Why is Black-and-White so Important in Color?

R. W. G. Hunt
City University, London, England

Abstract

In the visual system, the retina communicates with the brain by means of a black-white (achromatic) signal and two color-difference signals, a red-green and a yellow-blue. The existence of the black-white signal has important implications in imaging. First, if areas that are intended to be black. gray, or white are reproduced with even a slight tinge of hue, the defect is usually very noticeable. because these achromatic perceptions correspond to the color-difference signals being balanced at their null levels. Second the achromatic signal largely determines the apparent contrast of scenes, and their images only look correct if their gray scales are adjusted with due allowance for the effect of the surround on the black-white signal. Third, in luminance-chrominance television, important reductions in bandwidth are possible because of the lower sharpness required in the chrominance signals as a consequence of the greater number of cones necessary to generate the color-difference signals than to generate the black-white signal; the extent to which advantage is taken of this situation is discussed in connection with various forms of imaging.

Introduction

In the first half of this century, color vision was most often regarded as being fully explained in terms of the trichromacy provided by the three different types of cone. This way of thinking was promoted by the work of Thomas Young[1] and Herman von Helmholtz[2] in the nineteenth century, and is sometimes referred to as the Young-Helmholtz theory of color vision. But towards the end of the nineteenth century, the German physiologist Ewald Hering[3] advanced the view that, after the light had been absorbed by the cones, the responses were transformed into three opponent signals, a black-white (achromatic), a red-green, and a yellow-blue. This Hering theory of color vision lay overshadowed by the Young-Helmholtz theory for many years, although it had its place, in the first half of the twentieth century. in various zone theories that combined both the tri-receptor and opponent processes in successive stages; among such theories are to be found those of Von Kries,[4] Adams,[5] Schrodinger,[6] and Muller.[7] But it was not until 1955 when Dorothea Jameson and Leo M. Hurvich[8] published the first of a series of papers entitled *Some quantitative aspects of an opponent-colors theory* that Hering's ideas became prominent. This was followed by a series of papers from 1956 onwards by G. Svaetichin,[9] R. L. DeValois and his co-workers,[10] and others, in which black-white and color-difference signals were discovered in various species, particularly in fish and monkeys.

The Reproduction of Blacks, Grays, and Whites

Areas that are perceived to be black, gray, or white correspond to the color-difference signals being balanced at their null levels. Any deviation from these null levels is easily detected, and hence if blacks, grays, and whites are reproduced with even a slight tinge of hue, the defect is usually very noticeable. It is for this reason that overall *color balance* is so important in images; if an image has a color cast, for instance a magenta cast, blacks, grays, and whites will be tinged with a very obvious magenta hue, and most pale colors will undergo noticeable changes in hue. It is interesting that, in recently developed color difference formulae, the color spaces are more finely divided in areas near the gray scale than elsewhere; this is true of the CMC[11,12] and CIE94[13,14,15] color difference formulae and also of the OSA color space.[16]

The Effects of Surrounds on Contrast

If the apparent contrast of an image is too low or too high, it has the appearance of being either misty or harsh, respectively. For good reproduction in images, it is essential to produce the right contrast, and, if this is not done, no amount of alteration of the color content will remedy the defect.

The contrast is determined mainly by the relation between the original and reproduced luminance and these are mediated by the black-white visual signal. This signal is greatly affected by the nature of the surround; a dark surround, as occurs in projection, lowers the apparent contrast so much that an increase is necessary in the system gamma (the slope of the relationship between log reproduced and log original luminances) from 1.0 to about 1.5; in the case of dim surrounds, as commonly occur in television viewing, the gamma has to be increased from 1.0 to about 1.25.[17,18]

Luminance-Chrominance Systems

The Field- Sequential System

The vast majority of color reproductions do not attempt to reconstruct the spectral composition of the original colors, but only to elicit in the three different types of cone of the retina the same or similar responses. In television, these responses are produced by causing individually modulated beams of red, green, and blue light to excite the cones. Various ways of combining the effects of the three beams have been used including the projection of the three images in register on a screen, the superimposition of virtual images

in the three colors, the presentation of red, green, and blue areas whose images on the retina are too small to be resolved (the mosaic method), and the production of images in red, green and blue light in rapid succession at a frequency high enough for the light to blend together to give mixture colors (the field-sequential system).

