COLOR AND COLOR SPACES

Douglas A. Kerr, P.E.

November 8, 2005
Issue 8

ABSTRACT

"Color space" refers to a specific system of coordinates that allows us to describe a particular color of light. In this article we discuss the concept of the color of light and the concept of a "color model"—which if precisely specified is today most often described as a "color space". We then describe in some detail a number of specific color spaces used in a wide range of fields.

TABLE OF CONTENTS

BACKGROUND

Color

Color is the principal property of visible light by which a human observer can distinguish different “kinds” of light. It is a subjective property, and in general the color of a light source cannot be determined by simple measurement of fundamental physical properties of the light.

It has been ascertained that to describe a particular color of light we must state three values. Color is thus a “3-dimensional property” in the mathematical sense. In the case of another three-dimensional property, the location of a point in space, many different sets of three variables (coordinates) may be used.[1] Similarly, in the case of color, many different systems of three variables may be utilized. A particular one is traditionally called a *color model*, and its variables are said to be the coordinates of its three-dimensional *color space*.

In this article, we describe a number of color models which are important both in theoretical work and in the representation of color in a wide range of

---

[1] For example: rectangular (Cartesian) coordinates, cylindrical coordinates, spherical coordinates, geodesic coordinates.

technical fields, including the representation of still and moving images in digital form.

"Color model" and "color space"

In recent times, it has become common to use the term "color space" not in the sense just described but rather to mean a particular fully specified color model. In the interest of consistency, we will use the term *color space* here in that new sense.

Luminance and related metrics

Luminance is a measure of the "potency" of light emitted from a surface—generally as it relates to the light emitted in a particular direction (such as toward an observer) from any particular small region of the surface.

There are other potency metrics that are closely parallel to luminance. If we are interested in the totality of light emitted by some emitter, the potency metric is *luminous flux*. If we are interested in the light emitted in a particular direction from a "point source" (such as a distant star), the potency metric is *luminous intensity.* If we are interested in the light falling on a surface, the potency metric is *illuminance*. These metrics have different dimensionalities and are quantified with different units.

Luminous flux is rather akin to electrical power, and the other metrics follow this similarity. Accordingly, if we combine the output of two sources which exhibit certain luminances, the total emission has a luminance that is the sum of those two luminances.

In the definitions of color spaces, luminance is not discussed in absolute terms of the quantitative scientific unit but rather on a relative scale, perhaps considered to range from 0 to 1.

COLOR SPACE FAMILIES

Most color spaces with which we will be concerned fall into one of three families. We will discuss the basic concepts of these families in a generic way. Later, when we encounter specific spaces, this background will help us to grasp their principles.

Luminance-chromaticity color spaces

<u>The luminance-hue-saturation space</u>

One kind of color space that is well related to the intuitive human perception of color uses these three properties (color coordinates):

Luminance is the property that describes the "brightness" of the light[2]. (Many people are startled to learn that luminance is a part of color, but it is in the formal sense we are considering here.)

Hue is the property that distinguishes red from orange from blue from blue-green, and so forth.

Saturation is the property that distinguishes red from pink. It is sometimes said to describe the "purity" of the color.

The properties *hue* and *saturation* are said to jointly describe the *chromaticity* of the color of interest. Chromaticity is in fact the property that the average person thinks of as the "color" of light (not realizing that luminance is one aspect). Since chromaticity actually embraces two of our "color coordinates", it is a two-dimensional property (in the mathematical sense.

Because of this situation the luminance-hue-saturation space is said to be in the *luminance-chromaticity* family. In other members of that family, chrominance is characterized by pairs of coordinates other than hue and saturation. Sometimes luminance is described as indicating the *quality* of the light and luminance the *quantity*.

There are various scales by which hue and saturation may be quantified (given numerical values). An important scheme depends on a graphic presentation of chromaticity called the CIE chromaticity diagram. It is actually part of a different color space within the luminance-chromaticity family.

The CIE xyY color space and the CIE chromaticity diagram

Another important luminance-chrominance color space which is very useful in technical work is known as the *CIE xyY color space* (CIE are the initials of the French name of the International Commission on Illumination). It plots any chromaticity as a point on a graph (the *CIE chromaticity diagram*[3]*)* whose two axes correspond to two arbitrary variables called x and y.[4] These are defined so that the "mapping" of different chromaticities to points on this graph results in certain desirable properties. This description of chromaticity

---

[2] Rigorously, as we mentioned just previously, *luminance* only relates to the strength of light emerging from an area of finite size, such as a spot on a scene being photographed. Other metrics apply to other situations. We need not be concerned here with the distinction. So far as the description of color is concerned, the concept is the same in each case.

[3] Actually, the "*1931 CIE chromaticity diagram*"; two later chromaticity diagrams have been adopted by the CIE, one in 1960 (with variables **u** and **v**) and one in 1976 (with variables **u'** and **v'**, generally called today just **u** and **v**). The 1931 version is still the one most widely used for general discussions of chromaticity, and when we speak in this paper of the "*CIE chromaticity diagram*" without specifying the version, it is the 1931 version which is meant.

[4] See Appendix B for a description of the concepts of this coordinate system.

is accompanied by a description of luminance in the variable Y, giving of course the expected three “coordinates”.

This space, and the CIE chromaticity diagram itself, are discussed in detail in Appendixes A and B.

Tristimulus spaces

Tristimulus spaces describe the color on interest in terms of the amounts of light of three specified *primary chromaticities* that may be combined to produce that color. One important example of this class, the “RGB” space, is based on the use of the primaries which are the red, green, and blue emissions from a color cathode ray tube display system. Thus, color descriptions in this space can be readily used to drive such a display. The RGB space also relates well to the capture of color images by the typical television camera or digital still camera.

Other tristimulus spaces are used in the description of color in other contexts, such as theoretical scientific work. They involve the use of other primaries, some of them not having any physical realizations. One example is the *CIE tristimulus space*.[5] It describes a color in terms of the amounts of three such non-real primaries, called X, Y, and Z. This space is discussed in Appendix B (and is in fact the underlying basis for the CIE xyY color space).

Luminance-chrominance spaces

There is another family of color spaces, of interest to the representation of color in television signals and digital images, that describe color in terms of *luminance* (as before) and a two-dimensional property called *chrominance*. (One must be careful not to confuse *chrominance* and *chromaticity*—we’ll emphasize the distinction in a little bit.)

The concept is this. Suppose we first generate white light[6] whose luminance is the luminance of the color of interest. Now imagine that we add to that white light the amount and “flavor” of non-white light required to produce the color of interest. This “colorant dose” (as we would say when mixing paint) is described by the property *chrominance*. As we would expect,

---

[5] Also discussed in Appendix B. In fact, the term ”tristimulus space”, while applicable to a range of color spaces, is usually reserved for use in connection with this space.

[6] “White” does not automatically describe a specific chromaticity, but only an arbitrary conceptual range. Various chromaticities have been defined to be considered as white in different situations. Most color spaces involve a certain definition of white. This matter is discussed further in Appendix A.

chrominance is a two-dimensional property: its description requires the values of two coordinates.[7]

The reader may note an apparent paradox in this concept. If the white light component alone has the luminance which is the luminance of the color of interest, and we add further light (the "colorant" described by the chrominance property), won't the resulting light (the color of interest) have a greater luminance, a luminance greater than we have already stated for that very color?

The secret is that the "colorant" has zero luminance! Then how can it have any effect on the composite light? How can it even exist?

It doesn't exist physically; it is a mathematical fiction. That's all right, since we do not actually generate the described color by physically adding together white light and "colorant" light. The addition is done mathematically, with the white light and the colorant both typically described by their R, G, and B coordinates (under the RGB tristimulus space). The resulting RGB description is then typically fed to an RGB-based display system to render the color of interest.

Since the colorant component must have zero luminance, its own RGB description will involve negative values of the amount at least one of the RGB primaries (again, not physically possible, but fine for a mathematical fiction). However, for any color that can be represented by the RGB space, when this description is added to the RGB description of the white light component, R, G, and B will all have positive values.

We will discuss specific luminance-chrominance spaces later in the article, under "Color Spaces for Television Images" and "Modern CIE Color Spaces". (They are complicated, and I don't want to lose momentum here!)

Chromaticity vs. chrominance

The distinction between *chromaticity* and *chrominance* often eludes the reader at first. Here we will point out the fundamental distinction.

Imagine that we have described, under a luminance-chromaticity space, the color of the light emerging from a certain spot on a test object illuminated by two identical floodlights. If we turn one of the floodlights out, the *luminance* is reduced, but there is no change in *chromaticity*.

---

[7] Different "luminance-chrominance" color spaces use different pairs of coordinates for the purpose. We will see the details of several of these when we later discuss specific color spaces.

Now let us again consider our example of the color of a spot on an object illuminated by two floodlights. This time we describe the color in a luminance-chrominance space. When we turn off one floodlight, the *luminance* of the light is reduced. The magnitude ("potency") of the *chrominance* also drops correspondingly.

The following analogy may help to understand that latter situation. Remember, we can think of *chrominance* as describing a "colorant dose". Imagine that we first mix a gallon of paint of a certain custom color. The "recipe" defines a certain colorant dose to be added to a certain amount of "base" paint.

If we instead need to mix only a quart of paint, to produce the same color we must cut down the size of the colorant dose proportionately. Thus, to mix up a "batch of light" of a reduced luminance, we must decrease both the amount of the "base white" (the luminance) and the amount of "colorant" (the chrominance).

GAMMA

Before we can discuss specific color spaces used for the encoding of color for computer graphics and television images, we must discuss the concept of *gamma*.

In conventional photography, the density created on a negative (itself a logarithmic measure)[8] would ideally vary directly with the logarithm of exposure, the slope of the plot of density vs. the logarithm of exposure (the "D log E" curve) being 1.0. In reality the slope is usually less. The value of the slope is often designated by the lower-case Greek letter *gamma* (γ).

