Color Perception and Applications

Penny Rheingans

University of Maryland Baltimore County

1 Color Vision

Processing of visual stimuli by the human visual system begins in the eye and continues in the brain. Although the anatomical structures of the eye and the brain are physically separate, linked only by the optic nerve, the characteristics of the two structures are similar enough that it would be reasonable to consider the eyes to be satellite portions of the brain.

1.1. Physiology of the Human Visual System

Light enters the eye, is focused by the cornea, passes through the variable-diameter iris, is further focused by the lens, and strikes the light-sensitive receptors of the retina in the back of the eye. The photoreceptors of the retina can be divided into two basic categories: rods and cones. At normal light levels, cones dominate the initial response to light, resulting in photopic vision. Three different kinds of cones are active in the human visual system: short-wavelength (S) cones with a maximum response at 450 nm (violet-blue), medium-wavelength (M) cones with a maximum response at 530 nm (slightly yellowish-green), and long-wavelength (L) cones with a maximum response at 560 nm (slightly greenish-yellow). The three types are sometimes called the blue, green, and red cones, respectively, because of the general wavelengths to which they respond most strongly. The maximum response of the S cone is much smaller than either of the other two, limiting our sensitivity to small blue objects (such as blue letters on a black background). The response of cones to incoming light has a logarithmic relationship with the intensity of light, making small differences in intensity more apparent at low light levels than at high levels.

Judgments about the dominant wavelength, or hue, of light can be made by comparing the responses of the three cone types. Consider the response of the S, M, and L cones to light of 610 nm (orange). For light of 480 nm (blue), the response of the M cones would be similar, but that of the S and L cones would be vastly different. This mechanism gives rise to the Tristimulus Theory of Color, which states that any color sensation can be matched by an appropriate mixture of any three other colors.

Cones are most densely packed into the central region of the retina, or fovea, while rods are located almost exclusively in the outer regions of the retina, or periphery. When we look directly at an object, it will be imaged on the fovea where the densely packed receptors provide high spatial acuity. In the center-most section of the fovea, the foveola, S cones are virtually absent, further reducing our sensitivity to small blue objects.

From the photoreceptors, neural responses pass through a series of linking cells, called bipolar, horizontal, and amacrine cells, which combine and compare the responses from individual photoreceptors before transmitting the signals to the retinal ganglia cells. Retinal ganglia cells have concentric receptive fields with a center-surround organization. The receptive field (RF) of a ganglion cell is the area of the retina where a light stimulus will give rise to a response in that particular cell. In general, receptive fields are small in the fovea and grow larger with increasing eccentricity (distance into the periphery). The receptive field of a ganglion cell has two distinct regions: a circular center region and an annular surrounding region. Most ganglia cells are excited by stimulus in the center region and inhibited by stimulus in the surround, called on-center, off-surround organization, though some cells have off-center, on-surround organization. Other ganglia cells respond to color differences in the center and surround. Most wavelength sensitive ganglia cells are excited by red light in the center and inhibited by green light in the surround, or vice versa, though some cells respond to differences between blue and yellow light.

In the retinal ganglia cells, incoming signals indicating the responses of S, M, and L cones are re-encoded in terms of the opponent color channels [Hurvich81]. These channels describe light sensations in terms of their achromatic, R-G, and B-Y components. The achromatic channel represents the intensity of light, without regard to its wavelength characteristics. It is computed as the sum of the signals from the M and L cones. The contribution of the S cones is ignored, due to its relatively small magnitude. The R-G channel measures the chromatic content of the light on a scale from red to green. It is computed as the difference of the L and M signals. The Y-B channel encodes the chromatic content of the light on a scale from yellow to blue. Some reflection will reveal that the content of this channel is orthogonal to that of the R-G channel. It is computed as the difference between the sum of the M and L signals and the S signal.

In the retinal ganglia cells the visual system splits into two independent pathways [Livingtone88]. The magnocellular pathway detects objects and their boundaries, as well as providing a basis for the perception of depth and motion. This pathway begins in a subset of the retinal ganglia cells with large receptive fields and receives input from the achromatic opponent color channel. All processing along this pathway is independent of the wavelength of light received, depending only on its intensity. The parvocellular pathway is responsible for the perception of color and fine detail. This pathway receives input from all three opponent color channels. These two pathways function independently and the judgments of each are reconciled at a much later stage in visual processing.

