CHAPTER

The Human Visual System

2

CHAPTER OUTLINE

Communicating Pictures. http://dx.doi.org/10.1016/B978-0-12-405906-1.00002-7

A study of the structure, function, and perceptual limitations of the human visual system (HVS) provides us with clues as to how we can exploit redundancy in visual information in order to compress it with minimum degradation in perceived quality. The way humans view and perceive digital images and video relates strongly to how we perceive the world in our everyday lives. While a full understanding of the HVS is still a long way off, we do know enough to be able to create the illusion of high quality in an image sequence where, perhaps, 199 out of every 200 bits from the original are discarded.

This chapter is intended to provide the reader with a basic understanding of the human visual system and, where possible, to relate its characteristics to visual redundancy and ultimately to a means of compressing image and video signals. We first consider the physical architecture of the HVS and the constraints it imposes on the way we see the world. We then review perceptual aspects of vision related to brightness, contrast, texture, color, and motion, indicating the physical limits of key visual parameters. Visual masking is a key component in perception-based compression, so we end by looking at how certain spatio-temporal visual features can be masked by certain types of content.

This chapter is primarily a review of the work of others who are much more expert in the biological, neurological, and psychological aspects of the subject than the author. For further information, the reader is directed to many excellent texts on this topic including those by Snowden et al. [1], Mather [2], Wandell [3], and Marr [4].

For colour versions of Figures 2.1, 2.11, 2.14, 2.16 and 2.17 please refer to the electronic version or the website.

2.1 Principles and theories of human vision

The vision system is the most powerful and complex of our senses and we have only began to understand it since the pioneering work of Helmholtz in the 1890s [5].

Between 40% and 50% of the neurons in the human brain are estimated to be associated with visual information processing. This compares to approximately 10% for hearing, 10% for the other senses, and 20% for motor functions, leaving 10–20% for everything else—things like playing chess [1]. This is not entirely surprising because visual activity is linked to almost everything we do. In basic terms we interpret colors, classify shapes, detect and assess motion, estimate distances, and do a pretty good job at creating a three-dimensional world out of the 2-D images that fall on our retinas.

So what is the difference between what we see and what we perceive? Quite a lot actually. For example, we fill in the blind spots where the optic nerve passes through the back of the eye, we correct the distorted imperfect images that fall on the retina, and ignore things that get in the way like our nose and the blood vessels in the eye. At a higher level, it is clear that the visual system does not exist to faithfully record images, but instead to interpret the world and survive in it—to recognize friends, to assess threats, to navigate, to forage, and to find mates. Humans have also used visual stimuli to entertain, communicate, and inform, probably over most of our 200,000 year history—from cave paintings to modern media.

A typical retina contains some 130 million light-sensitive rod and cone cells. Its neural processors are an extension of the visual cortex and it is the only part of the brain visible from outside the body. Each of the optic nerves, which carries signals from the eye to the brain, consists of around one million fibers and it is these that transmit the signals from the retina to the rest of the visual cortex. This contrasts starkly with other senses—for example, each auditory nerve comprises about 30,000 fibers.

We will examine the structure of the retina and the visual cortex in Sections 2.2 and 2.3, but first let us take a brief look at the theories that exist about how our visual system operates.

2.1.1 Theories of vision

Visual perception is highly specialized and complex. Much of the research historically into the biological and anatomical aspects of the vision system has been based on primate brains, which appear to be very similar to the HVS. More recently, however, technologies such as fMRI have revolutionized neuroscience and neuropsychology, providing ever increasing insight into the workings of the human system.

Until the late 1970s there were two main competing theories of vision—from the *Empiricists* and the *Rationalists*. Their views on the workings of the visual system differed significantly. Empiricists believed in innate, low level, bottom-up processing, allocating less importance to higher level functions. Julesz [6] for example showed that depth perception can arise from random dot stereograms with no high level meaning, thus inferring that this could not be driven by a top-down mechanism.

Empiricist models were largely based on the extraction of features which facilitated the generation of connected patterns. The pioneering work by Hubel and Wiesel [7] showed the orientation and scale selectivity of the low level visual system and Uhr and Vossler in 1961 demonstrated character recognition from such low level features. Empiricists however did not consider context.

In contrast, Rationalists focused on high level control of vision and the construction of percepts by the brain. They hypothesized mental models that are used to influence our search for and verification of lower level visual features [8]. Early work in this area developed *picture grammars* that described 2-D properties using 2-D structures [9]. These, however, described scene syntax rather than scene semantics and again did not show how 3-D visual inputs could be represented. This in turn led to a *scene-grammar* approach which attempted to map 3-D scenes to 2-D descriptors. This was achieved using either high level shapes, such as cubes, or by using low level structural elements, such as vertices and edges. Such models still failed to describe how humans can generate perceptions of unfamiliar objects since the models rely on prior knowledge of "grammars" to describe the visual input in a high level fashion.

From the Rationalist approaches emerged a hybrid *connectionist* approach. It was not until Marr's contribution in 1982 [4] that this approach was formalized. Connectionist theories emphasized interconnections between processing units embodying both the bottom-up (data driven) extraction of features combined with the top-down (interpretative) influence that generates perceptions based on this low level information. This can be strongly modulated by processes such as attention. This connectionist distributed processing type of approach is still influential and draws heavily on biological evidence providing close correlations with experimental data.

2.2 Acquisition: the human eye

Photons from the sun or other sources of light reflect off physical objects, and some of these photons enter our eyes. Our visual system senses energy radiated in the part of the electromagnetic spectrum from about 380 to 720 nm (420–790 THz) as shown in Figure 2.1. There is nothing particularly special about this range except that our physiology has adapted to it, and it has adapted to it because it is useful for things

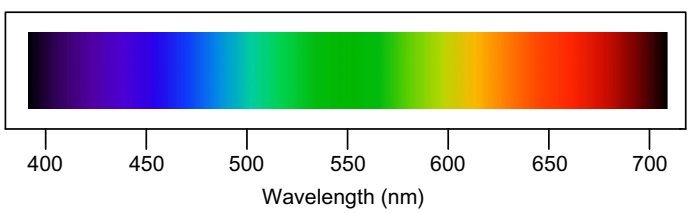

FIGURE 2.1

The visible spectrum.

like finding food and detecting threats and potential mates. Some animals see other ranges, for example many birds and insects see in the UV range as this enables them to see certain features in, for example, flowers. The human eye has a peak sensitivity around 555 nm in the green region of the spectrum—perhaps not a coincidence since the Earth's surface mostly reflects light in the green band from 550 to 560 nm due to foliage cover [1].

The human eye senses brightness approximately logarithmically over a broad range of luminance values and can also see colors that are not present in the spectrum. Unsaturated colors, such as pink, are not present and neither are colors like magenta which is formed from a combination of multiple wavelengths. A diagram of the human eye in cross-section is shown in Figure 2.2. Let us now look in a little more detail at its main components.

2.2.1 Retinal tissue layers

The eye is an approximately spherical volume filled with fluid, comprising four layers of tissue—the sclera, the retina, the choroid, and the ciliary body.

The sclera

The sclera is the layer of white tissue that provides the main strength of the eye's structure. It becomes transparent at the front of the eye where light enters through the cornea.

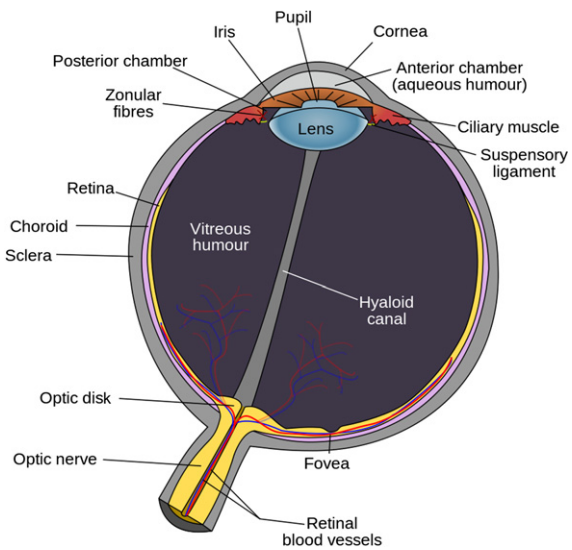

FIGURE 2.2
Cross-section of the human eye. (Public domain: http://commons.wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg.)

The ciliary body

The ciliary body is a ring of tissue that surrounds the eye. It contains the ciliary muscles that adjust the refractive power of the lens by changing its shape.

The retina

The retina contains the photoreceptors that are sensitive to light as well as several other types of neuron that process and combine signals from the photoreceptors and transmit these to the main brain areas via the optic nerve. A fundus photograph of a healthy human retina is shown in Figure 2.3. The macula is the darker area in the center and the optic disk, where the nerve fibers pass through the back of the eye, can be seen as the brighter area on the left. Interestingly, and despite its size, our vision system does a pretty good job of creating an impression of vision at this blind spot even though there are no photoreceptors present! Major nerve pathways are seen as white striped patterns radiating from the optic disk and blood vessels can also be clearly seen.

The choroid

The choroid is the capillary bed that provides nourishment to the photoreceptors. It also contains the light sensitive, UV absorbing pigment—melanin.

Let us now look at the front of the eye, where the light enters, in a little more detail.

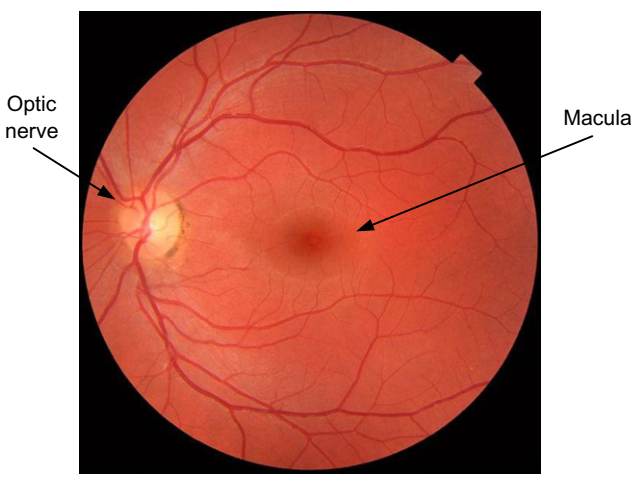

FIGURE 2.3

Fundus image of a healthy retina. (Public domain from: http://commons.wikimedia.org/wiki/File:Fundus_photograph_of_normal_right_eye.jpg.)

2.2.2 Optical processing

The cornea

The cornea is the sensitive transparent circular region at the front of the eye where light enters. It refracts the light onto the lens which then focuses it onto the retina.

The lens

The lens is the transparent structure located behind the pupil that, in combination with the cornea, refracts the incident light to focus it on the retina. For distant objects, the ciliary muscles cause it to become thinner and for close-up objects it becomes fatter. Interestingly the lens is only responsible for around 25% of the refractive power of the eye's optics, the rest is provided by the cornea. Although the lens has much less refractive power than the cornea, it enables changes in focal length that enable us to accommodate objects of interest at various distances.

The relationship between object distance d_o, retinal distance d_r and focal length f is illustrated in Figure 2.4 and equation (2.1). The focal length of the human eye is about 17 mm and the optical power is about 60 diopters.

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_r} \tag{2.1}$$

The iris

The iris is the individually colored circular region at the front of the eye containing the ciliary muscles that control the aperture of the pupil.