The first color television images to be seen anywhere were demonstrated by John Logic Baird in London in 1928 using the field-sequential system.[19] The first color television broadcast service was introduced in the U.S.A. in 1950; it also used the field-sequential system, employing 144 colored fields per second.[20] In the 1970s, closed-circuit field-sequential television was used for displaying positive images derived from color photographic negatives in order to facilitate the production of good prints from them.[21] Very recently, in one form of the Digital Mirror Device display system, field-sequential display is used at 150 or 180 fields per second.[22]

The Luminance-Chrominance Concept

The conceptual simplicity of the field-sequential system has kept it alive, but in 1953 the National Television Systems Committee (N.T.S.C.) of the U.S.A. was instrumental in launching luminance-chrominance television for broadcasting, and this method is now used almost universally.[23]

The seed thought for the luminance-chrominance method was sown in the mind of A.V. Loughren, a member of the N.T.S.C., when he came to Plate VI, facing page 144, of the book entitled *An Introduction to Color,* by Ralph M. Evans, published by Wiley in 1948.[24] This plate consisted of a full color image, together with a black-and-white image of the same scene, and another image of the scene in which, as nearly as possible, all the luminance differences between different parts of the scene had been removed, and only the chromaticity differences remained. The reason why Ralph Evans produced this plate was that, in the late 1940s, the projection of *Kodachrome* slides had become popular, and many who saw them commented on the apparent perception of three-dimensional effects in the projected images, an effect which was often attributed to the color content of the pictures. Ralph Evans was not convinced that the color was responsible for the apparent depth. and made this plate to investigate the situation. On asking people how much apparent depth they saw in the images of this plate, the universal response was that the full color image, and the black-and-white image, both exhibited considerable depth, but that the chromaticity-only image looked very flat. So Ralph Evans concluded that the apparent depth could not be attributed mainly to the color; the apparent depth was actually caused by the pictures being presented in isolation from their surroundings in large size in color, and therefore having much greater realism than had until then been generally experienced.

However, A.V. Loughren's interest was not in apparent depth issues. Because there were at that time some millions of black-and-white television receivers in use it was important that any new system incorporating color could be displayed on these monochrome receivers in black-and-white. With the field sequential system this was impossible, because the scanning speed had to be three times as fast as for black-and-white. What Ralph Evans's plate did was to suggest to A.V. Loughren that if color television signals were broadcast not as red, green, and blue signals, but as a luminance signal and two other signals that only carried the additional color information, then it should be possible to arrange for the black-and-white receivers to respond to the luminance signal only, while the color receivers responded to all three signals. In this way the color signals could be viewed in black-and-white on the existing monochrome receivers, a situation termed compatibility. With the current almost universal use of color in television, this type of compatibility is no longer an important issue in broadcasting with the existing systems (but compatibility with high-definition television systems would be desirable if practicable).

Reduction of Chrominance Bandwidth

Although the provision of compatibility was the driving force behind luminance-chrominance television another advantage of equal or even greater importance was achieved. It was found that the chrominance information can be much less spatially sharp than the luminance information without impairing the apparent sharpness of the composite picture. This made it possible to reduce the bandwidth of the chrominance signals to a quarter of that used for luminance and, with this reduced bandwidth, it was possible to interleave the chrominance information within the bandwidth used for the luminance information so that the total bandwidth for color transmission was no greater than that for monochrome. Furthermore, because the eye is not corrected for chromatic aberration, blue light is not sharply focussed on the retina, and it is therefore unnecessary for yellow-blue components of color differences to be displayed at as high a spatial resolution as red-green components; this effect is used in the N.T.S.C. system where one of the two chrominance signals has only one tenth the bandwidth of the luminance signals.

The reduction in chrominance sharpness that could be incorporated in imaging systems without impairment to the sharpness of the final composite picture was a very striking and, at first sight, surprising phenomenon. The surprise arose because, in the 1950s, the visual system was generally only thought of in terms of its retinal trichromacy, and not additionally in terms of its black-white (achromatic) signal and its red-green and yellow-blue color-difference signals. Thus, although the luminance-chrominance television system was a de novo invention as far as the inventors were concerned, a very similar system was in fact fully operational inside their own heads although they were quite unaware of it!

