Chapter 1. Basic Color Science

This is a brief introduction to basic color science. The emphasis is on mathematical models and different color spaces. The text was originally written as a course material for a course in graphical image technology. Most of the illustrations are taken from "Inverse Halftoning of Scanned Color Images". This text has been modified later on to be a part of a lab in a course in Graphic Arts. **Notice**: **Read this chapter thoroughly before you start the exercises in Chapter 2.**

1.1 What is color?

The human eye is able to detect light—i.e., electromagnetic radiation—with wavelengths in the interval between 380–780nm. The radiant flux of the observed light at each wavelength is expressed by a *Spectral Power Distribution* (SPD), such as in Fig. 1.1. We see color by means of light sensitive cells called cones on the eye's retina. There are three types of cones, sensitive to wavelengths approximately corresponding to red, green, and blue light. The response signals from the cones upon stimuli are processed by the brain, which associates the signal to a visual color sensation.

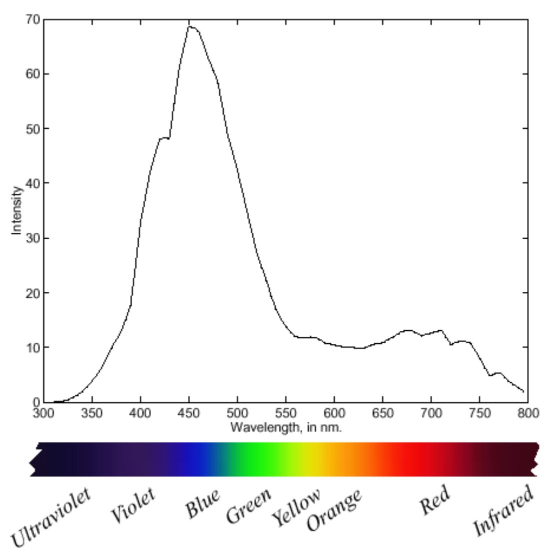


Fig 1.1: *An example of a spectral power distribution. This distribution will be perceived by the human observer as a bluish color.*

1.2 Measuring colors

The human eye has three different color receptors called cones. They are referred to as **L**, **M**, and **S** cones (also sometimes called, **r**, **g** and **b**), since they are sensitive to long, medium, and short wavelengths in the visible spectrum, respectively. In addition to the cones there are also receptors called rods. They are of one type only, are more sensitive to light, and are used for night vision.

The impression of color is related to how the human eye works, it is natural to use the eye's three sensitivity functions as a mathematical foundation. Furthermore, light with different spectral distributions yielding the same color impression should be measured as the same color.
Denote the incoming light's spectral photon distribution $E(\lambda)$, and the sensitivity functions of the cones as $L(\lambda)$, $M(\lambda)$, and $S(\lambda)$. This gives the following total stimulations of each cone type:

$$
\begin{aligned}
L &= \int_{\lambda} E(\lambda)L(\lambda)d\lambda \\
M &= \int_{\lambda} E(\lambda)M(\lambda)d\lambda \\
S &= \int_{\lambda} E(\lambda)S(\lambda)d\lambda
\end{aligned}
\tag{1.1}
$$

The values received by calculating such integrals over the incoming light and sensitivity functions are referred to as tristimulus values. $E(\lambda)$ can originate from a light source, or it can be reflected light from some object. The latter case can be described in the following way: denote the photon distribution from the light source illuminating the object as $I(\lambda)$, and the objects influence on the incoming light, the reflectance function, as $R(\lambda)$. Then make the following substitution in the equations above:

$$E(\lambda) = R(\lambda) \cdot I(\lambda) \tag{1.2}$$

The implications of Eq. 1.2 are that two objects may have the same color in one illumination and then look very different from each other in another, and that a single object is perceived as having different colors when viewed under different illuminations. These effects are called **metamerism**. Metamerism is a large problem when two colors are supposed to match each other. It is more or less unavoidable, unless the two objects have exactly the same reflectance functions.
Equations 1.1 and 1.2 would be a useful way of expressing colors, if only the sensitivity functions of the cone types were known exactly. Since they are not, a different approach has to be taken. In 1931, CIE (Commission Internationale de l'Éclairage) proposed that the sensitivity functions for the **L**, **M**, and **S** cones should be replaced by three other well defined sensitivity functions, called, $r(\lambda)$, $g(\lambda)$ and $b(\lambda)$. These were found by an experiment, described in the lectures. The conclusion was that not all of the reference wavelength colors were possible to create with the basis colors, which resulted in non-realistic color matching functions. Therefore a linear transformation was concocted:

$$
\begin{bmatrix} \bar{x}(\lambda) \\ \bar{y}(\lambda) \\ \bar{z}(\lambda) \end{bmatrix} = \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} r(\lambda) \\ g(\lambda) \\ b(\lambda) \end{bmatrix} \tag{1.3}
$$

These transformed functions can be seen in Fig 1.2. The transformation matrix may seem strange—why more decimals in the middle row? The answer is that $\bar{y}(\lambda)$ was desired to match the eye's spectral luminous efficiency curve for high light level (photopic) vision.

It simply took five decimals to achieve a good match.

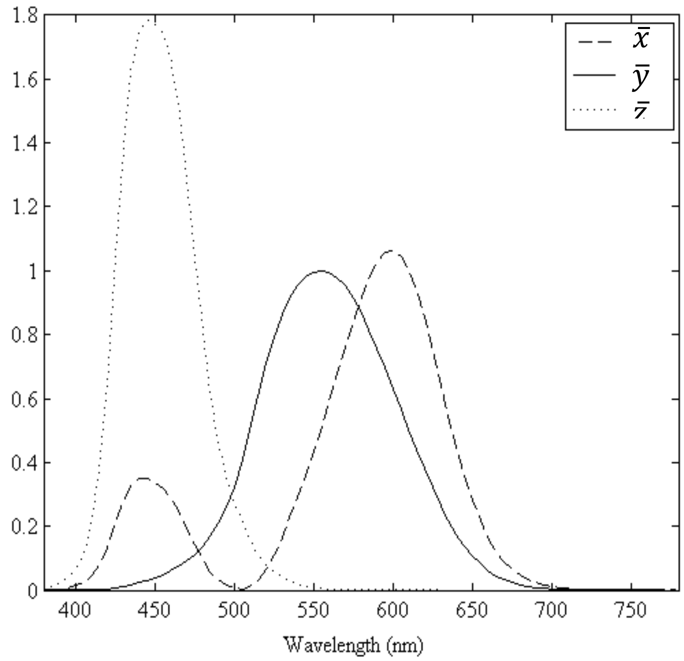


Fig 1.2: *CIE $\bar{x}\bar{y}\bar{z}$ color matching functions.*

From the color matching functions defined in Eq. 1.3, tristimulus values can be calculated. They are normalized for the current illumination, so that a completely white surface (reflectance function equal to one, for all wavelengths, i.e. $R(\lambda) \equiv 1$) always will give Y=100:

$$\begin{aligned} X &= k\int_{\lambda} I(\lambda)R(\lambda)\bar{x}(\lambda)d\lambda \\ Y &= k\int_{\lambda} I(\lambda)R(\lambda)\bar{y}(\lambda)d\lambda \\ Z &= k\int_{\lambda} I(\lambda)R(\lambda)\bar{z}(\lambda)d\lambda \end{aligned} \tag{1.4}$$

$$k = \frac{100}{\int_{\lambda} I(\lambda)\bar{y}(\lambda)d\lambda} \tag{1.5}$$

where, as before, $I(\lambda)$ is the photon distribution from the light source illuminating the object, and $R(\lambda)$ is the objects influence on the incoming light, the reflectance function. **SPDs are often only measured in sample intervals of 5nm, so the integrals above will in practice be replaced by sums (as in the lab exercises)**.
The illumination is often assumed to be one out of several standard illuminations, defined by CIE. Commonly either D50, or D65 is used. They have approximately the radiation characteristics of black bodies at the temperature (in Kelvin) given by the number after D multiplied by 100. Each illumination set has its own **white point**, for example:

- **Xn** = 96.42, **Yn** = 100, **Zn** = 82.49 for D50.
- **Xn** = 95.04, **Yn** = 100, **Zn** = 108.88 for D65.

