Understanding COLOUR

by Mike Bedford

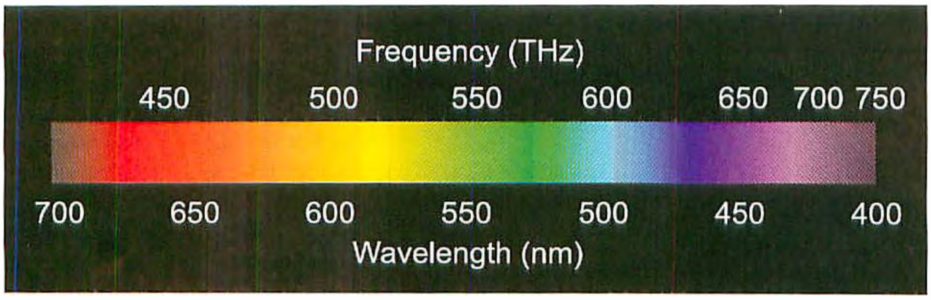

Figure 1.

Although it's a concept which most people think they understand, few people really have an in-depth appreciation of colour. Yet an understanding of certain elements of colour theory is essential in various areas. Traditionally it was artists, designers and printers who needed to appreciate colour. Increasingly, though, and of particular relevance to readers of Electronics & Beyond, it's in graphics programming and opto-electronic design where such knowledge is needed. Even power users of graphics and design packages will need this sort of appreciation of colour theory. My aim here is to provide a grounding in basic colour theory which can be applied in these application areas.

The Basics

Light is that portion of the electromagnetic spectrum which is visible to the human eye and is normally considered to lie in the approximate wavelength range 400nm to 700nm as shown in Figure 1. Normally, this, rather than frequency (actually in the range 430THz to 750THz), is quoted. The visible part of the spectrum is bounded by infra-red at the low-frequency end and ultra-violet at the high frequency end. Our eyes contains two kinds of receptors which are called rods and cones. The rods allows us to perceive shades of grey whereas the cones allow the brain to perceive different colours. There are actually three types of cones which are sensitive, respectively, to red-orange, green, and blue-violet light. When a single cone is stimulated, the brain perceives the colour associated with that cone. So, for example, if the eye receives light at a wavelength of 550nm, the green cones are stimulated and we see green. Clearly this is a simplification since we can actually perceive many millions of colours, not just the three which correspond to the different types of cone.

Additive Mixing

At this point, since it's a key to many of the methods of defining colour, let's think about colour mixing. By colour mixing, I'm talking about the sort of thing we were taught in school physics lessons, as opposed the type of mixing we learned in art lessons – we'll come onto this a bit later. In Figure 2 we see what happens if we project circles of red, green and blue light onto a screen in a darkened room. Where the red and green overlap we get yellow, the overlap between red and blue gives magenta, that between green and blue gives cyan, and in the central portion, where red, green and blue all overlap, we get white. In fact, by varying the amounts of red, green and blue we can produce any colour the eye can perceive so these three colours are referred to as primary colours. Similarly, the three colours cyan, magenta and yellow are called secondary colours. To be more precise, and to differentiate them from the subtractive primaries and the subtractive secondaries which we'll see shortly, we should really refer to red, green and blue as the additive primaries and cyan, magenta and yellow as the additive secondaries. Similarly, the type of mixing we've been looking at is called additive mixing.

Since you have a grounding in electronics, many of you are probably now considering why red plus green gives yellow and so forth. And that knowledge of electronics may well cause you to think in terms of frequency mixing like that employed, for example, in a superhetrodyne radio receiver. But a few quick calculations shows that this can't be the explanation. The mixing products of signals with frequencies f1 and f2 are f1+f2 and f1-f2. Since the frequencies of red and green light are, say, 462THz (650nm) and 577THz (520nm), if any appreciable frequency mixing were to take place, the mixing products would be at 1039THz and 115THz, neither of which are in the visible portion of the spectrum let alone the frequency which corresponds to yellow. In fact, mixing red and green to give yellow is purely an artefact of the way the brain processes the information – an optical illusion if you like. Although we can perceive an image with different colours in different parts of that image, we can only differentiate a single colour at a single point. So, when the brain is presented with signals indicating multiple colours at a particular point, it perceives this as a single different colour. In other words, the brain cannot differentiate a mixture of red and green light from yellow light. And in the case of mixing red, green, and blue light, this is clearly the case since, spectrally, there's no such thing as "white", it's all in our brains. Having said that the eye cannot differentiate yellow light from a mixture of red and green light there are exceptions. In the early days of high brightness LEDs, before the widespread availability of white LEDs, people experimented with a mixture of red, green and blue LEDs to produce white light for high reliability torches. Although the light appeared to be white, in much the same way as sunlight, the effect was very different when illuminating primary coloured objects. Since the light contained

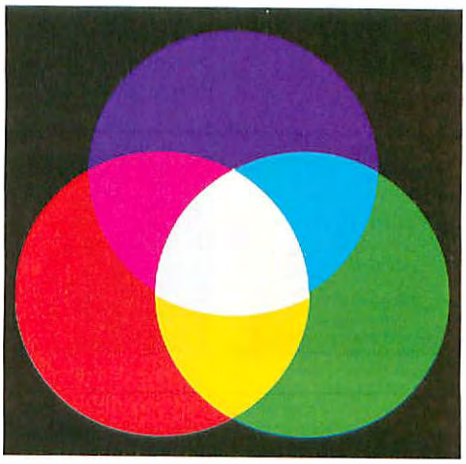

Figure 2.

just three spectral bands, any object of a *similar colour* to one of these bands would appear to glow in the light of the torch.

