CHAPTER FOUR

Color

In the summer of 1997, I designed an experiment to measure human ability to trace paths between connected parts in a three-dimensional diagram. Then, as is my normal practice, I ran a pilot study in order to see whether the experiment was well constructed. By ill luck, the first person tested was a research assistant who worked in my lab. He had far more difficulty with the task than anticipated—so much so that I put the experiment back on the drawing board to reconsider, without trying any more pilot subjects. Some months later, my assistant told me he had just had an eye test and the optometrist had determined that he was color blind. This explained the problems with the experiment. Although it was not about color perception, I had marked the targets red in my experiment. He therefore had had great difficulty in finding them, which rendered the rest of the task meaningless. The remarkable aspect of this story is that my assistant had gone through 21 years of his life without knowing that he was blind to many color differences. This is not uncommon, and it strongly suggests that color vision cannot be all that important to everyday life. In fact, color vision is irrelevant to much of normal vision. It does not help us determine the layout of objects in space, how they are moving, or what their shapes are. It is not much of an overstatement to say that color vision is largely superfluous in modern life; nevertheless, color is extremely useful in data visualization.

Color vision does have a critical function, which is hardly surprising because this sophisticated ability must surely provide some evolutionary advantage. Color helps us break camouflage. Some things differ visually from their surroundings only by their color. An especially important example is illustrated in Figure 4.1. If we have color vision, we can easily see the cherries hidden in the leaves. If we do not, this

Information Visualization. DOI: 10.1016/B978-0-12-381464-7.00004-1
© 2013 Elsevier, Inc. All rights reserved.

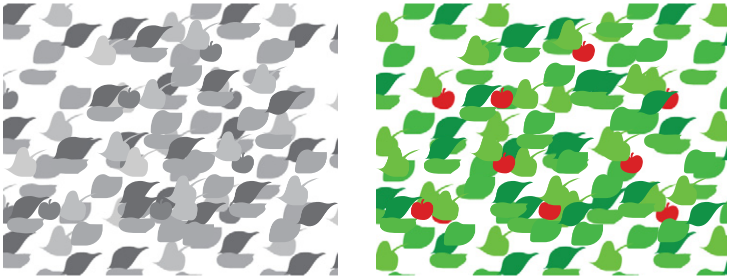

Figure 4.1 Finding the cherries is much easier with color vision.

becomes much harder. Color also tells us much that is useful about the material properties of objects. This is crucial in judging the condition of our food. Is this fruit ripe or not? It this meat fresh or putrid? What kind of mushroom is this? It is also useful if we are making tools. What kind of stone is this? Clearly, these can be life-or-death decisions. In modern hunter–gatherer societies, men are the hunters and women are the gatherers. This may have been true for long periods of human evolution, which could explain why it is mostly men who are color blind. If they had been gatherers, they would have been more than likely to eat poison berries—a selective disadvantage. In the modern age of supermarkets, these skills are much less valuable; this is perhaps why color deficiencies so often go unnoticed.

The role that color plays ecologically suggests ways that it can be used in information display. It is useful to think of color as an attribute of an object rather than as its primary characteristic. It is excellent for labeling and categorization, but poor for displaying shape, detail, or spatial layout. These points are elaborated in this chapter. We begin with an introduction to the basic theory of color vision to provide a foundation for the applications. The latter half of the chapter consists of a set of four visualization problems requiring the effective use of color; these have to do with color selection interfaces, color labeling, pseudocolor sequences for mapping, color reproduction, and color for multidimensional discrete data. Each has its own special set of requirements. Some readers may wish to skip directly to the applications, sampling the more technical introduction only as needed.

Trichromacy Theory

The most important fact about color vision is that we have three distinct color receptors, called *cones*, in our retinas that are active at normal light levels—hence *trichromacy*. We also have rods, sensitive at low light levels, but they are so overstimulated in all but the dimmest light that their influence on color perception can be ignored. Thus, in order to understand color vision, we need only consider the cones. The fact that there are only three receptors is the reason for the basic three-dimensionality of human color vision. The term *color space* means an arrangement of colors in a three-dimensional space. In this chapter, a number of color spaces, designed for different purposes, are discussed. Complex transformations are sometimes required to convert from one color space to another, but they are all three dimensional, and this three-dimensionality derives ultimately from the three cone types. This is the reason why there are three different colors of liquid crystal in a television screen—red, green, and blue—and this is the reason why we learn in school that there are three primary paint colors—red, yellow, and blue. It is also the reason why printers have a minimum of three colored inks for color printing—cyan, magenta, and yellow. Engineers should be grateful that humans have only three color receptors. Some birds, such as chickens, have as many as 12 different kinds of color-sensitive cells. A television set for chickens would require 12 types of differently colored pixels!

Figure 4.2 shows the human cone sensitivity functions. The plots show how light of different wavelengths is absorbed by the three different receptor types (S, M, L). It is evident that two of the functions, L and M, which peak at 540 nanometers and 580 nanometers, respectively, overlap considerably; the third, S, is much more distinct, with peak sensitivity at 450 nanometers. The short-wavelength S receptor absorbs light in the blue part of the spectrum and is much less sensitive, which is another reason (besides chromatic aberration, discussed in Chapter 2) why we should not show detailed information such as text in pure blue on a black background.

Because only three different receptor types are involved in color vision, it is possible to match a particular patch of colored light with a mixture of just three colored lights, usually called *primaries*. It does not matter that the target patch may have a completely different spectral composition. The only thing that matters is that the matching primaries are balanced to produce the same response from the cone receptors as the

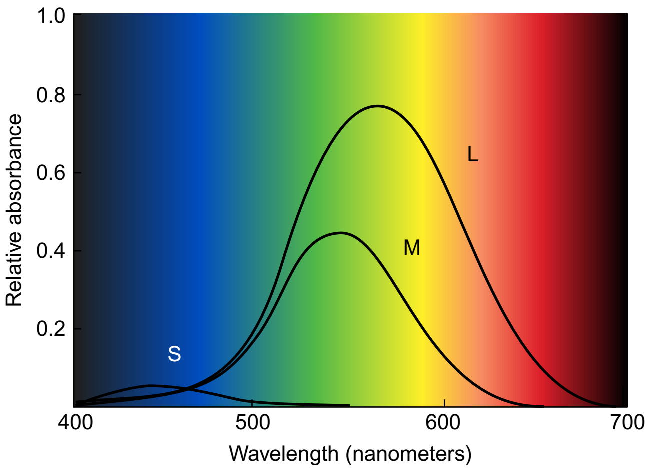

Figure 4.2 Cone sensitivity functions. The colors are only rough approximations to spectrum hues. *Abbreviations*: S, short-wavelength cone sensitivity; M, medium-wavelength cone sensitivity; L, long-wavelength cone sensitivity.

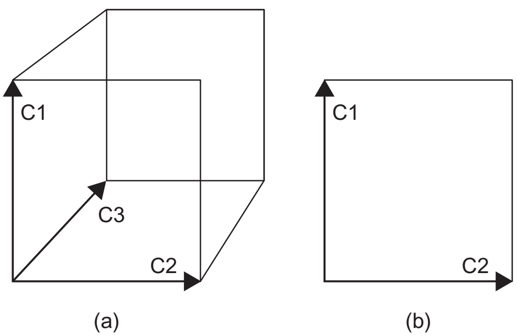

Figure 4.3 (a) Cone response space defined by the response to a colored light of each of the three cone types. (b) The space becomes two dimensional in the case of common color deficiencies.

patch of light to be matched. Figure 4.3(a) illustrates the three-dimensional space formed by the responses of the three cones.

Color Blindness

An unfortunate result of using color for information coding is the creation of a new class of people with a disability. Color blindness already disqualifies applicants for some jobs, such as telephone line maintenance workers, because of the myriad colored wires, and pilots, because of the need to distinguish color-coded lights. About 10% of the male population and about 1% of the female population have some form of color vision deficiency. The most common deficiencies are explained by lack of either the long-wavelength-sensitive cones (protanopia) or the medium-wavelength-sensitive cones (deuteranopia). Both protanopia and deuteranopia result in an inability to distinguish red and green, meaning that the cherries in Figure 4.1 are difficult for people with these deficiencies to see. One way to describe color vision deficiency is by pointing out that the three-dimensional color space of normal color vision collapses to a two-dimensional space, as shown in Figure 4.3(b).

Color Measurement

The fact that we can match any color with a mixture of no more than three primary lights is the basis of colorimetry. An understanding of colorimetry is essential for anyone who wishes to specify colors precisely for reproduction.

We can describe a color by the following equation:

$$C \equiv rR + gG + bB \tag{4.1}$$

where C is the color to be matched; R, G, and B are the primary light sources to be used to create a match; and r, g, and b represent the amounts of each primary light. The $\equiv$ symbol is used to denote a perceptual match; that is, the sample and the mixture of the red, green, and blue (*rR*, *gG*, *bB*) primaries look identical. Figure 4.4 illustrates the concept. Three projectors are set up with overlapping beams. In the figure, the beams only partially overlap so that the mixing effect can be illustrated, but in a color-matching experiment they would overlap perfectly. To match the lilac-colored sample, the projectors are adjusted so that a large amount of light comes from the red and blue projectors and a smaller amount of light comes from the green projector.

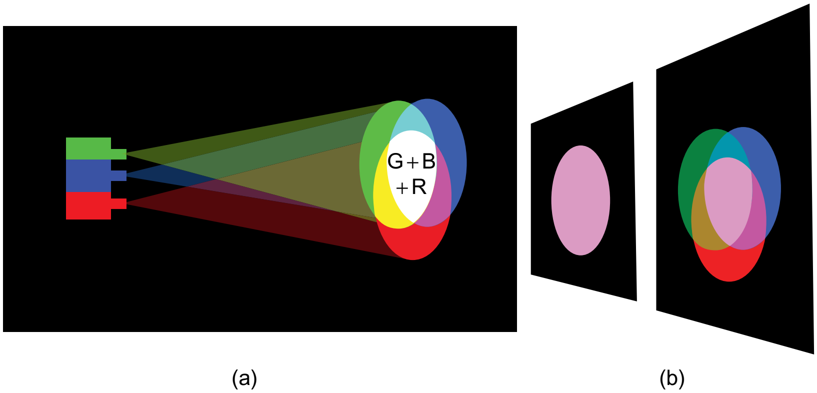

Figure 4.4 A color-matching setup. (a) When the light from three projectors is combined the results are as shown. Yellow light is a mixture of red and green. Purple light is a mixture of red and blue. Cyan light is a mixture of blue and green. White light is a mixture of red, green, and blue. (b) Any other color can be matched by adjusting the proportions of red, green, and blue lights.

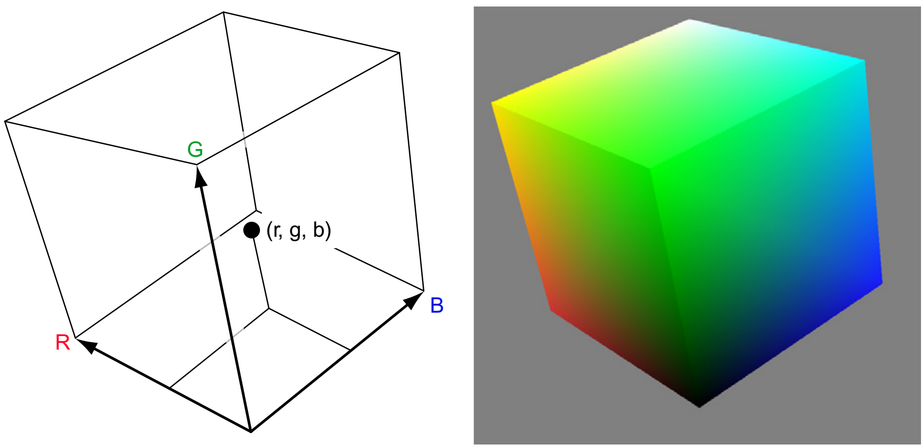

Figure 4.5 The three-dimensional space formed by three primary lights. Any internal color can be created by varying the amount of light produced by each of the primaries.

The *RGB* primaries form the coordinates of a color space, as illustrated in Figure 4.5. If these primaries are physically formed by the phosphor colors of a color monitor, this space defines the *gamut* of the monitor. In general, a gamut is the set of all colors that can be produced by a device or sensed by a receptor system.

It seems obvious that restrictions must be placed on this formulation. So far, we have assumed that the primaries are red, green, and blue, but what if we were to choose other primary lights—for example, yellow, blue, and purple? We have stated no rule saying they must be red, green, and blue. How could we possibly reproduce a patch of red light out of combinations of yellow, blue, and purple lights? In fact, we can only reproduce colors that lie within the *gamut* of the three primaries. Yellow, blue, and purple would simply have a smaller gamut, meaning that if we used them then a smaller range of colors could be reproduced.