The white point is calculated by assigning $R(\lambda) \equiv 1$ in Eq. 1.4 above. Note that the colors of all objects in the particular illumination are in the intervals, $0 \leq X \leq X_n$, $0 \leq Y \leq Y_n$ and $0 \leq Z \leq Z_n$, if fluorescence effect is neglected. They can not be negative since the color matching functions are positive, and they can not be higher than the white point, since no object can reflect incoming light better than an object with reflectance function equal to one, for all wavelengths.

1.3 Color spaces

From the XYZ tristimulus values defined in Eq. 1.4, several different color spaces can be derived, each suitable for different applications. We are going to look a little closer on the following spaces:

- **RGB, CMY, and CMYK**—simple device dependent color spaces for reproduction on computer monitors, or on paper.
- **Chromaticity spaces, such as CIE xyY**—color spaces based on the chromaticity characteristics. What this means will be explained later.
- **CIE 1976 (L* a* b*)**—device independent color space created for good perceptual uniformity. That is, colors at equal distance in any direction and at any location in the coordinate system, will be perceived as equally different by the human eye.

1.3.1 RGB, CMY and CMYK

In this section all variables are assumed to be in the interval [0, 1]. The most popular color spaces are probably RGB and CMYK. The first is an acronym for Red-Green-Blue and is used on computer monitors, since those utilize these three colors as primary colors. The second is an acronym for Cyan-Magenta-Yellow-blacK and is used for paper printing, since these are the ink colors utilized in the color printing process. The link between them is the CMY color space, Cyan-Magenta-Yellow. They are all device dependent spaces—the actual color is depending on the device's characteristics. The conversion between these formats that we will use is shown below, Eq. 1.6-1.8. Notice, however, that this equation is a simplification, assuming ideal primaries (RGB and CMY). An accurate conversion between RGB and CMY is generally more complex and depended on the specific devices.

$$\begin{bmatrix} C_3 \\ M_3 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1-R \\ 1-G \\ 1-B \end{bmatrix}; \qquad \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1-C_3 \\ 1-M_3 \\ 1-Y_3 \end{bmatrix} \tag{1.6}$$

When converting between CMY and CMYK, an index showing the number of color components in the space is present to distinguish the different components. Why the fourth (black) color is used will be explained later.

$$\begin{bmatrix} K \\ C_4 \\ M_4 \\ Y_4 \end{bmatrix} = \begin{bmatrix} \min(C_3, M_3, Y_3) \\ \frac{C_3 - K}{1 - K} \\ \frac{M_3 - K}{1 - K} \\ \frac{Y_3 - K}{1 - K} \end{bmatrix}$$

$$\begin{bmatrix} C_3 \\ M_3 \\ Y_3 \end{bmatrix} = \begin{bmatrix} C_4 \cdot (1-K) + K \\ M_4 \cdot (1-K) + K \\ Y_4 \cdot (1-K) + K \end{bmatrix} \qquad (1.7)\ \&\ (1.8)$$

The question that now remains is: how do we convert from XYZ tristimulus values to any of the three color spaces above? This is where the device dependency becomes obvious. The transformation from XYZ to RGB must depend on the device's characteristics, and primarily its white point. These characteristics must be measured, or obtained from the manufacturer of the device. Computer monitors often have a white point close to that of D65 and a suitable transformation matrix is proposed to be:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 41.2453 & 35.7580 & 18.0423 \\ 21.2671 & 71.5160 & 7.2169 \\ 1.9334 & 11.9193 & 95.0227 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad (1.9)$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = A^{-1} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \qquad (1.10)$$

where A is the matrix in Eq. 1.9. The connection to D65's white point is that the row sums in A are the white point coordinates for X, Y, and Z, respectively, for this particular illumination.