Subtractive Mixing

In the previous section I referred to another type of mixing, that which we were taught in art lessons – this is summed up in Figure 3. Here we see the effect of mixing three colours of paint – cyan, magenta and yellow (art teachers tend to talk about blue, red and green but this isn't strictly correct) – on a piece of paper. Now we see that mixing cyan and magenta gives blue, cyan and yellow gives green, magenta and yellow gives red, and that mixing all three gives black. The explanation lies in an understanding of how paints and pigments work. Clearly pigments don't actually emit light since they don't glow in the dark. Instead, they selectively reflect light by absorbing their complementary colour. Cyan paint, for example, absorbs red light so, when illuminated with white light, allows green and blue light to be reflected. And, as we already know, adding together green and blue light gives cyan. Cyan paint can, therefore, be thought of as minus red. Similarly, magenta is minus green and yellow is minus blue.

We can now return to mixing paints or inks and see that it makes perfect sense. Since cyan paint absorbs red and magenta paint absorbs green, when the two are mixed together we end up with something which absorbs both red and green. When the mixture is illuminated with white light, therefore, only blue is reflected. Because this form of mixing works by subtracting colours from the light illuminating it, the process is called subtractive mixing. Cyan, magenta and yellow are the subtractive primaries and red, green and blue are the subtractive secondaries.

Figure 3.

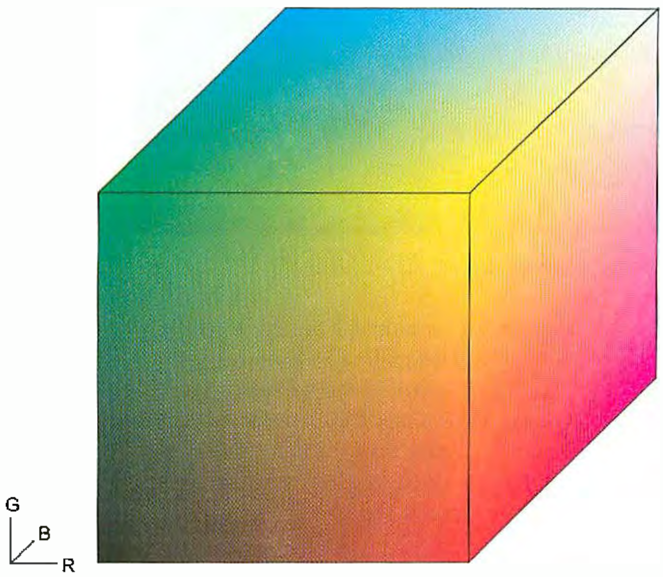

Figure 4.

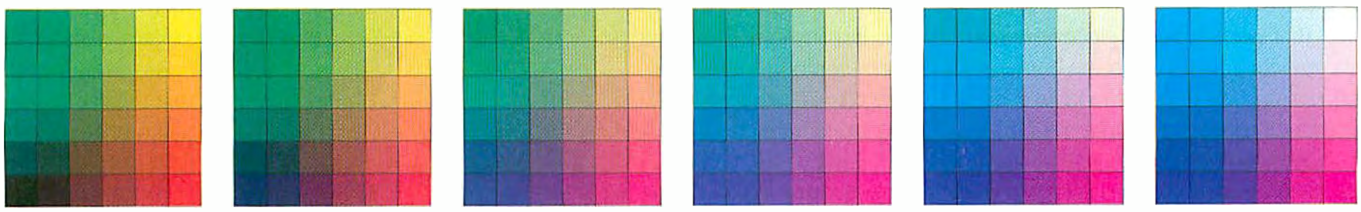

Figure 5.

Defining Colour

Officially, the colours of the rainbow are red, orange, yellow, green, blue, indigo and violet as in the well-know aide memoire "Richard of York gained battles in vain". This is interesting since it suggests that the rainbow contains just seven discrete colours whereas in reality, of course, it contains an infinite number of which many thousands can be differentiated by the human eye. And when there are so many colours, coming up with names for them all can be a daunting task. There are, of course, many names of colours in addition to the rainbow colours – cerise, turquoise, aquamarine, lemon, crimson, mauve, purple and scarlet for example – and many colours can take descriptive adjectives so we have lime green, bottle green, leaf green and so on. But these an imprecise and too few to name all the available colours. In the printing industry, books of standard colours are produced by companies specialising in colour matching. A popular colour matching scheme is Pantone – customers pick colours for logos etc. from a Pantone booklet and specify these colours to the printer using the corresponding Pantone number. This way, irrespective of which printing company is being used, the colours should be identical. Many PC graphics packages also allow Pantone colours to be selected. But although there's a clear benefit of using a colour matching method like this, the numbers assigned to the colours (e.g. Pantone 5415 CV) appear to be largely arbitrary. No doubt there is some sort of rationale to the numbers but this isn't something Pantone feel the customer needs to be concerned with.

In the early days of computer graphics, none of this was a problem. Since the first colour displays had just eight colours, the three additive primaries: red, green and blue, the three additive secondaries: cyan, magenta and yellow, plus black and white, using names was perfectly adequate. And even when the number of colours increased, perhaps, to 64, names were still used by some companies. So, the progression from yellow to green might be yellow, greenish yellow, yellow-green, yellowish green, and green and these would indicate 100% yellow, 75% yellow / 25% green, 50% yellow / 50% green, 25% yellow / 75% green, and 100% green respectively. Eventually we come to a point, though, at which attempting to use names just isn't feasible and we have to turn to numbers. Not only does using some mathematical definition of colour solve the problem of colour definition, but the numbers can also be manipulated directly by software.

