Color Imaging Systems And Color Theory: Past, Present And Future

John J. McCann
McCann Imaging, Belmont MA 02178 USA
mccanns @tiac.net

ABSTRACT

James Clerk Maxwell demonstrated the first color photograph in a lecture to the Royal Society of Great Britain in 1861. He used this demonstration to illustrate Thomas Young's idea that human vision uses three kinds of light sensors. This demonstration led to a great variety of color photographic systems using both additive and subtractive color.

Today, we have photographic, video, still digital, and scanning image-capture devices. We have electrophotographic, ink jet, thermal and holographic hard-copy systems, as well as, cathode ray tube, liquid-crystal display, and other light emission color devices. The major effort today is to get control of all these technologies so that the user can, without effort, move his color, digital image from one technology to another without changing the appearance of the image. The strategy of choice is to use colorimetry to calibrate each device. If all prints and displays sent the same colorimetric values from every pixel the images, regardless of the display, will appear identical.

The problem is that prints and displays have very different color gamuts. A more satisfactory solution is needed. In my view, the future emphasis of color will be in models of human vision that calculate the color appearance, rather than the color match. All the technologies listed above work one pixel at a time. The response at every pixel is dependent on the input at that pixel, regardless of whether the imaging system is chemical, photonic or electrical. Humans are different. The color they see at a pixel is controlled by that pixel and all the other pixels in the field of view. Human color vision is a spatial calculation involving the whole image. In the future, we will see more models that compute the color appearance from spatial information and write color sensations on media, rather than attempting to write the quanta catch of visual receptors.[1]

Keywords: color systems, history, colorimetry, color contrast and color constancy

THE FIRST COLOR SYSTEM

In 1861 James Clerk Maxwell made the first color reproduction system. He did it for a Friday Evening Discourse at the Royal Institution on Albemalre St. London. He wanted to illustrate the point that Thomas Young had made fifty years earlier, as a teacher at the Royal Institution. Namely, Young said there were only a limited number of different types receptors in the human retina responsive to different frequencies of light. Three different receptors were sufficient to produce the entire range of colors we see.

Maxwell's experiment was to make three color separation photographs of a bow made with a multicolored ribbon. The first photograph used a red filter that transmitted the long-wave visible light. The second photograph used a green filter that transmitted the middle-wave visible light, and the third used a blue filter that transmitted the short-wave visible light. At the lecture Maxwell superimposed these three photographs with three lantern projectors: one with a red, one with a green and one with a blue filter. The audience reported seeing a multicolor bow.

This is the first example of color photography in which the spectral information was recorded from the scene. At that time there were many examples of black and white photographs on which colors had been applied by a hand. For the 100th anniversary of this historic experiment, Ralph Evans of Kodak repeated the experiment.[2] What is of particular interest in Evans' paper is that Maxwell's experiment should not have worked. Maxwell's experiments were done with unsensitized emulsions. Ten years later, Vogel invented techniques for making silver halides respond to red light. In other words, the black and white film had very little sensitivity to the long-wave light transmitted by the red filter. However, Evans believed

Part of the IS&T/SPIE Conference on Human Vision and Electronic Imaging III
San Jose, California • January 1998 • SPIE Vol. 3299 • 0277-786X/98/$10.00

that the red filter had a leak in the blue-uv region and that light exposed the long-wave record. Regardless of the color fidelity, Maxwell's intended demonstration of human vision was the first color system reproduction.

IVES COLORIMETRY PATENT

In 1890 F. E. Ives patented the idea that photographic films should use colorimetric primaries for their spectral sensitivity (U.S.P. 432,530). The curves Ives used in the patent were the color mixture curves of Maxwell.[3]