The pupil

The pupil is the aperture that controls how much light enters the eye and becomes smaller in brighter environments. The pupil however only adjusts by a factor of around 16:1 (dilation from 2 mm to 8 mm). Most of the compensation for varying light levels

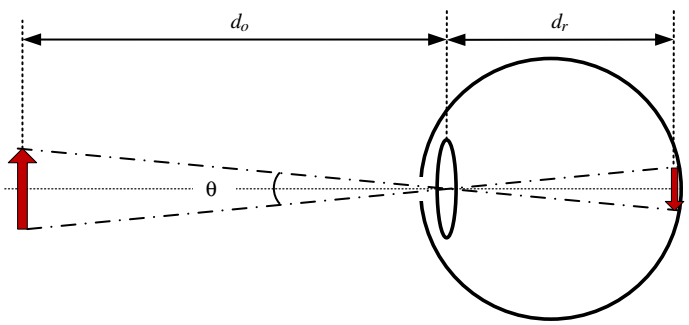

FIGURE 2.4

The focal length of the lens.

that allows us to adapt to some 8–9 orders of luminance magnitude is done in the retina by the photoreceptors and in other areas of the brain.

Now let us move to the transducer at the back of the eye—the retina.

2.2.3 Retinal photoreceptors and their distribution

When light hits the retina it is turned into electrical pulses by the responses of the photoreceptors. There are typically around 130 million sensor cells in the retina—125 million rods and 6–7 million cones. The distribution of these is as shown in Figure 2.5 [12]. Although the packing of cones is at its densest in the fovea, there are still several thousand per square millimeter in the periphery.

Rod cells

There are approximately 125 million rods in the retina. These are responsible for vision at low light levels (scotopic vision) and are so sensitive that they become saturated or bleached at mid light levels (mesopic vision). Rods do not mediate color vision and provide lower spatial acuity than foveal cone cells. The photopigment, rhodopsin, is most sensitive to green light of wavelength around 498 nm. The central fovea—the region of highest visual acuity—has a very low density of rods which is why, at night, we can see objects better by looking slightly away from them.

Cone cells

Cones operate at higher light levels than rods and provide what is known as photopic vision. In contrast to rods, they offer a range of different spectral sensitivity characteristics and they mediate color vision. Three types of cone exist and these

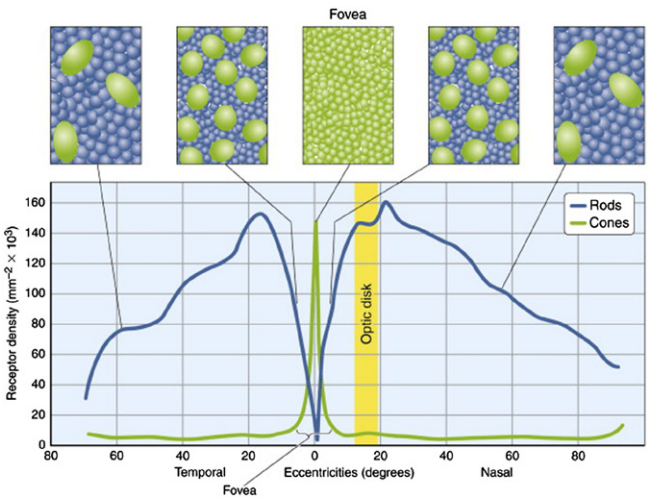

FIGURE 2.5

Photoreceptor distribution in the retina. (Reproduced with permission from: Mustafia et al. [12].)

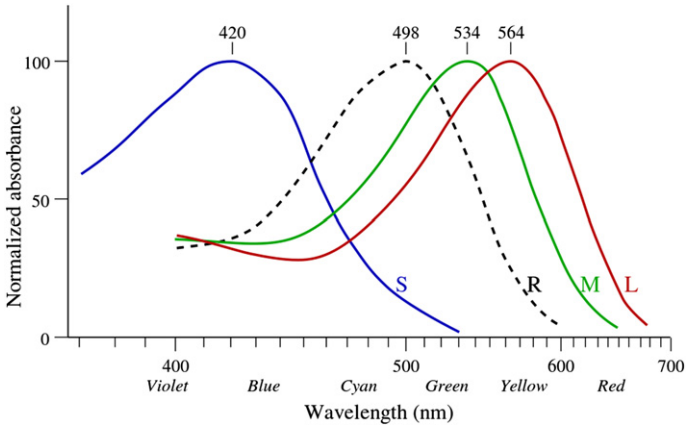

FIGURE 2.6

Normalized rod and cone responses for the human visual system. (Reproduced with permission from: Bowmaker and Dartnall [33]. //www.ncbi.nlm.nih.gov/pmc/articles/PMC1279132/. Avail Wikimedia Commons.)

are generally referred to as short wavelength (S) or blue; medium wavelength (M) or green; and long wavelength (L) or red. These have broad overlapping spectral responses with peaks at around 420, 534 and 564 nm respectively. Light at blue wavelengths tends to be out of focus since it is refracted more than red and green light. The normalized frequency response characteristics of the rods and cones is shown in Figure 2.6.

As shown in Figure 2.5, the central fovea has the highest cone density and thus provides the region of highest visual acuity.

Macula

The macula is a depressed region of diameter 5.5 mm near the center of the retina which surrounds the fovea. It is in this region that the number of cone cells starts to increase dramatically. It has two or more layers of ganglion cells, and at its center lies the fovea. Because of its yellow color the macula absorbs excess blue and ultraviolet light and acts as a natural sunblock.

Fovea

The fovea is the depression at the center of the macula about 1.5 mm in diameter that is responsible for central, high resolution vision. As shown in Figure 2.5, it has a very high spatial density of cones and very few rods. It also coincides with a region of the retina that is void of blood vessels and this further improves acuity as dispersion and loss are minimized. The center of the fovea, the foveola, is about 0.2 mm in diameter and is entirely formed of very compact, hexagonally packed, thin cones that are rod-like in structure. The fovea contains mostly M and L cones with around 5% of S cones.

The sampling resolution of the retinal image is highly non-uniform and falls off rapidly with increasing eccentricity. This creates a more blurred image in the peripheral vision which we do not perceive due to rapid eye movements and higher level visual processing. The effective angle subtended at the fovea is around 2°, hence we really do only "see" a small proportion of the visual scene sharply, yet we still perceive it all to be sharp. The average spacing of cone receptors in the fovea is 2–3 μm (in the periphery this increases to around 10 μm), and it is this that effectively limits our spatial resolution to approximately 1 arcmin. This is important when we assess the relative merits of high resolution TV formats in terms of viewing distance. The spacing also gives us the maximum spatial frequency that can be perceived. In practice the eye cannot resolve spatial frequencies of more than about 60 cycles per degree (cpd).

Optic disk and nerve

The optic disk is the portion of the optic nerve that can be seen on the retina, where the signals from the ganglion cells in the retina leave the retina along axons on their way to the visual cortex. About half of the nerve fibers in the optic nerve originate from the fovea with the remainder carrying information from the rest of the retina.

The retina contains a number of other cell types, which provide early vision processing and reduce the amount of information (HVS compression!) that is transmitted to the main visual cortex. These cells are briefly explained in Section 2.2.4.

2.2.4 Visual processing in the retina

The responses from the retina's photoreceptors are processed by a complex (and far from perfectly understood) network of neurons. This network comprises: bipolar, horizontal, amacrine, and retinal ganglion cells. Together, these provide early visual pre-processing, feature extraction and detection, resulting in significantly reduced information flow through the optic nerve. This structure is shown in Figure 2.7.

As depicted in Figure 2.7 bipolar cells connect to either rods or cones and are activated by an increase in the photons incident on the associated photoreceptors. The horizontal cells connect laterally across the retina to a greater spatial extent and, because of their stimulation from a neighborhood of photoreceptors, enable lateral inhibition. Whereas the response of bipolar cells in isolation is rather crude, the influence of the horizontal cells is to add an opponent signal where the response of one or more photoreceptors can influence the response of surrounding receptors, thus shaping the receptive field. The horizontal cells are also known to modulate the photoreceptor signal under different illumination conditions [10].

Relatively little is known about the roles of amacrine cells, except that they can have extensive dendritic trees and, in a similar way to horizontal cells, contribute feedback to the receptive fields of both bipolar and ganglion cells.

There are approximately 1.6 million retinal ganglion cells and these create the earliest receptive fields in the vision system, providing a basis for low level feature

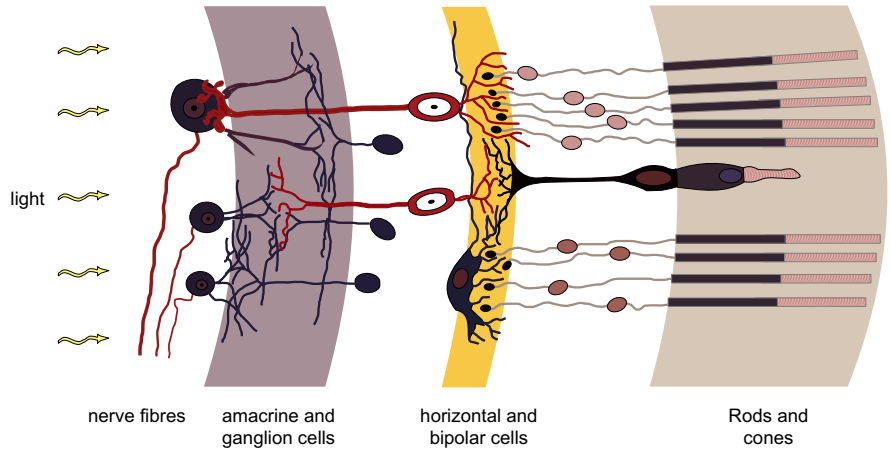

FIGURE 2.7
Retinal cell architecture (Public domain image adapted from http://commons.wikimedia.org/wiki/File:Retina-diagram.svg).

detection and the opponent coding of color (see Section 2.5). The axons from these form the optic nerve and transmit the electrochemical signals to the higher functions of the visual cortex. There are approximately 100 times more photoreceptors than ganglion cells in the eye—providing evidence of the compression performed by the retina.

The receptive field of a cell is a region within which a stimulus must occur if it is to influence the response of that cell. Receptive fields are formed using spatial opponency, i.e. responses from one part of the receptive field are opposed or inhibited by other parts of the receptive field. The interactions between the bipolar and the horizontal cells provide an orientation insensitive center-surround organization of the ganglion cell receptive field, where the center and surrounding areas act in opposition. This could comprise an excitatory center and an inhibitory surround as shown in Figure 2.8 or vice versa. Ganglion cells act as change detectors, responding only to differences in light intensities (or contrast) across their receptive field and not to absolute light levels. They are therefore sensitive to edges and can be modeled as a difference of Gaussian functions (DoG).

A scene is thus effectively encoded as intensity changes in the different color channels. However, it is clear that we see more than just edges, so this provides evidence that higher levels of the vision system perform in-filling between receptive field responses with averages in intensity levels. This is interesting as it can be viewed as a simple form of image synthesis, providing a biological analog to perceptual redundancy removal. A particularly striking example of this is the visual in-filling of our blind spot which contains no photoreceptors.

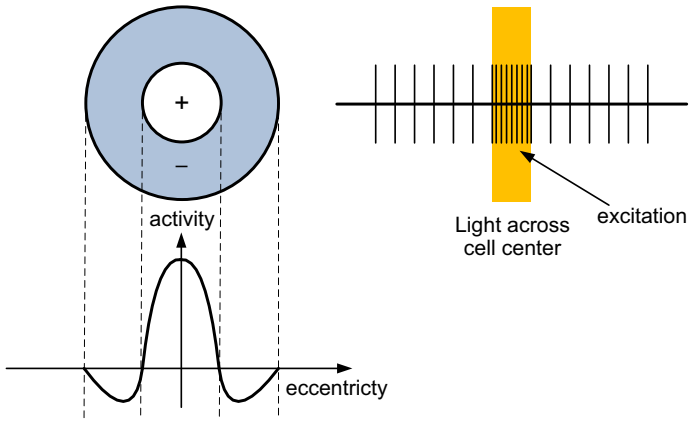

FIGURE 2.8
Spatial opponency, showing center-surround on cell and its firing pattern due to excitation.