1.3.2 Chromaticity

Sometimes the relative difference between the **X**, **Y**, and **Z** values are of more interest than the actual values. Then normalized versions, called chromaticity values, can be calculated:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \frac{1}{X+Y+Z} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \qquad (1.11)$$

Observe that these x, y, and z in Eq. 1.11 are not the same thing as the color matching functions, which also were denoted with lower-case letters. Since $x + y + z = 1$, one of the chromaticity values can be calculated from the other two, and can thus be ignored. Usually z is ignored. The chromaticity values give a possibility to plot colors in a two-dimensional subspace, with x and y as coordinates. By filling out with the Y component from the tristimulus values, a complete color representation is achieved, called CIExyY:

$$\begin{bmatrix} x \\ y \\ Y \end{bmatrix} = \begin{bmatrix} \dfrac{X}{X+Y+Z} \\ \dfrac{Y}{X+Y+Z} \\ Y \end{bmatrix}, \tag{1.12}$$

A plot of all monochromatic colors (pure colors of only one wavelength) in a chromaticity diagram is often referred to as *the spectral locus*. The spectral locus in the xy plane has the form of a horse shoe (see Fig. 1.3). Chromaticity diagrams are very common in color science literature, and are useful as long as one remembers that the colors exist in a three-dimensional space, and not two-dimensional as the chromaticity diagrams may imply.

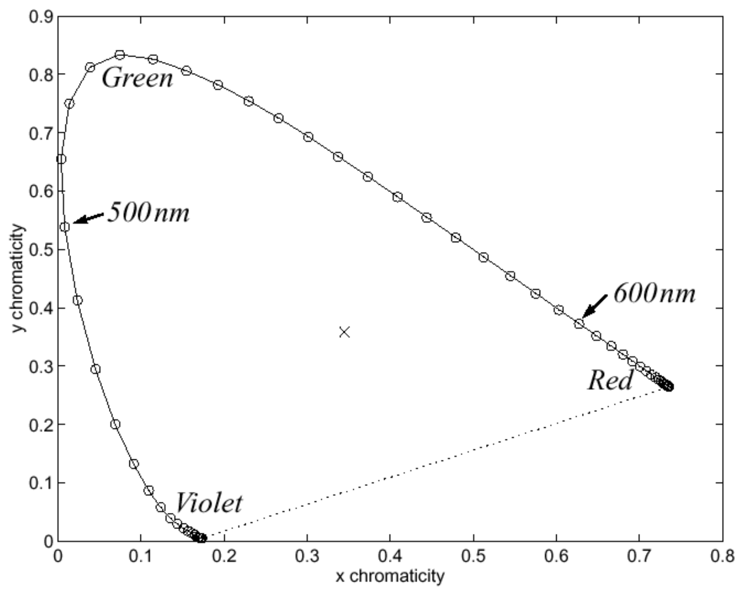


Fig 1.3: *The spectral locus in a chromaticity diagram. The purple boundary is dotted. The white point of D50 is marked with an "x". The distance between the circles is 5nm.*

A chromaticity diagram can be used to visualize the fact that not all colors can be reproduced on a specific display (see Fig. 1.4). It is surprising to see how small the color gamut of the display is, compared to the spectral locus. This CRT-display is especially bad on saturated green colors.

1.3.3 CIELAB

The CIE 1976 (L* a* b*) color space, usually written CIELAB, is derived from XYZ coordinates with aim on perceptual uniformity. Originally it was developed to give the textile industry an accurate way to describe colors. Now it serves as one of the most well-known device independent color spaces for all kinds of applications.