The relationship defined in Equation 4.1 is a linear relationship; consequently, if we double the amount of light on the left, we can double the amount of light on each of our primaries and the match will still hold. To make the math simpler, it is also useful to allow the concept of negative light. Thus, we may allow expressions such as

$$C \equiv -rR + gG + bB \tag{4.2}$$

Although this concept may seem nonsensical, because negative light does not exist in nature, it is, in fact, practically useful in the following situation. Suppose we have a colored light that cannot be matched because it is outside the gamut of our three primary sources. We can still achieve a match by adding part of one of the primaries to our sample. If the test samples and the *RGB* primaries are all projected as shown in Figure 4.4, this can be achieved by swiveling one of the projectors around and adding its light to the light of the sample.

If the red projector were redirected in this way, we would have

$$C + rR \equiv gG + bB \tag{4.3}$$

which can be rewritten as

$$C \equiv -rR + gG + bB \tag{4.4}$$

Once we allow the concept of negative values for the primaries, it becomes possible to state that any colored light can be matched by a weighted sum of *any* three primaries as long as each is distinctive in cone space. The primaries do not even have to match an actual color, and in fact the most widely used color standard is based on non-physical primaries, as we shall see.

Change of Primaries

Primaries are arbitrary from the point of view of color mixture—there is no special red, green, or blue light that must be used. Fundamental to colorimetry is the ability to change from one set of primaries to another. This gives us freedom to choose any set of primaries we want. We can choose as primaries the liquid crystal hues of a monitor, three differently colored lasers, or some hypothetical set of lamps. We can even choose to base our primaries on the sensitivities of the human cone receptors. Given a standard way of specifying colors (using a standard set of primaries), we can use a transformation to create that same color on any number of different output devices. This transformation is described in Appendix A.

We now have the foundations of a color measurement and specification system. To illustrate how color specification works, it is useful to think about how it might be done with real lamps, before moving toward more abstract concepts. Red, green, and blue lamps could be manufactured to precise specifications and set up in an instrument so that the amounts of red, green, and blue light falling on a standard white surface could be set by adjusting three calibrated dials, one for each lamp. Identical instruments, each containing sets of colored lamps, would be sent around the world to color experts. Then, to give a precise color specification to someone with the standard instrument, we would simply need to make a color match by adjusting the dials and sending that person the dial settings. The recipient could then adjust his or her own standard lamps to reproduce the color.

Of course, although this approach is theoretically sound, it is not very practical. Standard primary lamps would be very difficult to maintain and calibrate and they would be very expensive. But, we can apply the principle by creating a set of *abstract* primary lamps defined on the basis of the human receptor characteristics. This is how color specification systems work.

One of the basic concepts in any color standard is that of the standard observer. This is a hypothetical person whose color sensitivity functions are held to be typical of all humans. The idea assumes that everyone has the same receptor functions. In fact, although humans do not display exactly the same sensitivities to different colors, with the exception of the color deficiencies, they come close. Most serious color specification is done using the *Commission Internationale de l'Eclairage* (CIE) system of color standards. These are based on standard observer measurements that were made prior to 1931. Color measuring instruments contain glass filters that are derived from the specifications of the human standard observer. One advantage is that glass filters are more stable than lamps.

The CIE system uses a set of abstract observer sensitivity functions called *tristimulus values*; these can be thought of as a set of abstract receptors and they are labeled *XYZ*. They are transformations of actual measured sensitivities, chosen for their mathematical properties. One important feature of the system is that the *Y* tristimulus value is the same as luminance. More details of the way the system is derived are given in Appendix B.

Figure 4.6 illustrates the color volume created by the *XYZ* tristimulus functions of the CIE system. The colors that can actually be perceived are represented as a gray volume entirely contained within the positive space defined by the axes. The colors that can be created by a set of three colored lights, such as the red, green, and blue

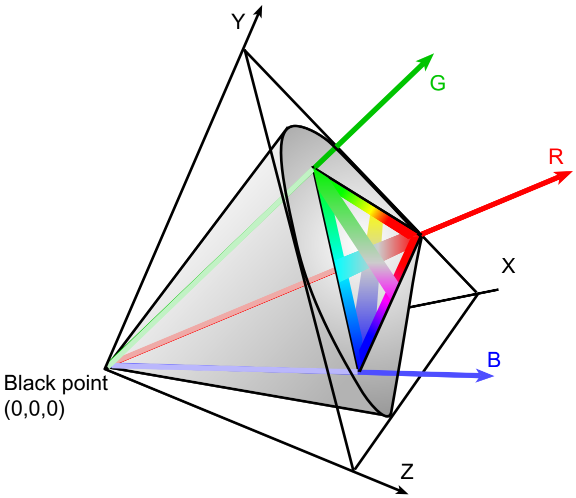

Figure 4.6 The *X*, *Y*, and *Z* axes represent the CIE standard virtual primaries. Within the positive space defined by the axes, the gamut of perceivable colors is represented as a gray solid. The colors that can be created by means of the red, green, and blue monitor primaries are defined by the pyramid enclosed by the *R*, *G*, and *B* lines.

monitor phosphors, are defined by the pyramid-shaped volume within the *RGB* axes, as shown. This is the *monitor gamut*. The *X*, *Y*, and *Z* axes are the CIE primaries, and they are outside the gamut of physically realizable colors.

The CIE tristimulus system based on the standard observer is by far the most widely used standard for measuring colored light. For this reason, it should always be used when precise color specification is required. Because a monitor is a light-emitting device with three primaries, it is relatively straightforward to calibrate a monitor in terms of the CIE coordinates. If a color generated on one monitor, such as a cathode ray tube (CRT), is to be reproduced on another, such as a liquid crystal display, the best procedure is first to convert the colors into the CIE tristimulus values and then to convert them into the primary space of the second monitor.

The specification of surface colors is far more difficult than the specification of lights, because an illuminant must be taken into account and because, unlike lights, pigment colors are not additive. The color that results from mixing paints is difficult to predict because of the complex way that light interacts with pigment. A treatment of surface color measurement is beyond the scope of this book, although later we will deal with perceptual issues related to color reproduction.

Chromaticity Coordinates

The three-dimensional abstract space represented by the *XYZ* coordinates is useful for specifying colors, but it is difficult to understand. As discussed in Chapter 3, there are good reasons for treating lightness, or luminance, information as special. In everyday speech, we often refer to the color of something and its lightness as different and independent properties. Thus, it is useful to have a measure that defines the hue and vividness of a color while ignoring the amount of light. *Chromaticity coordinates* have exactly this property through normalizing with respect to the amount of light.

To transform tristimulus values to chromaticity coordinates, use

$$\begin{aligned} x &= X/(X+Y+Z) \\ y &= Y/(X+Y+Z) \\ z &= Z/(X+Y+Z) \end{aligned} \tag{4.5}$$

Because $x + y + z = 1$, it is sufficient to use x, y values only. It is common to specify a color by its luminance (Y) and its x, y chromaticity coordinates (x, y, Y). The inverse transformation from x, y, Y to tristimulus values is

$$\begin{aligned} X &= Yx/y \\ Y &= Y \\ Z &= (1-x-y)Y/y \end{aligned} \tag{4.6}$$

Figure 4.7 shows a CIE x, y chromaticity diagram and graphically illustrates some of the colorimetric concepts associated with it. Some of the useful and interesting properties of the chromaticity diagram include the following:

1. If two colored lights are represented by two points in a chromaticity diagram, the color of a mixture of those two lights will always lie on a straight line between those two points.

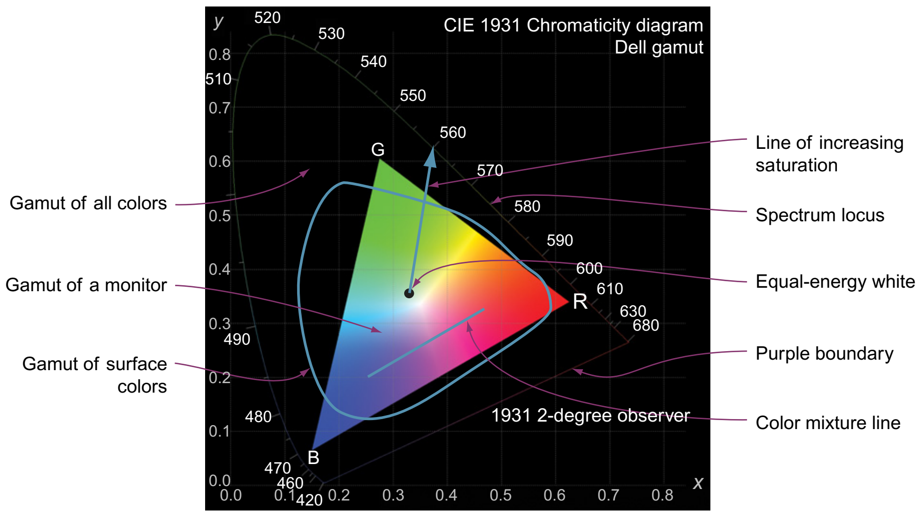

Figure 4.7 CIE chromaticity diagram with various interesting features added. The colored triangle represents the gamut of a computer monitor. Colors as shown are only approximate.

2. Any set of three lights specifies a triangle in the chromaticity diagram. Its corners are given by the chromaticity coordinates of the three lights. Any color within that triangle can be created with a suitable mixture of the three lights. Figure 4.7 illustrates this with typical monitor *RGB* primaries.
3. The *spectrum locus* is the set of chromaticity coordinates of pure monochromatic (single-wavelength) lights. All realizable colors fall within the spectrum locus.
4. The *purple boundary* is the straight line connecting the chromaticity coordinates of the longest visible wavelength of red light (about 700 nm) to the chromaticity coordinates of the shortest visible wavelength of blue (about 400 nm).
5. The chromaticity coordinates of equal-energy white (light having an equal mixture of all wavelengths) are 0.333, 0.333. But, when a white light is specified for some application, what is generally required is one of the CIE standard illuminants. The CIE specifies a number that corresponds to different phases of daylight; of these, the most commonly used is D65. D65 was made to be a careful approximation of daylight with an overcast sky. It also happens to be very close to the mix of light that results when both direct sunlight and light from the rest of the sky fall on a horizontal surface. D65 also corresponds to a black-body radiator at 6500 degrees Kelvin. D65 has chromaticity coordinates $x = 0.313$, $y = 0.329$. Another CIE standard illuminant corresponds to the light produced by a typical incandescent tungsten source. This is illuminant A (chromaticity coordinates $x = 0.448$, $y = 0.407$), and it is considerably more yellow than normal daylight.
6. *Excitation purity* is a measure of the distance along a line between a particular pure spectral wavelength and the white point. Specifically, it is the value given by dividing the distance between the sample and the white point by the distance between the white point and the spectrum locus (or purple boundary). This measure defines the vividness of a color. A less technical, but commonly used, term for this quantity is *saturation*. More saturated colors are more vivid.
7. The complementary wavelength of a color is produced by drawing a line between that color and white and extrapolating to the opposite spectrum locus. Adding a color and its complementary color produces white.

There is a widely used standard for the color of monitor primaries called *sRGB*. The chromaticity coordinates for sRGB are set out in Table 4.1.

When a computer display is used to generate a color, the CIE tristimulus values formed from some set of red, green, and blue settings can be calculated by the following formula:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \frac{x_R}{y_R} & \frac{x_G}{y_G} & \frac{x_B}{y_B} \\ 1 & 1 & 1 \\ \frac{z_R}{y_R} & \frac{z_G}{y_G} & \frac{z_B}{y_B} \end{bmatrix} \begin{bmatrix} Y_R \\ Y_G \\ Y_B \end{bmatrix} \tag{4.7}$$

Table 4.1 *Cromaticity Coordinates for the sRGB Standard*

	Red	Green	Blue
x	0.64	0.30	0.15
y	0.33	0.60	0.06

where (x_R, y_R, z_R), (x_G, y_G, z_G), and (x_B, y_B, z_B) are the chromaticity coordinates of the particular monitor primaries, and Y_R, Y_G, and Y_B are the actual luminance values produced from each phosphor for the particular color being converted. Notice that for a particular monitor the transformation matrix will be constant; only the Y vector will change.

To generate a particular color on a monitor that has been defined by CIE tristimulus values, it is only necessary to invert the matrix and create an appropriate voltage to each of the red, green, and blue electron guns of the monitor. Naturally, to determine the actual value that must be specified, it is necessary to calibrate the monitor's red, green, and blue outputs in terms of luminance and apply gamma correction, as described in Chapter 3. Once this is done, the monitor can be treated as a linear color creation device with a particular set of primaries, depending on its phosphors. For more on monitor calibration, see Cowan (1983). It is also possible to purchase self-calibrating monitors adequate for all but the most demanding applications.