2.3 The visual cortex

A simple diagram showing the major visual pathways from the retinas to the primary visual cortex is shown in Figure 2.9. The optic nerve axons from the retinal ganglion cells in each eye meet at the optic chiasm where all the fibers from the nasal side of each eye cross to join those from the temporal side of the other eye and form two optic tracts. The left visual field is thus processed by the right hemisphere and vice versa. The fibers in the optic tract then synapse at the lateral geniculate nucleus (LGN)

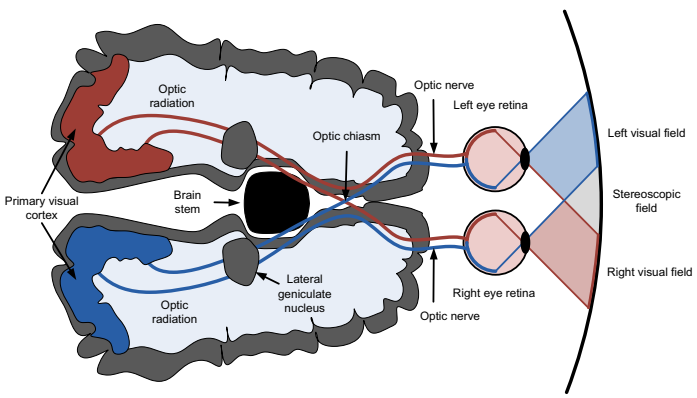

FIGURE 2.9
The visual cortex.

where they are distributed to other areas of the visual cortex. After the LGN, signals pass to the primary visual cortex, V1, located at the rear of the skull. They then pass to V2, and branch out to at least 20 other centers, each providing specialized functions such as detecting motion, recognising faces, or interpreting color and shape.

The HVS provides its functionality largely through the combined processes of inhibition, excitation, biased competition, input races, opponent coding, and adaptation. These are considered in more detail below.

2.3.1 Opponent processes

As discussed above and depicted in Figure 2.8, many neural functions in the vision system appear to encode data using opponent coding. Opponent coding is a process in which neurons encode opposite or opponent features in their single output. For example, a motion sensitive cell has a base firing rate which reduces when motion in one direction is detected and which increases with motion in the opposite direction. The opponent processes of inhibition and excitation appear crucial to function, with inhibitory responses often being used to control a normally excitatory state.

We also interpret color by processing signals from cones and rods in opponency. The L, M, and S cone spectral responses overlap in terms of their wavelengths so it is more efficient for the visual system to use differences between their responses rather than individual responses. This suggests that there are three opponent channels: red versus green, blue versus yellow, and black versus white. We will consider this further in Section 2.5.

2.3.2 Biased competition

Biased competition refers to the way in which information has to compete for limited resources in the visual pathways and cortices. Inputs from feature detectors in the visual system compete to be encoded into short term visual memory. The competition between such inputs is weighted according to higher level feedback, called attention, to ensure that those features relevant to the current task are prioritized. For example, during the Stroop task (subjects are asked to state the color of a series of words rather than read the words themselves) it has been observed that task-based attention causes enhancement of the relevant color feature sensitive areas of the brain while suppressing responses from those areas that perform word processing.

2.3.3 Adaptation processes

Adaptation is the process by which a cell's response reduces over time when its input is constant. Adaptation to varying degrees is common in most cortical cells and, when coupled with opponent coding processes, is responsible for the after-effects we experience related to brightness, color, or motion. After an opponent-coding feature-tuned cell adapts to a particular stimulus and that stimulus is removed, the cell reacts—briefly signaling the opposite feature to which it has adapted. Examples of such after-effects include the sensation of slowness after traveling at high speeds, and

the perception that static objects move in the opposite direction to that of a previously viewed moving object. Dramatic brightness and color after-effects are also frequently used as the basis for visual illusions. Even observing smiles for a prolonged period can make expressionless faces seem sadder!

2.3.4 V1—the primary visual cortex

The primary visual or striate (layered) cortex (otherwise known as V1) is probably the best understood area of the HVS. The role of V1 is to extract basic visual features from a scene and it has been shown that V1 contains receptive fields that are sensitive to line orientations, color, and spatial frequency. This effects visual data compaction through abstraction of visual stimuli into higher level constructs. The majority of V1 connections are from the retina but it is also known to receive feedback connectivity from higher functional areas, which influence its cell's receptive fields.

Hubel and Wiesel [7] identified three primary detectors in V1:

- **Simple cells:** Tuned to specific orientations of edges.
- **Complex cells:** Phase insensitive versions of simple cells, i.e. the response to an appropriately oriented stimulus is the same no matter where it falls within the receptive field.
- **Hypercomplex cells:** That show stimulus length sensitivity, i.e. the cell response increases as the stimulus length increases.

Neurons in V1 are generally grouped in columns that respond to a particular low level feature such as line orientation, spatial frequency, or color. V1 is known to preserve the retinotopic mapping of the ganglion cells on the retina in its organization. This is however not a linear or uniform mapping since V1 exhibits cortical magnification, dedicating a disproportionate percentage of its cells to processing foveal information. This emphasizes the importance of the higher resolution central part of the visual field.

2.3.5 V2—the prestriate cortex

V2 is strongly linked to V1 with both feedforward and feedback connections and it also connects strongly to V3, V4, and V5. Although less well understood than V1, V2 cells appear to operate in a similar manner, combining V1 features tuned to orientation, spatial frequency, and color as well as responding to more complex patterns. Damage in V2 can lead to poor texture and shape discrimination. There is also evidence of V2 cell modulation by binocular disparity.

2.3.6 Dorsal and ventral streams

Connections from V1 to the V2 area are generally considered to diverge into two streams—dorsal and ventral (otherwise referred to as "where" and "what" or "action" and "perception"). The dorsal stream relates to motion and depth perception, originating in V2 and passing dorsally through the brain to the motion sensitive parts of the visual cortex. The ventral stream on the other hand is associated with the perception

of shapes and object recognition, again originating in V2, but in this case passing ventrally through V4 to the inferior temporal cortex (IT).

2.3.7 Extrastriate areas

The areas of the visual cortex beyond V1 are generally referred to as the *extrastriate* areas. These possess larger receptive fields and generally compute more complex and abstract features. They also generally exhibit strong attentional modulation. Although the function of these areas is not well understood, their connectivity is informative.

V3 is strongly connected to V2 with additional inputs from V1. It processes both dorsal and ventral streams and has been shown to respond to global motions.

V4 lies on the ventral stream and has been shown to respond to shapes of relatively high complexity. It also shows strong attentional modulation and a significant amount of color sensitivity. Damage to V4 can cause impairments to shape and texture discrimination.

V5 contains cells that respond to complex motion patterns and is believed to be the main area dedicated to the perception of motion.

The *medial superior temporal* (MST) area is fed primarily from the medial temporal (MT) area, and is also sensitive to motion. There is evidence that the MST computes optic flow, especially in the context of global expansions and rotations.

The *visual short term memory* (VSTM) can be considered to be a short term buffer where prioritized data is stored for further processing. VSTM has been shown to persist across visual interruptions such as saccades and blinks and thus provides a link between the previously viewed scene and the new one after interruption.

The *inferior temporal cortex* (IT) provides complex shape recognition by parsing simpler shapes from lower levels in the HVS. It contains cells that trigger in response to specific shapes and patterns and hosts cells that are known to respond to faces. It fuses information from both halves of the visual field (that up to this point was separated across the hemispheres of the brain). At higher stages of visual processing, the receptive fields become larger and more complex in their connectivity. For example, in the IT some receptive fields are thought to cover the entire visual field.

Many other important areas of the visual cortex have been identified and investigated, including the *fusiform face area* (FFA), largely responsible for recognition of faces and other complex entities. The *frontal eye field* (FEF) is a further important area that controls voluntary eye movements. It has been proposed that this is based on some form of visual saliency map, encoded with behaviorally significant features.

2.4 Visual fields and acuity

2.4.1 Field of view

The field of view of the human eye is approximately 95° in the temporal direction (away from the nose), 60° in the nasal direction (toward the nose), 75° downward, and 60° upward, enabling a horizontal field of view in excess of 180°. Obviously,

when rotation of the eye is included, the horizontal field of view increases still further, even without head movement.

2.4.2 Acuity

It is not too difficult to estimate the spatial acuity of the retina based simply on the distribution of cones in the fovea. We illustrate this in Example 2.1. The measured Nyquist frequency for the HVS is normally between 40 and 60 cpd. This upper value has been attributed to the photoreceptor sampling grid pattern which is hexagonal rather than rectangular and to other local enhancement techniques. No aliasing effects are normally experienced by the HVS and this is due to the fact that the optics of the eye perform low pass filtering, acting as an anti-aliasing filter.

Example 2.1 (Visual acuity of the retina).
Assuming the photoreceptor densities indicated in Figure 2.5, and that the distance from the lens to the retina is 16.5 mm, estimate the acuity of the human visual system in the fovea.

Solution. We can see from Figure 2.5 that the cone density in the central fovea peaks at around $150{,}000$ cones/mm^2. If we assume a very simple packing arrangement this gives around 400 cones per linear mm. To be conservative let us take the distance between cone centers as 0.003 mm. The optics of the eye dictate [2] that, for a lens to retina distance of 16.5 mm, there are 0.29 mm/deg subtended. The angular resolution is thus $0.003/0.29 = 0.0103°$ between cone centers. Now, since we need at least two samples per cycle to satisfy the Nqyuist criterion, this means that there are $0.0206°$ per cycle. Hence the estimate for the maximum detectable spatial frequency is 49 cpd subtended.

2.4.3 Light, luminance, and brightness

Sensations of brightness are not direct analogs of photometric intensity (luminance). The number of photons that enter the eye and the perception of brightness are only related indirectly. For example, the apparent brightness of an achromatic target, under equi-luminant conditions, will depend on the frequency content of the electromagnetic radiation reflected from its surroundings. Thus, a gray target on a relatively dark background looks brighter than the same target on a lighter background. As a result of this context-based perception, it is generally thought that the visual system computes brightness using luminance ratios across the contrast boundaries in a scene. The influence of color on brightness also provides evidence that perceptions of luminance are related to electromagnetic frequency [11].

The influence of local contrast can be seen, in combination with the effects of lateral inhibition, in the Mach bands shown in Figure 2.10. Here the intensity is uniform over the width of each bar, yet the visual appearance is that each strip is

FIGURE 2.10

Mach band effect.

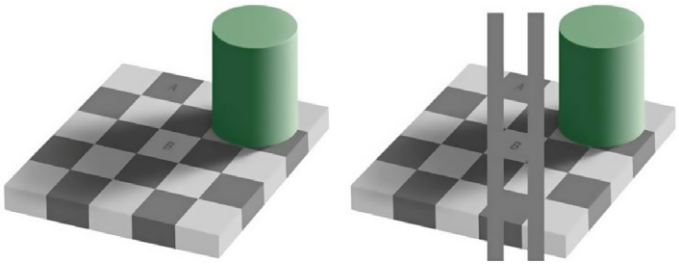

FIGURE 2.11

Adelson's grid. (Reproduced with permission from: http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html.)

darker at its right side than its left. Another striking example is provided by the Adelson grid as shown in Figure 2.11, where squares A and B both appear very different due to shadowing, yet are actually identical.

There is significant confusion about the terms used in connection with light radiation, so let us define some terms related to light and how we perceive it [13].

Radiant intensity and radiance

The radiant energy (joules) is the energy propagating from a source and the radiant flux Φ is the radiant energy per unit time (joules/sec or watts). Radiant intensity I is defined as the radiated light power in a given direction and is measured in terms of the amount of light passing through a solid angle (watts/steridian). Radiance L is defined as the light intensity that falls on a unit projected area (watts/steridian/m^2).