RGB and CMY

One of the most common ways of defining colour is by the proportion of each of the three additive primary colours red, green, and blue. This is called the RGB method. The proportion is normally defined either as a value in the range 0-1, a percentage or as a number in the range 0-255. This latter method is appropriate for typical colour displays which are capable of showing 16.7 million colours since this is the number of

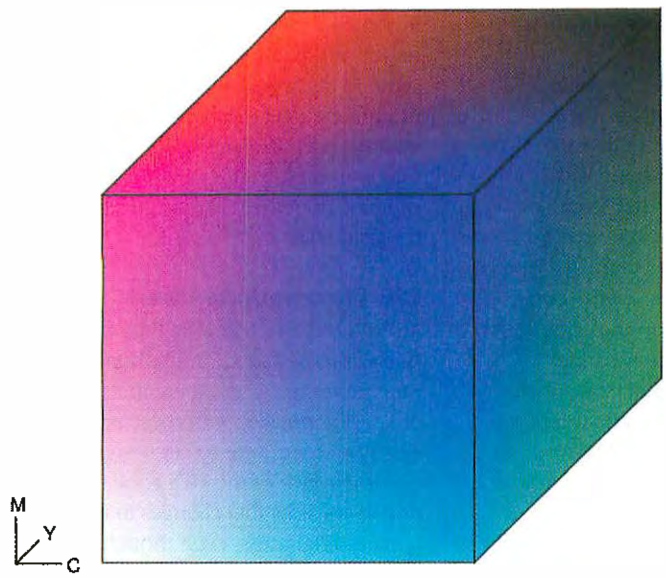

Figure 6.

possible combinations of 265 levels of red, green and blue. The RGB colours are often visualised as a cube – the RGB colour cube – which is shown in Figure 4. Here the amount of red, green and blue increases along the x, y and z-axes, respectively, starting at the bottom left hand corner. Sometimes this cube is shown differently, perhaps with the black corner at the back or the top right but the basic principle is the same. All possible colours can be found somewhere inside this cube. However, since this diagram only allows you to see the colours on three of the faces, and even if you could rotate it you'd still only see the colours on the other three faces, we show a number of slices through an RGB colour cube as Figure 5. Specifically the cube has been sliced from front to back as viewed in Figure 4. Note also that this particular version of the colour cube is far from the idealised one which contains all possible colours. Instead of the infinite number of shades in the theoretical colour cube, this one has six levels for each of red, green, and blue so includes just 216 colours. All practical imaging systems will be capable of displaying a limited number of colours.

Since they operate by additive mixing, RGB is a useful way of thinking about colour on a display screen, but less so colour printing which operates by subtractive mixing. A close relative to the RGB colour cube is the CMY colour cube. Now, instead of the x, y and z-axes representing red, green and blue, they represent the subtractive primaries cyan, magenta and yellow. The CMY colour cube is shown as Figure 6 and you'll notice that it is, in fact, an RGB colour cube with each point reflected about the centre of the cube. You won't be surprised to learn, therefore, that colours can easily be converted between the two colour models using the following simple relationships:

$$C = 1.0 - R$$
$$M = 1.0 - G$$
$$Y = 1.0 - B$$

Where C, M, Y, R, G and B are values of cyan, magenta, yellow, red, green and blue, respectively, in the range 0-1. If working with values in the range 0-100% or 0-255 rather than 0-1, each of the constants 1 in the formulae should be changed to 100 or 255, respectively.

CMY is important in the printing industry since inks combine according to the subtractive mixing model. Traditionally, a colour original would be split up into the three subtractive primaries by photographing it through coloured filters. Plates are then made for each of the primaries and these are used, one at a time with the appropriately coloured ink. The full colour image is, therefore, built up progressively. Today, the process of splitting a full colour image into the subtractive primaries (which is called colour separation) is normally done by software, especially if the artwork is being mastered

Figure 7.

digitally. In fact, I've over-simplified things by suggesting that colour printing is based on the CMY model; in fact full colour reproduction is referred to as four-colour printing and conforms to the CMYK model where K stands for blacK (because B would be confused with Blue in RGB). Although perfect cyan, magenta and yellow inks will overprint to produce black, most printing inks aren't perfect primaries so blacks would appear as muddy browns. However, by extracting the black content and using this information with a pure black ink, a better full colour image results. The conversion of a colour defined in CMY to CMYK (which should really be C'M'Y'K to differentiate the different cyan, magenta and yellow contents in the two models) is given by the following equations:

$$K = \text{minimum}(C, M, Y)$$
$$C' = C - K$$
$$M' = M - K$$
$$Y' = Y - K$$

Figure 7 is a full colour original and Figures 8a, b, c and d are the C, M, Y and K contents separated digitally. Then, in the parts of Figure 9 we see the image being built up as it would be in lithographic printing. Note that the printing normally takes place in order of increasing density, that is yellow, magenta, cyan and black.

HLS and HSB

Although the RGB and CMY colour models are, perhaps, the most obvious ways of

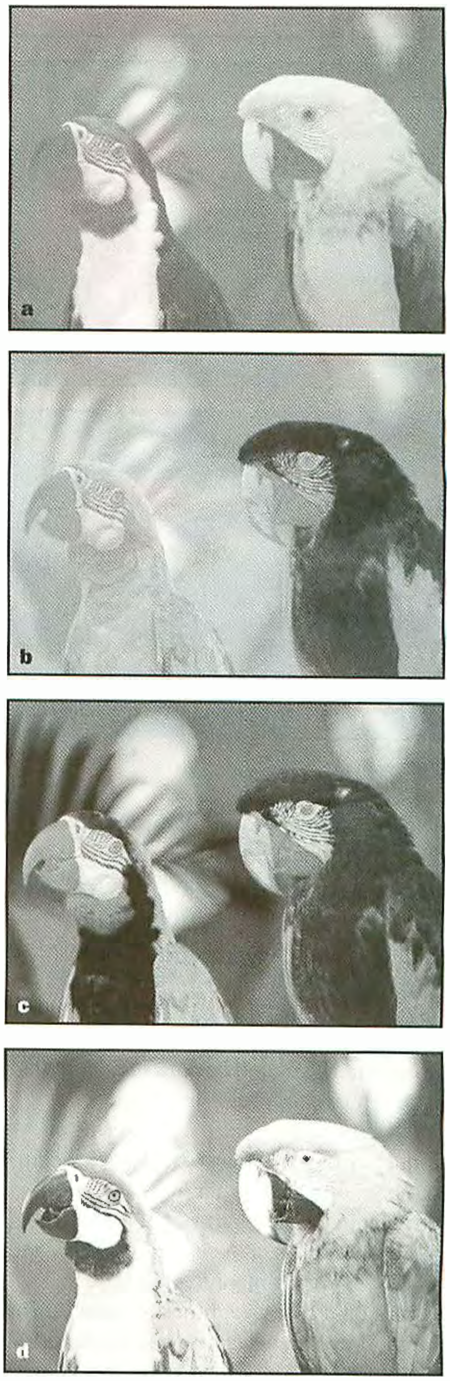

Figure 8.