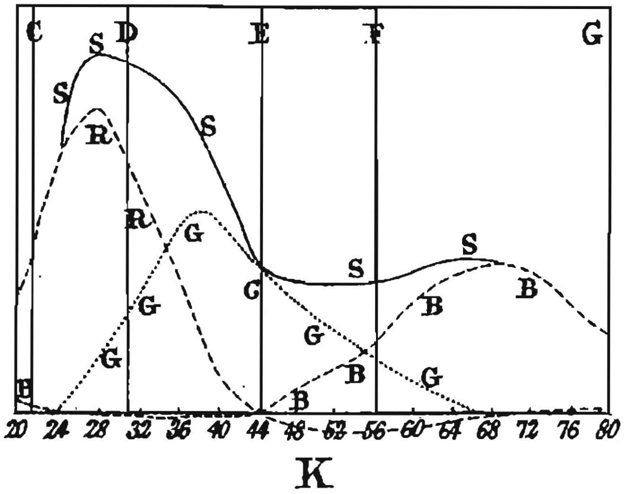

Figure 1. Diagram of the color matching functions reported by Maxwell[4]. The horizontal axis is calibrated in wavelength by interference lines created with two parallel glass plates and an air gap. Maxwell used 16 distinct points on the spectrum listed in the horizontal axis. The number 20 was reported as red on the left. The number 40 was green, 60 was blue and 80 was indigo. The lines C, D, ...G are Fraunhofer's spectral lines. The vertical axis is intensity measured by the width of the slits This graph shows the matches for observer K. The R, G, B curves show the intensities of R (24), G (44) and B (68) to match wavelengths in the spectrum. The S curve is the sum of the other three curves. This data shows that the red and green matching functions overlap considerably.

Ives's idea has dramatic advantages and disadvantages. The advantage is that a film with the same spectral sensitivity as the cones in the human eye would have the best color fidelity. With input sensors identical to cone sensitivities one can capture the information to make a perfect match. Of course there is a second demand of such a reproduction, namely that the contrast range from white to black also matches the contrast range of the original scene. If the black and white *input/output function* does not have a slope of 1.0, then the colors will be altered. If the slope of the film is higher than 1.0, the colors will be more saturated. If the slope is lower than 1.0 then the colors are less saturated than the original. To benefit from colorimetric sensitivity curves one has to use a slope 1.0 film.

The disadvantage of Ives Patent is that the spectral sensitivities of the cone pigments in the eye have enormous overlap. At the peak of the long-wave cone (Figure 2), the middle-wave cone is 50% as sensitive the long-wave cone. What this means is that the color separation is greatly reduced.

In 1973 Tom Taylor and I developed a computer program that calculated the best filter combination to alter a black and white film sensitivity to mimic the response of the human cone pigments. We began with Polaroid 55P/N film and found filter combinations that altered the films sensitivity to match one of the human cone pigment's sensitivity. The film plus Wratten #9 and CC20C approximates the long-; the film plus #8, #13 and CC20C approximates the middle-: the film plus #47 and #86A approximates the short-wave cones.5 In Figure 2, the solid lines show the sensitivities of the long, middle and short wave cones. The dip in the solid curve for the short-wave cone is the effect of the lens and macular filtration found in the eye. Thus, the solid lines represent the cone response function. As mentioned above, the overlap is substantial, particularly for the long- and middle-wave cones.