Luminance

None of the above measures take account of the composition of the light in terms of its wavelength or wavelengths. Luminance does this by weighting the radiance value according to the human visual response. The luminance, Y, is thus normally represented by equation (2.2), where $L(\lambda)$ is the incident light intensity as a function of wavelength at a given position and time and $\bar{y}(\lambda)$ is a wavelength-dependent weighting (or relative luminous efficiency) that reflects the perceived luminance of the incident

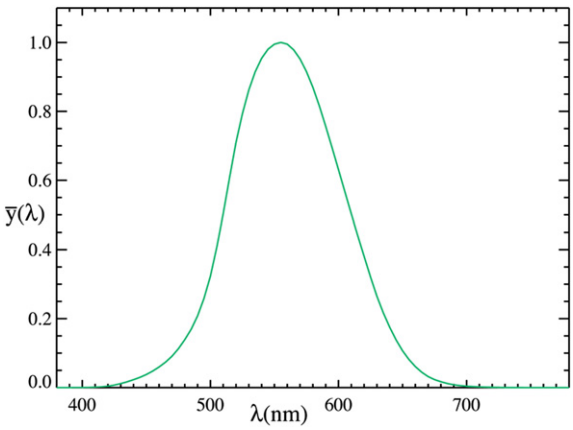

FIGURE 2.12
CIE luminous efficiency curve. (Public domain image: http://en.wikipedia.org/wiki/File:CIE_1931_Luminosity.png.)

flux. Luminance is measured in candelas/m^2 (cd/m^2):

$$Y = K \int_{320}^{720} L(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda \tag{2.2}$$

The constant $K = 685$ lumens/W. The function $\bar{y}(\lambda)$ will vary from person to person, but was measured by the CIE[1] in 1929 for a standard observer as shown in Figure 2.12. If in the integral of equation (2.2), the term $L(\lambda)$ is replaced by radiant flux $\Phi(\lambda)$ then the result is termed luminous flux and has the units of lumens.

Brightness

Brightness is the perception elicited by the luminance of a visual target and, as such, represents a non-quantitative indication of the physiological sensation of light. It is normally assessed subjectively by an observer with reference to a white area adjusted to have the same brightness as the target viewed.

Luma

An image sensor normally produces output values proportional to radiance, with the incident light typically filtered into R, G, and B bands. However, before these emerge from the camera, they are usually non-linearly mapped according to the response of the HVS, using a technique known as gamma correction. These outputs are normally labeled as R', G', and B'. Luma is the term used to denote the combination of these corrected components that produces the gamma-corrected luminance signal, according to the specification of the color space used. Gamma correction is considered further in Chapter 4.

[1]The International Commission on Illumination.

2.4.4 Light level adaptation

Our ability to adapt to a wide range of light levels is critical for function and survival. The photoreceptors, indeed the whole of the visual system, must remain sensitive as the ambient light intensity varies. This variation is quite dramatic as we are able to see in conditions from starlight (10^{-3} cd/m^2), through indoor lighting (10^2 cd/m^2) to bright sunlight (10^5 cd/m^2). The eye can thus function (at least to some degree) across some eight orders of magnitude in luminance level—a ratio of 100,000,000:1.

The eye, of course, cannot cope with this dynamic range instantaneously. If we consider our eye's instantaneous dynamic range (where the pupil opening is fixed), then this is around 10 f-stops. If we include rapid pupil adaptation this increases to about 14 f-stops. The pupil diameter typically ranges from about 2 mm in bright conditions to around 8 mm in dark conditions, giving a 16:1 dynamic range or 4 f-stops.

Adaptation also takes place in both rods and cones. The cones adapt quicker than the rods and this explains why it takes much longer to adapt to darkness (dark adaptation) than to brightness.

A typical dark adaptation characteristic is shown in Figure 2.13. This illustrates the variation over time of an individual's threshold after adaptation to a bright light. This figure shows that the dark adaptation process is quite slow, taking several minutes for complete adaptation, but also that the adaptation occurs in two stages. The first of these is due to the cones adapting and the second is due to the rods. The retinal photoreceptors adapt through pigment bleaching—but also by feedback from horizontal cells.

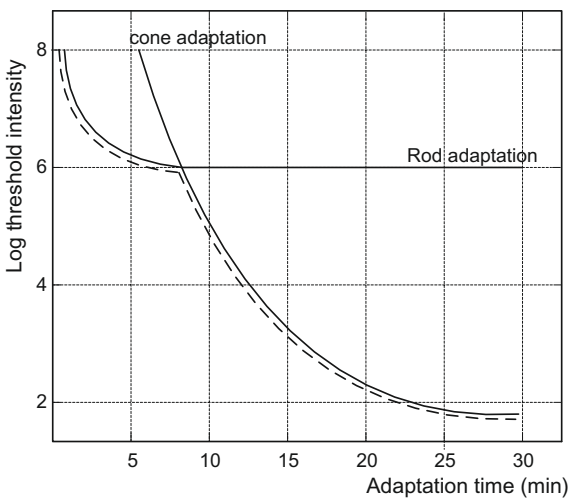

FIGURE 2.13

Dark adaptation of rods and cones.

2.5 Color processing

If you ask someone if they prefer to watch a color rather than a monochrome TV programme or movie, invariably they would say yes on the basis that it is closer to reality. This is obvious—consider Figure 2.14 which, on the right, shows a color image of Yellow Water in Kakadu, Australia and, on the left, a monochrome version of the same image. Most would agree that the color image is easier to interpret, providing improved differentiation between natural objects and features in the scene and a better sense of depth and engagement.

Yet, despite few people seeing the world in black and white, we readily accept monochrome images and are perfectly happy to suspend disbelief, for the sake of entertainment or information. In some cases (for example in Steven Spielberg's *Schindler's List*) directors actually choose to use monochrome imagery to effect mood and increase impact in films. This acceptance of, and occasional preference for, luminance-only signaling may be associated with the way in which we process our trichromatic sensors.

A final point worth highlighting here is that the choice of primary Red, Green, and Blue colors in displays is not based on the sensitivities of the S, M, and L cones. Instead they are selected to be well spaced spectrally, thus allowing a large gamut of colors to be produced.

Following from Section 2.4 and equation (2.2) we can generate average cone responses R, G, B to incident radiation as follows, where $\bar{y}_r(\lambda)$, $\bar{y}_g(\lambda)$, and $\bar{y}_b(\lambda)$ are the luminous efficiencies for the red, green, and blue color channels respectively:

$$R = K_r \int_{380}^{720} L(\lambda)\bar{y}_r(\lambda)\mathrm{d}\lambda$$

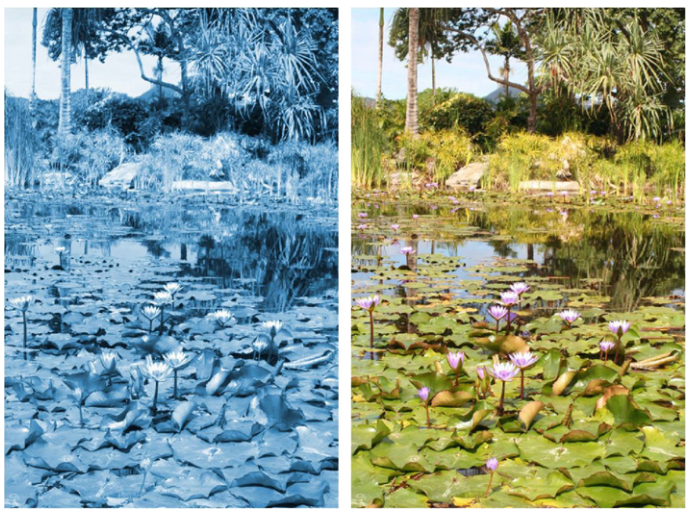

FIGURE 2.14

Increased immersion from color images.

$$G = K_g \int_{380}^{720} L(\lambda)\bar{y}_g(\lambda)\mathrm{d}\lambda$$
$$B = K_b \int_{380}^{720} L(\lambda)\bar{y}_b(\lambda)\mathrm{d}\lambda \tag{2.3}$$

2.5.1 Opponent theories of color

We have three types of cone. However, most mammals only have two types—evolved from a single yellow-green cone centered around 550 nm (corresponding to chlorophyll reflectivity maximum). A blue cone was then added centered around 430 nm—which assists with plant discrimination. Humans and other primates added a third cone type (actually the yellow-green cone probably split into two types)—at the red end of the spectrum. This enabled its owners to forage more effectively (red fruit can be more easily differentiated) and signal more dramatically. As a point of interest, some birds possess four channels and the mantis shrimp has 12!

As described earlier, color perception in the HVS employs an opponent-based approach. The reason for this becomes clear if we consider color constancy requirements. Clearly if the light reflected from a certain object varies, we would ideally like to perceive it as a brighter or darker version of the same color. This can be achieved if we consider the ratio of the outputs from the photoreceptors rather than their absolute outputs. This was first identified by Herring in 1878, who realized that certain hues could not co-exist in a single color sensation. For example, we can experience a red-yellow = orange and blue-green = cyan but cannot readily experience red-green or blue-yellow. Hence the pairs *red-green* and *blue-yellow* are termed opponent pairs. Herring argued that we could not experience both colors in a pair simultaneously as they are encoded in the same pathway. This observation has been validated from psychophysics experiments and it is easy to convince yourself of it. For example, if you stare at green long enough you get a red after-effect (and vice versa) and if you stare at blue long enough you get a yellow after-effect (and vice versa).

A diagram showing a possible opponent model is given in Figure 2.15. This shows a luminance signal generated from the sum of the signals from green and red cones and two chrominance channels generated from ratios of red to green and yellow to blue. Three things should be noted here. Firstly expert opinion varies on the nature of the combining processes and there is some evidence for the chrominance channels being based on color differences (i.e. *R-G* and *Y-B*). However, because the sensitivities of the cone spectral responses overlap, most wavelengths will stimulate at least two types of cone. Some tristimulus values therefore cannot physically exist as an additive color space and imply negative values for at least one of the three primaries. Secondly it is well known that color sensations vary significantly with context and spatial frequency (see Figure 2.16 and try this for yourself—if you move the page closer, the blue bars will turn green). Theories to explain aspects of this have been proposed by authors such as Shapley and Hawken [20]. Finally, the observant reader will notice that there is no contribution from blue cones in the luminance channel. The main reason for this is that there are relatively few blue cones in the fovea and almost

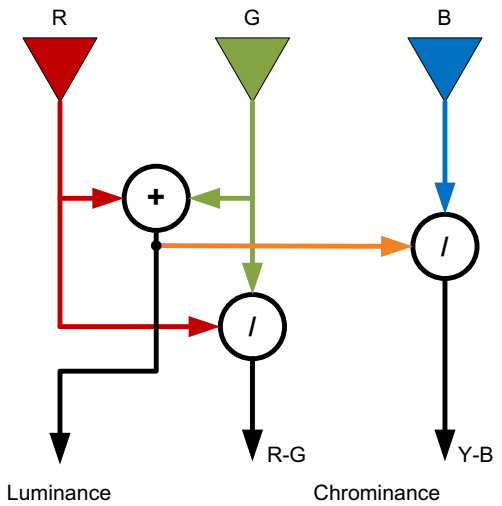

FIGURE 2.15

Opponent processing of color.

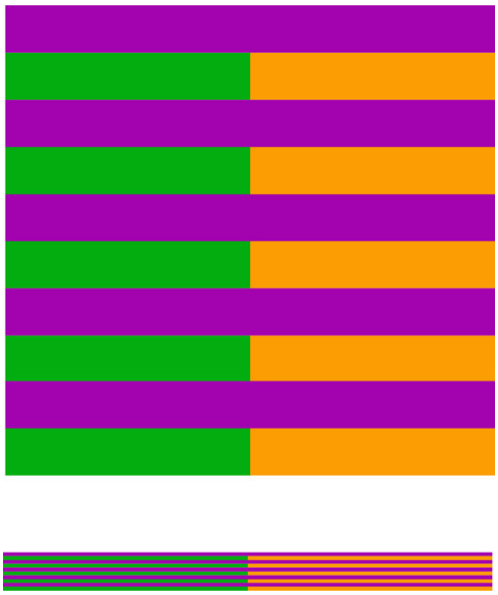

FIGURE 2.16

Color dependence on context. The bottom picture is just a squashed version of the top one yet the green stripes become blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this book.)