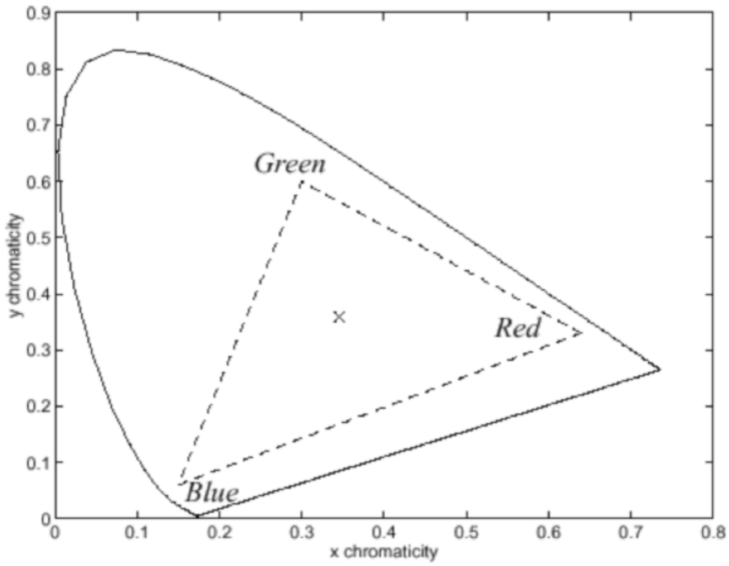


Fig 1.4: *The color gamut of a CRT-display. The corners of the triangle are the chromaticity coordinates for red, green, and blue phosphor.*

The transformation between XYZ values and CIELAB values is defined by:

$$L^* = \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16, & \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right), & \frac{Y}{Y_n} \le 0.008856 \end{cases}$$

$$a^* = 500\left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \tag{1.13}$$

$$b^* = 200\left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right)$$

$$f(x) = \begin{cases} x^{\frac{1}{3}}, & x > 0.008856 \\ 7.787x + \frac{16}{116}, & \frac{Y}{Y_n} \le 0.008856 \end{cases}$$

The constants X_n, Y_n, and Z_n are the XYZ values for the chosen reference white point. When working with color monitors good choices could be something close to D65's.
How shall the coordinates in CIELAB be interpreted? Figure 1.5 gives a hint. **L*** is the lightness, while **a*** is (approximately) corresponding to the greenness/redness and **b*** is (also approximately) the yellowness/blueness.

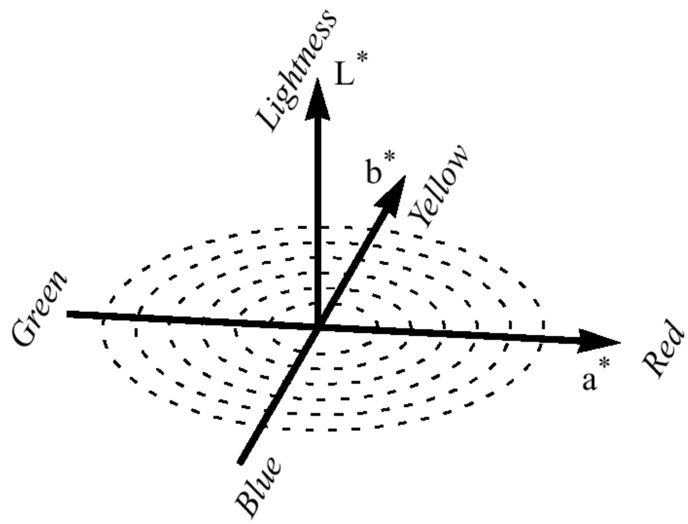


Fig 1.5: *The CIELAB color space.*

1.4 Hue and Saturation based spaces

The color spaces described earlier, although practical to use in technical applications, are not very similar to the human brain's way of classifying colors. Rather, terms like hue (H), saturation (S), lightness (L), brightness (B), intensity (I), or value (V), becomes interesting for these purposes. These characteristics could be grouped three by three to form "humanistic color spaces". Usual combinations are HSB, HSL, HSI, and HSV. The definitions of the important properties are as follows:

- **Hue**—the attribute of a visual sensation whether an area appears to be similar to one of the perceived colors red, yellow, green, and blue, or a combination of two of them. The hue is roughly the color of the dominant wavelength in the SPD.
- **Saturation**—the colorfulness of an area, judged in proportion to its brightness. The more an SPD is concentrated to one wavelength, the more saturation will be associated with the color. A color can be desaturated by adding light that contains power at all wavelengths.
- **Lightness**—the non-linear perceptual response to brightness or luminance (the subjective and objective measure of the same thing), defined as L* in the previous section.
- **Brightness**, **intensity**, or **value**—alternative technical definitions of lightness.