Color Differences and Uniform Color Spaces

Sometimes it is useful to have a color space in which equal perceptual distances are equal distances in the space. Here are three applications:

- **Specification of color tolerances.** When a manufacturer wishes to order a colored part from a supplier, such as a plastic molding for an automobile, it is necessary to specify the color tolerance within which the part will be accepted. It only makes sense for this tolerance to be based on human perception, because ultimately it is people who decide whether the door trim matches the upholstery.
- **Specification of color codes.** If we need a set of colors to code data, such as different wires in a cable, we would normally like those colors to be as distinct as possible so that they will not be confused. This can be accomplished by making them as far apart as possible in a uniform color space.
- **Pseudocolor sequences for maps.** Many scientific maps use sequences of colors to represent ordered data values. This technique, called *pseudocoloring*, is widely used in astronomy, physics, medical imaging, and geophysics. A uniform color space can be used to create perceptually equal steps in a sequence of colors.

The CIE XYZ color space is very far from being perceptually uniform; however, in 1978, the CIE produced a set of recommendations on the use of two uniform color spaces that are transformations of the *XYZ* color space. These are called the *CIElab* and the *CIEluv* uniform color spaces. The reason why there are two color spaces, rather than one, has to do with the fact that different industries, such as the paint industry, had already adopted one standard or the other. Also, the two standards have somewhat different properties that make them useful for different tasks. Only the *CIEluv* formula is described here. It is generally held to be better for specifying large color differences; however, one measurement made using the *CIElab* color difference formula is worth noting. Using *CIElab*, Hill et al. (1997) estimated that there are between two and six million discriminable colors available within the gamut of a color monitor.

The *CIEluv* equations are:

$$\begin{aligned} L^* &= 116(Y/Y_n)^{1/3} - 16 \\ u^* &= 13L^*(u' - u'_n) \\ v^* &= 13L^*(v' - v'_n) \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} u' &= \frac{4X}{X + 15Y + 3Z} \quad & u'_n &= \frac{4X}{X_n + 15Y_n + 3Z_n} \\ v' &= \frac{9Y}{X + 15Y + 3Z} \quad & v'_n &= \frac{4X}{X_n + 15Y_n + 3Z_n} \end{aligned} \tag{4.9}$$

u' and v' are a projective transformation of the x, y chromaticity diagram, designed to produce a perceptually more uniform color space. X_n, Y_n, and Z_n are the tristimulus values of a reference white. To measure the difference between colors, ΔE^*_{uv}, the following formula is used:

$$\Delta E^*_{uv} = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2} \tag{4.10}$$

The *CIEluv* system retains many of the useful properties of the *XYZ* tristimulus values and the x, y chromaticity coordinates.

The u', v' diagram is shown in Figure 4.8. Its official name is the *CIE 1976 Uniform Chromaticity Scale diagram*, or UCS diagram. Because u', v' is a projective transformation, it retains the useful property that blends of two colors will lie on a line between the u', v' chromaticity coordinates. (It is worth noting that this is not a property of the *CIElab* uniform color space.)

The u^*, v^* values change the scale of u', v' with respect to the distance from black to white defined by the sample lightness, L^* (recall from Chapter 3 that L^* requires Y_n, a reference white in the application environment). The reason for this is straightforward: The darker the colors, the fewer we can see. At the limit, there is only one color: black.

A value of 1 for ΔE^*_{uv} is an approximation to a *just noticeable difference* (JND).

Although they are useful, uniform color spaces provide, at best, only a rough first approximation of how color differences will be perceived. In complex environments, many factors influence how much difference is seen between two adjacent colors. Contrast effects can radically alter the shape of the color space. Small patches of light give different results than large patches. In general, we are much more sensitive to differences between large patches of color. When the patches are small, the perceived differences are smaller, and this is especially true in the yellow–blue direction. Ultimately, with very small samples, small-field tritanopia occurs; this is the inability to distinguish colors that are different in the yellow–blue direction. Figure 4.9 shows two examples of large patches of color on a white background

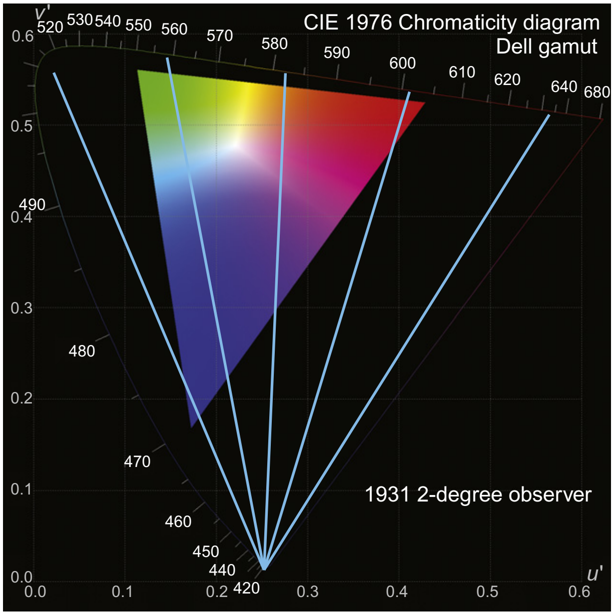

Figure 4.8 CIE *Lu′v′* UCS diagram. The lines radiating from the lower part of the diagram are called *tritanopic confusion lines*. Colors that differ along these lines can still be distinguished by the great majority of color-blind individuals.

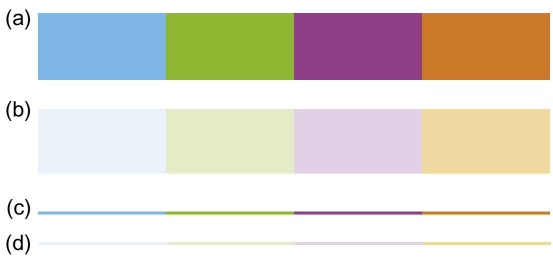

Figure 4.9 (a) Large samples of saturated colors. (b) Large samples of the same colors less saturated. (c) Small samples of the same saturated colors. (d) Small samples of the less saturated colors.

and the same set of colors in smaller patches. In the larger patches, the low-saturation colors are easy to distinguish. They are not so easy to distinguish in the small patches.

[G4.1] Use more saturated colors when color coding small symbols, thin lines, or other small areas. Use less saturated colors for coding large areas.

Opponent Process Theory

Late in the 19th century, German psychologist Ewald Hering proposed the theory that there are six elementary colors and that these colors are arranged perceptually as opponent pairs along three axes: black–white, red–green, and yellow–blue (Hering, 1920). In recent years, this principle has become a cornerstone of modern color theory, supported by a variety of experimental evidence (for a review, see Hurvich, 1981). Modern opponent process theory has a well-established physiological basis: Input from the cones is processed into three distinct channels immediately after the receptors. The luminance channel (black–white) is based on input from all the cones. The red–green channel is based on the difference of long- and middle-wavelength cone signals. The yellow–blue channel is based on the difference between the short-wavelength cones and the sum of the other two. These basic connections are illustrated in Figure 4.10. There are many lines of scientific evidence for the opponent process theory. These are worth examining, because they provide useful insights.

Naming

Opponent color theory predicts that certain color names should not occur in combination. We often describe colors using combinations of color terms, such as *yellowish green* or *greenish blue*. The theory predicts that people will never use *reddish green* or

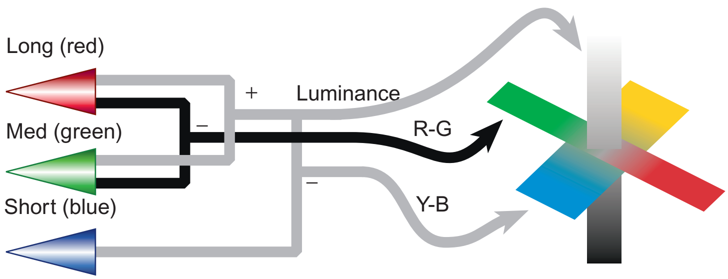

Figure 4.10 In the color opponent process model, cone signals are transformed into black–white (luminance), red–green, and yellow–blue channels.

yellowish blue, because these colors are polar opposites in the opponent color theory (Hurvich, 1981). Experiments have confirmed this.

Cross-Cultural Naming

In a remarkable study of more than 100 languages from many diverse cultures, anthropologists Berlin and Kay (1969) showed that primary color terms are remarkably consistent across cultures (Figure 4.11). In languages with only two basic color words, these are always black and white; if a third color is present, it is always red; the fourth and fifth are either yellow and then green, or green and then yellow; the sixth is always blue; the seventh is brown, followed by pink, purple, orange, and gray in no particular order. The key point here is that the first six terms define the primary axes of an opponent color model. This provides strong evidence that the neural basis for these names is innate; otherwise, we might expect to find cultures where lime green or turquoise is a basic color term. The cross-cultural evidence strongly supports the idea that certain colors—specifically, red, green, yellow, and blue—are far more valuable in coding data than others.

Unique Hues

There is something special about yellow. If subjects are given control over a device that changes the spectral hue of a patch of light and are told to adjust it until the result is a pure yellow, neither reddish nor greenish, they do so with remarkable accuracy. In fact, they are typically accurate within 2 nm (Hurvich, 1981).

Interestingly, there is good evidence for two unique greens. Most people set a pure green at about 514 nm, but about a third of the population sees pure green at about 525 nm (Richards, 1967). This may be why some people argue about the color turquoise; some people consider it to be a variety of green, whereas others consider it to be a kind of blue.

It is also significant that unique hues do not change a great deal when the overall luminance level is changed (Hurvich, 1981). This supports the idea that chromatic perception and luminance perception really are independent.

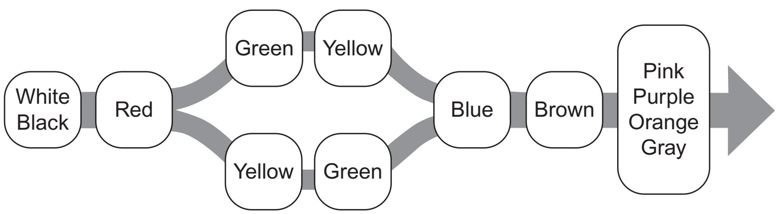

Figure 4.11 This is the order of appearance of color names in languages around the world, according to the research of Berlin and Kay (1969). The order is fixed, with the exception that sometimes yellow is present before green and sometimes the reverse is the case.

Neurophysiology

Neurophysiological studies have isolated classes of cells in the primary visual cortexes of monkeys that have exactly the properties of opponency required by the opponent process theory. Red–green and yellow–blue opponent cells exist, and other configurations do not appear to exist (de Valois & de Valois, 1975).

Categorical Colors

There is evidence that certain colors are canonical in a sense that is analogous to the philosopher Plato's theory of forms. Plato proposed that there are ideal objects, such as an ideal horse or an ideal chair, and that real horses and chairs can be defined in terms of their differences from the ideal. Something like this appears to operate in color naming. If a color is close to an ideal red or an ideal green, it is easier to remember. Colors that are not basic, such as orange or lime green, are not as easy to remember.

There is evidence that confusion between color codes is affected by color categories. Kawai et al. (1995) asked subjects to identify the presence or absence of a chip of a particular color. The subjects took much longer if the chip was surrounded by distracting elements that were of a different color but belonged to the same color category than if the chip was surrounded by distracting elements that were equally distinct according to the sense of a uniform color space but crossed a color category boundary.

Post and Greene (1986) carried out an extensive experiment on the naming of colors produced on a computer monitor and shown in a darkened room. They generated 210 different colors, each in a 2-degree (of visual angle) patch with a black surround. Perceiving colors in darkness and in isolation is a very special case that is very different from the usual way we see color, so the results should not be taken as generally applicable. Nevertheless they are interesting.

Figure 4.12 illustrates the color areas that were given a specific name with at least 75% reliability. A number of points are worth noting:

- The fact that only eight colors plus white were consistently named, even under these highly standardized conditions, strongly suggests that only a very small number of colors can be used effectively as category labels.
- The pure monitor red was actually named orange most of the time. A true color red required the addition of a small amount from the blue monitor primary.
- The specific regions of color space occupied by particular colors should not be given much weight. The data was obtained with a black background. Because of contrast effects, different results are to be expected with white and colored backgrounds.

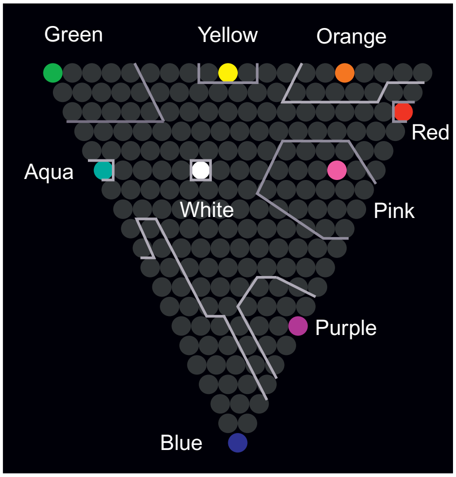

Figure 4.12 The results of an experiment in which subjects were asked to name 210 colors produced on a computer monitor. Outlined regions show the colors that were given the same name with better than 75% probability.