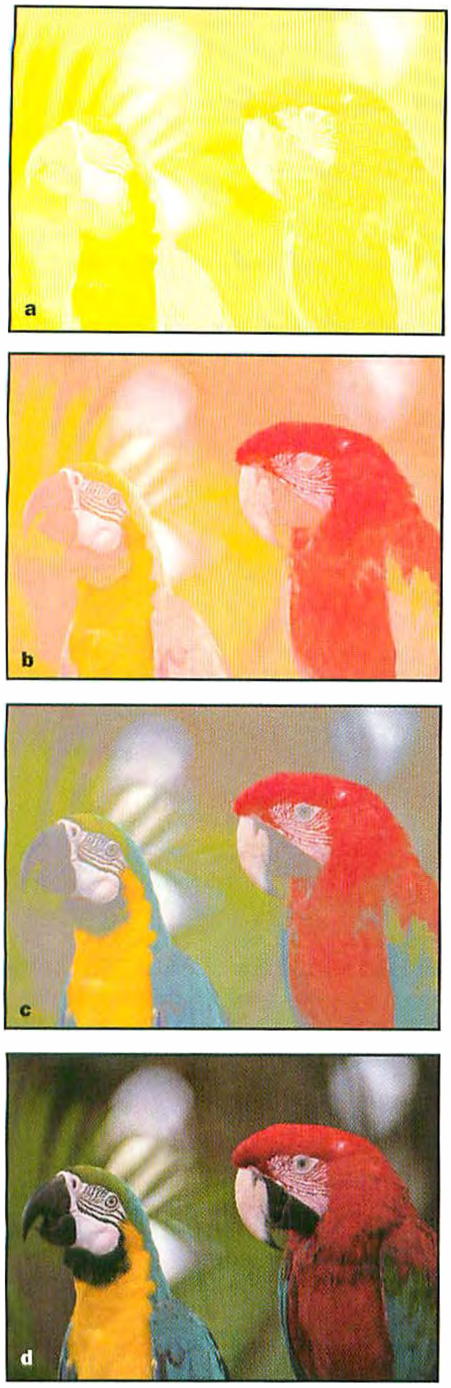

Figure 9.

defining colour in terms of mixing, they don't really conform to our perception of colour. The HLS model, on the other hand, uses concepts, if not the words themselves, which we can more easily relate to. H stands for hue and is best described as "colour", by which I mean the continuum from red through orange, yellow, green, cyan, blue, violet and back to red. Note that although red and violet are at opposite ends of the visible spectrum in terms of frequency or wavelength, the two do seem to merge into each other just as red does into orange or blue into violet. L stands for lightness and has, as its extremes, black and white. You can think of the brightness control on a TV as a lightness control. Turn it right down and everything is black, turn it right up and everything would be white if it wasn't for the fact that TV manufacturers place a limit before this point. S stands for saturation and is the property which is influenced by the colour control on a TV. Turn it right down and everything is a shade of grey, turn it right up and the colours are all vibrant. Hue takes values of 0°-360°, whereas lightness and saturation both take values of 0-1, 0-100% or 0-255. HLS colour space is represented as a double cone as shown in Figure 10. Hue is the angle around the spine of the cone, lightness is the distance up the cone with 0 at the bottom and 1 at the top, and saturation is the radial distance from the spine. At first you might find it surprising that it's this particular shape. However, as you study it and as you note, for example, that a lightness of 0 is always black irrespective of hue and saturation and that a lightness of 1 is always white irrespective of hue and saturation, it should make more sense. As with the RGB colour cube, we've also shown a few slices through the HLS colour cone in Figure 11 so that you can see some of the internal colours. This is particularly important here since you can only ever see the colours with a saturation value of 1 on the outer surface of the cone. The following formulae shows how HLS values (0-1) can be obtained from RGB values (0-1):

IF minimum colour is B then

$$H = 120 * \frac{G - B}{R + G - 2 * B}$$

ELSE IF minimum colour is R then

$$H = 120 * \left(1 + \frac{B - R}{B + G - 2 * R}\right)$$

ELSE

$$H = 120 * \left(200 + \frac{R - G}{R + B - 2 * G}\right)$$

ENDIF

$$L = \frac{MAX(R, G, B) + MIN(R, G, B)}{2}$$

IF L < 0.5 THEN

$$S = \frac{MAX(R, G, B) + MIN(R, G, B)}{MAX(R, G, B) - MIN(R, G, B)}$$

ELSE

$$S = \frac{MAX(R, G, B) - MIN(R, G, B)}{2 - MAX(R, G, B) - MIN(R, G, B)}$$

ENDIF

Another colour model, which obviously has quite a bit in common with HLS, is the HSB model which is represented as a single cone as shown in Figure 12. Here, H stands for hue as in HLS, similarly S stands for saturation, and B stands for brightness. Occasionally, this colour space is called HSV, in which case V stands for value but is defined just the same way as brightness. Hue is defined in exactly the same way as in the HLS model, but saturation is defined rather differently and brightness is not the same as lightness. In this model, the 100% brightness level is achieved if any of the additive primaries are at 100% so all the fully saturated colours are at a brightness value of 1.0 rather than a lightness value of 0.5 as in HLS.