The parvocellular pathway provides information about the color and small detail of objects in a scene. The anatomical components of this pathway appear to include one type of retinal ganglia cell, parvocellular layers of the LGN, layers 4Cb, 2, and 3 of the primary visual cortex, the stripes in visual area 2 responsible for color (and to some extent form) perception, visual area 4 which seems to be responsible for the higher level processing of color information, and perhaps the temporal-occipital region responsible for the identification of objects. Later stages of the pathway appear to be split into parts which are sensitive to wavelength differences, presumably for color perception, and parts which are not, presumably for perception of small detail. Characteristics of this pathway include small receptive fields, sensitivity to differences of both wavelength and brightness, and relatively slow response times.

1.2 Characteristics of Human Visual Perception

Several mechanisms of the human visual system enable the perception of stimuli over an enormous dynamic range of light level and stimulus magnitude. These include the contraction and dilation of the iris, the logarithmic response of photoreceptors, the inhibitory surround of retinal ganglia receptive fields, contrast effects, the phenomenon of adaptation, and the constancy of visual qualities. In general, such mechanisms optimize the judgment of relative quantities at the expense of absolute judgments, facilitating the detection of spatial and temporal change.

The perceived intensity or hue of an area can be significantly affected by nearby colors. This phenomenon is called *simultaneous contrast*. For example, a grey patch on a red background will seem slightly green, while the same patch on a green background will seem slightly red.. Simultaneous contrast seems to occur independently on each of the opponent channels and have effects of comparable magnitude [Ware88]. Ware observes that these contrast effects are strongest where smooth color gradients are present, i.e., where adjacent colors and color changes are similar.

Adaptation refers to the reduction in response to a constant stimulus over time. For example, a person walking from inside a dark building out into the bright sunlight will first experience the sunlight as extremely bright, reducing the world to an almost featureless field of glare. After a few minutes outside, the perceived brightness has diminished, allowing the world to be seen much more clearly. Through adaptation the visual system becomes more sensitive to small differences in stimulus value around the value of the adapting stimulus and less sensitive to small differences far from the value of the adapting stimulus. Essentially, the visual system dynamically adapts for optimum performance under the current conditions. This phenomenon of stable perception under differing absolute conditions is called *constancy*. In addition to adaptation to intensity level, the human visual system adapts to the wavelength of light, the spatial frequency of a pattern, and the direction and velocity of movement. A number of interesting after-effect illusions occur after a strong adapting stimulus is replaced by a neutral stimulus. In such situations, the viewer sees the opposite of the adapting stimulus, not the actual neutral stimulus.

2 Color Representation

2.1 Color Models

Color models provide a conceptual framework for thinking about color sequences by describing the ways in which colors can be defined. Specifically, a *color model* specifies the basic components used to describe a color. Components can be primary colors which are added or subtracted from each other, perceived qualities of the color, proposed perceptual mechanisms, or something more abstract. Color models for computer graphics and visualization can be classified into three categories: device-derived, intuitive, and perceptually uniform.

The components of a *device-derived color model* correspond directly to the signals used in the color display devices themselves. Because of this correspondence, no additional transformations need to be applied before displaying a color calculated in a device-derived model. Accordingly, the principal attraction of device-derived models is their ease of use for the applications programmer. The two most common video device-derived models are RGB, used in most color monitors, and YIQ, used in color television broadcast.

The family of *intuition-based color models* which define hue as a basic quality of color provide a more intuitive way for people to specify colors. The hue of a color associates it with a place in the spectrum. For a monochromatic light, the hue corresponds to the wavelength. Hue values are generally in the range [0, 360] and describe angular distance from red. The various hue-based models define two additional basic qualities of color : one describing its vividness (called saturation or chroma) and one describing the amount of light emitted (called lightness, brightness, value, or intensity). Some of the most common models are Hue-Saturation-Value (HSV), Hue-Saturation-Brightness (HSB), and Hue-Lightness-Saturation (HLS).