In a cathode-ray tube (CRT) visual display, the luminance of the spot on the screen generated by the electron beam from a beam gun is typically not proportional to the control voltage to the gun, but rather to some power of the voltage (often about 2.2).[9] This exponent is often designated *gamma* by parallel with the related concept for film.

In earlier eras, in one-to-many systems such as television broadcast we typically went to great extent to move as much complexity as possible from the "many" units (the TV receivers) to the "one" units (the TV studio-transmitter complex). In that vein, in analog TV transmission, to eliminate the

---

[8] *Density* for a point in a photographic negative is numerically defined as the common logarithm of the ratio of the intensity of the light falling on the point to the intensity of the light passing through the point. Thus, a point having a density of 2.0 allows only 1/100 of the light intensity to pass through.

[9] This should not really be a surprise; luminance has a dimensionality akin to electrical power, and power varies as the square (second power) of voltage or current. Thus from this alone we might well expect an exponent of 2.

need to put non-linear circuitry in the TV receiver to overcome the voltage-luminance nonlinearity of the display gun, the signal transmitted is precompensated for that nonlinearity ("gamma precompensated"). Thus typically the transmitted signal voltage goes as about the 0.45 power (1/2.2) of the luminance observed by the camera.

Actually, the human eye does not respond linearly to luminance. The response, as with many other areas of human perception (such as sound loudness), is more nearly logarithmic. The gamma-precompensated luminance signal involved in TV transmission is in fact a crude but still useful approximation to a logarithmic representation: the human eye responds nearly linearly to gamma-precompensated luminance. This affords many advantages in image manipulation.

There is a third motivation. The nonlinear nature of the transmitted signal is desirable in terms of the subjective impact to the viewer of noise introduced in transmission. The rationale for this is beyond the scope of this article.

In any case, as well will see shortly, in most color spaces of the RGB family (to be discussed next), the variables indicating the amounts of the three primaries needed to "make up" the color of interest are expressed as a nonlinear transform of the original values. The nonlinear function used is often a power function similar to (in fact, often identical to) that used for gamma precompensation in television systems (often with the same exponent, or "gamma"), and in fact its initial purpose was identical. Thus, even in RGB-family color spaces other than those use in television, the nonlinear transformation is often referred to as "gamma precompensation".

In connection with color spaces using a substantially different nonlinear function (as we will see in connection with the CIE L*a*b* color space), the term "gamma precompensation" is perhaps less justifiable, and is less frequently used.

COLOR SPACES FOR COMPUTER GRAPHICS

The onset of color graphic display capability for computers required the development of schemes for coding the colors of image elements. Here we will describe some of the most-widely used spaces encountered in defining the colors to be used for display elements, and also for defining the colors in images being composed or edited in image composition or editing software.

The RGB space

The RGB color space describes a color in terms of the potency of three light ingredients of different specified chromaticities which if added together will produce that color. These three "primaries" are described as red, green, and blue, and the variables R, G, and B describe the amounts of each in the "mix". This space is an example of the *tristimulus* color space family.

Figure 1 shows the chromaticity of the three primaries in a typical RGB color space on the CIE chromaticity diagram. (See Appendix A for a discussion of this diagram.)

In the actual representation of the color in terms of the variables R, G, and B, the values of these three variables are precompensated for the assumed gamma of the display mechanism. The scaling of r, g, and b is such that if R = G = B, the color represented has the chromaticity of the “reference white” for the color space.

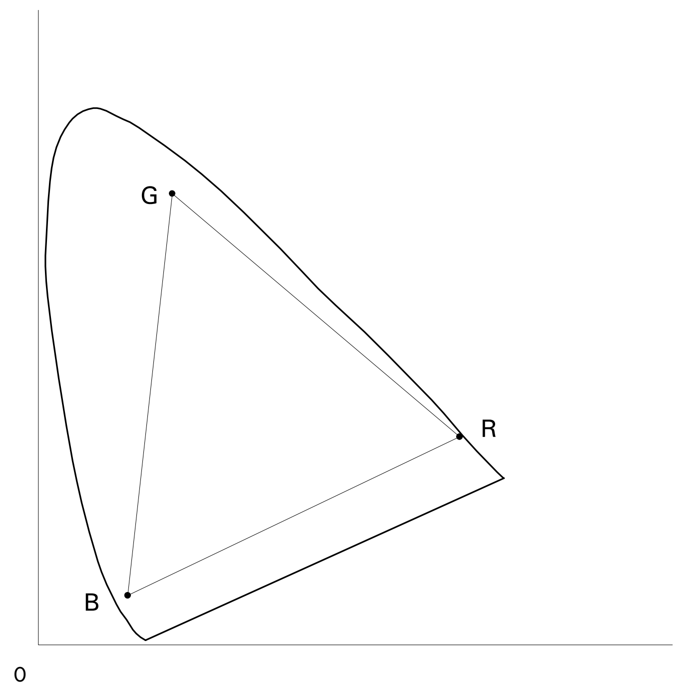


Figure 1. RGB primaries on the CIE chromaticity diagram

In formal mathematical work, the symbols R, G, and B are used for the “linear” (non-gamma-precompensated) form of the variables, and R′, G′, and B′ for the gamma-precompensated form. However, in connection with the RGB color space in practical use, the symbols R, G, and B always refer to the gamma-precompensated form of the variables. Thus, to avoid confusion, in mathematical work in this paper, we will use the symbols **r**, **g**, and **b** for the non-gamma-precompensated variables, and R, G, and B for the gamma-precompensated ones.[10]

Using a form of that notation, the nonlinear transformation used in many (but not all) the standardized RGB color spaces takes on this form:

[10] Note also that, while in equations we will follow the usual mathematical convention of showing the symbols for variables in italics, we do not do so in the text proper, since in many cases those variables identify the “coordinates” of a color where, by convention, they are not shown in italics.

$$C = c^{1/\gamma}$$

where C represents any of R, G, or B, **c** represents any of **r**, **g**, or **b**, and γ (lower case Greek gamma) is the exponent defining the nonlinear transform.

This space followed directly from the color display mechanism of most computers, which used a tricolor cathode-ray tube with guns controlling the emission of light of three primary chromaticities, “red”, “green”, and “blue” (just as was used for classical color television).

In the use of the RGB space in computer memory or in file, it is common to use an 8-bit number to represent each of these three intensities. Thus, at the human interface (where the operator or artist might choose a color, or where the color of a point on the image would be displayed to the artist), it is common to represent the range of values for each component as being 0-255. In other cases, a range of 0-100% is used at the human interface.

The *luminance* of the color rendered does not follow an absolute scale, but rather depends on the display mechanism and the setting of its “brightness” control.

We often plot the “chromaticity gamut”—the range of chromaticities that the color space can represent—of an RGB color space on the CIE chromaticity diagram. It turns out that it comprises all chromaticities that are enclosed by the triangle joining the points giving the chromaticity of the three primaries, R, G, and B. (See figure 1.)

But this can be misleading. Not all those chromaticities can be achieved for every luminance that can be represented by the RGB model.

For example, we can only have the a color with the chromaticity of the G primary itself by using “G” light alone; that is, the values R and B must be zero. With a “100%” dose of G (that is, for R,G,B=0,1,0), we will have a certain luminance—less than that we would have if R, G, and B were all 1, the maximum luminance of the color gamut of the RGB space.

If we wish to have a greater luminance than we get for R,G,B=0,1,0, we can only do it by adding “R” or “B” light. As soon as we do that, the point representing the resulting chromaticity on the chart moves toward the R and/or B primary point—the saturation declines. To reach maximum luminance, we have no choice but to use R,G,B=1,1,1, in which case by definition the saturation is zero.

The HSV (HSB) space

Although the RGB space relates well to the actual color display mechanism, it is not intuitive for the computer graphic artist wishing to indicate a desired color.[11] Thus, at the human interface, another space came into play, with its coordinates being *hue*, *saturation*, and *value* (a synonym for luminance): H, S, and V. It is sometimes called the HSB space (for *hue*, *saturation*, and *brightness*).

How is hue described in this system? In "theoretical" work, hue (for spectral hues) is often described in terms of the wavelength of the spectral (monochromatic) color having that hue. For the non-spectral purple colors, the hue is usually described as a fractional distance along the locus of non-spectral purples on the CIE chromaticity diagram.

Of course this means of describing hue would not be practical for a working artist (or even for a computer user setting the color of his Windows desktop!). Instead, in the HSV color space, hue is described in terms of a "color wheel", reminiscent of those we used in elementary school art classes.

The hues of the three primaries of the assumed display system are arbitrarily placed at equidistant azimuths around this circle, at 0° (red), 120° (green), and 240° (blue). Well-known names of three intermediate hues, *yellow*, *cyan*, and *magenta*, are placed at intermediate azimuths. The "non-spectral purples" fall in the azimuth range between magenta and red.

In effect, the circle is a transformation of the periphery of the CIE chromaticity diagram, embracing both the locus of spectral hues and the locus of non-spectral purple hues.

Then, in the HSV representation of a color, the hue parameter (H) would be given as the azimuth (in degrees) of the corresponding hue on the wheel, with 0° representing the hue of the red primary.

In HSV systems, brightness (V) is often defined as just $(R+G+B)/3$.[12] This primitive definition ignores the differing sensitivity of the eye to the different primaries, and the differing maximum available intensities of the three CRT primaries. In other cases, V is defined as the maximum value among R, G, and B.

Here we have the same dilemma as with the RGB space: we cannot achieve as much saturation for higher brightness as we can for lower brightness.

---

[11] We should perhaps note here that none of the spaces we will discuss here are really satisfying for the sophisticated graphic artist, who will likely wish to choose colors under systems developed years ago for use with pigments, such as the Munsell or Pantone systems. Most serious graphic art software packages afford the opportunity to choose colors based on one or more of these systems. The colors chosen are then usually converted by the software to an RGB representation for storage.