CIE Chromaticity Chart

Although the HLS and HSB colour models purport to be based on the human perception of colour, there are some very obvious anomalies with these schemes. For example, if we were to circumnavigate the circumference of the HLS cone, we'd notice considerable changes in the apparent lightness, even though all are colours we'd encounter have lightness values of 0.5. In particular, the colours in the yellow region appear very much lighter than those in the blue region. So although there's a clear theoretical basis to the colour models we've seen so far, we really need to adopt something else if we're genuinely interested in the human perception of colour.

A colour model which was devised back in 1931 but is still in widespread use is the CIE colour space which includes all the colours which can be perceived by the human visual system. It is based on the spectral power distribution of the light emitted from a coloured object and then calibrated by the sensitivity curves for the human eye. Effectively, colours are represented by the proportion of three primary colours, referred to as X, Y and Z, which correspond to the sensitivities of the three types of cone sensors in the human

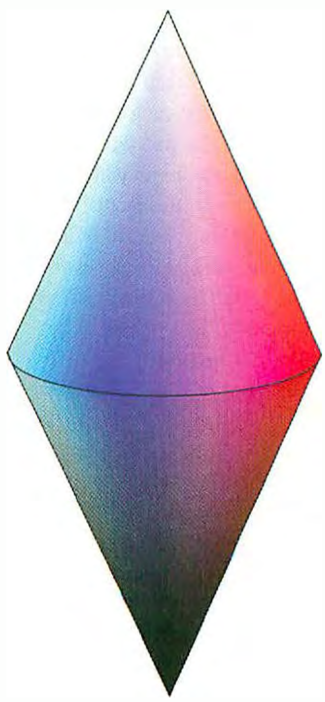

Figure 10.

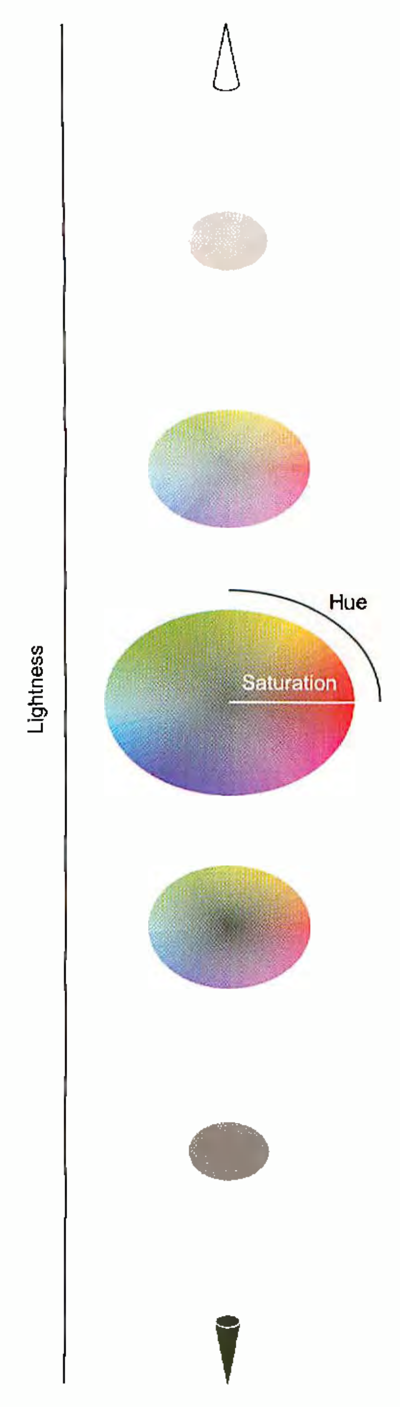

Figure 11.

eye. Furthermore, Y represents the perceived brightness and is referred to as luminance. This differs from the lightness in the HLS and HSB models, or the intensity of the three primaries in the RGB or CMY models, each of which are proportional to the physical power level measured in a unit such as watts per square metre. Figure 13 shows the spectral curves for the three CIE primaries. You'll notice that the blue curve is taller than the others because the eye is less sensitive to blue light. Rather than showing the full colour space, this model is normally shown as the CIE chromaticity diagram which appears as Figure 14. This is a slice through the colour space which has equal luminance – specifically the slice for X+Y+Z = 1. In other words, the X and Y primaries are normalised according to the following formulae to give the co-ordinates x and y which are shown on the diagram.

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

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

In fact, it turns out that all hue and saturation combinations can be represented in terms of x and y alone. Y can, therefore, be used to represent luminance alone so giving the Yxy colour space. It's interesting to note that the colours round the curved edge of the CIE chromaticity diagram are pure colours, that is spectral colours. The corresponding wavelengths are indicated on the diagram. However, the straight line section of the perimeter, the so-called purple line, includes those shades of magenta or purple for which there is no equivalent monochromatic light.

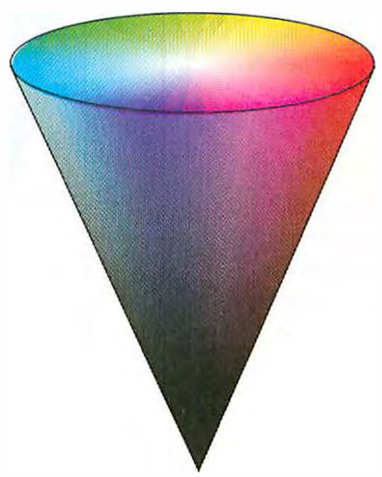
Figure 12.