1.5 Neugebauer's equations

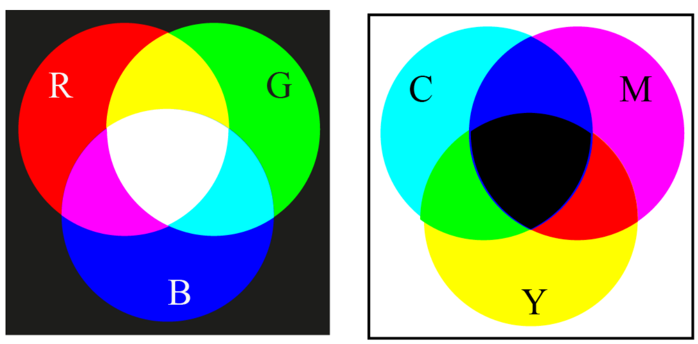
When color tones are to be reproduced, usually a set of basis colors are used. For example, the picture on a TV screen is formed by dots of the basis colors red, green, and blue. When all the three colors are present, white color is perceived, and when none of them is present, black is perceived. This is called additive color mixing and is illustrated in Fig. 1.6(a). The spectral distributions of the lights are added:

$$I_{tot}(\lambda) = I_1(\lambda) + I_2(\lambda) \tag{1.14}$$

Unlike the TV, a piece of paper does not add energy to the illuminating light. The printed inks can be seen as filters, which absorb some of the incoming light before it is reflected (see Eq. 1.15). The more colors printed on the same spot, the more filters will subtract energy at certain wavelengths from the illumination. This is called subtractive color mixing and is illustrated in Fig. 1.6(b), although the term multiplicative color mixing would be more mathematically correct:

$$I_{tot}(\lambda) = R_1(\lambda) \cdot R_2(\lambda) \cdot I(\lambda)\ , \tag{1.15}$$

where R_1 and R_2 are the reflectance functions of the two colors. In subtractive color applications, the basis colors cyan, magenta, and yellow, are usually used. When all the three colors are present, most of the light is absorbed and black is perceived.



a b

Fig 1.6: *(a) Additive color mixing. (b) Subtractive color mixing.*

When printing color images, four colors are usually used. The three basis colors above (CMY), and pure black ink (K). The reason for using black, despite the fact that it can be produced by the other three colors, is practical. There are three main reasons:

- It is cheaper to print large black areas with black ink than with cyan, magenta, and yellow ink. Also, one layer of ink dries faster than three, and smears less in the printing process.
- There is a large risk that small black letters etc. will be a bit blurred if printed in three colors, due to misregistration between the printing plates corresponding to the printing colors.
- Introducing the fourth color, the shadowed areas of images can be printed with more color depth, and a richer blackness. The color gamut is expanded.

Neugebauer's equations deal with an alternative additive color mixing, which arises when several small different colored areas are averaged together by the human eye. The simplest form of the equations is:

$$\begin{bmatrix} X_{tot} \\ Y_{tot} \\ Z_{tot} \end{bmatrix} = \sum_i a_i \begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix}, \tag{1.16}$$

where a_i is the fractional area covered with the color (X_i, Y_i, Z_i) and thus:

$$\sum_i a_i = 1 \tag{1.17}$$

When printing with four inks as described above it is possible to create 16 different colors (of which 9 are black), so the sum in Eq. 1.16 will have 16 terms. Neugebauer's equations may also be used in other applications where there are many small different colored areas next to each other.