[12] Again, brightness is not defined on an absolute scale, but rather on a relative scale, the absolute value depending on the display mechanism and the setting of its brightness control.

Thus, a color such as H = 60°, S = 100%, and B = 95% cannot be achieved. The HSB system, however, vainly allows such a color to be described.

The details of the HSB space definition—in fact, of any of the several different definitions encountered in practice—are beyond the scope of this article.

The HSL space

The HSL space (sometimes designated HLS) is an attempt to get around (or perhaps hide) the brightness-saturation conflict we just mentioned. In that space, *hue* (H) is defined in terms of the color circle just as in the HSB system. *Lightness* (L) is conceptually the same as brightness in the other space. It also runs from 0-100%, but is usually defined in an even more peculiar way with respect to its value for any RGB combination: the average of the highest and lowest value among R, G, and B.

The third parameter is again called *saturation*, and again has the symbol S, but is also defined in a rather peculiar way. Its value runs from 0-100%. It indicates not the actual saturation of the color of interest but rather the fraction that saturation is of the maximum saturation available **at the particular lightness** called for (considering the brightness-saturation conflict mentioned above).

For example, to achieve a color of hue 0° (red) with a lightness of 80% requires that we use not only a large value of R but substantial values of G and B as well. Because of the presence of G and B, the resulting color cannot achieve 100% saturation (to do so would require that R appear alone).

In particular, to get the highest available saturation for red for an L of 80%, we would need to use R = 100%, G = 60%, and B = 60%. The actual saturation we would get would be about 25%. Nevertheless, under the HSB system, that saturation would be called 100%, since it is the greatest attainable at that lightness for the stated hue.

The details of the HSL space definition—in fact, of any of its several different definitions encountered in practice—are beyond the scope of this article.

COLOR SPACES FOR TELEVISION IMAGES

Introduction

Several color spaces are in use for the encoding of moving images for analog television transmission purposes and for further encoding using such digital transmission schemes as MPEG.

In each case, we assume that the original image representation, from the camera or equivalent, is in RGB form, with the values R, G, and B precompensated for the assumed gamma of the ultimate display device.

Although the scope of this article is not intended to embrace modulation schemes and similar matters related to the application of color spaces to television transmission, we must venture a little into those electrical engineering topics in order to grasp the rationale behind the color spaces in this area.

The YUV color space

The color space identified as YUV is the basis for encoding a television signal in analog form for broadcast in the European PAL TV system and in the current version of the American NTSC TV system. [13]

The space is based on a description of a color in terms of R, G, and B, the amounts of three primaries whose chromaticities are intended to match those of the three primaries of a classical color CRT display. These are precompensated for an assumed display device gamma of 2.2 in this fashion:

$$R = \boldsymbol{r}^{0.45} \qquad (0.45 = 1/2.2)$$

$$G = \boldsymbol{g}^{0.45}$$

$$B = \boldsymbol{b}^{0.45}$$

The value *Y* is then determined as a weighted sum of *R*, *G*, and *B*:

$$Y = 0.299R + 0.587G + 0.114B$$

The object is to approach the situation in which we would have the same value of Y for colors of different chromaticities which nevertheless appear to the human viewer to be "equally bright".[14] However, this is only approximately achieved, since Y is derived from a linear combination of gamma-precompensated RGB values, whereas true luminance is best reckoned as a linear combination of the non-gamma-precompensated (linear) **rgb** values. Y is thus not a true indication of luminance (not even a gamma-precompensated one). As we will see later, this value is often given the name *luma*.

U and *V* represent blue and red "color differences", as follows:

$$U = 0.492(B\text{-}Y)$$

[13] The YUV color space should not be confused with the CIE L*uv color space, a different creature altogether.

[14] A "black-and-white" TV receiver operates only from the signal carrying *Y*.

$V = 0.877(R\text{-}Y)$

The combination of *U* and *V* are said to define the “chrominance” of the color[15].

In PAL and modern NTSC television transmission, the portion of the overall signal which conveys *U* and *V* is called the *chrominance* (or often *chroma*) signal, sometimes designated C. The portion which conveys *Y* is called the *luminance* (or *luma*) signal, designated Y. In addition to being shorter, the terms *chroma* and *luma* by convention remind us that these are gamma-precompensated values (and in the case of *luma*, reminds us that it really isn’t a luminance value at all—not even a gamma-precompensated one). These two short terms are borrowed for use in connection with other color spaces of a similar nature.

Note that this color space should not be confused with the CIE uv chromaticity diagram (part of the uvY color space) color space nor the CIE L*uv color space. In fact, the chrominance axes of the YUV space do not even approximately match the chromaticity axes of the uv chromaticity diagram nor the chrominance axes of the L*uv space. If anything, they are almost interchanged (u vs. v).

The YIQ color space

The color space identified as “YIQ” was until recently utilized in the encoding of television images for analog broadcast in the North American system (NTSC). It can today perhaps best be understood as a variant of the YUV space (although, interestingly enough, the YUV space had not been defined when the YIQ space came into use), and has essentially been replaced by the YUV space in modern analog television transmission.

Consider the chrominance plane defined by *U* and *V* (that is, by Y′-B′ and R′-B′ with the appropriate scaling). We define a new set of coordinate axes, *Q* and *I*[16], with the same origin but lying at an angle of 33° counterclockwise from the *U* and *V* axes, respectively. The *Q* and *I* values then describe the chrominance of the color. In television transmission, the portion of the overall signal which conveys *Q* and *I* is called the *chrominance* (or often *chroma*) signal, sometimes designated C. The portion which conveys *Y* is called the *luminance* (or *luma*) signal, designated Y. As in the case of the YUV space, the terms *chroma* and *luma* by convention remind us that these

---

[15] Recall that c*hrominance* differs from *chromaticity* in that if a color of a given chromaticity is increased in its luminance, the “magnitude” of its chrominance also increases.

[16] The designations Q and I are mnemonic for *quadrature* and *in-phase*, an allusion to the way in which these are transmitted in television transmission by quadrature amplitude modulation of a “chrominance subcarrier”. They are generally mentioned in the opposite order: “I and Q”; the order “Q and I” we used here is intended to reflect the parallelism with U and V, respectively.

are gamma-precompensated values (and that Y is not really a luminance value at all).

Why the new set of axes? The eye's *chrominance acuity*—its ability to discern fine detail carried by chrominance change—is highest for chrominance changes along a certain direction of the chrominance plane[17], and substantially lower for the direction at right angles to that. The Q axis is aligned with the "lowest acuity" direction. This allows, in television transmission, allocation of substantially less bandwidth to the Q component (transmitting it with reduced "resolution"), reducing the overall bandwidth required for the transmission of the image chrominance.

Curiously enough, the allocation of different signal bandwidth to the Q and I components of chrominance, an objective of the original design of the NTSC system and the YIQ space, is not exploited in most modern TV encoding schemes, such as YUV.

The actual development of Y is given by:

$$Y = 0.299R + 0.587G + 0.114B$$

The development of I and Q are given by:

$$I = 0.736(R\text{-}Y)\text{-}0.268(B\text{-}Y)$$

$$Q = 0.478(R\text{-}Y) + 0.413(B\text{-}Y)$$

(The more complex expressions for I and Q, compared to those for U and V in the YUV space, are a result of the rotated axes of the YIQ space.)

The YPbPr space

The scales for U and V (in the YUV space) and of Q and I (in the YIQ space) were chosen to produce, in television transmission, an appropriate amplitude ("voltage") range of the entire composite signal (Y + C) over the full gamut of colors. This is a requirement for proper performance of the overall modulation scheme used to convey the composite signal as a radio-frequency signal in TV broadcast. The range is not the same for *U* and *V* (nor for *Q* and *I*.)

If we wish to convey a video signal across an analog interface as three separate "baseband" (unmodulated) electrical signals, one for luminance and two for chrominance, it is attractive for the three signals to have the same voltage ranges. The use of voltages based on *Y*, *U*, and *V* would not meet that criterion.

The YPbPr space is conceptually identical to the YUV space, with Pb derived from the blue color difference value, B-Y (like U), and Pr derived from the red

[17] As seen on the CIE chromaticity diagram.

color difference value, R-Y (like V). The coefficients of Pb and Pr, however, are chosen so that both have the same range over the full gamut of colors that can be represented by the space.

The ranges of the variables, in abstract terms, are: for Y, 0-1 “unit” (by definition); for Pb and Pr, ± 0.5 “unit”. The actual electrical signal at an interface is ordinarily scaled such that the range for Y is 0-0.7 V and for Pb and Pb, ± 0.35 V (that is, one “unit” is 0.7 V).

Again, as in the case of the YUV and YIQ spaces, note that Y isn’t rigorously an indicator of luminance (not even gamma-precompensated luminance).

Unfortunately, “three-channel” electrical interfaces of this type are often (but incorrectly) labeled “YUV”.

The YCbCr space

The YCbCr space is encountered in one form in connection with the encoding of television images for digital representation or transmission. Another form is used for the digital encoding of still images in the JPEG encoding system. It is equivalent to the YPbPr space, except that its three variables are defined as being in 8-bit digital form.

In both forms, the value Y is derived from a weighted sum of R, G, and B (where these values are gamma-precompensated). The standard weighting is:

$$Y = 0.299R + 0.587G + 0.114B$$

where the range of R, G, and B is assumed to be 0-1.

The range of Y will then also be 0-1.

Note that Y does not represent the luminance of the color; luminance is reckoned as a weighted sum of the non-gamma-precompensated R, G, and B values. Y is not even gamma-precompensated luminance. Y here is sometimes spoken of as “pseudo-luminance”. It is also often called *luma*, a term drawn from television signal practice.

Two “color difference” values are then derived:

$$Cb = 0.564(B - Y)$$

$$Cr = 0.713(R - Y)$$

The coefficients in those expressions ensure that (if R, G, and B are within the range of 0-1) Cb and Cr will lie in the range -0.5 to $+0.5$.

(We can see that the YCbCr space is essentially a rescaled form of the YUV space.)