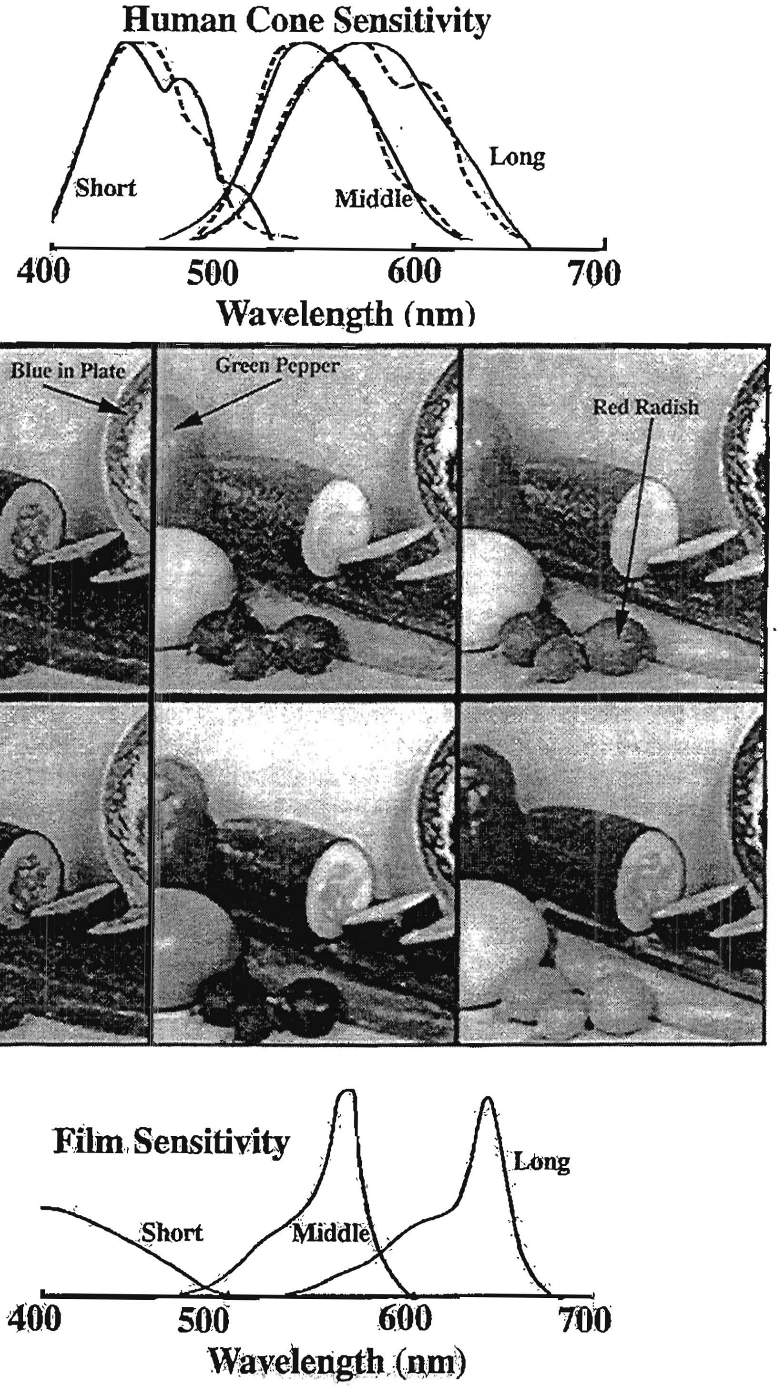

Figure 2. This figure compares the sensitivity functions of the human eye with those of a typical three-color integral film. Further, it compares the images made with these sensitivity curves. At the top of the figure is the long-, middle-, and short-wave sensitivity function of the three cones in the eye corrected for lens and macular absorptions (solid lines). Below that are short- (left), middle-(middle), and long-wave (right) separation photographs taken with a film-filter combination that fits closely the cone sensitivities (dashed lines). Below that, the second row of photographs shows the normal color separation made with non-overlapping spectral sensitivities. Below that is the sensitivity function of a typical integral color film.

The dotted lines in Figure 2 show the film-filter combination that approximates quite closely the cone response function. This allows us to make photographic images that record the differences in cone response to real life objects. Three color separation records are shown below the sensitivity curves. The short-wave or blue record is on the left. It is noticeably different from the middle-wave record in the middle. What is of greater interest, is that the middle- and long-wave records are almost identical. All of the objects have been recorded with the same lightnesses as in the middle wave-record. There are two exceptions, the radish at the bottom of the image is slightly lighter in the long- than in the middle-wave record. On careful inspection, the green peppers are a very small amount darker in the red record than the green one. The cone pigment separation photos have a normal range of lightnesses from black-gray-white axis, but as we would expect from the overlap of spectral sensitivities, there are almost no changes along the red-gray-green axis of color space.

In the original color image the radish was a bright red, the background was gray and the pepper was a bright green. In the cone response color space, a bright red is characterized by the amount that the radish is darker in the green record than in the red record. A gray is characterized by the fact that they are the same. A bright green is characterized by the amount that the green record is lighter than the red record. These are very small differences. The only record with easily visible differences is the blue record. Here the blue in the plate is darker in both the red and green records than in the blue record. Here the greater separation in sensitivity curves leads to easily recognizable lightness shifts in the photographs.