The opponent nature of the color-difference signals of color vision provides an explanation of the reduction of sharpness possible in chrominance signals. If we denote the strengths of the three cone outputs as ρ, γ, β, we can represent the three signals as $2\rho + \gamma + (1/20)\beta$ for black-white r- g for red-green and $\rho + \gamma - 2\beta$ for yellow-blue. The factors 1/20 and 2 in the black-white signal are included to allow for the fact there are many fewer β cones in the retina, and perhaps rather more ρ cones than γ cones. It is assumed that for achromatic colors (whites, greys, and blacks) the ρ, γ, and β signals are equal; the two color opponent signals then become zero for these colors.

Consider now a horizontal grey line with a small gap in it of a lighter gray; the color-difference signals will be

zero throughout so, for the presence of this gap to be detected, there must be a change in the black-white signal. Because the black-white signal collects from all three types of cone an adequate change will occur if the gap corresponds on the retina to a distance equal to at least one cone diameter (if the cone happens to be a β type the change in the signal will be rather small, but this will be a rare event because of the small number of β cones). If now we consider a horizontal green line with a red gap in it of the same lightness, it will have to be detected by the red-green signal; it is thus necessary, in this case, to have at least two cones in the gap, a ρ, and a γ. This would lead to a requirement for chrominance signals to be half as sharp as luminance signals. But the different types of cone are distributed randomly in the retina and, when this is allowed for,[25] it turns out that, on average, there must be at least four cones in the gap to detect a red-green change; and hence a four to one difference in the sharpness required for luminance and chrominance is to be expected, and this is the ratio found to be acceptable in practice. Considering now a horizontal blue line with a brown gap in it of the same luminance this must be detected by the yellow-blue signal so that there must be at least one ρ, one γ, and one β cone in the gap, and, because of the paucity of β cones, a much larger number of cones is now required in the gap; hence a yellow-blue chrominance signal can be reduced by a factor of about ten to one. The factor is not greater than this, such as twenty to one, because the non-linearities introduced by gamma-correction (to be described in the next section) prevent changes in the strength of the yellow-blue chrominance signal from being equivalent to changes in the β-cone response only, for all colours.

The Effects of Gamma Correction in Television

The luminance signal used in broadcast television is, in fact, not a true luminance signal. This is because the transfer characteristic of the display device most commonly used in 1953 was a power function with an exponent of about 2.2* referred to as a gamma of 2.2 (the gamma in this case is the slope of the relationship between the input signal and the amount of light produced on a log-log plot); the signals from the camera, E_R, E_G, E_B, were therefore *gamma-corrected* by being raised to the power of 1/2.2 (or 0.45). The luminance signal, E_Y', is then composed as:

$$E_Y' = 1E_R^{1/2.2} + mE_G^{1/2.2} + nE_B^{1/2.2}$$

l, *m*, and *n* having values in the N.T.S.C. system of 0.299, 0.587, and 0.114, respectively: these factors are determined by the chromaticities of the display phosphors and of the reference white used in the system. Because the chromaticities of modern phosphors are significantly different from those used in the original N.T.S.C. system and modem systems usually use as reference white Standard Illuminant D_{65} instead of Standard Illuminant C (as used in the N.T.S.C. system), there has been some discussion about changing the factors to correspond to current systems. However, so long as the signal is not a true luminance signal, there seems little justification for making changes. If, at some time, systems using true luminance signals were adopted, then adjustment of the factors would be desirable. Another fact that should not be forgotten however, is that real observers show a considerable spread of spectral luminous efficiency (V(λ)) curves, so that attaining true Standard Observer luminance will not achieve exact luminance for most real observers.