Cb and Cr collectively are said to express the *chrominance* of the color; this is often called *chroma* (again, a term drawn from television practice).

In the digital television version, Y is expressed in 8-bit form, with a range from 16 to 235 (decimal). Cb and Cr are expressed in 8-bit form, both with a range of 16-240, thought of as being ± 112 about a center value of 128.

The restricted range of Y accommodates two aspects of television production and transmission practice. For one, at both the "light" and "dark" extremes, there is additional numeric range[18] available to accommodate outputs from the camera accidentally lying outside the nominal full range. Thus we avoid "clipping" in such circumstances. The additional range available at the "dark" end also accommodates the concept of a "blacker-than-black" representation used for "blanking" the inactive parts of the picture. This assures reliable rendering of these as black at the receiver. It also accommodates the concepts of a "really-blacker-than-black" representation through which synchronizing pulses (to synchronize the horizontal and vertical scanning of the image) are conveyed in transmission.

However, in the forms of this color space used in such applications as the JPEG representation of photographic images, Y, Cb, and Cr are all scaled so that they occupy the range 0-255[19]. For Cb and Cr the center point in this case (representing a "zero" value) is again 128.

Note that since this is a luminance-chrominance (not luminance-chromaticity) space, if we begin with some color and "attenuate" it (such that its chromaticity does not change). Y, Cb, and Cr all decrease.

Often (through editorial carelessness) this space is referred to as "YCrCb" (presumably through the assumption that Cr and Cb would be in the same order as "R" and "B" in "RGB").

We will encounter the YCbCr color space again (in its "sYCC" form) in the section on Color Spaces for Photographic Images.

COLOR SPACES RELATED TO COLOR PRINTING

The color spaces we have discussed so far are directly related to an assumed technique for displaying the image based on the emission of light of three primary chromaticities. There is another family of spaces used in a computer context and related to another technique of image production: color printing with pigmented inks.

---

[18] Called "headroom" and "footroom".

[19] At one time the coefficients of the conversion to 8-bit form were such that for the maximum should be 256, but of course that is not possible with an 8-bit representation. This little "oops" has been gently squeezed out of the current version of the controlling specification, and the coefficients are now scaled to suit a range of 255 units.

This is a very complex field, involving sophisticated science, art, and craft. We will here take a very simplified view of the area.

Three-color printing

In the basic technique of "three-color" printing, three different inks are used to print an image. They are said to be of three different "colors". Rigorously, however, an ink does not have a color. Rather, it has a *reflectance spectrum*, a curve of the fraction of the light falling on the ink that is reflected as a function of the wavelength of the light. If we take the spectrum of the incident light and multiply it by the reflectance spectrum of the ink, we will get the spectrum of the reflected light. That light, to a human observer, will exhibit a particular color—it will have a certain brightness (luminance) and a certain chromaticity.

To the user, that color is thought of as "the color of the ink". However, if we change the spectrum of the incident light, the spectrum of the reflected light will also change, and thus its apparent color. Thus, to have a sample of a certain ink exhibit a consistent color, it must be viewed under light of a consistent spectrum.

Note that that this is not as simple as just calling for a consistent **color** of incident light. We can have incident light of two different spectra which nevertheless exhibit the same color to an observer. But the light reflected by a certain ink illuminated by those two kinds of light may not exhibit the same color to an observer.[20]

All that having been said, from here on we will nevertheless, for conciseness, refer to the "color" of ink and to the "color" achieved by the use of the ink on paper.

Traditionally, the "three-color" printing process has used three kinds of ink whose reflectance spectra have been fairly well standardized. We describe these qualitatively, in terms of the hue which they exhibit when illuminated by light of some fairly-standard spectrum (such as "sunlight"), as *cyan*, *magenta*, and *yellow*. Formerly (and to some extent yet today), they were described as "process blue", "process red", and "process yellow", the word "process" of course being an allusion to the "three-color printing process".

In "emissive" color-generating techniques (such as that of the CRT-based systems used in computer and television displays), the spectra of the three

---

[20] This is of course a large problem in the design of not only printing inks but also of paints for products. The manufacturer would like for a refrigerator in "sea blue" to look the same whether it is illuminated by incandescent or fluorescent light. The problem is complicated by the fact that the human perception of chrominance is actually relative, and the person viewing the refrigerator is also seeing surrounding objects whose apparent chrominances also depend on the type of illumination.

primaries (weighted by the relative brightness of each) add to produce the spectrum of the emitted light, which determines its visible color.

In an ink printing process, each ink acts as a filter, reflecting the various wavelengths of the incident light in accordance with the ink's reflectance spectrum. Conversely, we can think of the ink as absorbing various wavelengths of the incident light in inverse accordance with the ink's reflectance spectrum.

When two or three kinds of ink are applied, it is as though we have filters in cascade: each ink absorbs the various wavelengths of the incident light as appropriate. A certain wavelength may have a certain fraction of its energy absorbed by the blue ink at a certain spot, and then another fraction of its energy is absorbed by the yellow ink at that spot.

For that reason, the ink "primaries" (such as cyan, magenta, and yellow) are often said to be *subtractive primaries*, in contrast to the emissive primaries (such as red, green, and blue), said in this context to be *additive primaries*. Color spaces related to the ink printing process are often distinguished as "reflective spaces".

To achieve the gamut of image colors needed in printing, we must be able to control the "density"[21] of each of the colors of ink. In the most common ink printing process, we cannot do this directly—at a given spot on the paper, the ink is either deposited or not. We however achieve the effect of different "densities" of an ink by the use of the *halftone process*. In that process, the ink is actually applied in a grid of tiny dots, usually on a fixed grid. The diameters of the dots are varied, thus changing the fraction of the paper area affected by that ink.

There are some resulting subtleties in the way the absorption of the different inks interact. Thus the phenomenon of the production of a "color" of the printed image is not as simple as we make it appear in this article. Our intent here is merely to give an understanding of the context in which print-related color spaces operate.

The CMY space

The simplest color space related to the three-color printing process is, not surprisingly, the "CMY" (cyan-magenta-yellow) space. Its three variables represent the relative density of the cyan, magenta, and yellow ink that would be needed to produce the color of interest (and remember, that all assumes that the ink image is illuminated by "white" light not just of a certain chromaticity but in fact of a particular spectral distribution).

[21] We do not use "density" in this section to mean the logarithmic density used in photographic technology, nor to any specific quantitative definition.

In the simplest application of this space, the C, M, and Y variables for a particular color are related to the RGB representation of the color in this simple way (all variables being stated with a range from 0 to 1):

$$C = 1\text{-}R$$

$$M = 1\text{-}G$$

$$Y = 1\text{-}B$$

The CMY space is also relevant to color photography, where the image on a color print or positive transparency is produced by three dyes operating in a "subtractive" mechanism.

Four-color printing

The three-color printing process suffers from a number of practical shortcomings. For one, even if the greatest practical "density" of all three inks is applied to an area, not all the incident light is absorbed, and the area will not appear black to the observer (but rather a muddy brown). Other low-reflectance colors are similarly unsatisfactory. In addition, the use of three "dense" layers of ink can make the printed paper wetter than is desirable.

The solution is the introduction of a fourth ink, black. Obviously, to make a black portion of the image, we can use only the black ink. But for other low-reflectance (dark) colors, we also use some density of black in connection with a reduced density of the other three inks. In effect, we "factor out" the common absorption of all three "colored inks" and replace it with absorption by the black ink. This is often called "gray component removal" (GCR) or "undercolor removal" (UCR)[22].

It is not, however, necessarily most effective to "factor out" all of the common absorption of the three colored inks. There are many empirical formulas used by printers (or mechanized "pre-press" processing software) to determine how much black to use (and how to correspondingly reduce the density of the other inks) to best achieve a certain "color" of the image.

The CMYK color space

The widespread use of the "four-color" process in printing has led to the use of a computer-oriented color space directly related to it. The CMYK space adds a fourth variable, K, to reflect the amount of black used in the eventual recipe for a color. (Evidently "K" was chosen for "black" since "B" was already in use for "blue".)

By representing colors in this form in a computer system, the actual printing process can be given explicit instructions by the computer as to how the

[22] The two terms refer to slightly different aspects of the process as practiced.

undercolor removal should be done for each area of the image. Thus, a sophisticated artist or photo editor, familiar with the practical subtleties of the four-color printing process, can “tune” the undercolor removal process for best final printed result.

Most graphic arts and photo editing software packages allow the operator to adjust the parameters of the default algorithm for the automatic application of undercolor removal. One parameter of the algorithm in effect tells the system how aggressive to be in “factoring out” the common absorption of the three additive primaries for any given color, the range being from “don’t do it” to “take it all out”. In many cases, the algorithm varies for different ranges of overall “brightness”.

MODERN CIE COLOR SPACES

We earlier spoke of the CIE Yxy color space and the related CIE chromaticity diagram. (These are described in detail on Appendixes B and A.)

An important practical issue in commercial color work is that of matching the color of light (or the “color” of a reflective surface) to some established specification. A measure of “color difference”, as perceived by a human observer, is needed to quantify this concept. The matter of chrominance difference is especially important. Unfortunately, the distance between two chrominances on the original CIE chromaticity diagram (the “1931” diagram) does not consistently correspond with perceived chromaticity difference.

The CIE thus subsequently defined other color spaces in which distance on the chromaticity plane more nearly corresponds to perceived chromaticity difference. A number of these figure in the description of color in modern image coding systems.

The new “gamma”

The basic intent of the non-linear representation of R, G, and B, and thus of values derived from them, such as Y, is to accommodate the expected nonlinear transfer function of the assumed display device. In the original YIQ space, for example, this assumed device is a 1948-vintage cathode ray tube (CRT).

As a result of changes in the design of CRT’s, many of them today exhibit substantially different gamma values. And of course other display mechanisms, such as color LCD panels, have entirely different transfer characteristics.