Properties of Color Channels

From the perspective of data visualization, the different properties of the color channels have profound implications for the use of color. The most significant differences are between the two chromatic channels and the luminance channel, although the two color channels also differ from each other.

To display data on the luminance channel *alone* is easy; it is stimulated by patterns that vary only from black to white through shades of gray. But, with careful calibration (which must be customized to individual subjects), patterns can be constructed that vary only for the red–green or the yellow–blue channel. A key quality of such a pattern is that its component colors must not differ in luminance. This is called an *isoluminant* or *equiluminous* pattern. In this way, the different properties of the color channels can be explored and compared with the luminance channel capacity.

Spatial Sensitivity

According to a study by Mullen (1985), the red–green and yellow–blue chromatic channels are each capable of carrying only about one-third the amount of detail carried by the black–white channel. Because of this, purely chromatic differences are not suitable for displaying any kind of fine detail. Figure 4.13 illustrates this problem with colored text on an equiluminous background. In the part of the figure where there is only a chromatic difference between the text and the background, the text becomes very difficult to read.

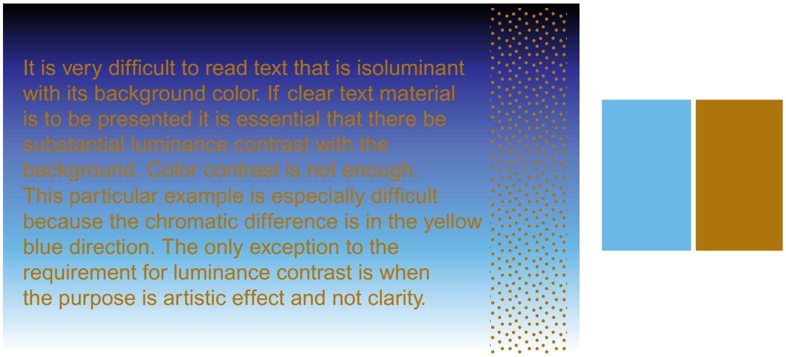

Figure 4.13 Brown text on a blue gradient. Notice how difficult it is to read the text where the luminance is equal, despite a large chromatic difference. Brown is a dark yellow so these colors differ on the blue–yellow channel.

[G4.2] When small symbols, text, or other detailed graphical representations of information are displayed using color on a differently colored background, always ensure luminance contrast with the background. This guideline is a variation of G3.4.

Stereoscopic Depth

It appears to be impossible, or at least very difficult, to see stereoscopic depth in stereo pairs that differ only in terms of the color channels (Lu & Fender, 1972; Gregory, 1977). This is because stereoscopic depth perception is based primarily on information from the luminance channel.

[G4.3] Ensure adequate luminance contrast in order to define features important for perceiving stereoscopic depth.

Motion Sensitivity

If a pattern is created that is equiluminous with its background and contains only chromatic differences, and that pattern is set in motion, something strange occurs. The moving pattern appears to move much more slowly than a black against white pattern moving at the same speed (Anstis & Cavanaugh, 1983). Motion perception appears to be primarily based on information from the luminance channel.

[G4.4] Ensure adequate luminance contrast in order to define features important for perceiving moving targets.

Form

We are very good at perceiving the shapes of surfaces based on their shading; however, when the shading is transformed from a luminance gradient into a purely chromatic gradient, the impression of surface shape is much reduced. Perception of shape and form appears to be processed mainly through the luminance channel (Gregory, 1977).

> **[G4.5]** When applying shading to define the shape of a curved surface, use adequate luminance (as opposed to chromatic) variation. This is a supplement to G2.1.

Even though small shapes should not be defined by purely chromatic boundaries, this does not apply to large shapes, such as the R in Figure 4.14, which can be seen clearly. Nevertheless, the shape will be more clearly perceived if a luminance difference border is added, however thin. This also helps distinguish the color of the shape.

> **[G4.6]** If large areas are defined using nearly equiluminous colors, consider using thin border lines with large luminance differences (from the colors of the areas) to help define the shapes.

To summarize this set of properties, the red–green and yellow–blue channels are inferior to the luminance channel in almost every respect. The implications for data display are clear. Purely chromatic differences should never be used for displaying object shape, object motion, or detailed information such as text. From this perspective, color would seem almost irrelevant and certainly a secondary method for information display; nevertheless, when it comes to coding information, using color to display data categories is usually the best choice. To see why, we need to look beyond the basic processes that we have been considering thus far.

Figure 4.14 Even large shapes are seen more clearly if a luminance contrast boundary is provided.

Color Appearance

Color (as opposed to luminance) processing, it would appear, does not help us to understand the shape and layout of objects in the environment. Color does not help the hunter aim an arrow accurately. Color does not help us see shape from shading and thereby shape a lump of clay or bread dough. Color does not help us use stereoscopic depth to guide our hands when we reach out to grasp something. But color is useful to the gatherer of food. Fruits and berries are often distinguished by their color.

Color creates a kind of visual attribute of objects: This is a red berry; that is a yellow door. Color names are used as adjectives because colors are perceived as attributes of objects. This suggests a most important role for color in visualization—namely, the coding of information. Visual objects can represent complex data entities, and colors can naturally code attributes of those objects.

Monitor Surrounds

The *XYZ* tristimulus values of a patch of light physically define a color, but they do not tell us how it will appear. Depending on the surrounding colors in the environment and a whole host of spatial and temporal factors, the same physical color can look very different. If it is desirable that color appearance be preserved, it is important to pay close attention to surrounding conditions. In a monitor-based display, a large patch of standardized reference white will help ensure that color appearance is preserved. When colors are reproduced on paper, viewing them under a standard lamp will help preserve their appearance. In the paint and fabric industries, where color appearance is critical, standard viewing booths are used. These booths contain standard illumination systems that can be set to approximate daylight or a standard indoor illuminant, such as a typical tungsten lightbulb or halogen lamp.

Color Constancy

The mechanisms of surface lightness constancy, discussed at some length in Chapter 3, generalize to trichromatic color perception. Both chromatic adaptation and chromatic contrast occur and play a role in color constancy. Differential adaptation in the cone receptors helps us to discount the color of the illumination in the environment. When there is colored illumination, different classes of cone receptors undergo independent changes in sensitivity; thus, when the illumination contains a lot of blue light, the short-wavelength cones become relatively less sensitive than the others. The effect of this is to shift the neutral point at which the three receptor types are in equilibrium, such that more blue light must be reflected from a surface for it to seem white. This discounting of the illumination, of course, is exactly what is necessary for color constancy. A piece of everyday evidence that adaptation is effective is the fact that not many people are aware of how much yellower ordinary tungsten room lighting is than daylight. The consequence for adaptation is that we cannot see absolute colors, and when colored symbols appear on differently colored backgrounds their apparent hue will be altered.

Color Contrast

Chromatic contrast occurs in a way that is similar to the lightness contrast effects discussed and illustrated in Chapter 3. Figure 4.15 shows a color contrast illusion. It has been shown that contrast effects can distort readings from color-coded maps (Cleveland & McGill, 1983; Ware, 1988). Contrast effects can be theoretically accounted for by activity in the color opponent channels (Ware & Cowan, 1982). However, as with lightness contrast, the ultimate purpose of the contrast-causing mechanism is to help us see surface colors accurately by revealing differences between colored patches and background regions.

From the point of view of the monitor engineer and the user of color displays, the fact that colors are perceived relative to their overall context has the happy consequence of making the eye relatively insensitive to poor color balance. Try comparing an image on a computer screen with that same image printed. Individual colors will undoubtedly be very different, but the overall impression and the information conveyed will be mostly preserved. This is because relative color is much more important than absolute color.

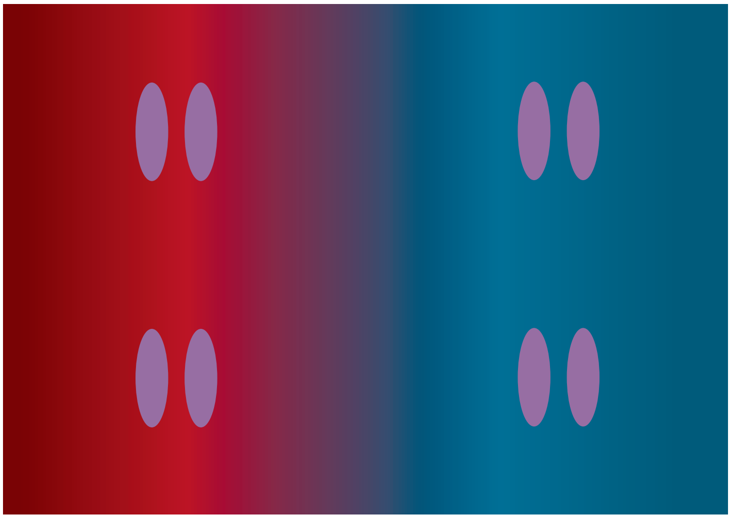

Figure 4.15 A color contrast illusion. The ellipses are all the same color but seem pinker on the right and bluer on the left.

Saturation

When describing color appearance in everyday language, people use many terms in rather imprecise ways. Besides using color names such as *lime green*, *mauve*, *brown*, *baby blue*, and so on, people also use adjectives such as *vivid*, *bright*, and *intense* to describe colors that seem especially pure. Because these terms are used so variably, scientists use the technical term *saturation* to denote how pure colors seem to the viewer. A high-saturation color is vivid, and a low-saturation color is close to black, white, or gray. In terms of the color opponent channels, high-saturation colors are those that give a strong signal on one or both of the red–green and yellow–blue channels.

Equal-saturation contours have been derived from psychophysical experiments (Wyszecki & Stiles, 1982). Figure 4.16(a) shows a plot of equal-saturation values in a CIE chromaticity diagram. These contours, derived from studies of human perception, show that it is possible to obtain much more highly saturated red, green, and blue colors on a monitor than yellow, cyan, or purple values. Figure 4.16(b) shows equal-saturations contours (not derived from perception) in the popular hue, saturation, and value (HSV) transformation commonly used in computer graphics (Smith, 1978). Comparing the two diagrams, it is striking that two colors having equal HSV saturations will not have close to equal perceptual saturation. In particular, pure red, green, and blue on a monitor will be more perceptually saturated than pure cyan, magenta, or yellow. To obtain a set of perceptually equally saturated colors we would have to restrict our color gamut to contour 6 in Figure 4.15(a), but this would mean giving up a large amount of useful RGB color space, including the most vivid colors, so this is usually inadvisable.

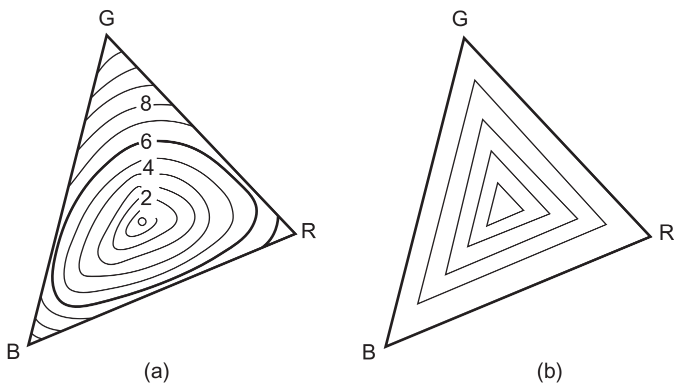

Figure 4.16 (a) The triangle represents the gamut of colors obtained using a computer monitor plotted in CIE chromaticity coordinates. The contours show perceptually determined equal-saturation contours. (b) Equal-saturation contours created using the HSV color space, also plotted in chromaticity coordinates.

Using the general principle that stronger visual effects should be used to show greater quantities (G1.3), we can establish a guideline for the use of saturation in color coding. Because there are few discriminable steps in saturation, and because of contrast effect that may occur if the background is variable, only a few saturation levels can be reliably judged.

> **[G4.7]** If using color saturation to encode numerical quantity, use greater saturation to represent greater numerical quantities. Avoid using a saturation sequence to encode more than three values.

Brown

Brown is one of the most mysterious colors. Brown is dark yellow. Whereas people talk about a light green or a dark green, a light blue or a dark blue, they do not talk about dark yellow. When colors in the vicinity of yellow and orange yellow are darkened, they turn to shades of brown and olive green. Unlike red, blue, and green, brown requires that there be a reference white somewhere in the vicinity for it to be perceived. Brown appears qualitatively different from orange yellow.

There is no such thing as an isolated brown light in a dark room, but when a yellow or yellowish orange is presented with a bright white surround, brown appears. The relevance to visualization is that, if color sets are being devised for the purposes of color coding—for example, a set of blues, a set of reds, a set of greens, and a set of yellows—in the case of yellows, brown may not be recognized as a set member.

Applications of Color in Visualization

So far, this chapter has been mainly a presentation of the basic theory underlying color vision and color measurement. Now we shift the emphasis to applications of color, for which new theory will be introduced only as needed. We will examine four different application areas: color selection interfaces, color labeling, color sequences for map coding, and color reproduction. Each of these presents a different set of problems, and each benefits from an analysis in terms of the human perception of color. We will use these applications to develop guidelines and continue to develop theory.