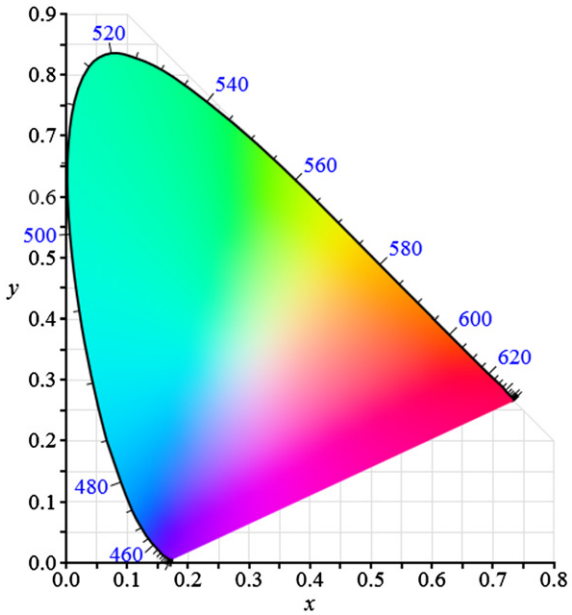

FIGURE 2.17

The CIE 1931 chromaticity chart. (Reproduced with permission from Ref. [21].)

none at its center. This is probably because of the effects of chromatic aberration which would significantly reduce acuity.

2.5.2 CIE 1931 chromaticity chart

CIE, the International Commission on Illumination, in 1931 defined a color-mapping function based on a standard observer, representing an average human's chromatic response within a 2° arc, to primaries at $R = 435.8$ nm, $G = 546.1$ nm, and $B = 700$ nm. Figure 2.17 shows the CIE 1931 chromaticity chart. The boundary represents maximum saturation for the spectral colors, and the diagram forms the boundary of all perceivable hues. The colors that can be created through combinations of any three primary colors (such as *RGB*) can be represented on the chromaticity diagram by a triangle joining the coordinates for the three colors. Color spaces are discussed in more detail in Chapter 4.

2.6 Spatial processing

Spatial resolution is important as it influences how sharply we see objects. As discussed in Section 2.4, the key parameter is not the number of pixels in each row or column of the display, but the angle subtended, θ, by each of these pixels at the

viewer's retina. We thus use the term *spatial* here to indicate things that are not moving as opposed to implying that there is some sense of spatial constancy, regardless of viewing distance.

2.6.1 Just noticeable difference, contrast, and Weber's law

Contrast is the visual property that makes an object distinguishable from other objects and from a background. It is related to the difference in the color and brightness of the object and other objects within the same field of view. Because the HVS is more sensitive to contrast than absolute luminance, we perceive the world similarly regardless of changes in illumination. It is well known that the perception of a constant increment in illumination is not uniform, but instead varies with illumination level. This is captured by Weber's law which identifies that, over a reasonable amount of our visible range, the following expression for contrast, C, holds true, for a target T and a background B:

$$\triangle C = \frac{\triangle Y}{Y} = \frac{\text{JND}(Y_T - Y_B)}{Y_B} = \text{constant} \tag{2.4}$$

Weber's law implies that the just-noticeable difference (JND) between two stimuli is proportional to the magnitude of the stimuli. This is sometimes confused with Fechner's law that states that a subjective sensation is proportional to the logarithm of the stimulus magnitude. Figure 2.18 shows an example test chart for JND testing. For most observers the JND value is between 0.01 and 0.03.

Weber's law fits well over 2–3 decades of $\log\left(Y\right)$ and is illustrated in Figure 2.19. As background illumination is increased above or decreased below this region the slope will change. As we will see in later chapters, this subjective assessment of luminance sensitivity can be important in image and video quantization.

2.6.2 Frequency-dependent contrast sensitivity

One of the most important issues with HVS models concerns the relationship between contrast sensitivity and spatial frequency. This phenomenon is described by the

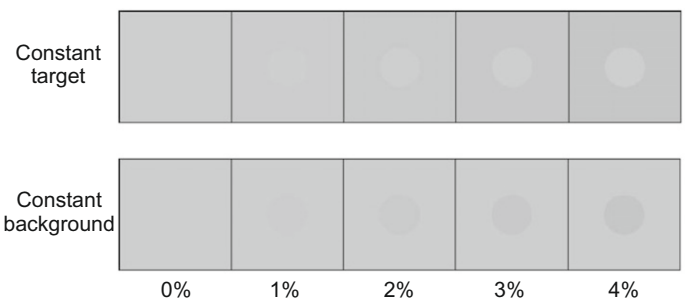

FIGURE 2.18

Just-noticeable differences at different contrast increments.

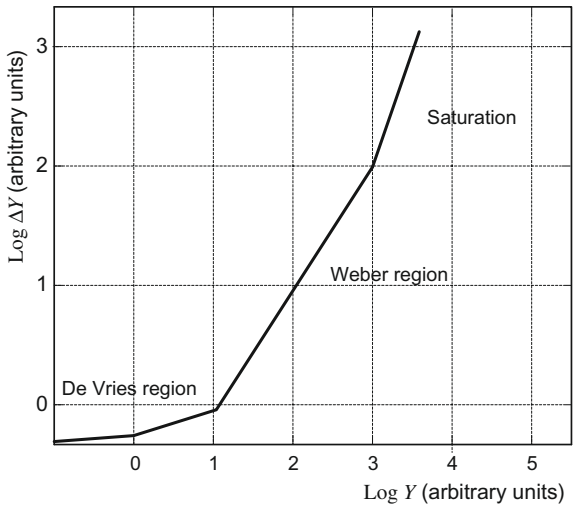

FIGURE 2.19

JND curve for human vision.

contrast sensitivity function (CSF). The CSF model describes the capacity of the HVS to recognize differences in luminance and chrominance as a function of contrast and spatial frequency.

A spatial contrast sensitivity chart or sine wave grating is shown in Figure 2.20. Typical contrast sensitivity responses for luminance and chrominance are shown in

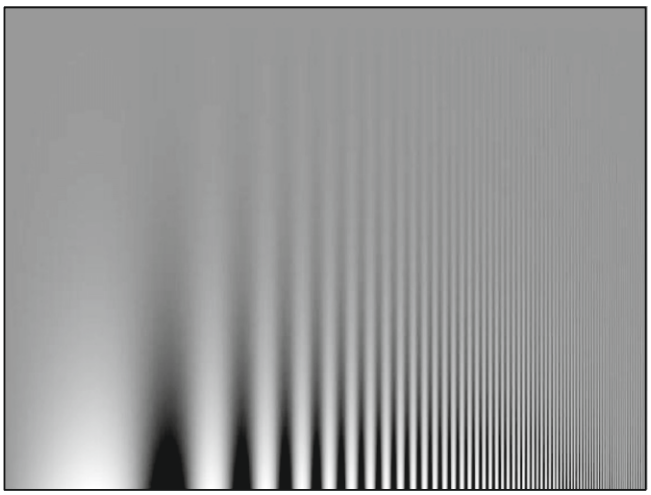

FIGURE 2.20

Contrast sensitivity chart.

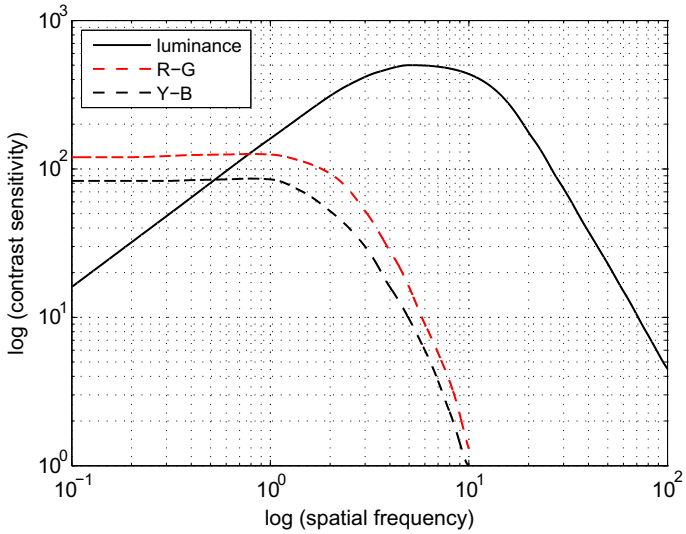

FIGURE 2.21

Luminance and chrominance CSF responses.

Figure 2.21 where contrast sensitivity is defined as:

$$C(f) = \frac{Y_B(f)}{\text{JND}(Y_T(f) - Y_B(f))} \tag{2.5}$$

This illustrates the sensitivity of the visual system to spatial sinusoidal patterns at various frequencies and contrast levels. The bandpass characteristic is primarily due to the optical transfer function of the eye together with the retinal (lateral inhibition) processing and some properties of region V1. Its shape is influenced not only by the spacing and arrangement of photoreceptors, but also possibly by the limitation on the number of spatial patterns representable in V1.

Essentially, the HVS is more sensitive to lower spatial frequencies and less sensitive to high spatial frequencies. Mannos and Sakrison [22] presented a CSF model for grayscale images as a non-linear transformation followed by a Modulation Transfer Function (equation (2.6)):

$$C\left(f\right) = 2.6\left(0.0192 + 0.114f\right)\mathrm{e}^{-(0.114f)^{1.1}} \tag{2.6}$$

where spatial frequency is:

$$f = \sqrt{f_h^2 + f_v^2}$$

f_h is the horizontal frequency component and f_v is the vertical frequency component. It is often preferable to work in terms of a spatial frequency normalized to the display

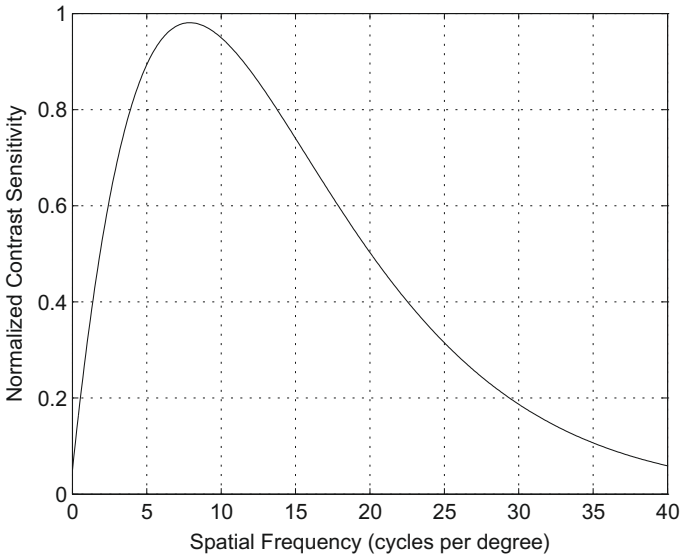

FIGURE 2.22

Luminance contrast sensitivity function.

characteristics in terms of cycles per pixel, thus:

$$f_d = \frac{f}{f_e}$$

where f_e is the number of pixels per degree for the experiment being conducted, which depends on the resolution of the display, the dimensions of the display and the viewing distance. This fits well with measured results and its CSF response function (normalized) is shown in Figure 2.22.

The contrast sensitivity of human vision is plotted here against spatial frequency. We can observe that:

1. The sensitivity to luminance information peaks at around 5–8 cycles/deg. This corresponds to a contrast grid with a stripe width of 1.8 mm at a distance of 1 m.
2. Luminance sensitivity falls off either side of this peak and has little sensitivity above 50 cycles/deg.
3. The peak of the chrominance sensitivity curves occurs at a lower spatial frequency than that for luminance and the response falls off rapidly beyond about 2 cycles/deg. It should also be noted that our sensitivity to luminance information is about three times that for *R-G* and that the *R-G* sensitivity is about twice that of *B-Y*.