As a result, a TV receiver may well contain circuitry to mediate between the gamma assumed by the signal and the characteristics of the actual display mechanism.

There are two other advantages of the nonlinear representation of luminance and chrominance values. Firstly, when these are to be transmitted as signals (as in the case of television), the nonlinear representation produces a superior perceived image quality for any given signal-to-noise ratio (SNR) in the transmission channel.

Secondly, the nonlinear representations follow somewhat crudely the non-linear response of the human eye. As a consequence, various image manipulation tasks, such as the superimposition of images, can be preformed in a more straightforward way. Unfortunately, the exponent used for gamma precompensation, typically 0.45, is not ideal from the standpoint of correspondence with the response of the eye.

In the color space we are about to review, the CIE L*a*b* space, there is no concern with display device characteristics—the space is totally "device independent". However, correspondence with the nonlinear response of the eye (not accommodated by earlier CIE spaces) **is** an objective, and thus again the color parameters of this space are expressed in nonlinear form. The exponent is optimized for correspondence with human response, a value of 1/3 being specified. Although not apt, some workers speak of inverse of this value (3.0) as the "gamma" of the L*a*b* space.

The CIE L*a*b* ("CIELAB") space

A field of large practical concern is that of specifying and measuring the "color" of a reflective surface, such as an area in color printing or an object which is painted or made of colored plastic. The CIE color spaces we have discussed so far are intended to relate to emissive light sources. In 1976, the CIE published a color space specifically intended to relate to reflective "color".

It also introduced some concepts that had emerged from ongoing research into human perception of color. It recognized that human response to luminance was not linear with the "power-like" definition of luminance which was the basis for earlier CIE spaces (as we discussed in the previous topic).

Additionally, research had indicated that, although the eye as a "camera" sensed color on an RGB basis, at a higher level of image processing in the brain the human perception of chromaticity appeared to follow a space in which the two axes were "redness-*vs*.-greenness" and "yellowness-*vs*.-blueness".[23] The chrominance plane of the new space followed this concept.

Although this new color space was intended to deal with reflective "color", it was soon adapted for use in describing "luminous" (light) colors as well. It has come into widespread use in various graphics software packages and for other "color management" purposes. It is a member of the "luminance-chrominance" family. The traditional designation of the space is "L*a*b*,

[23] This had been mentioned previously in connection with the Y-I-Q color space.

where the asterisks remind us of the nonlinear nature of its three variables. We will omit the asterisks from here on.

The luminance-like aspect of the reflective color is actually described by the variable *lightness*, indicated as L. It is nonlinearly related to the reflectance of the surface so as to more closely follow the human eye's perception; the asterisk in its designation reminds us of its nonlinear nature. L is defined thus:

$L = 116\ Y_n^{1/3} - 16$ range: 0-100

where *Yn* is in effect the reflectance of the color of interest, defined by:

$Yn = Y/Y_0$

where Y the traditional CIE luminance value of the light reflected by the surface of interest and Y_0 is the CIE luminance of the Illumination. (Note that, curiously enough, the same value Y is both the luminance and the value of one of the three CIE tristimulus values, Z, *Y*, and *Z*! [24]) The exponent 1/3 provides the nonlinearity mentioned above.

Chrominance is described by two variables **a** and **b**, representing the red-*vs*.-green and yellow-*vs*.-blue axes respectively. They are defined as:

$a = 500(Xn^{1/3} - Yn^{1/3})$ range: -500 through +500 *

$b = 200(Yn^{1/3} - Zn^{1/3})$ range: -200 through +200 *

* but only the range from -128 through +127 is usually used

where *X*n, *Y*n, and *Z*n are the reflective CIE tristimulus values, defined as:

$Xn = X/X_0$

$Yn = Y/Y_0$ (yes, this is the same as the reflectance!)

$Zn = Z/Z_0$

where X, Y, and Z are the CIE tristimulus values (amounts of the "fictional primaries" X, Y, and Z) for the light reflected by the surface and X_0, Y_0, and Z_0 are the tristimulus values for the illumination.[25] A positive value of **a** represents the red direction on the red-green axis, while a positive value of **b** represents the yellow direction on the yellow-blue axis.

In common practice, the full range of L* (0-100 units) is encoded in the digital representation, usually mapped unto 0-255 in an 8-bit context,

[24] How this comes about is described in Appendix B.

[25] When used for luminous rather than reflective color, X_0, Y_0, and Z_0 the tristimulus values of the illuminance, are replaced by X_r, Y_r, and Z_r, the tristimulus values of the reference white for the system.

0-65280 (256 X 255) in a 16-bit context. But usually for **a** and **b**, only the range -128 through +127 units is actually retained (mapped onto 0-255 in an 8-bit context, 0-65280 in a 16-bit context [26]).

For all the above expressions, if any of the ratios X/X_0, Y/Y_0, and Z/Z_0 is less than or equal to 0.008856, then that ratio is replaced by:

$$\frac{903.3F+16}{116}$$

where F represents the affected ratio. [27]

The purpose of the linear piece of the nonlinear function is that with the traditional function (stated just above), when the inverse function is used (to convert from Lab to XYZ) the slope of the function becomes infinite at the origin, thus leading to implementation difficulties. The linear piece avoids this complication.

When this space is applied to light as such (not to the "reflective color" of a surface), the parameters X_0, Y_0, and Z_0 become the description of the "reference white" to be used.

The coordinates **a** and **b** together describe the chrominance of the color.

Note that since this is a luminance-chrominance (not luminance-chromaticity) space, if we begin with some color and "attenuate" it (such that its chromaticity does not change). L, **a**, and **b** all decrease.

With regard to the designation "CIELAB" for this space, many people mistakenly believe that "LAB" is short for the word "laboratory", and we often read fanciful "translations" of this space's designation based on that fallacy.[28]

---

[26] With the "CIELAB" encoding, this is done on a "signed integer" (twos-complement) basis, where (in an 8-bit context) 0 means 0, 1 means 1, and 255 means "-1"; with the "ICCLAB" encoding, this is done on an "offset" basis, where 128 means 0, 127 means -1, and 129 means +1.

[27] The constants 903.3 and 0.008856 are given here in decimal form, as has been the case so far in the formal CIE standard. However, decimal values of these constants, no matter to how many significant figures they are expressed, produce a discontinuity in the function at the "joint" between the two parts of the curve, leading to some problems in execution. As a result of a campaign by noted color space maven Bruce Lindbloom, the latest version of the CIE standard will express these constants as the ratios of integers, which produces a precise "joint".

[28] But adding to the confusion is the color space, very similar to CIELAB, established by Hunter Associates Laboratory, Inc., and called the "Hunter Lab color scale".

The CIE L*uv ("CIELUV") space

The CIE L*uv color space was developed to approach the situation in which differences in the distance between the points representing two colors was generally proportional to the perceived difference between the colors. Like the CIE L*a*b* space, It is a luminance-chrominance space. Unlike the L*a*b* space, the chrominance values here (u and v) are linear (not gamma-precompensated).

The traditional CIE variable Y*, representing *luminance*, is replaced by the variable L (lightness) in this space, and the chrominance is given by two variables u and v. As before, the asterisk in L* remind us of the non-linear nature of the three variables. We will omit it from here on.

In the equations that follow, we will also use the symbols U and V (rather than u and v) for the chrominance variables to avoid confusion with the variables u and v of the CIE 1960 chromaticity diagram and the variables u′ and v′ of the CIE 1976 chromaticity diagram, upon the latter of which in fact the definition of U and V are defined (as we will see shortly).[29]

The variables of this new space are defined as follows:

$L = 116 y_r^{1/3} - 16$ for $y_r > 0.008856$ range: 0-100

$L = 903.3 y_r$ for $y_r \leq 0.008856$

$U = 13L(u' - u'_r)$

$V = 13L(v' - v'_r)$

where y_r is the luminance of the color of interest, normalized to the luminance of the reference white for the color space, u' and v' are the coordinates of the chromaticity of the color of interest in the CIE 1976 chromaticity diagram system, and u'_r and v'_r are the coordinates in that system of the reference white chromaticity for the color space.

The variables are further defined as follows:

$$y_r = \frac{Y}{Y_r}$$

$$u' = \frac{4X}{X + 15Y + 3Z}$$

$$v' = \frac{9Y}{X + 15Y + 3Z}$$

$$u'_r = \frac{4X_r}{X_r + 15Y_r + 3Z_r}$$

[29] But note that this U and V are not the U and V of the YUV color space1

$$v'_r = \frac{9Y_r}{X_r + 15Y_r + 3Z_r}$$

where X, Y, and Z are the CIE tristimulus values of the color of interest and X_r, Y_r, and Z_r are the tristimulus values of the reference white for the system.

Note that the chrominance axes of this space essentially correspond in their orientation to the chromaticity axes of the CIE uvY color space.

The CIE L*CH space

As we noted at the outset, a coordinate system which most intuitively relates to the human perception of color defines a color in terms of luminance, hue, and saturation. In the *a-b* plane of the CIE L*a*b* color space, the angular position of the chrominance point (with respect to the set of axes) in fact corresponds to hue, and the radius from the origin to the point, **divided by L***, is indicative of saturation.[30]

To produce a more intuitive set of coordinates, we can recast the *a-b* chrominance plane from Cartesian (rectangular) coordinates to polar coordinates, using the variable C to represent the radius to the chrominance point and the variable H to represent the angle in degrees to the point, measured counterclockwise from the positive *a* axis. C essentially indicates the product of saturation and luminance[31], and H essentially indicates hue. The resulting space is called the CIE LCH space.

Note that H in this space essentially defines a position on a "hue wheel", but it is quite different from the hue wheel of the HSB and HSL spaces.

As we said at the outset, just as in the case of the Lab space, this is a luminance-chrominance (not luminance-chromaticity) space. The coordinates H and C collectively describe the chrominance. If we begin with some color and "attenuate" it (such that its chromaticity does not change). L and C both decrease. (H, being essentially an angle, does not change.)