Application 1: Color Specification Interfaces and Color Spaces

In data visualization software, drawing applications, and CAD systems, it is often essential to let users choose their own colors. There are a number of approaches to this user interface problem. The user can be given a set of controls to specify a point in a three-dimensional color space, a set of color names to choose from, or a palette of predefined color samples.

Color Spaces

The simplest color interface to implement on a computer involves giving someone controls to adjust the amounts of red, green, and blue light that combine to make a patch of color on a monitor. The controls can take the form of sliders, or the user can simply type in three numbers. This provides access, in a straightforward way, to any point within the RGB color cube shown in Figure 4.5; however, although it is simple, many people find this kind of control confusing. For example, most people do not know that to get yellow you must add red and green. There have been many attempts to make color interfaces easier to use.

Many of the most widely used color interfaces in computer graphics are based on the hue, saturation, and value (HSV) color space (Smith, 1978). This is a simple transformation from HSV coordinates to *RGB* monitor coordinates. *Hue*, in Smith's scheme, represents an approximation to the visible spectrum by interpolating in sequence from red to yellow (= red + green) to green to cyan (= green + blue) to blue to purple (= blue + red) and back to red. *Saturation* is the distance from neutral monitor values, on the white–gray–black axis, to the purest hue possible given the limits of monitor primaries. Figure 4.17 shows how hue and saturation can be laid out in two dimensions, with hue on one axis and saturation on the other, based on the HSV transformation of monitor primaries. As Figure 4.16(b) shows, HSV creates only the crudest approximation to perceptually equal saturation contours. *Value* is the name given to the black–white axis. Some color specification interfaces based on HSV allow the user to control hue, saturation, and value variables with three sliders.

Because color research has shown the luminance channel to be very different from the chromatic (red–green, yellow–blue) channels, it is a good idea to separate a luminance

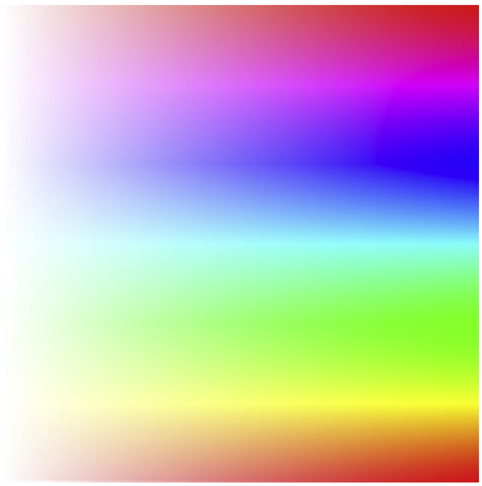

Figure 4.17 This plot shows hue and saturation, based on Smith's (1978) transformation of the monitor primaries.

(or lightness) dimension from the chromatic dimensions in a color specification interface. In addition, because the chromatic channels are perceived integrally, it is usually best to lay out the various hue and saturation choices on a plane, but not as shown in Figure 4.17, as this devotes far too much space to neutral colors and does not reflect the perceptual structure of color space derived from the color opponent channels. Figure 4.18 provides a selection of much better layouts. All are compromises among the constraints of colors produced by computer monitors, the desire to produce a neat geometric space, and the goal of producing a perceptually meaningful representation of a color plane orthogonal to the luminance channel.

> **[G4.8]** In an interface for specifying colors, consider laying out the red–green and yellow–blue channel information on a plane. Use a separate control for specifying the dark–light dimension.

A common interface method is to provide a single slider control for the black–white dimension and to lay out the two opponent color dimensions on a chromatic plane. The idea of laying out colors on a plane has a long history; for example, a color circle is a feature of a color textbook created for artists by Rood (1897). With the invention of computer graphics, it has become far simpler to create and control colors, and many ways of laying out colors are now available.

Figure 4.18(a) shows a color circle with red, green, yellow, and blue defining opposing axes. Many such color circles have been devised over the past century. They differ mainly in the spacing of colors around the periphery.

Figure 4.18(b) shows a color triangle with the monitor primaries, red, green, and blue, at the corners. This color layout is convenient because it has the property that mixtures of two colors will lie on a line between them (assuming proper calibration); however,

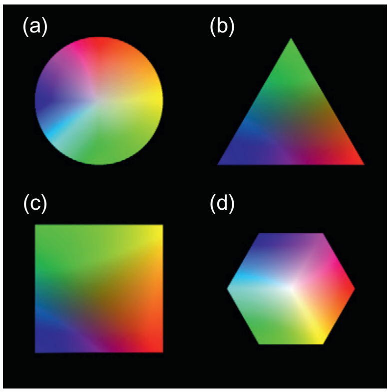

Figure 4.18 A sampling of four different geometric color layouts, each of them embodying the idea of a chromatic plane. (a) Circle. (b) Triangle. (c) Square. (d) Hexagon.

because of linear interpolation, only a very weak yellow occurs between the red and green corners (50% red, 50% green). The strongest yellow on a monitor comes from having both red and yellow at full strength.

Figure 4.18(c) shows a color square with the opponent color primaries, red–green and yellow–blue, at opposite corners (Ware & Cowan, 1990).

Figure 4.18(d) shows a color hexagon with the colors red, yellow, green, cyan, blue, and magenta at the corners. This represents a plane through the single-hexcone color model (Smith, 1978). The hexagon representation has the advantage that it gives both the monitor primaries (red, green, and blue) and the print primaries (cyan, magenta, and yellow) prominent positions around the circumference.

To create a color interface using one of these color planes, it is necessary to allow the user to pick a sample from the color plane and adjust its lightness with a luminance slider or some other control. In some interfaces, when the luminance slider is moved, the entire plane of colors becomes lighter and darker according to the currently selected level. For those interested in implementing color interfaces, algorithms for a number of color geometries can be found in Foley et al. (1990).

Another valuable addition to a color design interface is a method for showing a color sample on differently colored backgrounds. This allows the designer to understand how contrast effects can affect the appearance of particular color samples.

[G4.9] In an interface for designing visualization color schemes, consider providing a method for showing colors against different backgrounds.

The problem of the best color selection interface is by no means resolved. Experimental studies have failed to show that one way of controlling color is substantially better than another (Schwarz et al., 1987; Douglas & Kirkpatrick, 1996). Douglas and Kirkpatrick, however, have provided evidence that good feedback about the location of the color being adjusted in color space can help in the process.

Color Naming Systems

The facts that there are so few widely agreed upon color names and that color memory is so poor suggest that choosing colors by name will not be useful except for the simplest applications. People agree on red, green, yellow, blue, black, and white as labels, but not much more; nevertheless, it is possible to remember a rather large number of color names and use them accurately under controlled conditions. Displays in paint stores generally have a standard illuminant and standard background for sample strips containing several hundred samples. Under these circumstances, the specialist can remember and use as many as 1000 color names, but many of the names are idiosyncratic; the colors corresponding to *taupe*, *fiesta red*, and *primrose* are imprecisely defined for most of us. In addition, as soon as these colors are removed from the standard booth, they will change their appearance because of illumination-induced adaptation and contrast effects.

The Natural Color System (NCS), a standardized color naming system, has been developed based on Hering's opponent color theory (1920). NCS was developed in Sweden and is widely used in England and other European countries. In NCS, colors are characterized by the amounts of redness, greenness, yellowness, blueness, blackness, and whiteness that they contain. As shown in Figure 4.19, red, green, yellow, and blue lie at the ends of two orthogonal axes. Intervening "pure" colors lie on the circle circumference, and these are given numbers by sharing out 100 arbitrary units; thus, a yellowish orange might be given the value Y70R30, meaning 70 parts yellow and 30 parts red. Colors are also given independent values on a black–white axis by allocating a blackness value between 0 and 100. A third color attribute, intensity (roughly corresponding to saturation), describes the distance from the grayscale axis. In NCS, for example, the color *spring nymph* becomes 0030-G80Y20, which expands to blackness 00, intensity 30, green 80, and yellow 20 (Jackson et al., 1994). The NCS system combines some of the advantages of a color geometry with a reasonably intuitive and precise naming system.

In North America, other systems are more popular than NCS. The Pantone® system is widely used in the printing industry, and the Munsell system is an important reference for surface colors. The Munsell system is useful because it provides a set of standard color chips designed to represent equal perceptual spacing in a three-dimensional mesh. (Munsell color chips and viewing booths are available commercially, as are Pantone products.) The NCS, Pantone, and Munsell systems were originally designed to be used with carefully printed paper samples providing the reference colors, but computer-based interfaces to these systems have been developed as part of illustration

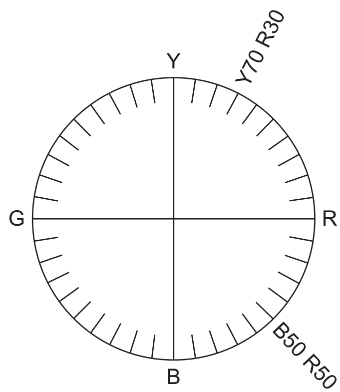

Figure 4.19 The Natural Color System (NCS) circle, defined midway between black and white. Two example color names are shown in addition to the "pure" opponent color primaries. One is an orange yellow and the other is purple.

and design packages. Rhodes and Luo (1996) described a software package that enables transformations between the different systems using the CIE as an intermediate standard.

Color Palettes

When the user wishes to use only a small set of standardized colors, providing a color palette is a good solution to the color selection problem. Often, color selection palettes are laid out in a regular order according to one of the color geometries defined previously. It is useful to provide a facility for the user to develop a personal palette. This allows for consistency in color style across a number of visualization displays.

> **[G4.10]** To support the use of easy-to-remember and consistent color codes, consider providing color palettes for designers.

Sometimes a color palette is based on one of the standard color sets used by the fabric industry or the paint industry. If this is the case, the monitor must be calibrated so that colors actually appear as specified and it must be placed in a standardized viewing environment.

Application 2: Color for Labeling (Nominal Codes)

Suppose we wish to create a visualization where colored symbols represent companies from different industrial sectors—red for manufacturing, green for finance, blue for retail, and so on. The technical name for this kind of labeling is *nominal information coding*. A nominal code does not have to be orderable; it simply must be remembered and recognized. Color can be extremely effective when we wish to make it easy for someone to classify visual symbols into separate categories; giving the objects distinctive colors is often the best solution. One of the reasons why color is often preferred is that the alternatives are generally worse. For example, if we try to create grayscale codes that are easily remembered and unlikely to be confused, we find that four is about the limit: white, light gray, dark gray, and black. Given that white will probably be used for the background and black is likely to be used for text, this leaves only two. In addition, using the gray scale as a nominal code may interfere with shape or detail perception. Chromatic coding can often be employed in a way that only minimally interferes with data presented on the luminance channel. Many perceptual factors must be considered when choosing a set of color labels.

Distinctness. A uniform color space, such as *CIEluv*, can be used to determine the degree of perceived difference between two colors that are placed close together. It might be thought that an algorithm based on *CIEluv* could be used to simply choose a set of colors that are most widely separated, but most color scheme design problems are too complex for this; background colors, symbol sizes, and application-specific requirements all must be taken into account. Also, when we are concerned with the ability to distinguish a color rapidly from a set of other colors, different rules may apply. Bauer et al. (1996) showed that the target color should lie outside the convex hull of the surrounding colors in the CIE color space. This concept is illustrated in Figure 4.20.

Unique hues. The unique hues—red, green, yellow, and blue, as well as black and white—are special in terms of the opponent process model. These colors are also special in the color vocabularies of languages worldwide. Clearly, these colors provide natural choices when a small set of color codes is required.

[G4.11] Consider using red, green, yellow, and blue to color code small symbols.

Contrast with background. In many displays, color-coded objects can be expected to appear on a variety of backgrounds. Simultaneous contrast with background colors can dramatically alter color appearance, making one color look like another. This is one reason why it is advisable to have only a small set of color codes. A method for reducing contrast effects is to place a thin white or black border around the color-coded object. This device is commonly used with signal lights; for example, train signals are displayed on large black background discs. In addition, we should never display codes using purely chromatic differences with the background. There should be a significant luminance difference in addition to the color difference.

[G4.12] For small color-coded symbols, ensure luminance contrast with the background as well as large chromatic differences with the background.

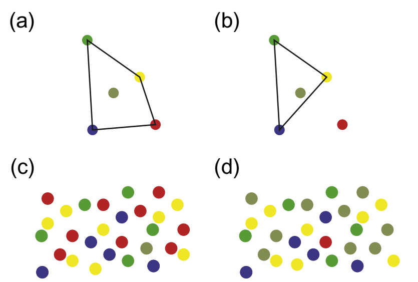

Figure 4.20 The convex hull of a set of colors is defined as the area within a rubber band that is stretched around the colors when they are defined in CIE tristimulus space. Although illustrated in two dimensions here, the concept can easily be extended to three dimensions. (a) Gray is within the convex hull of red, green, yellow, and blue. (b) Red lies outside the convex hull of green, blue, yellow, and gray. (c) The gray dot is difficult to find in a set of red, green, yellow, and blue dots. (d) The red dot is easy to find in a set of green, blue, yellow, and gray dots.