The contrast sensitivities of the HVS lead us to our first basic means of compressing images—The perceptual CSF model can be used for reduction of imperceptible information and, if we use luminance and color difference signals as the basis of our

representation, then we can sample the chrominance signals at around half the rate of the luminance signal, without any loss of perceived quality. Furthermore, both luminance and chrominance signals can be more coarsely quantized at higher spatial frequencies due to their reduced contrast sensitivity. These mechanisms for achieving this will be explored in Chapter 4.

2.6.3 Multiscale edges

Edge localization models of early vision have generally featured some arrangement of tuned spatial filters at multiple image scales followed by a feature extraction process and finally integration of these features into a map (such as Marr's primal sketch [4]). The orientation and frequency selectivity of the visual system in many ways resembles multiscale transforms such as the wavelet or complex wavelet transform and these have been employed in various HVS models. As we will see later in the book, these tools are also used extensively in image and video compression, where the transformed image data in frequency, orientation, and scale can be exploited to remove invisible content by, for example, filtering with a CSF characteristic.

2.6.4 Perception of textures

Texture is an important visual cue that we exploit in tasks such as edge localization, depth perception, and general object recognition. The visual system appears to operate under the assumption of high entropy in a scene. For example, the distribution of pebbles on a beach provides good depth cues because they appear to get smaller the further they are from the viewer. It has been proposed that texture perception employs a filter–rectify–filter (FRF) process, which first applies a non-linear transform to the output of the initial multiscale spatial filters, followed by a second filtering stage to provide a low frequency surface on which to do edge localization. This model implies that humans should be less accurate at determining texture defined edges compared to those associated with luminance gradients and this has been confirmed experimentally.

This also suggests that the HVS would experience some degree of change blindness associated with textures and again this has been confirmed from subjective trials. Consider the two images in Figure 2.23. They look identical but on closer inspection are actually significantly different. Again this observation reinforces the notion that a texture analysis–synthesis approach to compression may offer potential.

2.6.5 Shape and object recognition

There is evidence from lesion studies that visual processing can be divided into three broad stages:

1. Extracting local features.
2. Building shapes and surfaces from these features.
3. Creating representations of objects in the scene.

There is also some evidence that the HVS uses other properties from a scene to help group features, for example proximity, color and size similarity, common fate (regions

FIGURE 2.23

Texture change blindness. (Images courtesy of Tom Troscianko.)

with similar direction and speed properties are likely to be part of the same object) and continuity. Marr [4] proposed that features are grouped using average local intensity, average size, local density, local orientation, local distances between neighboring pairs of similar items, and local orientation of the line joining neighboring pairs of items.

To actually recognize objects, further integration of features over larger spatial distances must take place. A number of explanations for this exist including feedback from higher functional areas of the visual cortex that modulate edge detector responses so as to enhance contours [2]. Surfaces are also likely to be important grouping cues as features on the same surface are likely to be part of the same object.

An object representation is thus likely to be formed from a combination of low level image analysis with more symbolic higher level abstract representations coupled through significant feedback between layers and regions. Object recognition is complex because objects have both intrinsic and extrinsic properties. The extrinsic factors, such as change of viewpoint, occlusion or illumination changes, can have a major impact on recognizability. These extrinsic properties can provide context but can also confound recognition.

2.6.6 The importance of phase information

It has been widely reported that the phase spectrum is highly significant in determining the visual appearance of an image; distortions to phase information, especially over larger areas, can result in poor image quality. See for example Figure 2.24. Equally, it has been shown that it is possible to obtain the phase indirectly from local magnitude information. There is evidence that complex cells in the human visual cortex V1 respond to local magnitude information rather than phase. It has thus been suggested that the HVS might use local magnitude information and local phase to determine an image's appearance. Vilankar et al. [14] conducted an experiment to quantify the contributions of local magnitude and local phase toward image appearance as a function of spatial frequency using images distorted using the dual tree complex wavelet transform. They confirmed that both local magnitude and local phase do

FIGURE 2.24
The importance of phase information in visual perception. Left: original. Right: phase distorted version using the complex wavelet transform (Reproduced with permission from Vilankar et al. [14]).

indeed play equally important roles and, in some cases, local phase can dominate the image's appearance. While we are still at the early stages of understanding these processes, there is no doubt that they may in future provide a better understanding of the effects of image compression on the human visual system.

For the above reasons, it is widely accepted that signal processing operations on images should preserve phase information. Linear phase FIR digital filters are therefore (almost) exclusively used in compression operations, both for pre- and post-processing and for interpolation.

2.7 Perception of scale and depth

2.7.1 Size or scale

Size and scale are complex topics in vision. When a spatial pattern is viewed at increasing distance, the image it projects on the retina moves further from the retina; the image it projects stimulates a response reflecting the higher spatial frequency. Thus, at this level changes in scale simply relate to changes in spatial frequency. It has been shown that adaptation to a certain frequency at one distance does affect the perception of a different frequency at a different distance, even if both project the same spatial frequency at the retina. So, even though distance is not integrated into the frequency information from the scene, depth information must be incorporated at higher levels to compensate and allow humans to assess the size of an object at different distances. This ability to judge the size of objects regardless of distance is called size constancy and can lead to misleading and entertaining results, especially when combined with other depth cues such as perspective (Figure 2.25).

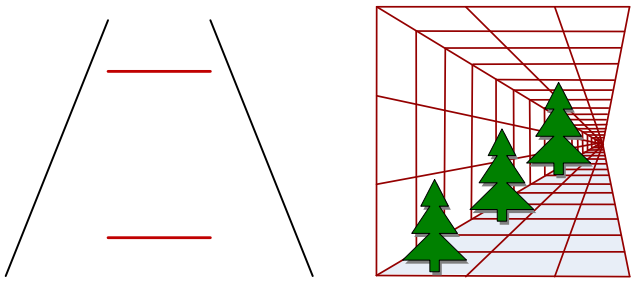

FIGURE 2.25

Perspective-based depth cues can be very compelling and misleading.

2.7.2 Depth cues

There is no doubt that depth assessment is a dominant factor in our ability to interpret a scene; it also serves to increase our sense of engagement in displayed image and video content. Stereopsis, created through binocular human vision, is often credited as being the dominant depth cue, and has indeed been exploited, with varying degrees of success, in creating new entertainment formats in recent years. It is however only one of the many depth cues used by humans—and arguably it is not the strongest. A list of depth cues used in the human visual system is given below:

- **Our model of the 3-D world:** Top-down familiarity with our environment enables us to make relative judgements about depth.
- **Motion parallax:** As an observer moves laterally, nearer objects move more quickly than distant ones.
- **Motion:** Rigid objects change size as they move away or toward the observer.
- **Perspective:** Parallel lines will converge at infinity—allowing us to assess relative depths of oriented planes.
- **Occlusion:** If one object partially blocks the view of another it appears closer.
- **Stereopsis:** Humans have two eyes and binocular disparity information, obtained from the different projections of an object onto each retina, enables us to judge depth.
- **Lighting, shading, and shadows:** The way that light falls on an object or scene tells us a lot about depth and orientation. See Figure 2.26.
- **Elevation:** We perceive objects that are closer to the horizon as further away.
- **Texture gradients:** As objects recede into the distance, consistent texture detail will appear to be finer scale and will eventually disappear due to limits on visual acuity.
- **Accommodation:** When we focus on object our ciliary muscles will either stretch (a thinner lens for distant objects) or relax (a fatter lens for closer objects). This provides an oculomotor cue for depth perception.
- **Convergence:** For nearer objects, our eyes will converge as they focus. This stretches the extraocular muscles, giving rise to depth perception.

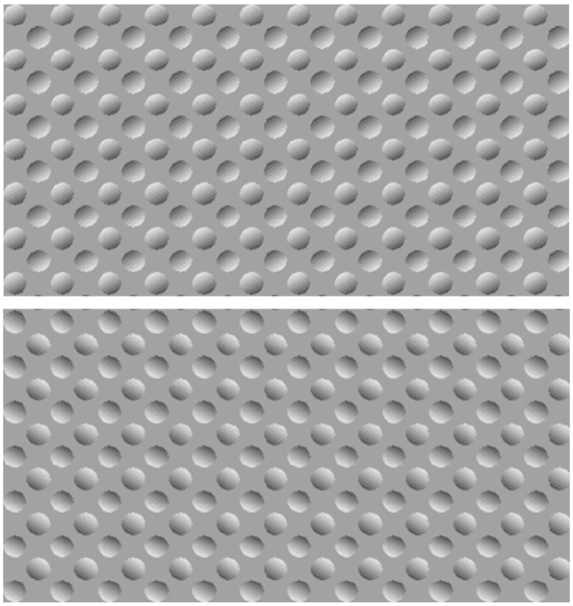

FIGURE 2.26

Pits and bumps—deceptive depth from lighting.

It should be noted that certain depth cues can be confusing and can cause interesting illusions. For example, we have an in-built model that tells us that light comes from above (i.e. the sun is above the horizon) and our expectation of shadows reflects this top-down knowledge. Figure 2.26 shows exactly this. The top diagram clearly shows alternating rows of bumps and pits starting with a row of bumps at the top. The bottom diagram is similar except that it starts with a row of pits at the top. In reality, the only difference with these diagrams is that the bottom one is the top one flipped by 180°. Try this by turning the page upside down. This provides an excellent example of how hard wired certain visual cues are and how these can lead to interesting illusions. Another example of an illusion driven by top-down processes is the *hollow mask*. In Figure 2.27, we can see a normal picture of Albert Einstein. In reality this is a photograph of a concave mask. Our visual system is so highly tuned to faces that, even when our disparity cues conflict with this, we cannot help but see this as a convex face. The right picture shows that, even when we rotate the mask to an angle where the features are distorted and you can see it is clearly concave, it still looks like a convex face!

2.7.3 Depth cues and 3-D entertainment

3-D video technology exploits stereopsis generated through binocular vision. Stereopsis is produced using two cameras spaced at the interocular distance, to create two views which can be displayed in a manner so that the left view is only fed to the left eye and the right view is fed to the right eye. While this does indeed create a sensation

FIGURE 2.27
The hollow mask illusion.

of depth, and can be effective if used well for certain types of content, there are several issues that cause problems. These relate to conflict between the accommodation and the convergence of the eyes.

When we watch a 3-D movie we view it on a 2-D screen so that is where we naturally focus. However, 3-D objects projecting out of the screen cause our eyes to converge because they appear nearer. This can lead to confusing depth cues causing fatigue and, in some cases, nausea. In addition, it has been reported that up to 20% of the population do not see stereoscopically. Of the remainder, a large proportion do not like to watch movies in 3-D and choose to see them in 2-D. There are a range of reasons given for this, including the wearing of glasses, the reduced light levels (up to 50%) due to current 3-D glasses or the general feeling of forced, over-emphasized or inconsistent depth cues.

Equally, directors, cinematographers and technologists are only gradually beginning to understand the complex interactions between their domains. For example, directors and cinematographers use shots to support narrative and the average shot length in a movie is surprisingly short (around 4 s). It is now known that the HVS takes longer to adapt to the transitions between shots in 3-D than in 2-D, so shot lengths should be correspondingly longer in the 3-D case (more of this later). The rules of cinematography have had to change to cope with 3-D.

2.8 Temporal and spatio-temporal response

The temporal response of the HVS is dependent on a number of internal and external factors. Essentially the impulse response of photoreceptors defines the maximum

temporal sampling of the retinal image. However, external influences such as ambient lighting, display brightness, and viewing distance also cause variations. If an object is moving across the visual field and not tracked, its image will blur (motion blur) due to the fact that photoreceptors take more that about 100 ms to reach their peak response. These observations have significant impact, not only on camera and display design but also on cinematography (e.g. rules are often imposed to limit camera pan speeds) [16,17].