1.6 Applications

One interesting application of the color science is color halftoning. That is, how to approximate a continuous tone image with one or several binary images (one for each ink color). When two different colored inks are printed on top of each other, there is a typical subtractive (multiplicative) color mixing. Let $R_p(\lambda)$ be the paper's reflectance function. Let $C(\lambda)$, $M(\lambda)$, $Y(\lambda)$, and $K(\lambda)$ be the absorption filters of the four inks, and let $I(\lambda)$ be the illuminant of the total area. Then the blue color (cyan and magenta on top of each other) will have the spectrum:

$$I_{blue}(\lambda) = R_P(\lambda) \cdot C(\lambda) \cdot M(\lambda) \cdot I(\lambda) \tag{1.18}$$

This should be interpreted as that both the paper and the cyan and magenta inks remove energy of certain wavelengths from the incoming light. By using this in the tristimulus equations, **X**, **Y**, and **Z** values can be calculated for the blue color. The same can be done for all the 16 colors that are possible when printing with four inks. They may then be used in the Neugebauer equations, defined in the previous section.

We may know the fractional coverage for each of the printing colors separately, but how do we know for example the fractional area covered with both cyan and magenta inks? Won't it differ a lot if one of the printing colors is translated a bit (misregistration)? Yes, it would if all printing colors had the same screen angles. Then only a small translation of one of the printing plates would yield a large color shift. **This is one of the main reasons why the printing industry does not print all inks with the same screen angle** (see Fig. 1.7). If different angles are used, a resistance from color shifts due to misregistration is achieved—approximately the same fractional coverages remain despite translation.

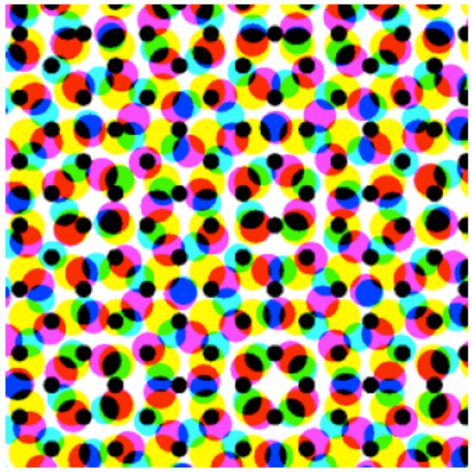


Fig 1.7: *A color halftone with 30% cyan, 40% magenta, 50% yellow, and 10% black.*

The approximation of the fractional coverage is calculated in the following way, assuming a semi-stochastic overlap behavior, and known as the Demichel equations:
Let **c**, **m**, **y**, and **k** be the fractional areas covered with cyan, magenta, yellow and black ink, respectively. The fractional area not covered with each particular ink is 1-**x**, where **x** represents one of the four letters above. Then the fractional area covered with cyan and magenta, but not yellow or black (yielding blue) is:

$$a_{cm} = cm(1-y)(1-k) \tag{1.19}$$

In the same manner, for example the area solely covered by yellow is:

$$a_Y = (1-c)(1-m)y(1-k) \tag{1.20}$$

In this way 16 different fractional areas can be calculated for the current CMYK color. These values can then be used in the Neugebauer's equations, to calculate the color perceived by the human eye. The accuracy of the values calculated above depends on halftone characteristics, such as dot shape and screen angles.

1.7 Light sources

Figure 1.8 shows the spectra for a number of light sources, of which *CIED65*, *plank90k* and *Tungsten60W* are being used in this lab. As seen in this figure, for example the tungsten light source has the most concentration in the longer and medium wavelength, meaning a yellow-brownish color, while plank90k is bluish and CIED65 color neutral (representing the day-light).