Mees and Pledge did an extensive study on film spectral sensitivity and did not to use film sensitivities that mimicked human sensitivities as suggested by Ives. Instead, they decided to emphasize color analysis incorporating non-overlapping spectral sensitivities[6] In integral photography the dye control and color isolation is achieved by keeping the silver and dye or dye precursors separate. Each layer has to work independent of the other layers to ensure saturated colors. In the early days of subtractive photography, invented by du Huron and C. Clos, each color separation was made on a separate piece of film. Then each dye was applied to the image substrate in three steps. In integral layered films, spacer layers are introduced so that the chemical reactions developing the red sensitive silver halide controlling the cyan dye are keep these separate from those developing the green sensitive silver halide controlling the magenta dye.

The type of image processing found in humans is difficult to incorporate in integral color films. Let us consider an example of red, gray and green areas in a photographic image. Both the red and green areas are bright and fully saturated. The gray is neutral. To make a neutral gray scale from white to black requires that the all three subtractive dyes (cyan, magenta and yellow) are laid down in the print in equal proportions. This is achieved by matching the responses of the red-, green- and blue- sensitive emulsions. In three color separation photographs a gray area will have the same lightness in each separation. Since the red area is a full saturation color, the print needs to have the maximum amount of both yellow and magenta dyes. Of still greater importance, there must be no cyan dye in the red. Even a trace of cyan will make the red dark and muddy. Similarly, the fully saturated green area must have the maximum amount of yellow and cyan dyes. Any trace of magenta will make it dark and muddy.

The bottom half of Figure 2 shows the sensitivities and separations of a photographic system. We can compare the difference responses of photography and the eye. If we look at the radishes in Figure 2 [lower row], we see that the film sensitization generates a white in the red, and blacks are in the green and blue records. This is exactly what is required for dye control. A white in the red separation means that all of the cyan dye is kept away from the print. A black in the green and blue separations means that maximum amounts of magenta and yellow dyes will be in the print.

Photographic film designers often achieve greater color saturation by increasing the *input/output function* or slope, so as to make gray values near white become white and gray areas near black become black. This leads to more saturated colors everywhere, but reduces the range of intensities that the film can see.

Let us return to the separations with the same response as the cone pigments in the eye (Figure 2 top row). The lightness of the radishes in red is middle gray: the lightness in the green is slightly darker. If we make a color print with this separation we have middle gray radishes with a hint of red. A satisfactory print requires that the red-green difference be altered from nearly indistinguishable to maximally different. Remember we cannot tolerate even a trace of cyan. That means the red record has to become white, and the green record has to become black. In digital systems that mimic the eye this is possible because red-green interactions that are independent of white gray black axes are possible. Just as in the human eye, we can stretch the color difference without changing the gray scale. In color separation photography with separate color printing this is also possible. Graphic Arts have a long tradition of adding the green negative film to the red positive film to enhance colors. In integral films it is more difficult because the spacer layers that keep the chemical reaction of the red and green layers separate make it difficult for red green specific responses. If the red equals green, then nothing happens. If the red and green are different, then make them 100 times more different. What is easy to do in digital images and possible in three separations very difficult in integral films.

It is quite reasonable to see that integral color films have been very carefully designed to give pleasing color renditions that are not colorimetric. In such films, the introduction of colorimetric sensors would destroy the usefulness of the process. We plotted a collection of the most saturated Munsell papers in a color space based on the light absorption by cone pigments. The space was normalized to the maximum in each waveband and grays fell along the axis from white to black. The cone absorption was scaled by a cube root lightness function to correct for scattered light in the eye.[7] For the long-, middle-, and short-wave axes the range from the maximum white to black is 10 for each axis. In the middle plane of this color space the most saturated red in the Munsell Book is separated from the most saturated green by a distance of 2 units. Along the yellow-blue axis the distance is 4 units, while the distance from white to black is 17 units. The cone pigments significantly compress the information in the saturation plane of their color space.