This particular space is today sometimes called the LCH(ab) space, a reminder that the reference axis for definition of the hue angle, H, is the positive **a** axis of the **a**-**b** plane of the parent Lab space.

There is a variant, called the LCH(uv) space in which the reference axis for definition of the hue angle, H, is the positive **u** axis of the **u**-**v** plane of the Luv space.

---

[30] Remember than in "luminance-chrominance" spaces, for a given chromaticity (hue and saturation), the chrominance variables scale with luminance.

[31] This quantity is called "chroma", thus the choice of the symbol. However, in some contexts the whole vector on this plane (described by both H and C) is called "chroma", consistent with the terminology used in color spaces applicable to television signals. Under that viewpoint, C is the **magnitude** of the chroma.

COLOR SPACES FOR PHOTOGRAPHIC IMAGES

The sRGB color space

A color space that is widely used for the interchange of digital images is identified as the "sRGB" (standard RGB) color space. As its name suggests, it is an RGB type of space. It is defined by international standard IEC 61966-2-1.

Today, many scanners, digital cameras, display systems, and printers inherently operate on the basis of the sRGB color space (or can be set to do so).

An advantage of the sRGB color space its that its gamut is a good match to the gamut of a typical computer CRT display.

There are disadvantages of the use of the sRGB color space as this "standard of interchange". One is that its gamut (the range of colors that it can directly represent) is relatively limited compared to the gamut that can actually be achieved on many modern output devices (such as certain types of printer.[32] Accordingly, in some situations, various alternative color spaces are utilized in representing a digital camera image or the like. We will look into that presently.

The defining properties of the sRGB space are:

- A specific set of the primary chromaticities, R, G, and B, defined in terms of their CIE x and y coordinates. These essentially correspond to a common set of the "phosphor" chromaticities of a typical computer CRT display.
- A particular "white point", which is the chromaticity of any color whose coordinates follow R=G=B. This is the daylight chromaticity known as "D65". Its correlated color temperature is 6500 K.
- A particular gamma-precompensation curve, slightly more complicated that the simple power curve described earlier in connection with the generalized RGB space.

The gamma precompensation curve used in the sRGB color space is defined by the following expressions. Here, ***c*** represents any of the "linear" variables, **r**, **g**, or **b**, while C represents the corresponding gamma-precompensated variable, R, G or B.

$$C = 12.92c \qquad \text{for } 0 < c \leq 0.0031308$$

$$C = 1.055c^{1/2.4} + 0.055 \qquad \text{for } 0.0031308 < c$$

[32] The gamut of the sRGB color space is essentially that of the type of display device contemplated by the original RGB space.

This two-piece curve can be broadly approximated by:

$$C = c^{1/2.2}$$

which as we said earlier is the typical curve used for generalized RGB spaces.

The purpose of the linear piece of the sRGB gamma precompensation function is that with the traditional function (stated just above), when the inverse function is used (to convert from R, G, and B to **r**, **g**, and **b**) the slope of the function becomes infinite at the origin, thus leading to implementation difficulties. The linear piece avoids this complication.

The Adobe RGB color space

Adobe Systems introduces an alternate to the sRGB color space, generally called "Adobe RGB (1998)". It differs from sRGB in these ways:

- Its green primary is different from the green primary in the sRGB color space. being located farther from the white point. This provides a larger chromaticity gamut.

- The gamma precompensation function is the traditional one: $C = c^{1/2.2}$.

The Adobe RGB color space is widely used as an alternative to sRGB in digital camera outputs and as a "working space" for the processing of digital camera images. The attraction is its larger chromaticity gamut. A disadvantage (given that the same "bit depth" is used for the variables) is worse "granularity" of the color representation (the same number of distinct code representations is spread over a larger gamut).

The sYCC color space

Perhaps the most commonly used output image file format for digital cameras is the Exif (Exchangeable Image File Format) file, in particular, its JPEG version. (JPEG refers to a system of "compressing" the information in an image file to take represent the image in fewer bits than originally.) It is defined by international standard IEC 61966-2-1, Annex G.

The JPEG coding system actually operates upon image data in YCbCr form, and in the form of JPEG prescribed for use in the Exif format, that YCbCr form is specified to be based on an underlying sRGB color space.

The resulting color space is often described as the "sYCC" color space (YCC being a shorthand for YCbCr), but that designation has an implication beyond the use of the YCbCr representation.

We encountered the YCbCr color space before in the section on Color Spaces for Television Images, and its detailed definition is given there. It is the "still images" version that is relevant here.

The “code space” of a YCbCr representation has the following ranges:

Y: 0 to1

Cb, Cr: -0.5 to +0.5

This is a larger code space than is required to accommodate the possible range of RGB values that are legitimate under the basic definition of the sRGB space (in which R, G, and B can only take on values in the range 0-1). Said another way, the YCbCr color space supports a gamut considerably larger than that supported by the basic sRGB space. In fact, for parts of the luminance range, the gamut implied by the YCbCr code space embraces chromaticities outside the CIE chromaticity diagram, that is, “impossible” (or “invisible”) chromaticities.

The sYCC colors lying outside the sRGB gamut have RGB representations in which at least one of the variables R, G, and B have values lying lie outside the range normal range of 0 to 1—values greater than 1, or negative values. Negative values, in particular, are at first hard to visualize. Nevertheless, such values can readily be treated mathematically.

Thus, if we fully utilize the sYCC color space (allowing the variables to take on any values in their prescribed range) we can represent with valid YCbCr coordinates some realizable colors lying outside the sRGB gamut, while still operating within the basic sRGB encoding concept.

Only a small change in the sRGB color space definition is required for it to participate as the underlying layer in the sYCC color space: the gamma precompensation curve must be extended to accommodate negative values of R, G, and B. This is done in the obvious way, with the added third quadrant of the curve being symmetrical with the original first quadrant. This symmetrical curve is defined by the following expressions:

$$C' = -1.055(-C)^{1/2.4} + 0.055 \qquad \text{for } C < 0.0031308$$

$$C' = 12.92C \qquad \text{for } -0.0031308 \le C \le 0.0031308$$

$$C' = 1.055C^{1/2.4} - 0.055 \qquad \text{for } 0.0031308 < C$$

While sYCC images containing “out-of-sRGB-gamut” colors can be processed by various image processing software applications that are “sYCC aware”, and can be passed on over various interfaces to the drivers for suitable rendering devises, a problem arises if these applications need to deliver an output in orthodox sRGB form.

In that case, the application must “map” such out-of-gamut colors to valid sRGB colors. This can be done in several ways. One is to just force them barely inside the sRGB chromaticity gamut, leaving the “in-gamut” colors alone. However, this approach can result in a range of different colors in the sYCC image ending up with the same sRGB color, leading to unpleasant results in the affected areas of the final image—“flattening” of the colors (much as we see in the "clipping" of colors falling outside the dynamic range of the camera).

Another approach is to uniformly "squeeze" the entire gamut of the image to fit within the sRGB chromaticity gamut. This avoids the visible anomalies of the other approach, but does not preserve the accuracy of any of the colors.

These different strategies of "accommodation" are spoken of as different "rendering intents", the term supposedly reflecting that the choice among them is typically determined by the type of image and the use to which the final output is to be put.

The details of these processes and of their implications on image results are beyond the scope of this article.

The e-sRGB color space

The International Imaging Industry Association ("I3A") has standardized a variant of the sRGB color space having a substantially expanded gamut compared to that of sRGB. The basic principle is to recognize and allow values of **r**, **g**, or **b** (the non-gamma-precompensated form) that lie in the range -0.53 to +1.68. The result is that the gamma-precompensated forms (R, G, and B) then can have values in the range -0.75 to +1.25 (a range of 2.0 "units" altogether). (The symmetrical gamma-precompensation function already described in connection with the sYCC color space is used.)

Before these values are placed in digital form, they are divided by 2 and then 0.375 is added, bringing their range to 0 to 1.

That adjusted value is then converted to digital form, normally with a precision of either 10, 12, or 16 bits. Thus the range (expressed in decimal form) becomes 0-1023, 0-4095, or 0-65535.

There is of course a YCbCr representation of this color space (called e-sYCC). It is derived in exactly the same was as the sYCC representation of sRGB, but with a further wrinkle.

Following the usual equations defining Cb and Cr, those variables can also attain the range -0.75 to +1.25 (a range of 2.0 "units" altogether). As with the variables R, G, and B in the e-sRGB color space, these variables are divided by two and 0.375 is added (making their range 0-1) before they are converted to digital form (again, in either 10, 12, or 16 bits).

Again following the usual equation, Y it can also take on the range -0.75 to +1.25. Here, however, that is fit into a final range of 0-1 before conversion to digital form by merely clipping values below 0 and above 1 (forcing them to 0 and 1, respectively).

This clipping of course effectively discards some of the potential expanded gamut attained by this overall color space, but the loss does not turn out to be significant—the areas lost have restricted chromaticity gamuts (much of their theoretical gamuts lies outside the range of visible colors). The plus side of the clipping approach is that the resolution of Y is improved (for any given bit depth), since the available number of distinct code values is stretched over a smaller range of Y).

Essentially these same two color spaces are also defined by IEC 61966-2-1, Annex G, where they are spoken of as “bg-sRGB” and “bg-sYCC”.[33]

The PCS color space

In modern times, a sophisticated scheme of transforming color representations from one color space to another has been defined by the International Color Consortium. Thus scheme depends on “color management profiles”, definitions of the transformation of color representation between the color space of some input or output device and a certain standard reference color space. That color space, called the Profile Color Space (PCS). It is based on either the CIE XYZ or CIE L*a*b space. Although that would seem to be the definition of the space almost trivial, in fact the matter is complicated by issues of accommodation of the white and black points of the source and destination media, “chromatic adaptation” of the viewer, and other arcane colorimetric matters.

In any event, the PCS color space is ordinarily never used for an “external” description of an image—only as the intermediate language between two links of a transformation chain.