Figure 15 is another version of the CIE diagram onto which the colours of red, green and blue LEDs and of the red, green and blue phosphors on a CRT are indicated. In each case, the triangles bounded by the three "primary" colours contain all the colours which can be produced by mixing the three colours. The colours within the triangle are referred to as the gamut of a LED-based display or of a CRT. The LED gamut is the smallest and quite clearly there is a significant number of colours which cannot be mixed from red, green and blue LEDs – basically because they're not perfect primaries. Most notably, the traditional green LED is a very yellowish green. You'll notice, though, that a so-called "pure green" LED is also shown. This is a newer type of LED which uses a similar semiconductor to that which made blue LEDs possible and is much closer to a true primary green. Now the range of colours is much improved, even over the CRT. Also shown on this diagram are the gamuts of four-colour printing and photographic film. You'll notice that the gamut of four-colour (CMYK) printing is particularly limited. For this reason, for high quality work, additional colours are sometimes used in addition to cyan, magenta, yellow and black – the so-called process colours. One common approach is to add a further two "primary" colours, specifically an orange and a green to give six-colour printing. Another option is to add one or more "spot" colours. A spot colour isn't used in mixing, instead it allows a single colour, which cannot faithfully be mixed from the primaries, to be printed. This might be used, for example, in a glossy brochure showing a red car. If four-colour printing wouldn't do justice to the red of the car, the exact shade would be printed as a spot colour after the four process colours.

YUV Colour Space

So far we've looked at five colour models in some detail and there are dozens of others. Many are effectively obsolete, though, or applicable only to very specialised areas so we're not going to go through them all here. However, there is one other we should look

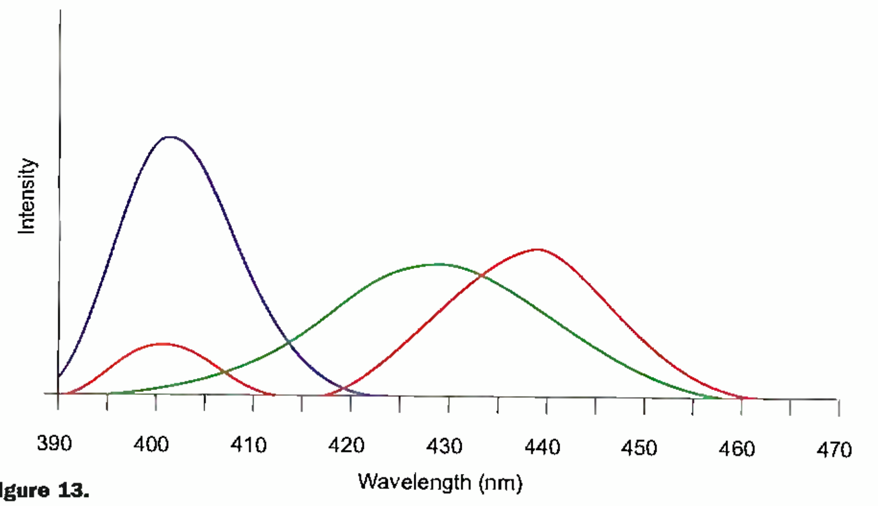

Figure 13.

at because of its importance in colour TV and video. This is the YUV colour model.

The YUV colour model separates luminance from colour. I'll explain later why this is useful for broadcast TV and in MPEG encoding. Y represents the luminance but, despite numerous reports to the contrary, it is not the same as Y, the luminance in the CIE Yxy. In fact, the YUV model was originally referred to as the Y'UV model in order to make the distinction but the prime has now been dropped. Y in YUV is a gamma corrected luminance value and is sometimes referred to as luma in order to reduce confusion. Gamma correction involves applying a non-linear transformation to the luminance value to take into account the non-linear relationship between the voltage applied to a CRT and the perceived lightness which results. Failure to apply this gamma correction before transmitting a video signal results in the inefficient use of bandwidth. Luma is determined by gamma correcting each of the additive primary colours, R, G and B, and then combining them with different weightings which correspond to the sensitivity of human vision to each of the additive primaries standardised for video. This is shown in the following formulae:

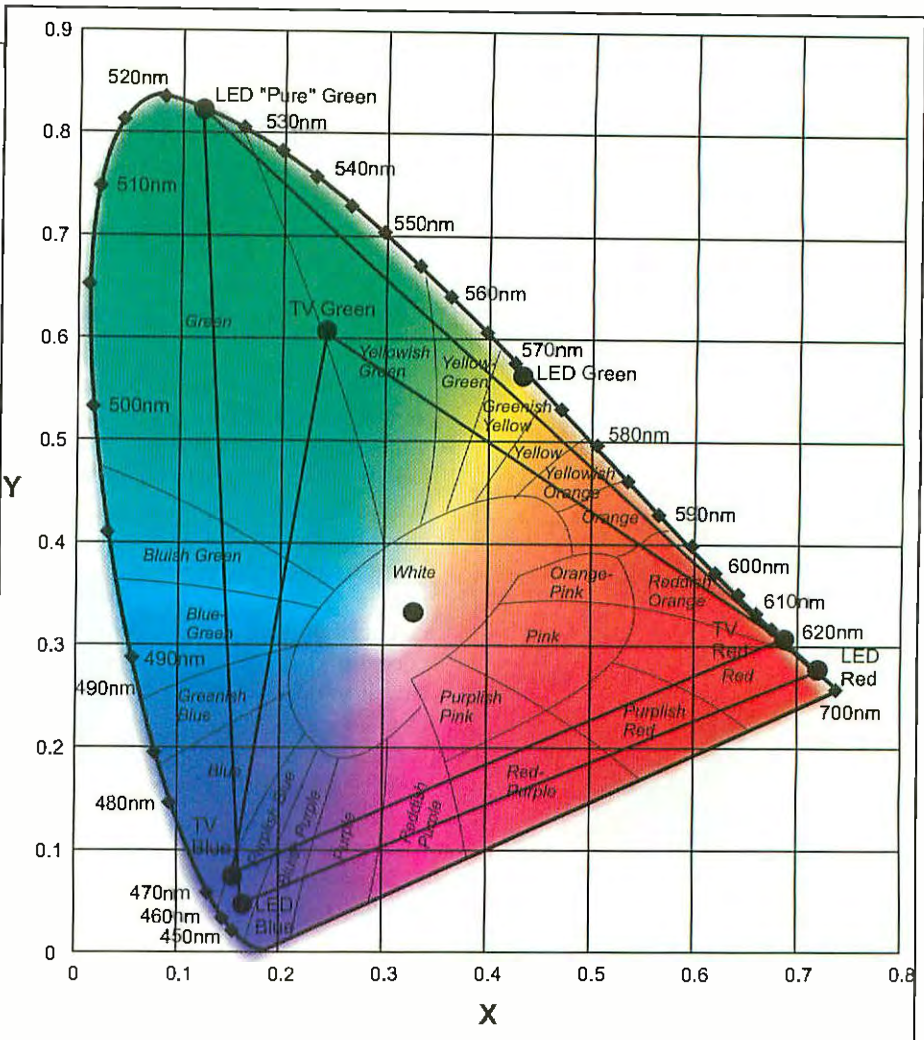

Figure 14.

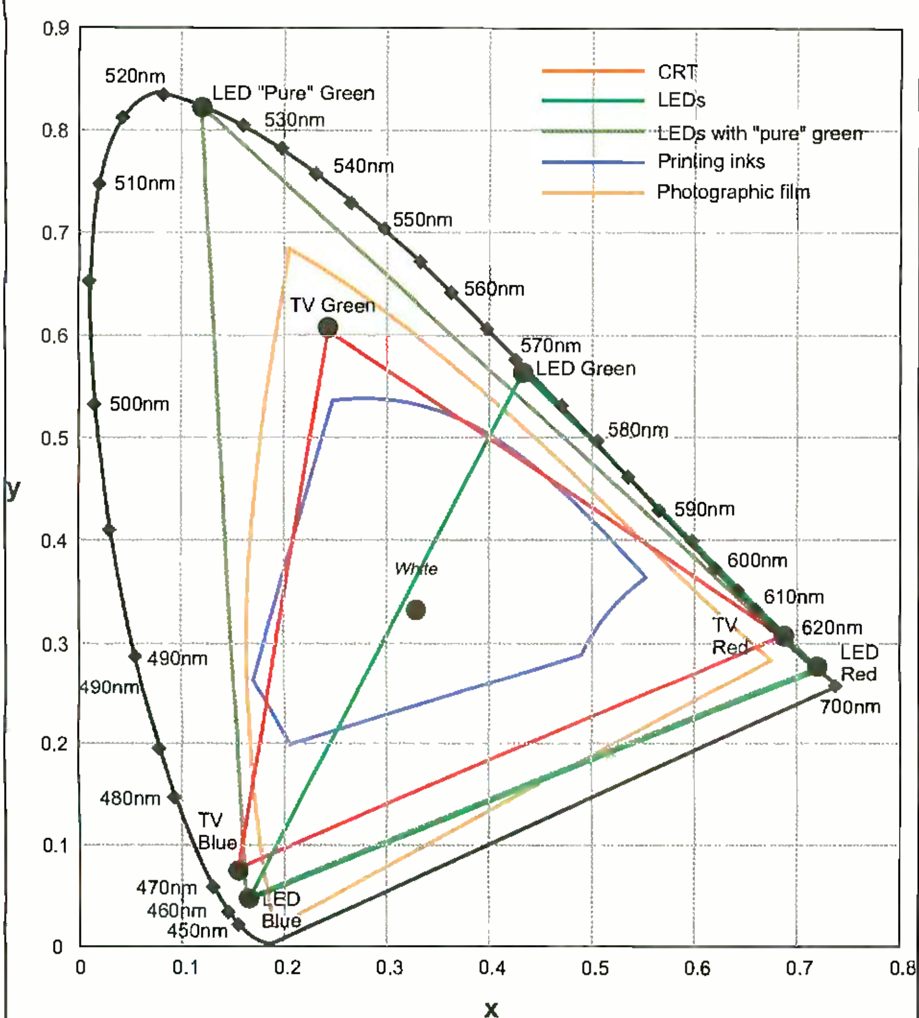

Figure 15.

$$R_{gamma} = 1.099 \times R^{0.45} - 0.099$$

$$G_{gamma} = 1.099 \times G^{0.45} - 0.099$$

$$B_{gamma} = 1.099 \times B^{0.45} - 0.099$$

$$Y' = 0.288\, R_{gamma} + 0.587\, G_{gamma} + 0.114\, B_{gamma}$$

The U and V values, which are, in fact, colour difference signals (i.e. the difference in intensity between two of the primaries and the luma value) are now calculated according to the following formulae:

$$U = 0.493 * (B - Y')$$

$$V = 0.877 * (R - Y')$$

If the signal is being transmitted as component video, such as from a DVD player to a TV on separate cables, no more encoding is necessary. However, for analogue broadcast TV the U and V values are combined to produce a chroma signal, C, by quadrate modulation. The reason for the strange looking coefficients in the above equations is so that the composite

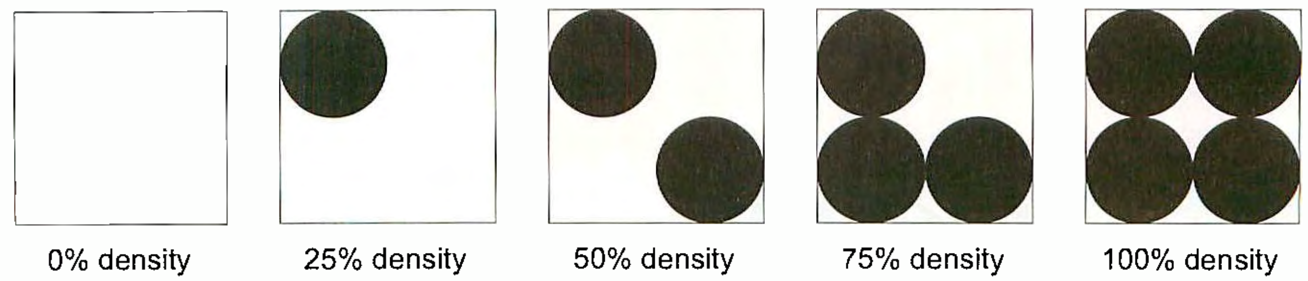

Figure 16.

