Optics*: The incident light field enters the pupil of the eye and a spectral irradiance image is formed on the retina. The* retinal image *is blurred relative to the displayed image, and the spectral irradiance is affected by lens and macular pigment absorptions.* Cone mosaic*: The retinal image is spatially sampled by the L-, M-, and S-cone mosaics.* Cone excitations*: The retinal image irradiance, spectrally weighted by each cone photopigment absorptance function, is integrated within the cone's aperture and temporally integrated over the exposure duration to produce a pattern of cone excitations. This figure should be viewed in color. The color version is available at https://color.psych.upenn.edu/supplements/earlyencoding/computationsColorFig.pdf. We thank Nicolas Cottaris for the figure.*

the more general case of modeling the formation of the retinal spectral irradiance, given a description of the light sources and objects in a three-dimensional scene (Lian *et al.*, 2019).

The optics of the eye collect the incident light field and focus the rays to produce the spectral irradiance arriving at the retina. Factors such as diffraction and aberrations in the eyes optics mean that this image is blurred relative to the displayed image. In addition, wavelength-selective absorption of short-wavelength light by the lens and inert macular pigment also affect the spectral irradiance. Of note (but not illustrated in Figure 1.2), the density of the macular pigment is high in the central area of the retina and falls off rapidly with increasing eccentricity.

Photoreceptors spatially sample the retinal image. Excitations of photopigment molecules in these photoreceptors provide the information available to the visual system for making perceptual inferences about the scene. Here we consider the cone photoreceptors, which operate at light levels typical of daylight. There are three spectral classes of cones, each characterized by its own spectral sensitivity. That there are three classes leads to the trichromatic nature of human color vision. Figure 1.2 illustrates a patch of cone mosaic from the central region of the human retina. The properties of the mosaic are quite interesting. For example, there are no S-cones in the very center of the retina, and many properties of the mosaic (e.g., cone density, cone size, cone photopigment optical density) vary systematically with eccentricity (Brainard, 2015; Hofer & Williams, 2014).

Not considered here is a separate mosaic of highly sensitive rod photoreceptors that is interleaved with the cone mosaic. The rods mediate human vision at low light levels (Rodieck, 1998). We also ignore the melanopsin containing intrinsically sensitive retinal ganglion cells (Gamlin *et al.*, 2007; Hattar *et al.*, 2002; Van Gelder & Buhr, 2016). The principles we develop, however, also apply to modeling the excitations of these receptors.

Modeling of the initial visual encoding is well understood, and we explain the key linear systems principles next, using a simplified representation of the light stimulus. Advanced modeling of the subsequent neural processes includes non-linearities; the mathematical principles and computational methods we introduce are a fundamental part of the full description. After explaining the mathematical principles, we illustrate how to extend them through computational modeling that harnesses the power of computers to characterize biological reality in more detail than is possible with analytic calculations alone.

1.3 Mathematical Principles

1.3.1 Linear Systems

Linear systems and the tools of linear algebra are the most important mathematical methods used in vision science. Indeed, when trying to characterize a system, the scientist's and engineer's first hope is that the system can be approximated as linear. A system, L, is linear if it follows the superposition rule:

$$L(x+y) = L(x) + L(y). \tag{1.4}$$

Here x and y are two possible inputs to the system and $x + y$ represents their superposition. The homogeneity rule of linear systems follows from the superposition rule. Consider that

$$\begin{aligned} L(x+x) &= L(2x) \\ &= L(x) + L(x) \\ &= 2L(x). \end{aligned}$$

This is easily generalized for any integer m to show that:[4]

$$L(mx) = mL(x). \tag{1.5}$$

4 It is an exercise for the reader to show that a system that follows the superposition rule also obeys the homogeneity rule, not just for integers, but for any real scalar. If x is a real-valued scalar, homogeneity also implies superposition. When $\mathbf{x}$ is a real-valued vector with entries x_n, however, a system can obey homogeneity but not superposition. For example, $f(\mathbf{x}) = \sqrt[3]{\sum x_n^3}$ satisfies homogeneity but not superposition. The reader may find it of interest to consider why we used an exponent of three rather than two for this example.

No physical system can be linear over an infinite range – if you put enough energy into a system it will blow up! But many systems are linear over a meaningful range of input values.

1.3.2 Linearity Example: Cone Excitations and Color Matching

Vision is initiated when a photopigment molecule absorbs a photon of light. The absorption can cause the photopigment, a protein, to change conformation, an event we refer to as a photopigment excitation. The excitation initiates a molecular cascade inside the photoreceptor that changes the ionic currents at the photoreceptor membrane. The change in current modulates the voltage at the photoreceptor synapse and causes a release of neurotransmitter (Rodieck, 1998).

The transformation from the spectral energy of light, $E(\lambda)$, incident upon a cone to the number of photopigment excitations, n, produced by that light is an important, early vision, linear system. Consider two different spectra, denoted by $E_1(\lambda)$ and $E_2(\lambda)$. Let L represent the system that describes the transformation between spectra and excitations. This system obeys the superposition rule:

$$L(E_1 + E_2) = L(E_1) + L(E_2). \tag{1.6}$$

This linearity holds well over a wide range of light levels typical of daylight natural environments (Burns *et al.*, 1987).

An important feature of photopigment excitations is that their effect on the membrane current and transmitter release does not differ with the wavelength of the exciting photon. Such differences might have existed because different wavelengths are preferentially absorbed at different locations within the cone outer segment, or because photons of different wavelengths carry different amounts of energy. The observation that all excitations have the same impact is called the Principle of Univariance. As Rushton wrote:

> The output of a receptor depends upon its quantum catch, but not upon what quanta are caught. (Rushton, 1972)

The color-typical human retina contains three distinct classes of cones, which are referred to as the L (long-wavelength sensitive), M (middle-wavelength sensitive), and S (short-wavelength sensitive) cones. While the effects of photopigment excitations are univariant, the probability of a photopigment excitation is wavelength-dependent. The wavelength-dependent probability that an incident photon leads to an excitation is characterized by the pigment's spectral absorptance.[5] The absorptance depends on the density of the photopigment within the

5 The absorptance spectrum is the probability that a photon is absorbed. Not all absorbed photons lead to an excitation, so an additional factor specifying the quantal efficiency (probability of excitation given absorption) needs to be included in the calculation. Current estimates put the quantal efficiency of human cone photopigment near 67%. In addition, the calculation of cone excitations from spectral irradiance requires taking into account the size of the cone's light-collecting aperture.