2.8.1 Temporal CSF

The visual system is sensitive to temporal variations in illumination. This characteristic is well known and exploited in the design of camera and displays in determining their temporal update rates. The temporal limit is imposed by the rate of response of the photoreceptors, the retinal circuitry and the response times in the visual cortex [2]. The eye retains the sensation of the incident image for a short time after it has been removed (persistence of vision). It has until recently been thought that an update rate of 60 Hz is sufficient to convey the perception of smooth motion. However, recent investigations are questioning this in the context of higher spatial resolution and larger displays. This is discussed in more detail in Chapter 13.

Kelly [19] showed that the visual sensitivity response to a temporally varying pattern at different frequencies was band-pass in nature, highly dependent on display brightness and peaking at around 7 or 8 Hz for dimmer displays, becoming diminishingly small at around 24 Hz. However, for brighter displays, the peak rises to around 20 Hz, disappearing at around 80 Hz. The frequency at which the temporal variation becomes unnoticeable is important as it relates to the flickering we experience due to the refresh of a display. The variation with display brightness explains why we can (just about) get away with 24 fps sampling for cinema but need higher rates for TV screens and even higher rates for computer monitors (as we sit much closer to them).

2.8.2 Spatio-temporal CSF

A diagram showing the spatio-temporal characteristics of the HVS is shown in Figure 2.28 [18]. This shows the band-pass temporal characteristic described above for low spatial frequencies. However, as spatial frequency increases, the temporal response becomes more low-pass in nature and similarly, with higher temporal frequencies, the spatial response tends to a low-pass characteristic. Thus for faster moving objects, we cannot easily assimilate their spatial textures. This trade-off between spatial and temporal update rates has been exploited in interlaced scanning as we will examine in Chapter 4.

2.8.3 Flicker and peripheral vision

The flicker fusion threshold is the frequency at which an intermittent light stimulus appears steady to the observer. This depends on a range of factors including

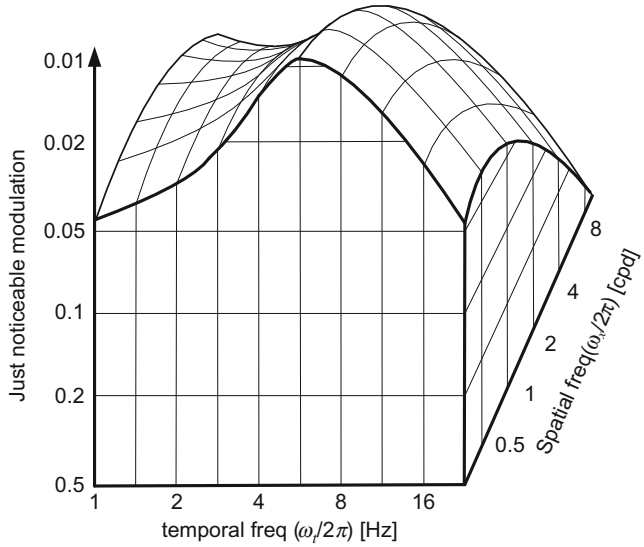

FIGURE 2.28

Spatio-temporal CSF. (Adapted from Kelly [18].)

the frequency and depth of the modulation, the wavelength of the illumination, the position on the retina where the stimulation occurs, and the degree of light or dark adaptation. Other factors such as fatigue can also have an influence.

Tyler [23] demonstrated that the flicker threshold (or critical flicker frequency (CFF)) varies with brightness (higher for a brighter source) and with retinal location (rods respond faster than cones). As a consequence flicker can be sensed in peripheral vision at higher frequencies than in foveal vision. This is important as all of our assumptions about flicker are based on central vision and small experimental screens. This explains why, as screens have become brighter and their sizes have become larger, flicker has become more noticeable. This is why TV manufacturers have had to upsample to 300 or even 600 Hz.

Figure 2.29 [23] shows the visual-field CFF contours for square-wave modulation (in cycles per second) as a function of eccentricity and meridian, with the field sizes scaled to stimulate a constant number of cones at each eccentricity, for two observers. This demonstrates flicker frequencies up to 90 Hz in the periphery.

It is also important to note that CFF generally refers to unstructured stimuli. If the interaction with spatial characteristics is of interest, the temporal contrast sensitivity function (CSF) is more informative. Interestingly the temporal CSF has been observed to be independent of eccentricity up to 30° on the nasal horizontal meridian [24]. In reality of course with natural scenes, complex interactions exist across different interacting patterns and motions in a scene.

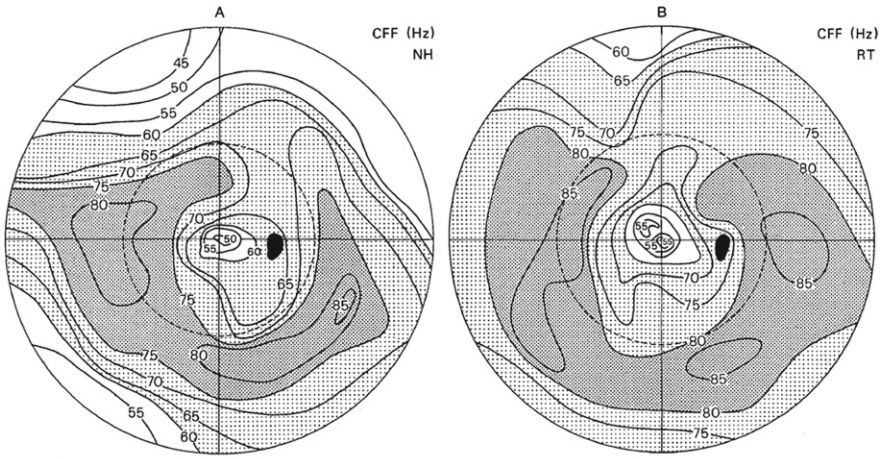

FIGURE 2.29

Variation of critical flicker frequency (Reproduced with permission from Tyler [23]).

2.9 Attention and eye movements

The HVS creates for us an illusion of fully perceiving a scene while in reality only a relatively small amount of information is passed to the higher brain levels. In this way we perceive only what is considered absolutely necessary, ignoring all other information, in order to compress the huge amount of visual data that continuously enters the visual system. The process that achieves this is referred to as attention and is supported at all cortical levels, but perhaps foremost through foveation and eye movements.

2.9.1 Saliency and attention

Attention modulates the responses of other parts of the visual system to optimize the viewing experience and to minimize the amount of additional information needed to achieve this. Attention can be bottom-up (exogenous) or top-down (endogenous) or a combination of both. In the first case this happens when lower-level responses attract attention (e.g. through detection of motion in the periphery). This process is often referred to as saliency and is based on the strength of the low level HVS response to the specific feature. Top-down or endogenous attention is guided (more) consciously by the task at hand.

There has been a substantial amount of research into models of saliency, perhaps most notably by Itti et al. [25] where top-down attention modulates bottom-up features. Eye movements are used to enable us to foveate on rapidly changing areas of interest and we will consider these briefly next. Attention and foveation have been proposed as a means of allocating bits optimally according to viewing in certain video compression scenarios [26,27].

2.9.2 Eye movements

Apart from when we compensate for head movements, our eyes move in two distinct ways:

1. **Saccades:** As rapid involuntary movements between fixation points. The visual system is blanked during saccading, so we do not experience rapid motion effects due to scanning. Saccades can take two forms:
   a. **Microsaccades:** Small involuntarily movements used to refresh cell responses that can fall off due to adaptation.
   b. **Foveation:** The eyes also move rapidly in photopic vision, to allow exploration of new areas of a scene providing an impression of increased acuity across the visual field.
2. **Smooth pursuit:** These are attention-guided smoother voluntary movements which happen, for example, when tracking moving objects. This process keeps the object of interest centered on the fovea so as to maintain high spatial acuity and reduce motion blur. Under the conditions of smooth pursuit, the spatio-temporal limits discussed earlier change dramatically. Girod [29] recomputed these characteristics and demonstrated that this type of eye movement has the effect of extending the temporal limit of our visual system, showing that frequencies of several hundred hertz can be perceived.

Eye tracking is used extensively in vision research [32] to assess the gaze and fixations of an observer in response to a stimulus. Some of the earliest and most famous research on this was published by Yarbus [28]. The results of Yarbus's experiment, where observers were provided with a range of viewing tasks associated with the Visitor painting, are shown in Figure 2.30. It is interesting to observe how the saccade patterns relate to the specified task.

2.10 Visual masking

Visual masking is the reduction or elimination of the visibility of one brief stimulus, called the "target," by the presentation of a second stimulus, called the "mask." An overview of the influences of masking and other psychovisual effects is presented by Girod [30a].

2.10.1 Texture masking

The visibility threshold for a target increases when the background is textured rather than plain. Spatial (or texture) masking causes the contrast threshold for a given spatial frequency to rise when a high contrast mask is present. The inhibitory nature of interacting receptive fields causes the HVS sensitivity to decrease to certain spatial patterns when they are viewed in the context of other patterns. The influence of the mask is related to the similarity between the spatial frequency content of the mask and the target. This effect can be explained by recognizing that the mask pattern elicits

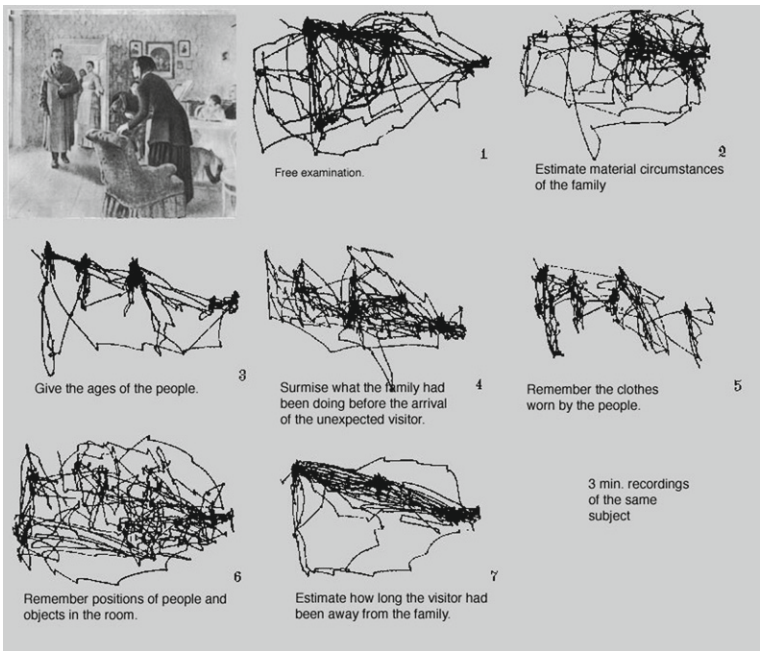

FIGURE 2.30

Eye movements in response to task (Public domain image from: http://commons.wikimedia.org/wiki/File:Yarbus_The_Visitor.jpg).

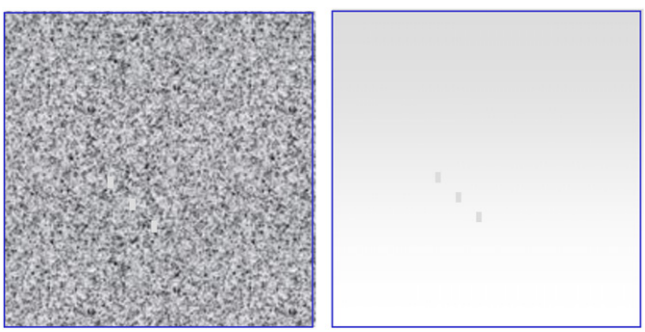

FIGURE 2.31

Example of texture masking.

responses in the same spatial filters as stimulated by the target test pattern, making them indistinguishable at later visual processing stages [2]. An example of texture masking is shown in Figure 2.31. This shows an identical target in both sub-figures (three rectangles). The target is highly visible on the right with a plain background but is very difficult to detect on the left.