It is interesting to compare the cone color space with the frequently used CIE Uniform Color Space [L*a*b* 1976] . L*a*b* begins with colorimetric X, Y, Z values, which are analogous to Maxwell's fundamentals, but based on W. D. Wright's experiments incorporated in a 1931 CIE standard. X, Y, Z are linear transforms of the cone sensitivity curves.[8] It uses cube roots to correct for scatter in the eye. L* is the cube root of normalized Y. The term a* is scatter corrected, normalized X-Y multplied by a color saturation stretch factor. The term b* is scatter corrected, normalized Y-Z multplied by a color saturation stretch factor. The stretch factor for a* is 500; for b* is 200. This color space has a red-to-cyan axis that is roughly twice as big as the black to white axis. The same is true for yellow-to-blue. This space is intended to be isotropic, that is uniformly spaced in appearance. The important idea is that the red-green axis of colors we see has been stretched hundreds of times from the relationship measured by the sensitivity of the cones. This is an essential role for the opponent color processing proposed by Hering[9].

There are many uses for color spaces. The simplest is the color match. Since a match happens when all the differences are zero, it does not matter what space you use. If you want to measure differences in color appearance, human cone space is a very poor choice. It compresses the saturated colors by a factor of hundreds compared to the neutral gray axis. The corollary of this observation is that cone color spaces are only valuable in making color matches. In all other cases, the color stretching from opponent color processing changes the saturation of the color to make other uses inappropriate.

So far, no industrial product has used colorimetric primaries as sensors to record the light from real life scenes. The overlap in spectral sensitivities of human cone pigments is so great that it allows almost no color isolation. By skillfully selecting film sensitivities, color photography successfully records virtually all colors. In fact, as an active photographer over many years and with a very wide range of wide of imaging systems, I cannot recall any specific examples in which color sensitivity made the image unacceptable. There have been numerous images in which details in whites and blacks have been lost. There have been over- and under-exposures. There have been many occasions in which the high contrast of films have distorted the colors in fine art reproductions. There have been problems in reproducing yellow because of gamut restrictions. I can remember a lecture, thirty years ago, that described a particular genus of morning glory was rendered pink-blue, rather than blue-pink. I also recall an experimental sensitizing dye for Polavision emulsion that made orange pumpkins to be rendered too yellow. This sensitizing dye was never used in a commercial film. In summary, careful system design has led to three minimally overlapping spectral sensitivities that do a remarkably good job, despite their departure from cone sensitivities.

THE ROLE OF INPUT / OUTPUT FUNCTIONS IN COLOR

As we described above, human cone color space is best used in color match. Color match in a reproductive color system requires the use of both cone spectral sensitivities and an input/output function with a slope of 1.0. We have seen from color quality experiments that observers prefer more saturated colors. Color system designers, both chemical and digital have moved away from accurate color reproductions. It is very interesting to look at the properties that systems designers have selected. Here we measured the response function of commercial photographic systems. We did this by exposing and printing photographs of a calibrated black and white target. We measured the prints with a reflectance colorimeter, integrating under the CIE Y sensitivity function. The photographic response from white to black describes how the print compares to the original. First, we plot this function as percent reflectance. The first curve is the input/output response of a conventional positive-negative chemical photograph. A gray scale was photographed on 35 mm negative film and printed on a minilab printer. The second curve is the input/output response of a digital camera record printed with a thermal inkjet printer. The third curve represents a theoretical perfect copy film in which output equals input. The chemical and digital response functions are very similar. The whites are at the top right of the graph. The white in both prints are darker than the white in the original gray scale. The input/output functions for photographs are not usually plotted on reflectance axes. The 18% middle gray falls in the lower left corner of the graph. All of the dark-grays and black fall on top of each other.