Because the E_Y' signal is not a true luminance signal some of the luminance information is carried by the two chrominance signals:

$$E_R^{1/2.2} - E_Y' \quad \text{and} \quad E_B^{1/2.2} - E_Y'$$

There is, consequently some loss of definition in the final picture as a result of restricting the bandwidth of the chrominance signals. It is for this reason that in more recent systems, such as those used in Photo CD and those proposed for high-definition television, the chrominance signals are only restricted to half the bandwidth of the luminance signal. This restriction is, however, applied both horizontally and vertically, as in the PAL and SECAM systems, and not only horizontally as in the N.T.S.C. system.

Luminance-Chrominance in Other Forms of Imaging

Color fax[26] uses signals corresponding to *L**, *a** and *b** of the CIELAB system. In this case a true luminance signal is used so that the *a** and *b** signals could be reduced to one quarter of the bandwidth used for the *L** signal, but a factor of only a half is used, because a half has become fashionable for modern systems. No advantage is taken of the possibility of reducing the bandwidth of the *b** signal further, but the reduction in bandwidths are applied both horizontally and vertically.

In electronic cameras using CCD arrays behind mosaics of red, green, and blue filter areas, it is common practice to use twice as many green areas as red or blue, because the green signal has the greatest contribution to the luminance signal. It is not usually possible to reduce the number of blue areas below the number for the red because of difficulties with color fringing at edges.

The advantage of the luminance-chrominance system is hard to achieve in film photography. Because silver halide emulsions of camera speed have natural blue sensitivity the arrangement in integral tripack films is usually to have the blue sensitive layer on top with a yellow filter underneath it, below which the green and red sensitive layers are situated. Because the light that exposes the blue layer has not been diffused by any other light-sensitive layers, it is sharper than that which exposes the other two layers: but the blue layer is the one which least needs to be sharp. In camera films, the best that can be done is to make the blue sensitive layer have as low a turbidity as possible. Films that are used for making prints from camera originals need not be so sensitive and in this case emulsions having much reduced blue sensitivity can be used, and then the green sensitive layer is at the top, and the blue sensitive layer at the bottom, of the tripack. In the case of papers intended for making prints from camera originals the yellow layer is at the bottom so that nonuniformity caused by the structure of the paper base results in irregularities in the yellow image which are much less noticeable than such changes in the magenta or cyan images.

* original text read 2.27

In graphic-arts printing the use of scanners to derive electronic signals from film makes it possible to manipulate the signals at will and a technique known as *grey component replacement* (*GCR*) is often used. In this technique, wherever all three inks would normally have been printed to produce a grey or a black component of a color GCR results in only black ink being printed instead of the three inks, with whatever amount of the remainder of the one or two inks being printed in addition. The black ink image corresponds to a luminance signal but some of the luminance is also carried by the remainder amounts of ink so that a true luminance-chrominance system is not achieved. However, it is found that improvements in sharpness do accrue from using GCR, although it is usually used only to a partial extent, because the black ink on its own is not usually able to produce as a good a black as when all four inks are printed.

Black ink on white paper is preferred for text, because this provides the maximum difference in luminance for the black-white visual signal with its high resolving power. Yellow text on white paper has very poor visibility because of the small luminance difference; and red text on green paper may exhibit no luminance difference at all with the result that the legibility is very poor unless the lettering is quite large.

Conclusion

So why is black-and-white so important in color? First, because black, grays and white colors correspond to the color-difference signals being at their null levels, any slight departure from the null condition is very noticeable. Second, images never look right unless their contrasts are correct and, because dark and dim surrounds affect the contrast of the black-white visual signal, images have to be adjusted to take these effects into account. Third, the sharpness of images depends much more on the luminance than on the chrominance content of the image, and if this is exploited it can lead to very useful economies in the information content necessary in transmitted and displayed signals.