[G4.13] If colored symbols may be nearly isoluminant against parts of the background, add a border having a highly contrasting luminance value to the color, for example, black around a yellow symbol or white around a dark blue symbol.

Figure 4.21 illustrates this principle with a variety of colors against a variety of backgrounds.

Color blindness. Because there is a substantial color-blind population, it may be desirable to use colors that can be distinguished even by people who are color blind. Recall that the majority of color-blind people cannot distinguish colors that differ in a red–green direction. Almost everyone can distinguish colors that vary in a yellow–blue direction, as shown in Figure 4.8. Unfortunately, this drastically reduces the design choices that are available.

[G4.14] To create a set of symbol colors that can be distinguished by most color-blind individuals, ensure variation in the yellow–blue direction.

Figure 4.8 shows the lines defining colors that can be discriminated by most color-blind individuals.

Number. Although color coding is an excellent way to display category information, only a small number of codes can be rapidly perceived. Estimates vary between about five and ten codes (Healey, 1996).

[G4.15] Do not use more than ten colors for coding symbols if reliable identification is required, especially if the symbols are to be used against a variety of backgrounds.

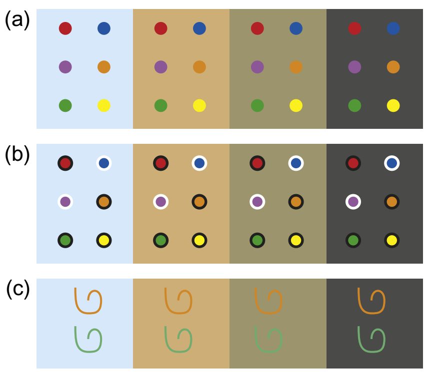

Figure 4.21 (a) Note that at least one member of the set of six symbols lacks distinctness against each background. (b) Adding a luminance contrast border ensures distinctness against all backgrounds. (c) Showing color-coded lines can be especially problematic.

Field size. To avoid the small-field color blindness illustrated in Figure 4.9, do not use very small color-coded areas. In general, the larger the area that is color coded, the more easily colors can be distinguished. Small objects that are color coded should have strong, highly saturated colors for maximum discrimination as already stated in G4.1. When large areas of color coding are used (for example, with map regions), the colors should be of low saturation and differ only slightly from one another. This enables small, vivid color-coded targets to be perceived against background regions.

[G4.16] Use low-saturation colors to color code large areas. Generally, light colors will be best because there is more room in color space in the high-lightness region than in the low-lightness region.

[G4.17] When color coding large background areas overlaid with small colored symbols, consider using all low-saturation, high-value (pastel) colors for the background, together with high-saturation symbols on the foreground.

Figure 4.22 shows two examples, one that follows these guidelines and one that contradicts them.

The goal of highlighting is to make some small subset of a display clearly distinct from the rest, and the same principles apply to the highlighting of text or other features in a display.

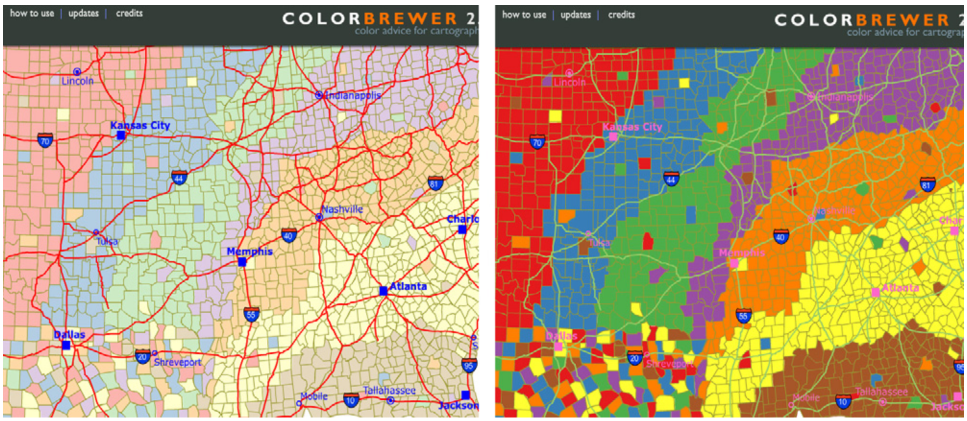

Figure 4.22 On the left is a map using low-saturation light colors for the area coding and high-saturation dark colors for the town and city symbols and linear features. On the right, a much worse solution shows high-saturation coding for areas and low-saturation symbols and linear features. Maps were generated using ColorBrewer2 (http://colorbrewer2.org).

[G4.18] When highlighting text by changing the color of the font, it is important to maintain luminance contrast with the background. With a white background, high-saturation dark colors must be used to change the font color. Alternatively, when changing the background color, low-saturation light colors should be used if the text is black on white.

Figure 4.23 illustrates these two alternatives.

> ***Conventions.*** Color-coding conventions must sometimes be taken into account. Some common conventions are red = hot, red = danger, blue = cold, green = life, and green = go. It is important to keep in mind, however, that these conventions do not necessarily cross cultural borders. In China, for example, red means life and good fortune, and green sometimes means death.

The following is a list of 12 colors recommended for use in coding: red, green, yellow, blue, black, white, pink, cyan, gray, orange, brown, purple. They are illustrated in Figure 4.24. These colors have widely agreed upon category names and are reasonably far apart in color space. The first four colors, together with black and white, are chosen because they are the unique colors that mark the ends of the opponent color axes. The entire set corresponds to the 11 color names found to be the most common in the cross-cultural study carried out by Berlin and Kay (1969), with the addition of cyan.

(a) Higlighting text by changing the characters must be done
using high saturation colors that contrast with the background.

(b)

```
import java.applet.Applet;
import java.awt.Graphics;
import java.awt.Color;

    public class ColorText extends Applet
{
    public void init()
    {
            red=100;
            green=255;
            blue=20;
    }
    public void paint (Graphics g)
    {
            Gr.setColor(new Color(red, green, blue));

            Gr.drawString("ColoredText". 30,50);
    }

    private int red;
    private int green;
    private int blue;
}
```

Figure 4.23 Two different methods for highlighting black text. (a) Change text itself using a relatively dark, high-saturation color. (b) Change text background using low-saturation light colors. Both maintain luminance contrast.

The colors in the first set of six would normally be used before choosing any from the second set of six.

Sometimes it is useful to group color codes into families. This can be done by using hue as a primary attribute denoting family membership, with secondary values mapped to a combination of saturation and lightness. Figure 4.25 illustrates some examples. Generally, we cannot expect to get away with more than two different color steps in each family. The canonical red, green, and blue hues make good categories for defining families. Yellow is not so good because dark yellow is perceived as belonging to a different family and yellow has few discriminable saturation steps. Family members then can be distinguished from one another by saturation, as in Figure 4.25(a), or, even better, by saturation and lightness, as in Figure 4.25(b). Interior designers often consider a family of warm colors (nearer to red in color space) to be distinct from a family of cool colors (nearer to blue and green in color space), although the psychological validity of this is questionable.

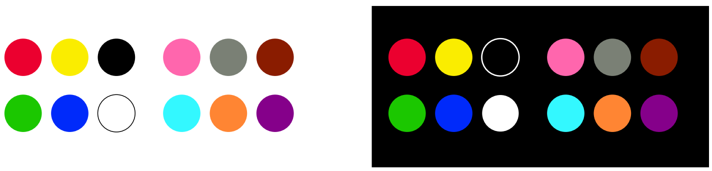

Figure 4.24 A set of 12 colors for use in labeling. The same colors are shown on a white and a black background.

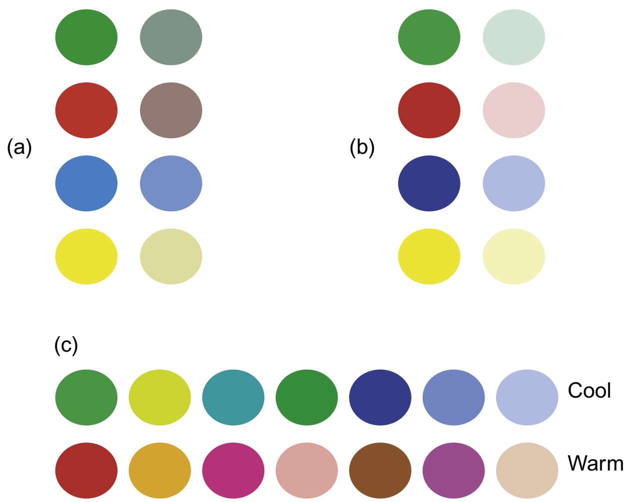

Figure 4.25 Families of colors. (a) Pairs related by hue; family members differ in saturation. (b) Pairs related by hue; family members differ in saturation and lightness. (c) A family of cool hues and a family of warm hues.

Application 3: Color Sequences for Data Maps

Somewhere in almost every newspaper and on every weather website is a map where regions are colored differently to show the forecast temperatures. Red is used to show hot weather, blue is used to show cold weather, and other colors are arranged in between, often using the colors of the rainbow, blue–cyan–green–yellow–orange–red.

Pseudocoloring is the technique of representing continuously varying map values using a sequence of colors. The result is sometimes called a *choropleth* map. Pseudocoloring is used widely for astronomical radiation charts, medical imaging, and many other scientific applications. Geographers use a well-defined color sequence to display height above sea level—lowlands are always colored green, which evokes vegetation, and the scale continues upward, through brown, to white at the peaks of mountains.

The most common coding scheme used in data visualization is a color sequence that approximates the physical spectrum, like that shown in Figure 4.26(b). Although this sequence is frequently used in physics and other disciplines and has some useful properties, it is not a perceptual sequence. This can be demonstrated by the following test. Give someone a series of gray paint chips and ask them to place them in order. They will happily comply with either a dark-to-light ordering or a light-to-dark ordering. Give the same person paint chips with the colors red, green, yellow, and blue and ask them to place them in order, and the result will be varied. For most people, the request will not seem particularly meaningful. They may even use an alphabetical ordering. This demonstrates that the whole spectrum is not perceptually ordered, although short sections of it are. For example, sections from red to yellow, yellow to green, and green to blue all vary monotonically (they continuously increase or decrease) on both the red–green and yellow–blue channels. Figure 4.27 shows seven different color sequences, but which is best and why?

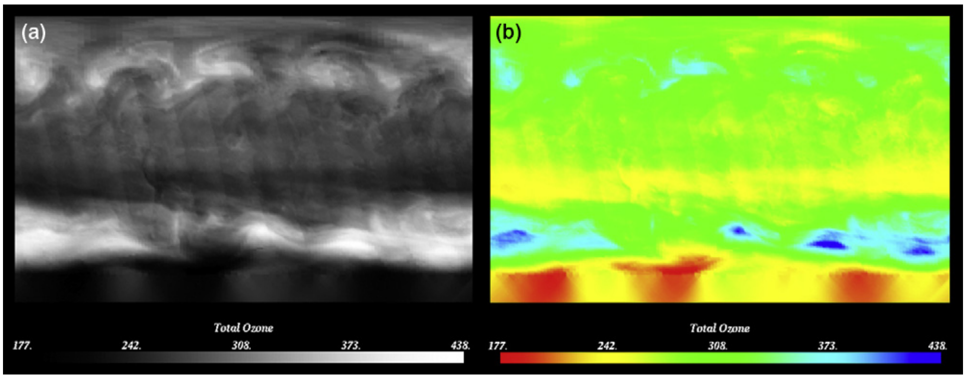

Figure 4.26 The same data showing ozone concentrations in the southern hemisphere is represented using (a) grayscale and (b) spectrum approximation pseudocolor sequences. (*Images courtesy of Penny Rheingans (Rheingans, 1999).*)

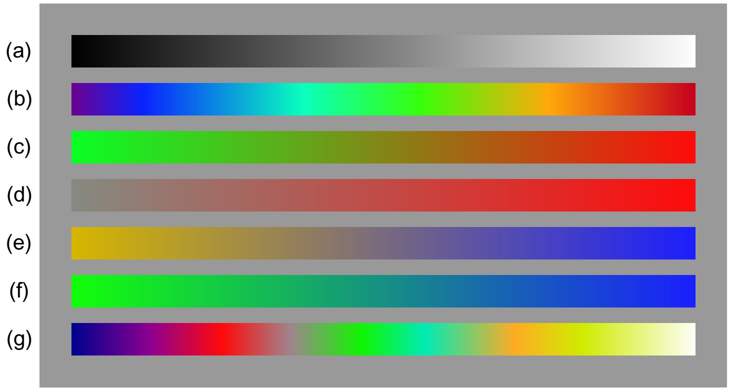

Figure 4.27 Seven different color sequences: (a) Grayscale. (b) Spectrum approximation. (c) Red–green. (d) Saturation. (e, f) Two sequences that will be perceived by people suffering from the most common forms of color blindness. (g) Sequence of colors in which each color is lighter than the previous one.