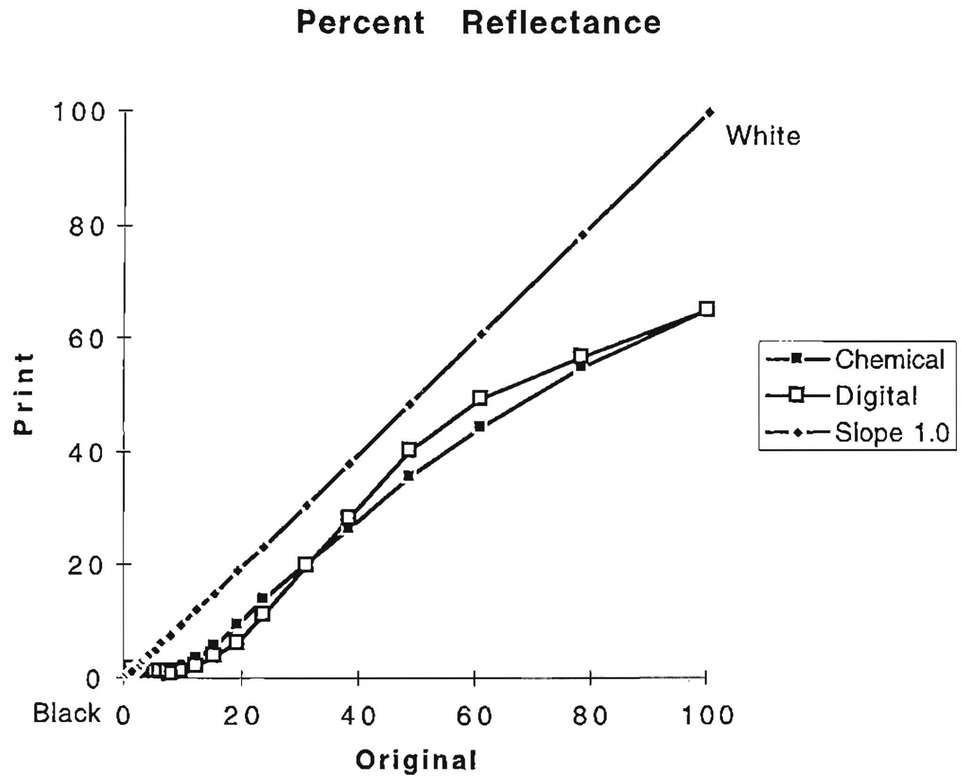

Figure 3. The plot of Print % reflectance vs. Original reflectance. The solid squares are the input/output response of a conventional positive-negative chemical photograph.. The open squares are input/output response of a digital camera record printed with a thermal inkjet printer. The third curve is the solid diamonds that represents a theoretical perfect copy film in which the output equals input.

The input-output characteristic curve for film is often called an H&D curve, named after Hurter and Driffeld, two English photographic scientists.[10] It is the plot of log exposure versus optical density (OD) of the resulting photograph. Optical density is the log of the reciprocal of reflectance. If a film were designed to exactly reproduce the original scene, it would have a slope 1.0 H&D curve. Input light would equal output reflectance for all values. When the slope is higher than 1.0, the image has more contrast between similar exposures than the original. Here, there are bigger differences in density compared to the original. When the slope is less than 1.0, the image has more compressed gray values than the original. Here there are smaller differences in similar densities compared to the original.

All color print films and digital systems use similar H&D curves. In the region of light gray, or Caucasian and Asian skin tones, the slope is 1.0. Colors in this region are reproduced accurately, compared with the original. Colors lighter than skin tones have a much lower slope. Whites and specular highlights are compressed together in the color print. Middle grays are expanded. The print makes middle grays darker than the original. Dark grays have a slope of 1.0. Blacks are compressed. All color print film and digital systems use very similar H&D curves. In the region of light gray, or Caucasian and Asian skin tones, the slope is 1.0. Colors in this region are reproduced accurately, compared with the original. Colors lighter than skin tones have a much lower slope. Whites and specular highlights are compressed together in the color print. Middle grays are expanded. The print makes middle grays darker than the original. Dark grays have a slope of 1.0. Blacks are compressed.

The high slope in the middle-tone is key to making very colorful pictures. A red has a very light tone in long-wave, or red light. Red is middle-to-dark gray in green and blue light. High-slope films that increase the darkness of the middle tones increase the saturation of colors. The exact slope of color film H&D curves have been determined by many consumer acceptance tests. Consumers never select slope 1.0 films as their favorite. Nevertheless, slope 1.0 without any loss of saturation is the ideal film for art reproduction.