General Reference

R. W. G. Hunt, *The Reproduction of Colour,* fifth edition, Fountain Press, Kingston-upon-Thames, London, (1995).

References

1. T. Young, *Philos. Trans.* **1802**, p. 12 and 387 (1820).
2. H. von Helmholtz, *Handbuch der Physiologischen Optik* Voss, Hamburg (1896).
3. E. Hering, Beitrag zur Lehre vom Simultankontrast, *Zeitschrift fur Psychologie und Physiologie den Sinnesorgane,* **1**, 18-28 (1890).
4. J. von Kries, Die Gesichtsempfindungen, in *Handbuch der Physiologie des Menschen., Vol.* III, Braunschweig: Vieweg, 109-282 (1905).
5. E. Q. Adams, A Theory of Color Vision, *Psychol Rev.,* **30**, 56-76 (1923).
6. E. Schrodinger, Grundlinien diner Theorie der Farben metrik im tagessehen, *Ann. der Physik* **63**, 397-426, 427-456, 481-520 (1920).
7. G. E. Muller, Uber die Farbenempfindungen, Bd. 2. *Zeits. Psychol. Physiol. Sinnesorg.,* Erganzungsbd, **17**, 1-430, **18**, 435-647 (1930).
8. D. Jameson and L. M. Hurvich, Some Quantitative Aspects of an Opponent-Colors Theory. I. Chromatic Responses and Spectral Saturation. *J. Opt. Soc. Amer.,* **45**, 546-552 (1955).
9. G. Svaetichin, Spectral Response Curves from Single Cones, *Acta Physiol. Scand.,* **39**, Suppl. 134, 17-47 (1956).
10. R. L. De Valois, I. Abramov, and G. H. Jacobs, Analysis of Response Patterns of LGN Cells, *J. Opt. Soc. Amer.*, **56**, 966-977 (1966).
11. F. J. J. Clark, R. McDonald, and B. Rigg, Modifications to the JPC79 Colour-Difference Formula, *J. Soc. Dyers Col.,* **100**, 128-132 and 281-282 (1984).
12. R. McDonald, Acceptability and Perceptibility Decisions Using the CMC Colour-Difference Formula, *Text Chem. Color,* **20** (6), 31-31 (1988).
13. R. S. Berns, D. H. Alman, L. Reniff, G. D. Snyder, and M. R. Balonon-Rosen, Visual Determination of Suprathreshold Color-Difference Tolerances Using Probit Analysis, *Color Res. Appl.,* **16**, 297-316 (1991).
14. D. H. Alman, CIE Technical Committee 1-29 Industrial Color-Difference Evaluation Progress Report, *Color Res. Appl.,* **18**, 137-139 (1993).
15. K. Witt, Modified CIELAB Formula Tested Using a Textile Pass/Fail Data Set. *Color Res. Appl.,* **197,** 273-285 (1994).
16. D. Nickerson, OSA Uniform Color Scale Samples: A Unique Set, *Color Res. Appl.* **6**, 7-33 (1981).
17. C. J. Bartleson and E. J. Breneman, Brightness Perception in Complex Fields, *J. Opt. Soc. Am.* **57**, 953-957 (1967).
18. R. W. G. Hunt, The Effect of Viewing Conditions on Required Tone Characteristics in Colour Photography, *Brit. Kinematog. Sound Tel.,* **51**, 268-275 (1969).
19. R. Herbert, J. L. Baird's Colour Television 1937-46, *J. Roy. Television Soc.,* **27/1**, 23-29 (1990).
20. R. W. G. Hunt, *The Reproduction of Colour,* fifth edition, p. 46, Fountain Press, Kingston-upon Thames, London (1995).
21. *Brit. J. Photogr.,* Gretag-Ferrex Colorverter and Translator, **119**, 166-169 (1972).
22. J. M. Younse and D. W. Monk, The Digital Mirror Device (DMD) and its Transition to HDTV, *Image Technology,* **76**, March, 32-33 & 42 (1994).
23. *Proc. Inst. Radio Engnrs.,* **39**, 1124-1331 (1951), **41**, 838-858 (1953), **42**, 5-344 (1954), **43**, 742-748 (1955).
24. R. M. Evans, *An Introduction to Color,* Wiley, New York (1948).
25. R. W. G. Hunt, The Strange Journey from the Retina to the Brain, J. *Roy. Television Soc.* **11**, 220-229 (1965).
26. A. H. Mutz and D. T. Lee, An International Standard for Color Facsimile, *IS&T and SID's 2nd Color Imaging Conference: Color Science, Systems and Applications,* 52-54, IS&T, Springfield, Va., U.S.A. (1994).

☆ This paper was previously published in *IS&T's 4th Color Imaging Conference Proc.,* p. 54 (1996).