Form and Quantity

Sometimes we want to see the forms in a data set. Where are the highs and lows, the ridges and spirals, in a map of ozone in the atmosphere? Sometimes we want to be able to read the quantities. What is the temperature going to be in my part of the world tomorrow?

Color theory predicts that different color sequences have very different properties in this regard (Ware, 1988). Because the luminance channel helps us see forms, a grayscale sequence should allow us to see forms much better than pure color sequences (no luminance variation). See Figure 4.26(a). The highs are white, the lows are black, and complex swirling patterns can be seen in the ozone concentrations. Look at Figure 4.26(b). Here red, green, and blue areas clearly stand out, but this visual segmentation is meaningless; it is not clear which areas are high and low, and much less detail is seen overall.

Experimental studies have confirmed that grayscale maps are much better for form perception (Ware, 1988; Kindlmann et al., 2004). In spite of this, a recent survey of papers containing pseudocolored maps found that more than 50% used an approximation to the physical spectrum—a rainbow as a color sequence (Borland & Taylor, 2007). The same paper argued that this color sequence "*hinders this task* [of effectively conveying information] *by confusing, obscuring, and actively misleading.*"

Nevertheless, there are advantages to the spectrum approximation color sequence. The first is that it results in much lower errors in reading values from a key. Ware (1988) found 17% scale errors with a grayscale map and only 2.5% error with a spectrum approximation. There are two reasons for this. A spectrum color sequence can convey significantly more legible steps than a simple blue-to-red sequence (that is sometimes used for coding temperature).

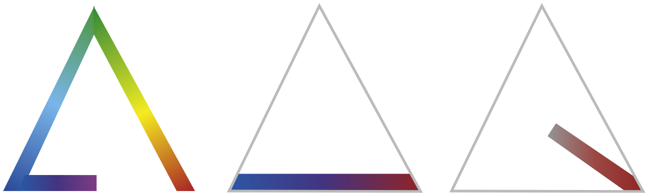

Figure 4.28 Sequences on a chromaticity diagram. (a) Spectrum approximation. (b) Blue–red sequence. (c) Saturation sequence.

Consider the sequences illustrated in Figure 4.28. These are based on the *RBG* triangle in the *CIEluv* uniform color space. A typical spectrum sequence actually starts with purple (not a true spectrum hue) and cycles clockwise through the spectrum colors to red. This path has more than two and a half times as many discriminable steps according to *CIEluv* than the red–blue sequence, and it has four times as many legible steps as a saturation sequence. Also, according to the *CIEluv* and *CIElab* color difference standards, there are about twice as many discernable steps in a traversal across the chromatic dimensions of color space as there are traversing the luminance dimension (Mahny, 1994). Another cause of errors in reading map values using a key is simultaneous contrast between parts of the display (Cleveland & McGill, 1983; Ware, 1988; Brewer, 1996b). These errors may be reduced in the spectrum sequence scale because the colors surrounding a particular point (and inducing contrast effects) are likely to partially cancel each other.

There are also sometimes semantic reasons for using a spectrum approximation sequence. Consider the case of the display of temperature in weather maps. The color blue is associated with cold. The color red is associated with heat. Having green and yellow in between provides a convenient method for conveying intermediate temperatures. In the case of a weather map, and many other visualizations, the ability to read quantity as well as see patterns in the data is essential. We may need to be able to read temperatures to better than 5 degrees over a 50-degree or greater range. This requires ten or more discriminable steps, something that is impossible to achieve with the blue-to-red sequence. Finally, the use of spectrum approximations for temperature maps is deeply embedded in large parts of our culture. Using this well-established standard can have huge efficiencies because it eliminates the need to learn something new.

[G4.19] Use a spectrum approximation pseudocolor sequence for applications where its use is deeply embedded in the culture of users. This kind of color sequence can also be used where the most important requirement is reading map values using a key. If this sequence is used, the spacing of the colors should be carefully chosen to provide discriminable steps.

Still, if it is important to show detail in the data, then it is essential to make that detail stand out using the luminance (black–white) channel because of the capacity of this channel to convey high-spatial-frequency information (Ware, 1988; Rogowitz & Treinish, 1996). Also, if form perception is the primary consideration, a sequence that trends upward or downward in luminance will be better.

Some authors have recommended that, for clarity, color sequences should constitute a straight line through a perceptual color space, such as *CIEluv* or *CIElab* (Robertson & O'Callaghan, 1988; Levkowitz & Herman, 1992). This would rule out the spectrum approximation sequence. Further, Spence et al. (1999) found that a color sequence combining variation in brightness, saturation, and hue was the most effective in a task requiring the rapid detection of low and high points in an image.

A better choice may be to design a sequence that cycles through a variety of colors, each one lighter than the previous. Sometimes this is called a *spiral color sequence*, because it can be thought of as spiraling upward in color space. Such a sequence can combine the advantages of monotonicity in luminance, so as to show form and detail, as well as reduce contrast-induced errors and enable accurate readings from a color key (Ware, 1988; Levkowitz & Herman, 1992; Kindelmann et al., 2004).

[G4.20] If it is important to see highs, lows, and other patterns at a glance, use a pseudocolor sequence that monotonically increases or decreases in luminance. If reading values from a key is also important, cycle through a variety of hues while trending upward or downward in luminance.

The designer of such a sequence can take advantage of the fact that monitor blue has much lower luminance than monitor red, which in turn has lower luminance than monitor green. Yellow, being the sum of red and green, has a very high luminance, almost equal to white. This is the basis for the sequence design shown on the right in Figure 4.29.

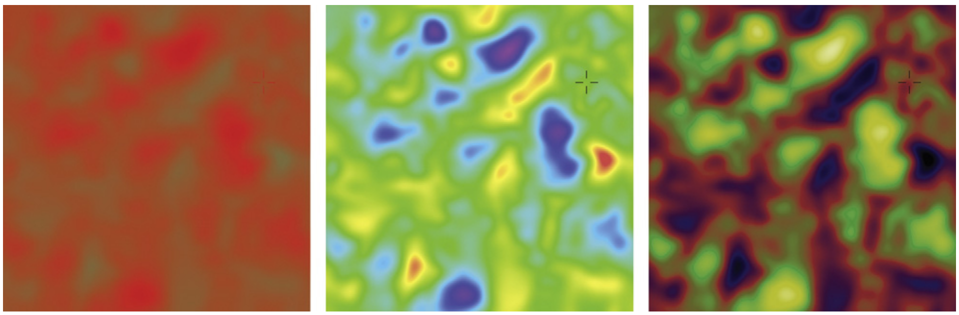

Figure 4.29 The same data represented with saturation, spectrum, and spiral color sequences. The spiral sequence makes it possible to easily see both the highs and lows, as well as read values accurately from a key.

Interval Pseudocolor Sequences

An interval sequence is one in which each unit step of the sequence represents an equal change in magnitude of the characteristic being displayed across the whole range of the sequence. In terms of color, this suggests using a uniform color space in which equal perceptual steps correspond to equal metric steps (Robertson & O'Callaghan, 1988). Using a contour map, not a color sequence, is the traditional way to display an interval sequence. Isovalue contour maps show the pattern of equal heights or other physical attributes with great precision, but using them to understand the overall shape of a terrain or an energy field takes considerable skill and experience. To support unskilled map readers, contours can be usefully combined with pseudocoloring, as shown in Figure 4.30(a). Even better may be a stepped pseudocolor sequence as shown in Figure 4.30(b).

Ratio Pseudocolors

A ratio sequence is an interval sequence that has a true zero and all that this implies: The sign of a value is significant; one value can be twice as large as another. Expressing this in a color sequence is a tall order. No known visualization technique is capable of accurately conveying ratios with any precision; however, a sequence can be designed that effectively expresses a zero point and numbers above and below zero. Brewer (1996a) called such sequences *diverging sequences*, whereas Spence and Efendov (2001) called them *bipolar sequences*. Such sequences typically use a neutral value on one or more

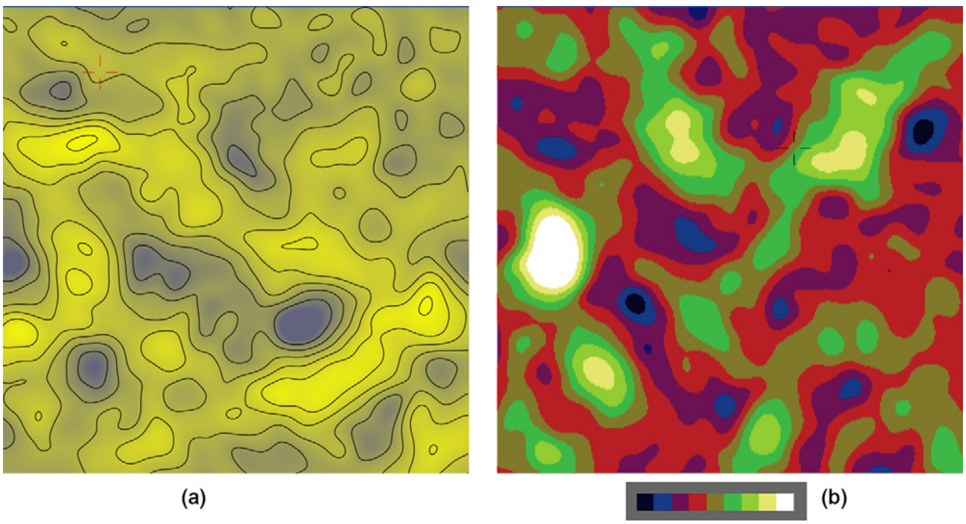

Figure 4.30 (a) Contours can show equal intervals in the data although numerical labels must be added for most applications. (b) A sequence of colors in discrete steps may be more reliably read using a key than a smoothly blended sequence.

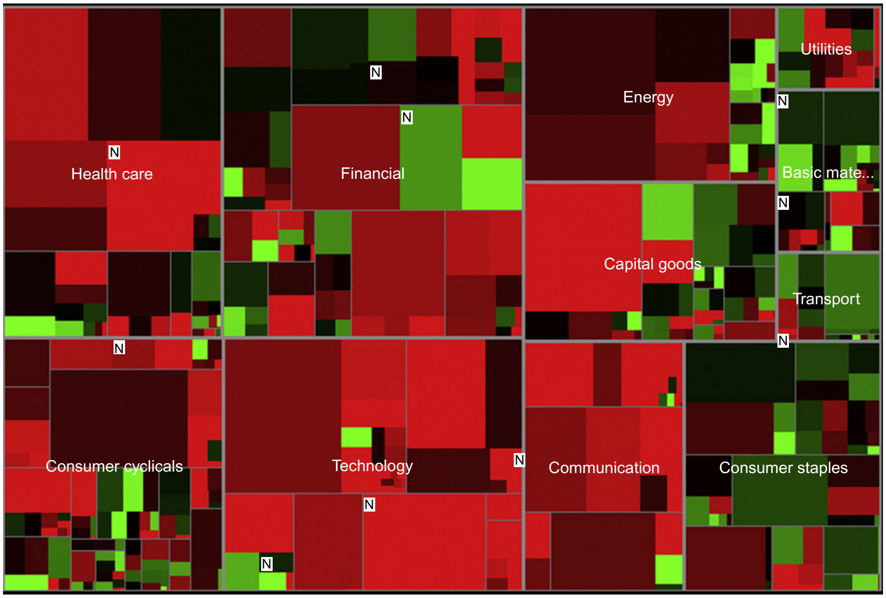

Figure 4.31 A color sequence with black representing zero. Increasing positive values are shown by increasing amounts of red. Increasing negative values are shown by increasing amounts of green. The map itself is a form of treemap (Johnson & Shneiderman, 1991). (*Courtesy of SmartMoney.com.*)

opponent channels to represent zero, and diverging colors (on one or more channels) to represent positive and negative quantities. For example, gray may be used to represent zero, increasing redness to represent positive quantities, and increasing blueness to represent negative quantities. In a target detection study, Spence and Efendov (2001) found that a red–green sequence was most effective, confirming the greater information-carrying capacity of this channel compared to the yellow–blue channel.

The example in Figure 4.31 shows a map of the stock market provided by SmartMoney.com. Market capitalization is represented by area, luminance encodes the magnitude of value change in the past year, and green–red encodes gains and losses. The website also gives users the option of a yellow–blue coding, suitable for most color-blind individuals.

Sequences for the Color Blind

Some color sequences will not be perceived by people who suffer from the common forms of color blindness: protanopia and deuteranopia. Both cause an inability to discriminate red from green. Sequences that vary mainly on a black-to-white scale or on a yellow-to-blue dimension (this includes green to blue and red to blue) will still be clear to color-blind people. Two sequences that will be acceptable to these individuals are shown in Figure 4.27(e, f). Meyer and Greenberg (1988) provided a detailed analysis of color sequences designed for common forms of color blindness.