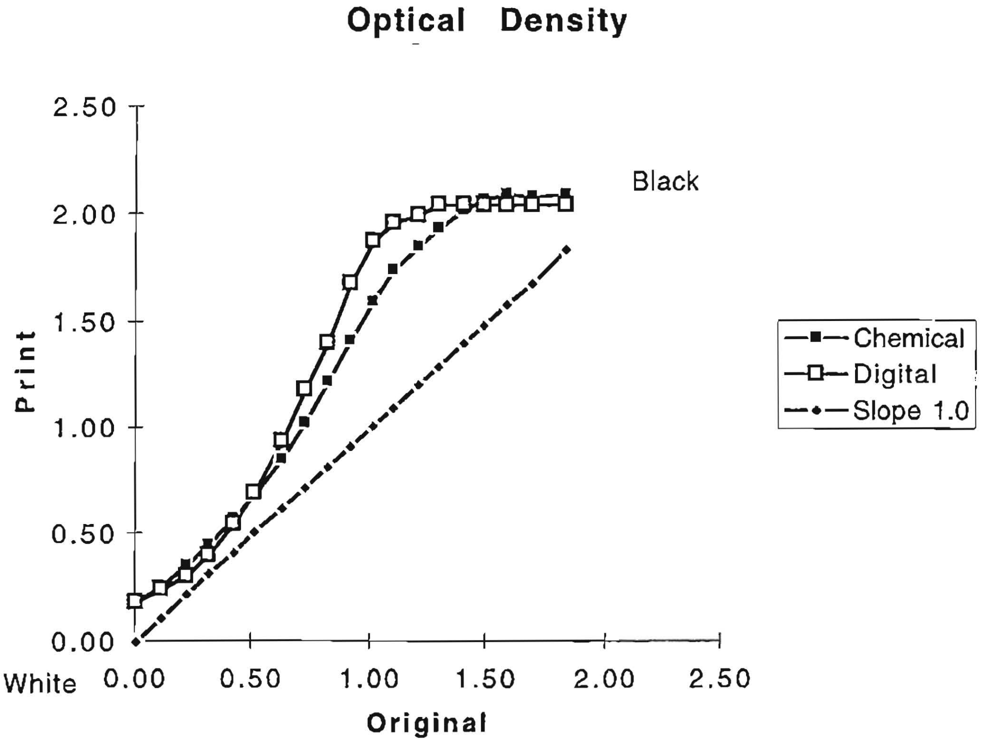

Figure 4. The plot of Print Optical Density vs. Original Optical Density. The solid squares are the input/output response of a conventional positive-negative chemical photograph.. The open squares are input/output response of a digital camera record printed with an thermal inkjet printer. The third curve is the solid diamonds that represents a theoretical perfect copy film in which the output equals input.

H. deRidder, et al's recent measurements show that observers prefer images that are more saturated than realistic images[11]. Their experiments showed observers a sequence of images that varied from low saturation, through actual to high saturation. They asked observers to select the most "natural" image and in a separate experiment they asked observers to select the preferred image. The data showed that observers preferred a more saturated image than the one that they selected as most "natural". These experiments show that people select pictures of slightly higher saturation than realistic pictures. Although people prefer a boost in saturation for pictures, the same boost distorts a reproduction of fine art. The problem for making the most accurate reproduction is to remove the film's boost in saturation put in at the factory.

These bring us back to a point made earlier in the paper. If we truly want to reproduce the colors we see in the world, we need to use sensors that generate exactly the same information as the cones. We cannot afford to ignore the black to white axis. If we want true reproduction we must use slope 1.0 systems. The issue becomes more interesting when we ask about whether we mean slope 1.0 at the image or slope 1.0 on the retina. The human eye exhibits considerable scatter. Since Fechner's early experiment on the appearance of middle gray we have known that gray scales that appear to the eye to be equally spaced in lightness are not proportional to the light coming from the grays. That is why a photographic gray card has a reflectance of 18 %. Munsell and Glasser showed that equally spaced grays from white to black are fit by a cube root of the radiance coming to the eye. That is why we see cube root functions in L*a*b* and CIECAM.

This psychophysical observation is in conflict with physiological data indicating that cone response is proportional to log radiance.12 Experiments that calculated the intensity at each pixel on the retina after scatter resolved this paradox. In part one, observers picked samples that appeared equally spaced in lightness. In part two, the experimenter measured the radiance coming to the eye (cube root of radiance). In part three the experimenters calculated the radiances at the retina after scatter. The result was that log radiance at the retina is proportional to appearance. The cube root response is caused by scatter in the human eye.[7] The scatter makes dark objects appear lighter. When the observer is instructed to make the dark steps appear equally different as light ones, he must pick samples that are much darker without scatter.