2.10.2 Edge masking

It has been observed that there is a distinct masking effect in the vicinity of a spatial edge [30a]. This has also been demonstrated by Zhang et al. [31] for the case of higher dynamic range content. An example of this, showing the experimental setup of the target and edge and the noise visibility variations according to edge contrast and distance of the target from the edge, is shown in Figure 2.32.

2.10.3 Temporal masking

Girod [30a] demonstrated the masking effects of video in the presence of temporal discontinuities (such as those that occur at shot cuts or scene changes in a movie).

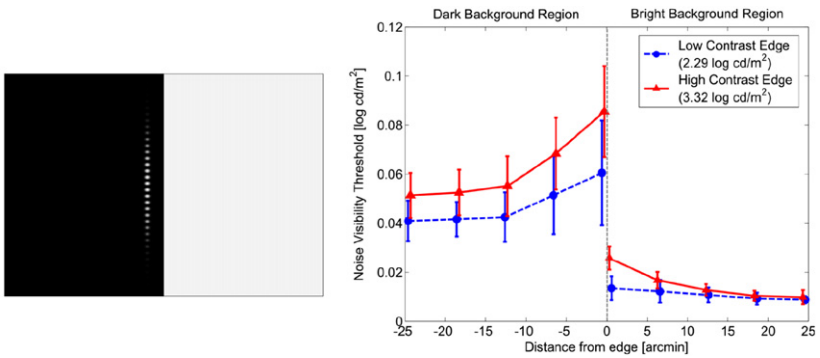

FIGURE 2.32

Edge masking for high and low dynamic range content.

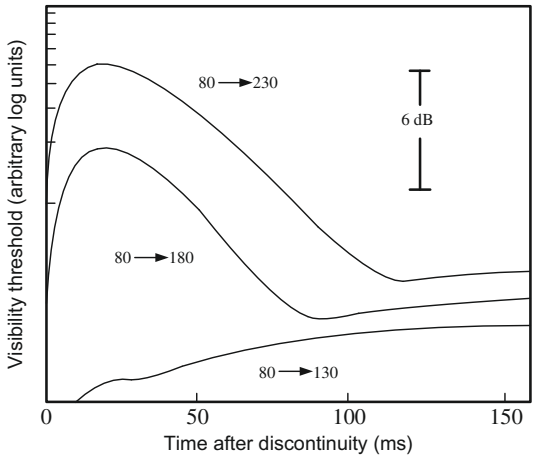

FIGURE 2.33

Temporal masking effects for various edge step sizes. (Reproduced with permission from Girod [30b].)

This effect is shown in Figure 2.33 for a gamma predistorted video signal. It can be observed that, as the extent of temporal discontinuity increases (shown here for an 8 bit edge), the temporal masking effect extends in duration over approximately 100 ms.

2.11 Summary: a perceptual basis for image and video compression

2.11.1 Influential factors

As we have seen throughout this chapter, many factors influence the way we see things: ambient light, scene or display brightness, spatio-temporal content of the scene, viewing distance, attention, task, expectation, physiological variations between subjects and many other subtle environmental factors.

Many researchers have tried to understand these influences on perception, using various abstract psychophysics experiments. While these provide very valuable insights into the limits and variations of our visual responses, few deal with real scenes and the complex interactions within them. However, we can take away several important guidelines and some of these have been used in the design of acquisition, display devices, and for optimizing signal representations and compression methods. Others hold potential for the future. These are summarized in the next subsection.

2.11.2 What have we learnt?

A summary of the observed characteristics of our visual system alongside ways in which these characteristics can be exploited in the design of vision-based technology are provided below.

Color processing

HVS characteristics	Implications
The HVS perceives color using three color channels in the visible part of the electromagnetic spectrum	Acquisition and display technology and visual signal representations for transmission and storage can be based on a tristimulus framework
Based on opponent theory, the HVS processes visual signals as a luminance channel and two chrominance channels	Luminance-only content is readily accepted as natural. Visual signal representations can be based on color spaces that split luminance and chrominance
The HVS has higher acuity in its luminance channel than its chrominance channels	Chrominance channels can be allocated lower bandwidth than the luminance channel

Spatial response

HVS characteristics	Implications
The HVS response to spatial frequencies typically peaks at 5–8 cpd and then falls off, becoming diminishingly small beyond 40–60 cpd	Frequencies beyond the upper range need not be processed. Higher frequencies can be coded more coarsely with implications for quantization. The lower frequency luminance roll-off is not normally exploited
Edges are an important aspect in human vision system processing. Sensitivity increases with contrast level	Edges should be preserved. Artificial edge artifacts introduced by quantization are very noticeable and should be avoided
The HVS can detect noise more readily in plain areas than in textured areas	Spatial (contrast) masking of quantization noise can be exploited in textured regions

Temporal response

HVS characteristics	Implications
The HVS temporal contrast sensitivity peaks at around 5–20 Hz and diminishes above 20–80 Hz (brightness dependent)	Visual processing technology should be capable of sampling at twice the visual threshold and should take account of viewing conditions
Our ability to see individual frames as a continuum (without flicker) occurs at approximately 50 Hz	Frame rates for acquisition and display should normally be at least 50 Hz
Critical flicker frequency is higher in the periphery than in the fovea	Higher frame refresh rates are required for larger screens and closer viewing

Spatio-temporal response

HVS characteristics	Implications
The eye can resolve spatial frequencies better at lower temporal frequencies while at higher spatial frequencies the temporal response becomes low-pass	This trade-off between temporal and spatial acuity can be exploited in interlaced scanning to provide the perception of increased frame rates

Processing textures

HVS characteristics	Implications
There is evidence that texture in-filling occurs in the HVS	Analysis–synthesis approaches to image and video coding hold potential
Texture variations and changes are more difficult to detect than shape changes	As above

Depth cues

HVS characteristics	Implications
The HVS exploits many and varied depth cues at all stages of processing	Stereoscopic 3-D content may not be the only way to achieve more immersive content

Dynamic range

HVS characteristics	Implications
Our visual response depends on the mean brightness of the display	Cinema frame rates can be lower than those for TV or computer monitors. Higher dynamic range displays offer better perceptions of reality and depth
There is a linear relationship between JND and background brightness over a wide range of brightness levels, with further increases at low and high values	This can be exploited in applying intensity-dependent quantizations, allowing step-sizes to increase more rapidly for low and high level signals
There is a non-linear relationship between luminance and perceived brightness	Implications for signal coding in that a non-linear (gamma) function is applied to camera outputs prior to coding

Attention, foveation, and eye movements

HVS characteristics	Implications
Our high acuity vision is limited to a 2 degree arc	Bit allocation strategies could target those regions of an image where attention is focused (if known!)
The spatio-temporal frequency response of the HVS is altered significantly by eye movements. Under smooth pursuit conditions, much higher temporal acuity is possible	Especially with larger and brighter screens, displays must be able to cope with temporal frequencies of many hundreds of hertz

Physiological variations between subjects and environments

HVS characteristics	Implications
The responses of individuals will vary naturally and with age	Subjective quality assessment experiments should be based on results from a large number of subjects (typically > 20)
Human responses vary significantly according to viewing conditions	Subjective tests should be conducted under tightly controlled viewing conditions, to enable cross referencing of results

The implications and exploitation of these characteristics are explored in much more detail in the remainder of this book.

References

[1] R. Snowden, P. Thompson, T. Troscianko, Basic Vision, Oxford University Press, 2006.
[2] G. Mather, Foundations of Sensation and Perception, second ed., Psychology Press, 2009.
[3] B. Wandell Foundations of Vision, Sinauer Assoc, 1995.
[4] D. Marr, Vision, Freeman, 1982. Reprinted MIT Press (2010).
[5] H. von Helmholtz, Handbook of Physiological Optics, first ed., Hamburg and Leipzig, Voss, 1896.
[6] B. Julesz, Binocular depth perception of computer-generated patterns, Bell System Technical Journal 39 (1960) 1125–1162.
[7] D. Hubel, T. Wiesel, Receptive fields of single neurones in the cat's striate cortex, Journal of Physiology 148 (1959) 574–591.
[8] R. Gregory, Eye and Brain, fifth ed., Princeton University Press, 1997.
[9] M. Clowes, Transformational grammars and the organization of pictures, in: A. Grasselli (Ed.), Automatic Interpretation and Classification of Images, Academic Press, 1969.

[10] H. Kolb, How the retina works, American Scientist 91 (2003) 28–35.
[11] R. Lotto, D. Purves, The effects of color on brightness, Nature Neuroscience 2 (11) (1999) 1010–1014.
[12] D. Mustafia, A. Engela, K. Palczewskia, Structure of cone photoreceptors, Progress in Retinal and Eye Research 28 (4) (2009) 289–302.
[13] A. Netravali, B. Haskell, Digital Pictures: Representation, Compression and Standards, second ed., Plenum Press, 1995.
[14] K. Vilankar, L. Vasu, D. Chandler, On the perception of band-limited phase distortion in natural scenes, in: Human Vision and Electronic Imaging XVI. Proceedings of the SPIE, vol. 7865, 2011, p. 78650C.
[15] L. Cormack, Computational models of early human vision, in: A. Bovic (Ed.), Handbook of Image and Video Processing, Academic Press, 2000.
[16] A. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1998.
[17] A. Watson, A. Allhumada, J. Farell, Windows of visibility: psychophysical theory of fidelity in time sampled visual motion displays, Journal of the Optical Society of America A 3 (3) (1986) 300–307.
[18] D. Kelly, Adaptation effects on spatio-temporal sine wave thresholds, Vision Research 12 (1972) 89–101.
[19] D. Kelly, Visual responses to time-dependent stimuli, Journal of the Optical Society of America 51 (1961) 422–429.
[20] R. Shapley, M. Hawken, Neural mechanisms for color perception in the primary visual cortex, Current Opinion in Neurobiology 12 (4) (2002) 426–432.
[21] <http://en.wikipedia.org/wiki/File:CIE1931xy_blank.svg>.
[22] J. Mannos, D. Sakrison, The effects of visual error criteria on the encoding of images, IEEE Transactions on Information Theory IT-20 (1974) 525–536.
[23] C. Tyler, Analysis of visual modulation sensitivity. III. Meridional variations in peripheral flicker sensitivity, Journal of the Optical Society of America A4 (8) (1987) 1612–1619.
[24] V. Virsu, J. Rovamo, P. Laurinen, R. Nasanen, Temporal contrast sensitivity and cortical magnification, Vision Research 22 (1982) 1211–1217.
[25] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.
[26] D. Agrafiotis, C. Canagarajah, D. Bull, J. Kyle, H. Seers, M. Dye, A perceptually optimised video coding system for sign language communication at low bit rates, Signal Processing: Image Communication 21 (7) (2006) 531–549.
[27] S. Davies, D. Agrafiotis, C. Canagarajah, D. Bull, A multicue bayesian state estimator for gaze prediction in open signed video, IEEE Transactions on Multimedia 11 (1) (2009) 39–48.
[28] A. Yarbus, Eye Movements and Vision, Plenum, New York, 1967.
[29] B. Girod, Motion compensation; visual aspects, accuracy and fundamental limits, in: M. Sezan, R. Lagendijk (Eds.), Motion Analysis and Image Sequence Processing, Kluwer, 1993, pp. 126–152.
[30a] B. Girod, Psychovisual aspects of image communication, Signal Processing 28 (3) (1992) 239–251.

[30b] B. Girod, The information theoretical significance of spatial and temporal masking in video signals, in: Proceedings SPIE/SPSE Conference on Human Vision, Visual Processing and Digital Display, Los Angeles, CA, USA, 1989, pp. 178–187.
[31] Y. Zhang, D. Agrafiotis, M. Naccari, M. Mrak, D. Bull, Visual masking phenomena with high dynamic range content, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 2284–2288.
[32] S. Liversedge, I. Gilchrist, S. Everling, Oxford Handbook of Eye Movements, Oxford University Press, 2011.
[33] J. Bowmaker, H. Dartnall, Visual pigments of rods and cones in a human retina, Journal of Physiology 298 (1980) 501–511.