Thus, if in some image management software application aware of the ICC color management scheme, we cite the files carrying the profiles for the source device (say, a particular digital camera) and the destination device (say, a particular printer), the application, in effect, first transforms the color descriptions in the image source file from the source color space to the PCS color space, and then transforms the descriptions from the PCS color space to the destination color space.

Of course, this is not necessarily done in two steps. The application may, in effect, first “multiply” the two space transforms to get a “net transform”, which it then uses to convert the color descriptions directly from the source color space to the destination color space.

This is a highly complex field, with many ramifications, and its details are beyond the scope of this article.

#

[33] Perhaps “bg” is intended to be evocative of “big gamut”.

APPENDIX A

The CIE Chromaticity Diagram

An important color space which is very useful in technical work is known as the *CIE xyY color space,* a member of the luminance-chromaticity family*.* (CIE are the initials of the French name of the International Commission on Illumination). It plots any chromaticity as a point on a graph (the *CIE chromaticity diagram*[34]*)* whose two axes correspond to arbitrary variables called just *x* and *y*.[35] These are defined so that the "mapping" of different chromaticities to points on this graph results in certain desirable properties which we will hear about later. We will see this diagram shortly.

The CIE xyY color space is derived from the CIE XYZ color space, a tristimulus space. It is described in Appendix B.

The underlying principles behind the CIE XYZ and xyY spaces and the CIE chromaticity diagram are described in Appendix B. Here, though, appendix we will deal with the practical significance of the CIE chromaticity diagram.

Before we discuss the CIE chromaticity diagram, we need to do a little review of some other underlying technical concepts.

Wavelength

Light is electromagnetic radiation, identical in physical nature to radio waves, except that the frequency of light is immensely greater than the frequency of even the highest-frequency radio waves in use today. When the science of optics was first developed, there was no radio, no electrical engineering, and no understanding of the wave nature of light. Even if there had been, there were certainly no instruments for directly measuring the frequency of light directly (there still aren't). Thus, the frequency of a light source was described in terms of the corresponding inverse property *wavelength*. Visible light comprises light whose wavelengths range from approximately 400 to 700 nanometers. (A nanometer is a millionth of a millimeter.)

---

[34] Actually, the "*1931 CIE chromaticity diagram*"; two later chromaticity diagrams have been adopted by the CIE, one in 1960 (with variables *u* and *v*) and one in 1976 (with variables *u*' and *v*'). The 1931 version is still the one most widely used for general discussions of chromaticity, and when we speak in this paper of the "*CIE chromaticity diagram*" without specifying the version, it is the 1931 version which is meant.

[35] See Appendix B for a description of the concepts of this coordinate system.

The Spectrum

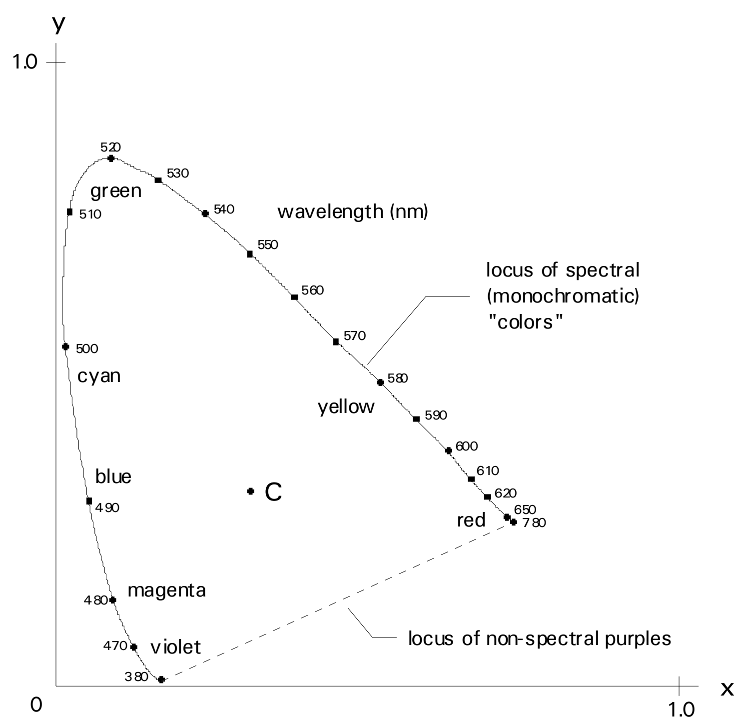
If we send a stream of sunlight through a triangular glass prism and direct the emerging beam onto a screen, we see the familiar "rainbow" band of continuously-changing hue. This is known as a *light spectrum*.

In a more technical sense, a *spectrum* is a plot of the distribution of the power content in a signal (such as an electromagnetic wave, of which light is an example) as a function of frequency (or wavelength).

If we take an opaque plate with a very narrow slit and put it between the prism and the screen, it will only pass the light of a particular part of the spectrum. As the width of the slit approaches zero, we find that we approach the situation in which the light passing through contains only one frequency (wavelength), Such a light source is called *monochromatic* (meaning "one color", not really a technically-correct definition). These chromaticities are called the *spectral colors* (again, not really technically correct).

Back to the CIE

If we plot the range of visible spectral chromaticities on the CIE chromaticity diagram, we get the horseshoe-shaped curve shown with a solid line on the figure below.



This is sometimes called the *locus of spectral colors,* or *locus of monochromatic colors*. ("Locus of spectral chromaticities" would be more accurate.) Any point on the locus corresponds to the chromaticity of monochromatic light of a particular wavelength. A scale of wavelength in that connection is shown on the curve.

The curve is also labeled with the familiar names of various hues. However, such a name (such as "red") does not actually imply a specific wavelength nor even a specific hue—it is a broad, qualitative designation.

We have not yet talked about the actual measures used to define hue and saturation. For the moment, note that spectral (monochromatic) chromaticities have the highest possible saturation (usually described as 100%); they are sometimes said to be the "pure colors" [36].

Chromaticities which lie along the straight line joining the open ends of the locus of spectral chromaticities (they are various "purple" hues) do not appear in the light spectrum, and accordingly are not monochromatic. They are called *the non-spectral purples*, and are arbitrarily considered to have 100% saturation.

In the next part of our discussion, we will have to utilize the concept of *white light*. In fact, the chromaticity we call "white" is not defined by a unique set of physical or subjective properties; the choice of a chromaticity to call white is arbitrary. Various chromaticities (defined in terms of their position on the CIE diagram) have been standardized for different uses. One often referenced is called "illuminant C"[37]. Its location is shown on the figure (as "C").

Any point in the interior of the region bounded by the locus of spectral chromaticities and the locus of non-spectral purples represents a chromaticity which is not 100% saturated.

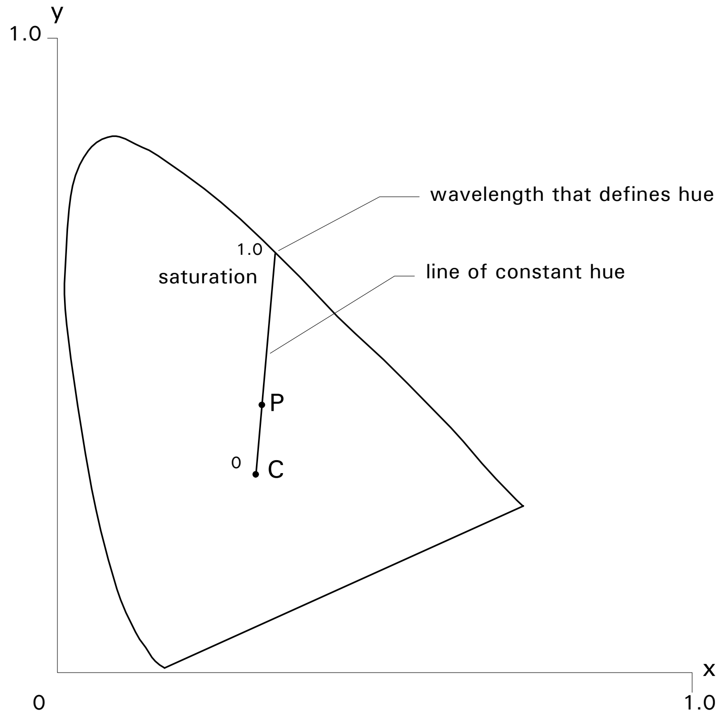
As we move closer to illuminant C, the chromaticities become more "pastel": their saturation becomes less. When we reach "C", we see a "colorless" light. Rigorously, this is a chromaticity whose saturation is said to be 0% (and thus whose hue is undefined).

In fact, the quantitative definition of any chromaticity, as seen on the CIE x-y diagram, in terms of hue and saturation can be done this way.

---

[36] Note that in many of phrases used in this field, the term "color" is used where "chromaticity" is actually the property of interest. In this paper, we will be tediously rigorous, and always use the term "chromaticity" where that is what is meant.

[37] It is intended to correspond to a certain type of daylight. Another reference white, "D65", is also often utilized today. Its chromaticity is that of a "black body" radiator at a temperature of about 6500 K.



On the diagram, draw a line from point "C" through the point representing the chromaticity of interest (P) and prolong it until it intersects the locus of spectral chromaticities (or the locus of non-spectral purples). If it hits the locus of spectral chromaticities, the hue is defined by the wavelength at the point of intersection (said to be the *dominant wavelength* of the light). If the line hits the locus of non-spectral purples, we have no tidy numerical way to state the hue.

In either event, the saturation of a color is defined as the ratio of (a) the distance from point "C" to the point representing the chromaticity of the color to (b) the total distance from point "C" to the locus. Thus the point halfway along a line from point "C" to the locus represents a chromaticity with 50% saturation.[38]

Imagine that we have light that is made by combining two light streams, each having a monochromatic (spectral) chromaticity and both of equal luminance. The chromaticity which the observer will perceive can be located on the CIE diagram merely by drawing a line between the points representing the two spectral chromaticities and locating its midpoint. This is a useful property of the CIE diagram[39].