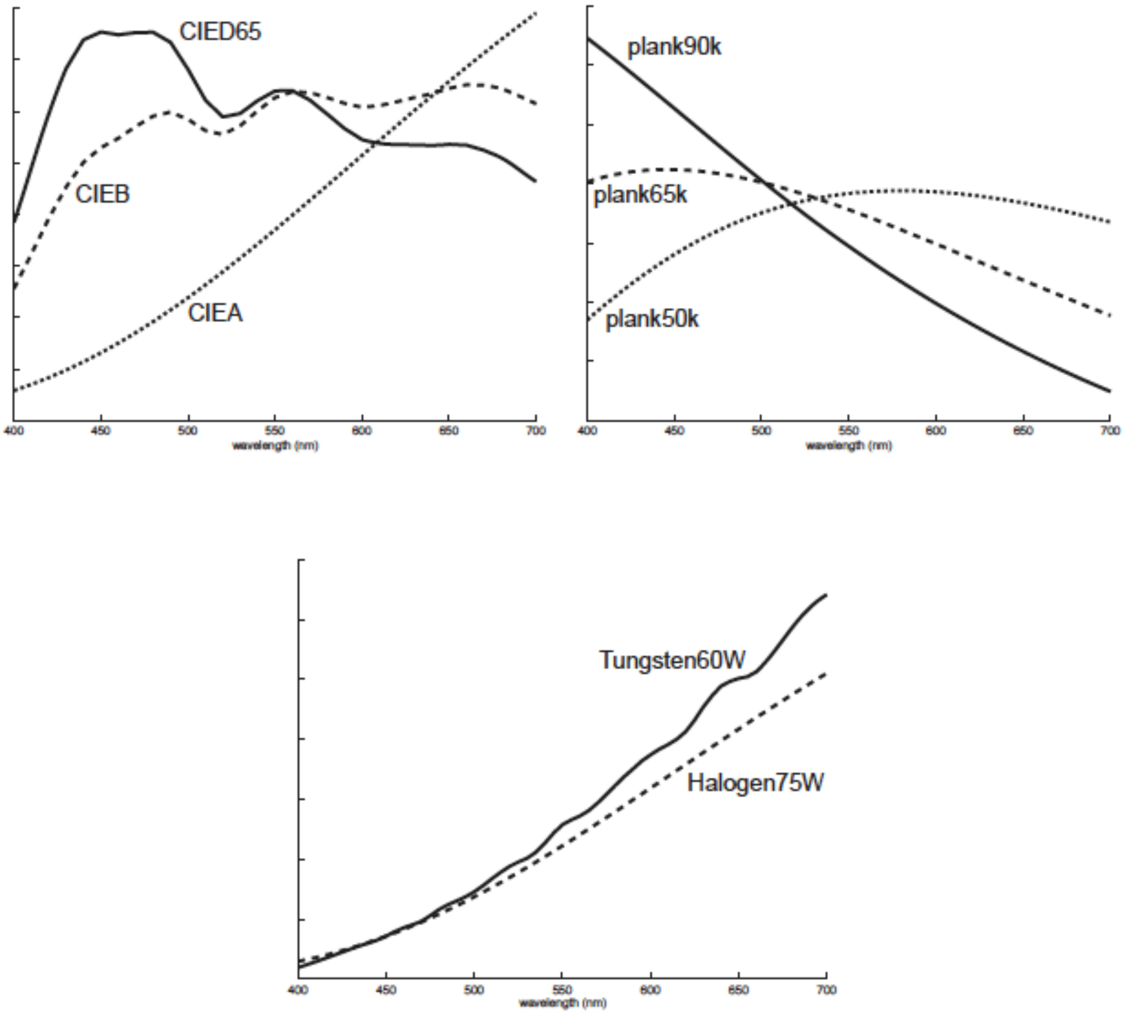


Fig 1.8: *The spectra for a different light sources*.

Glossary

- **Color gamut:** Färgomfång.
- **Cones:** Tappar (på näthinnan).
- **Continuous tone image:** Bild med kontinuerligt varierande (så när som på fin kvantisering) färg eller gråskala.
- **Distribution:** Fördelning.
- **Halftoning:** Rastrering.
- **Misregistration:** De olika tryckfärgernas tryckplåtar råkar hamna snett i förhållande till varandra under tryckprocessen. Kallas även för misspass.
- **Retina:** Näthinnan.
- **Rods:** Stavar (på näthinnan).

Jörgen Rydenius (1997), Martin Solli (2008), Modified by Sasan Gooran (2017)