For the definition of color sequences for the display of quantitative data, all of the color spaces described so far have one serious drawback. The Euclidean distance between two colors in the color spaces says little about the perceived color difference of those colors. If differences in color are meant to correspond to differences in the value of interval- or ratio-valued variables, precise interpretation of data displayed using these spaces is difficult. The introduction of perceptually uniform (or perceptually linear) color models addresses this problem. A *perceptually uniform color model* is one in which the perceptual distance between two colors is proportional to the Euclidean distance between their positions in the color space. In practice, even color spaces which claim to be perceptually uniform are only uniform under certain locality conditions. For very large color differences, the linear relationship between geometric distance and perceived difference breaks down. Examples of device independent, perceptually uniform color models include CIELUV, CIELAB, and Munsell.

2.2 Some Color Scales

Univariate color scales map a single value into a color. The most basic group of univariate color scales are those formed by increasing (or decreasing) values of one component of a color model. The most common of these scales are the grey scale, the spectrum scale, and the saturation scale.

Levkowitz [88] introduces the term *optimal color scale* to describe a scale which maximizes the total number of JNDs (just noticeable differences) while preserving a natural order. Such a color scale is subject to restrictions including: discretization into a fixed number of equidistant increasing value, extreme points of black and white, monotonically changing saturations, and monotonically increasing components of red, green, and blue.

Conceptually, a *double-ended color scale* is created when two monotonically increasing scales are pasted together at a shared end point. For instance, a scale from grey to red and a scale from grey to cyan can be stitched together to form a single scale from red to grey to cyan. Such color scales have three distinct groups of colors, representing the high, low, and middle values. The basic advantage of a double-ended scale is the clear visual classification of values as either high, low, or middle.

Banded color scales combine the capabilities of color scales with those of contour lines to show the locations of certain isolevels more clearly than continuous color scales. Such color scales can show high frequency detail of the data distribution more clearly than other scales.

Multivariate color scales map two or more values into a color. The most obvious multivariate color scales map one variable to one color component (such as red) and the other to a second color component (such as green). A variant of this scheme maps one variable on one color component (such as hue) and the other to a combination of the remaining two (in this case, lightness and saturation).

Choosing display parameters which are complementary hues can facilitate the perception of correlation among data variables. H is constrained to a single hue and its complement. L and S run over their entire ranges. The space is scaled so that is spans a square with one hue in the upper left, its complement in the lower right, and the greys running along the minor diagonal. This sequence has the desirable property that displayed values are easily divisible into three classes: the colors along the diagonal (greys), those above it (one hue), and those below it (complementary hue). In an image representing two scalar fields, points where the two variables have similar values will appear grey, those where one variable is significantly larger will appear to be one hue, and those where the other variable is significantly larger will appear to be the complementary hue. Points will be lightest when both variable values are large and darkest when both variable values are small.

2.3 Designing Effective Color Scales

Designing colormaps is as much an art as a science. Although well-chosen color maps can emphasize important features of a dataset, color maps can also exaggerate unimportant details or create visual artifacts due to unforeseen interactions between the choice of colors, the expectations of the viewer, the data, the purpose of the visualization, and human physiology. A colormap should be appropriate to the characteristics of the data, calling attention to the most important features. The common spectrum color scale maps the middle values to yellow, a particularly striking color. In applications where the location of middle values is of particular interest, this is appropriate. Such applications are not very common, however. More often, the high or low values are of greatest interest, and middle values are of least interest. In such situations, a double-ended colormap might be appropriate. Such a colormap would map the value zero to some unobtrusive color, such as grey, while mapping the high and low values to distinct and more prominent colors.