Bivariate Color Sequences

Because color is three dimensional, it is possible to display two or even three dimensions using pseudocoloring (Trumbo, 1981). Indeed, this is commonly done in the case of satellite images, in which invisible parts of the spectrum are mapped to the red, green, and blue monitor primaries.

Although this mapping is simple to implement and corresponds to capabilities of the display device (which usually has red, green, and blue phosphors), such a scheme does not map the data values to perceptual channels. In general, it is better to map data dimensions to perceptual color dimensions. For example:

Variable one → hue

Variable two → saturation

or

Variable one → hue

Variable two → lightness

Figure 4.32 gives an example of a bivariate color sequence from Brewer (1996a) that maps one variable to yellow–blue variation and the other to a combination of light–dark variation and saturation. It suffers from the usual problem that the low-saturation colors are difficult to distinguish.

As a word of caution, it should be noted that bivariate color maps are notoriously difficult to read. Wainer and Francolini (1980) carried out an empirical evaluation of a color sequence designed for U.S. census data and found that it was essentially unintelligible. One approach to a solution is to apply a uniform color space, and Robertson and O'Callaghan (1986) discussed how to do this. But, distinctness may not lead to something that is interpretable. We do not seem to be able to read different color dimensions in a way that is highly separable.

Pseudocoloring is not the only way to display a two-dimensional scalar field. Generally, when the goal is to display two variables on the same map, it may be better to use visual texture, height difference, or another channel for one variable and color for the other, in this way mapping data dimensions to more perceptually separable dimensions. Mapping the scalar field to artificial height and shading the resulting surface with an artificial light source using standard computer graphics techniques is another alternative. These methods are discussed later in the book.

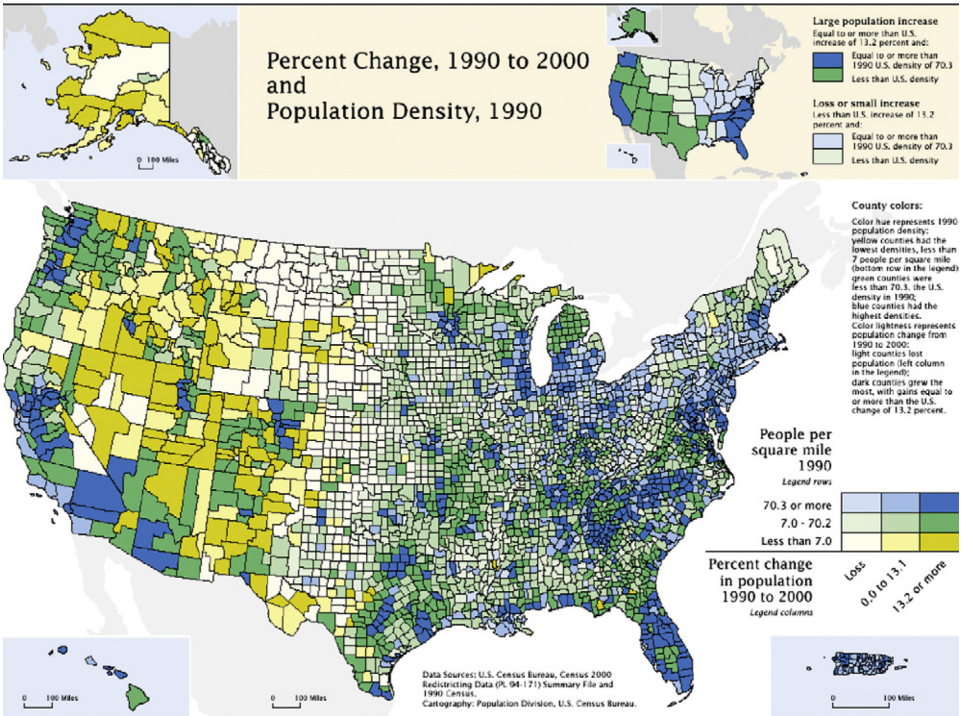

Figure 4.32 A bivariate coloring scheme using saturation and lightness for one variable and yellow–green–blue hue variation for the other. (*Courtesy of Cindy Brewer.*)

Many considerations go into making a color sequence that displays quantities without significant distortions, and this makes it unlikely that any predefined set of colors will exactly suit a particular data set and visualization goal. To show both overall form and detail and to provide the ability to read values from a key, it is often desirable to emphasize certain features in the data by deliberately using a nonuniform sequence; assigning more variation in color to a particular data range will lead to its visual emphasis and better discrimination of those values. Generally, the best way to achieve an effective color sequence is to place a good color editing tool in the hands of someone who understands both the data display requirement and the perceptual issues of color sequence construction (Guitard & Ware, 1990).

Application 4: Color Reproduction

The problem of color reproduction is essentially one of transferring color appearances from one display device, such as a computer monitor, to another device, such as a sheet of paper. The colors that can be reproduced on a sheet of paper depend on such factors as the color and intensity of the illumination. Northern daylight is much bluer than direct sunlight or tungsten light, which are both quite yellow, and is prized by artists for this reason. Halogen light is more balanced. Also, monitor colors can be reproduced only within the range of printing inks; therefore, it is neither possible nor meaningful to reproduce colors directly using a standard measurement system such as the CIE XYZ tristimulus values.

As we have discussed, the visual system is built to perceive relationships between colors rather than absolute values. For this reason, the solution to the color reproduction problem lies in preserving the color relationships as much as possible, not the absolute values. It is also important to preserve the white point in some way, because of the role of white as a reference in judging other colors.

Stone et al. (1988) described a process of gamut mapping designed to preserve color appearance in a transformation between one device and another. The set of all colors that can be produced by a device is called the *gamut* of that device. The gamut of a monitor is larger than that of a color printer (roughly the gamut of surface colors shown in Figure 4.7). Stone et al. described the following set of heuristic principles to create good mapping from one device to another:

1. The gray axis of the image should be preserved. What is perceived as white on a monitor should become whatever color is perceived as white on paper.
2. Maximum luminance contrast (black to white) is desirable.
3. Few colors should lie outside the destination gamut.
4. Hue and saturation shifts should be minimized.
5. An overall increase of color saturation is preferable to a decrease.

Figure 4.33 illustrates, in two dimensions, what is in fact a three-dimensional set of geometric transformations designed to accomplish the principles of gamut mapping. In this example, the process is a transformation from a monitor image to a paper hard copy, but the same principles and methods apply to transformations between other devices.

- **Calibration.** The first step is to calibrate the monitor and the printing device in a common reference system. Both can be characterized in terms of CIE tristimulus values. The calibration of the color printer must assume a particular illuminant.
- **Range scaling.** To equate the luminance range of the source and destination images, the monitor gamut is scaled about the black point until the white of the monitor has the same luminance as the white of the paper on the target printer.
- **Rotation.** What we perceive as neutral white on the monitor and on the printed paper can be very different, depending on the illumination. In general, in a printed image, the white is defined by the color of the paper. Monitor white is usually defined by the color that results when the red, green, and blue monitor primaries are set to their maximum values. To equate the monitor white with the paper white, the monitor gamut is rotated so as to make the white axes colinear.

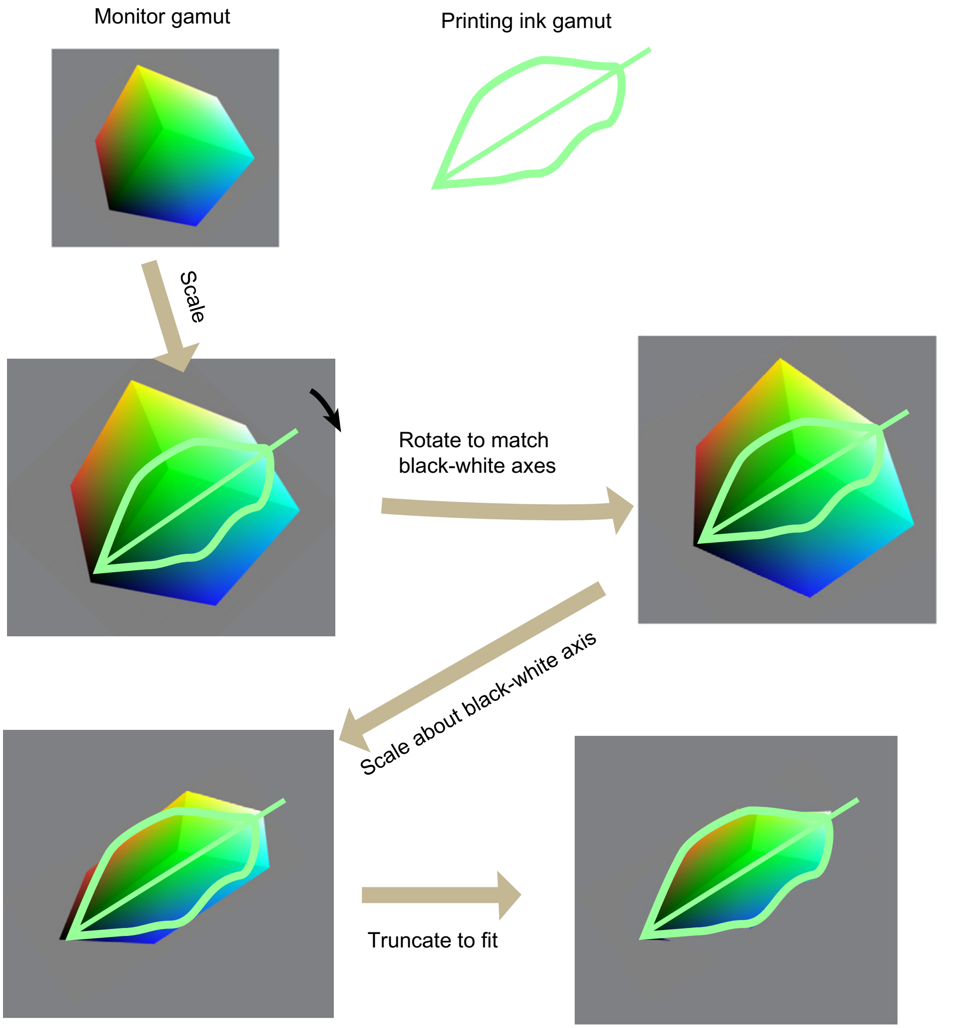

Figure 4.33 Illustration of the basic geometric operations in gamut mapping between two devices, as defined by Stone et al. (1988).

- **Saturation scaling.** Because colors can be achieved on a monitor that cannot be reproduced on paper, the monitor gamut is scaled radially with respect to the black–white axis to bring the monitor gamut within the range of the printing gamut. It may be preferable to leave a few colors outside the range of the target device and simply truncate them to the nearest color on the printing-ink gamut boundary.

For a number of reasons, it may not always be possible to apply these rules automatically. Different images may have different scaling requirements; some may consist of pastel colors that should not be made too vivid, whereas others may have vivid colors that must be truncated.

The approach adopted by Stone et al. (1988) is to design a set of tools that support these transformations, making it easy for an educated technician to produce a good result; however, this elaborate process is not feasible with off-the-shelf printers and routine color printing. In these cases, the printer drivers will contain heuristics designed to produce generally satisfactory results. They will contain assumptions about such things as the gamma value of the monitor displaying the original image and methods for dealing with oversaturated colors. Sometimes, the heuristics embedded in devices can lead to problems. In our laboratory, we usually find it necessary to start a visualization process with somewhat muted colors to avoid oversaturated colors on videotape or in paper reproduction.

Another issue that is important in color reproduction is the ability of the output device to display smooth color changes. Neural lateral inhibition within the visual system tends to amplify small artificial boundaries in smooth gradients of color as Mach bands. This sensitivity makes it difficult to display smoothly shaded images without artifacts. Because most output devices cannot reproduce the 16 million colors that can be created with a monitor, considerable effort has gone into techniques for generating a pattern of color dots to create the overall impression of a smooth color change. Making the dots look random is important to avoid aliasing artifacts (discussed in Chapter 2). Unless care is taken, artifacts of color reproduction can produce spurious patterns in scientific images.

Conclusion

There has been more research on the use of color in visualization than any other perceptual issue. Nevertheless, the important lessons are relatively few, and mostly they can be derived from opponent process theory. There are two chromatic channels (red–green and yellow–blue) and a luminance channel. Because of the low spatial resolution of the chromatic channels, small symbols should have high-saturation colors. Because of chromatic contrast in the opponent channels, we can only expect to have a few color symbols reliably identifiable. Contrast effects also make it desirable that larger regions should be less strongly colored in general.

It is impossible to keep a discussion of color entirely segregated in one chapter. Color affects every aspect of visualization and is mentioned in many other chapters, especially Chapter 5, which places color in the context of other methods for coding information.