---

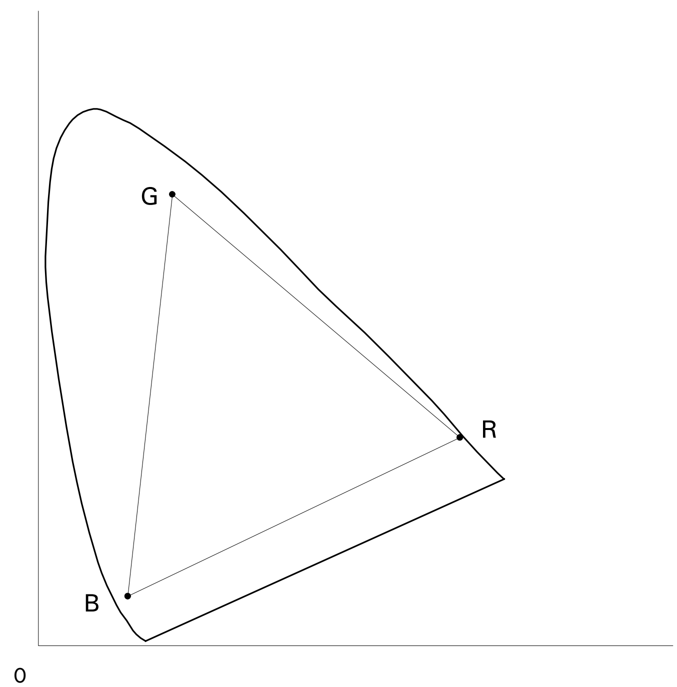
[38] Note that this quantitative definition of saturation is dependent on the choice of a particular chromaticity for "white".

[39] This property also obtains for the 1960 and 1976 CIE chromaticity diagrams.

If the two components do not have equal luminances, we can still use this technique. Suppose one spectral component has twice the luminance of the other. Then we still draw the line between the two components' points, but we mark a point on the line which is one-third the distance from the "stronger" point to the "weaker" one.

Does that mean that any point in the interior of the diagram represents a chromaticity which can only be made by one combination of spectral chromaticities (that is, by combining monochromatic light of only two specific wavelengths)? No, as one can readily visualize, we could make a particular "point" by many combinations of two or more spectral components, or by many combinations of two or more non-spectral chromaticities (ones inside the locus).

One conceptual way to make light of any desired chromaticity is to take spectral light (or non-spectral purple light) of the appropriate hue and "dilute it down" with some amount of "white light" (often meaning illuminant "C") to the desired saturation.



In a "tristimulus" color space, we define the color of light in terms of three "primaries": light sources of specified chromaticities. Appropriate amounts of these three primaries are added together to produce the desired color. The absolute amounts of the three primaries determine the luminance of the resulting color. Their relative amounts determine the chromaticity.

Above we see the CIE chromaticity diagram with an illustrative set of primaries—those of a typical form of the RGB space.

If we add together various proportions of, say, B and R, we can construct any chromaticity lying along the line B-R. If we then take that mix and add some amount of G, we will get a new result which lies in the interior of triangle R-G-B. In fact, by adding the proper proportions of R, G, and B, we can construct light whose chromaticity lies anyplace within triangle R-G-B, but never outside it.

Thus the triangle R-G-B bounds the set of chromaticities that can be defined in an orthodox manner by an RGB space based on that particular set of primaries. This is said to be the "chromaticity gamut" of that particular space. Any chromaticity lying outside that triangle would imply a negative value of the quantity or one or more of the primaries. This is of course not possible if we actually think of manufacturing the color with physical sources of the three primaries.

Note however that in an abstract frame of reference we could define chromaticities outside the triangle by accepting such negative values.

#

APPENDIX B

The CIE XYZ and xyY color spaces (and why the symbol for luminance is "Y")

After extensive study of human color perception, various researchers (including the renowned James Clerk Maxwell) concluded that the human eye contained three separate sets of light detectors, each of which had a different curve of response to the different wavelengths over the visible spectrum. The peaks of these three curves lie in the red, green, and blue portions of the spectrum.

A we saw earlier in our discussion of the 1931 CIE chromaticity diagram, from a set of three "primaries" we can make up any color within a certain gamut of chromaticities. Conversely, any color within that gamut can be described by the amounts of these three primaries that would need to be added to produce that color. A particular set of three monochromatic (spectral) primaries in the red, blue, and green regions (falling generally at the peaks of the three color receptor response curves) were adopted as a standard for this purpose. The three primary chromaticities are called R. G, and B, and these letters also serve as the names of the variables describing the amount of each primary used to make up a particular color. However, for consistency with the convention used elsewhere in this article, we will use the lower-case letters (**r**, **g**, and **b**) for these variables, reserving the upper-case letters for "gamma-precompensated" forms of the variables.

It turns out that, for any color which is so composed, the perceived luminance of the color follows a certain weighted summation of the values **r**, **g**, and **b**. The chromaticity of the color is determined by the ratios among **r**, **g**, and **b**. (Note that only two values are needed to describe the ratios among three values.) The set of these three values are the coordinates of one particular three-dimensional "color space".

Again, as we saw from out earlier study of the CIE diagram, with three monochromatic primaries we cannot produce all possible visible chromaticities, only those lying within the triangle formed by the primaries. We could describe a chromaticity which lies outside that triangle by utilizing negative values of at least one of the variables **r**, **g**, or **b**. That concept, however, is intellectually untidy (but is in fact embraced by certain "expanded gamut" color spaces in the RGB family).

To avoid this difficulty, the CIE decided to define a linear transform of the set of three variables **r**, **g**, and **b** into another set of three arbitrary variables, X, Y, and Z, such that for any visible chromaticity none of the variables would need to have negative values.

Remember that the variables **r**, **g**, and **b** represent amounts of three primary chromaticities of light (R, G, and B) to be added to make the color of interest. Similarly, the new variables X, Y, and Z also represent the amounts of three new primary chromaticities (also called X, Y, and Z) to be added together. The wrinkle here is that these are not physically-realizable, visible chromaticities; they are fictional creatures of mathematical convenience.

There would of course be an infinite number of linear transformations between the **r**, **g**, **b** coordinate system and an alternative system (X, Y, Z) which would meet our objectives. The CIE cleverly chose to make the transform function for the variable Y the very same weighted combination of **r**, **g**, and **b** that yields the luminance of a color. Thus the new variable Y was in fact the luminance of the light described in the XYZ space.[40] But Y also still represented an amount of a specific (although fictional) chromaticity of light (primary chromaticity "Y")![41]

Now the CIE chromaticity diagram is not intended to show all three variables (it would have to be a three-dimensional "solid" diagram to do that), but to show chromaticity only (in two dimensions). Its two coordinates are derived from *X*, *Y*, and *Z* this way:

$$x = \frac{X}{X+Y+Z}$$

$$y = \frac{Y}{X+Y+Z}$$

Note that in any color space, if we start with a particular color, with certain values of three primaries, and increase its luminance (without changing its chromaticity), the proportions of its three primary components remain unchanged. Thus, in the CIE system, for an unchanging chromaticity, X and Y increase by the same ratio that (X+Y+Z) does, and so x and y remain unchanged. Therefore, the point x,y in fact represents the chromaticity of a color.

Keeping in mind that the variable *Y* indicates the luminance of the color of interest, the entire 1931 color space is often designated the "xyY" (or sometimes "Yxy") space.

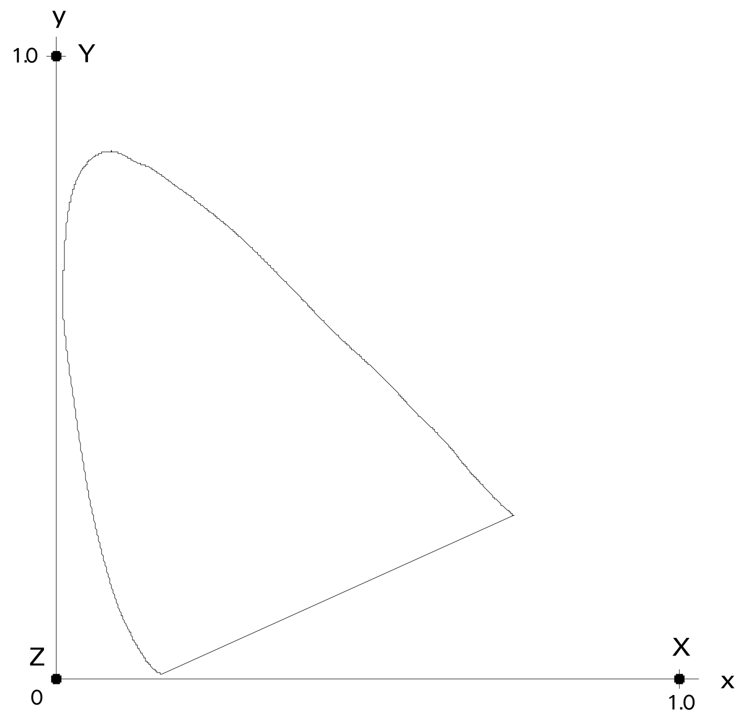
The primaries X, Y, and Z, though fictional and non-visible, have chromaticities which can be plotted on the CIE *x*-*y* plane. From the equations

---

[40] For this reason, in fact, "*Y*" is the symbol for *luminance* in most color spaces (also the symbol for *luma*, a property that is something like luminance).

[41] Most written descriptions fail to alert the reader to the fact that "Y" is both the name of a (fictional) primary chromaticity and the name of a variable (the amount of light of that chromaticity), two different concepts.

above, we can determine that the chromaticity of X lies at (1,0), of Y at (0,1), and of Z at (0,0), as we see in this figure.



We can see that the gamut triangle of X-Y-Z embraces the entire visible chromaticity gamut, which as we recall was an objective of the choice of those three primaries.

#