PAL signal (luma plus modulated chroma) is contained within the range -1/3 to +4/3. These limits reflect the capability of the composite video recording or transmission channel.

And finally, before leaving the topic of the YUV colour model I need to explain the rationale for separating the colour information from the luma. The eye is more sensitive to detail in colour than in luminance. In PAL TV, therefore, the U and V values are sampled at half the horizontal frequency as the luma. This provides a useful decrease in the bandwidth. A similar technique is employed in video compression, in the MPEG-2 standard, for example.

Practical Colour Mixing

We started off by looking at colour mixing and this is also where we'll conclude. For although it's possible to continuously alter the amount of the additive primaries red, green and blue on a CRT tube by controlling the voltage on the tube, this is not possible on paper. In most forms of printing a dot of ink is either present or it isn't. So the only way of altering the perception of darkness is to alter the size of the dot and/or its separation from its neighbours. Since lithographic printing really doesn't have much of a bearing on electronics and computing, I won't describe the solution adopted here. However, this is something which is relevant in desktop publishing so you may like to read up elsewhere on the subject. You should look for a book on printing or DTP, or search the Web for information on half-toning or screening. In most colour printers intended for use with PCs (e.g. inkjet printers), though, a technique called dithering is used. If a printer has a 1200dpi x 1200dpi resolution it can deposit dots of ink on a 1/1200th of an inch grid. But since it only has four colours of ink (cyan, magenta, yellow and black) and since it can overprint to produce another three (red, green, and blue) it can only print at this resolution in seven colours plus the white of the paper. Fortunately, for a photographic image, the eye generally won't notice if the actual resolution is much less than 1200dpi. So a block of printer dots can be used to represent one pixel in the image. Figure 16 shows how this can be useful and, for simplicity, I've used a small 2 x 2 dot block to represent a pixel. Since this contains four dots, 0, 1, 2, 3 or 4 dots can be printed thereby giving an overall density of 0%, 25%, 50%, 75% or 100%. And at a normal viewing distance, the individual dots will not be resolved so the eye will just see the shade of grey represented by the density. By increasing the size of the block, the number of shades of grey is increased at the expense of resolution. But this just relates to greyscale printing. Since we can do the same for each of the subtractive primaries, a 2 x 2 dot block actually allows 125 colours to be printed, 3 x 3 dots allows 1,000 colours and with 4 x 4 dots this rises to almost 5,000. In Figure 17 a few colours mixed by dithering are shown. The scale is large enough to allow you to see how they've been made up from individual dots in one of just 8 colours but from a distance you'll be able to perceive the colours.

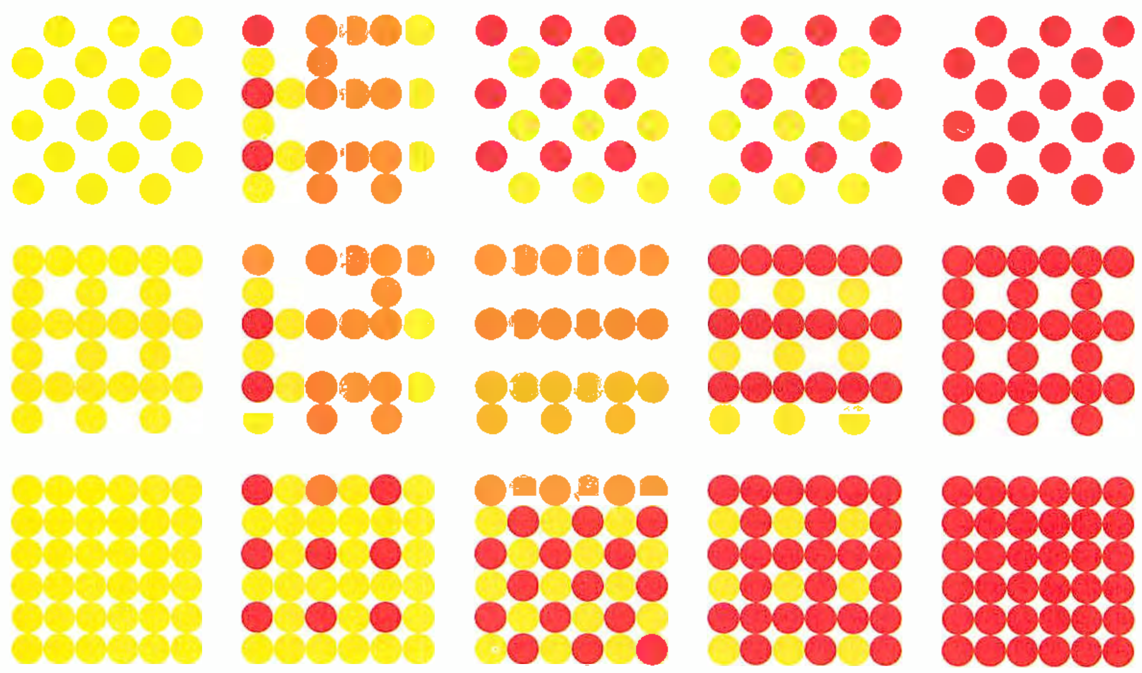

Figure 17.

This is the way that inkjet printers have operated in the past but some of the newer so-called photographic quality printers have less of a trade-off between the number of colours and the resolution. Some printers achieve this by having an extra two colours of ink – pale cyan and pale magenta. HP have an alternative technology which, for the first time, actually allows the size of the ink droplet deposited on the paper to be altered. For this reason, comparing headline resolutions alone isn't a good way of judging the quality you're likely to get from a printer for photographic work.