An effective colormap should also be well-suited to the primary purpose of the visualization. One fundamental division in visual analysis tasks arises from the dichotomy between quantitative and qualitative information display. For quantitative information, a color scale which is monotonically increasing in any of the opponent channels will tend to cause contrast effects and can encourage errors in mapping from a displayed color back to the represented value. In qualitative display, where judgments about the shape or surface properties of a structure are of primary importance, simultaneous contrast does not seem to pose a difficulty. The human visual system is experienced at identifying surface tendencies from luminance gradients in the presence of contrast effects. This suggests that for tasks which require reading metric values from a representation, a color sequence which does not vary monotonically with any opponent channel (such as the spectrum scale) is superior to one which does (such as the grey scale). Cues about surface properties, however, are best judged from lightness differences presented by scales such as the grey scale. This design guideline has been supported by experimental evidence [Ware88]. Ware's findings also suggest that a color scale which varies in both luminance and hue can be used to accurately represent both metric and surface properties by minimizing the effects of simultaneous contrast.

Trumbo [81] presents four basic principles important in the selection of colors for the representation of quantitative information. The *Order principle* requires that if data value levels are ordered then the colors chosen to represent them should be perceived as ordered. The *Separation principle* requires that significantly different levels of variables be represented by distinguishable colors. The *Rows and Columns principle* states that if preservation of univariate information is important, then the display parameters should not obscure one another. The *Diagonal principle* states that if detection of positive association of variables is a goal, then the displayed colors should be easily identified as belonging to one of three classes : those near the minor diagonal, those above it, and those below. Violating one or both of these principles does not necessarily mean that a color scheme is not useful, only that it might not be appropriate for some representation tasks. For example, a hue and lightness scheme would not be the best choice for a representation primarily designed to show positive association between variables because it violates the Diagonal principle. On the other hand, it would be a reasonable choice for a representation where the goal is perception as a class of colors representing similar values of one variable across differing values of the other variable.

2.4 Avoiding Unwanted Interactions with Color

The visual characteristics judged by the human visual system are not orthogonal quantities; that is, the perception of one characteristic may be influenced by the value of another.

Color-size interactions. Some visual experiments have suggested that the color of an object can influence the perceived size of that object. Tedford, Berguist, and Flynn [Tedford77] conducted observer experiments under precisely controlled conditions and found a significant color-size effect. Specifically, rectangles of the same size, saturation, and brightness appeared to have different sizes when colored red-purple, yellow-red, purple-blue, or green (in order of decreasing apparent size). At high saturations, this effect was statistically significant for all color pairs except yellow-red and purple-blue. At low saturations, only the difference between yellow-red and green rectangles was significant. In trials where hue was held constant and saturation varied, rectangles with higher saturations were consistently judged to be smaller than less saturated rectangles.

Cleveland and McGill [Cleveland83] investigated the implications of the color-size illusion for statistical maps. Subjects were shown a map of Nevada in which counties where colored either red or green with the total area of red and green nearly equal. Subjects were asked to judge which color, if any, represented the larger land area. Each subject was shown ten maps. On the average, subjects judged that the red areas were larger more often than they judged the areas the same or the green areas larger. When the experiment was repeated using low-saturation tones of red and green (formed by adding yellow), no such bias was observed. Their results suggest that the color of a region influences the perceived size of the region and that the effect is strongest for very saturated colors. This interaction has obvious ramifications on visualization effectiveness. If perceptions about object size are important, the visualization designer should take care in assigning distinct colors to objects, either explicitly or through a colormapping process. If color coding of objects is desirable, desaturated colors are likely to cause less pronounced interaction effects.

Interactions between color components. A common strategy of multivariate color schemes is to map each variable to a different component of color. These components could be intuitive (hue, saturation, brightness), physiological (opponent-color channels), or device-derived (red, green, blue). One would expect that these color model components would be perceptually orthogonal. Perceptual studies suggest that this is not entirely the case. Interactions have been observed between hue and brightness and between saturation and brightness. A saturated color is perceived as brighter than a desaturated color when the two are related in brightness (Helmholtz-Kohlrausch effect) [Yaguchi83]. Yaguchi and Ikedo hypothesized that a cancellation of hues in the chromatic channels was resulting in decreased perceived brightness. The Bezold-Brucke Phenomenon, describing the changes in perceived hue with increasing illumination levels, has been observed in experiments where subjects are asked to match the hues of patches with differing luminances [Hurvich81]. As the luminance of the brighter patch was increased, perceived hue shifted away from green and toward blue and yellow. While these effects may not be strong enough to make color schemes based on color components impractical, they can be expected to create slightly distorted perceptions.