Figure 5 shows the input output function of Lightness (L*). Here we see the curve has flipped again with the whites in the upper right. Just as with the OD plot in Figure 4, we see whites that are compressed, skin tones at slope 1.0 and expanded values in the middle grays. Lightnesses below 35 show expansion characteristic of low slope response functions. The high slope gives the customer-preferred increase in color saturation. The low slope functions provide discriminability between whites and blacks using the smaller response range of film compared to the real world.

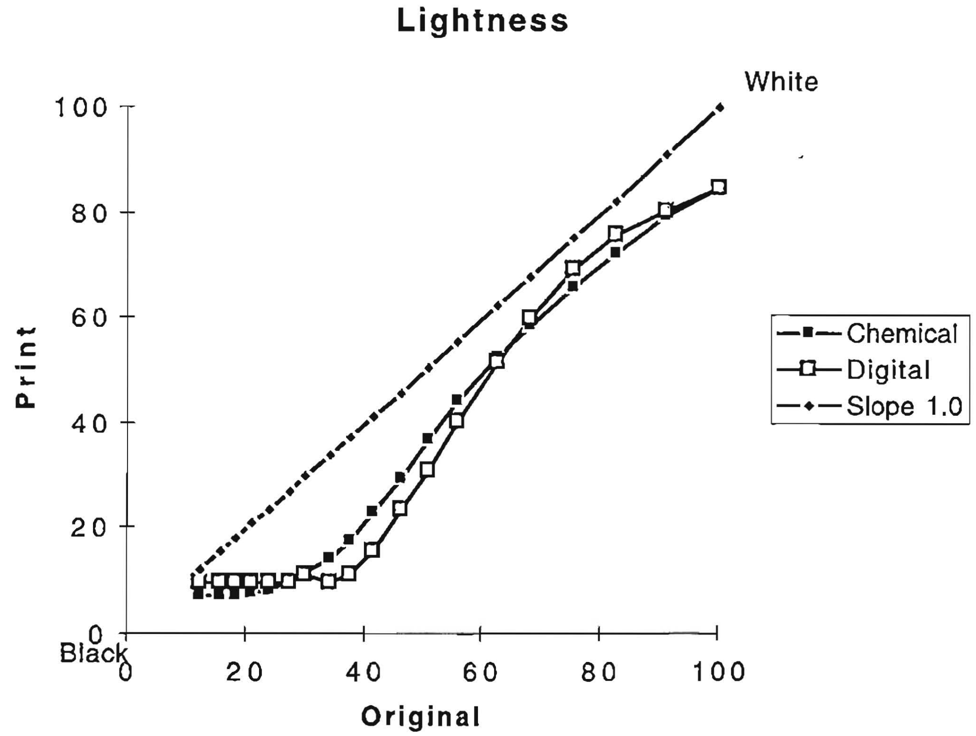

Figure 5. The plot of Print Lightness vs. Original Lightness. The solid squares are the input/output response of a conventional positive-negative chemical photograph.. The open squares are input/output response of a digital camera record printed with an thermal inkjet printer. The third curve is the solid diamonds that represents a theoretical perfect copy film in which the output equals input.

These curves in lightness are the best insight into the preferred treatment of a scene for a photograph. They tell us the departure from color match preferred by observers and film system designers. It portrays this information most accurately because it displays the data on the retina after scatter. As we mentioned above, the use of sensors that are metameric with cone pigments serves little purpose if we do not use a slope 1.0 function from black to white. Any nonlinearity in the gray axis will introduce color shifts in saturation. This ideal film would be optimal for reproducing fine art. It will look identical to the original. It would not be a good imaging system to do the job of photography because it would have prints with very limited dynamic range and will be too bland compared to today's digital and chemical photographic systems.

REFERENCES

[1] The comprehensive text of this lecture will be published in the J. Imaging Sci Technol., **42**, Jan-Feb., 1998. To avoid redundancy this paper is a new expansion of a part of the paper.

[2]R. M. Evans, Sci. Amer., **205** (Nov.): p. 118,1961.

[3] J. S. Friedman, History of Color Photography, The American Photographic Publishing Company, Boston, p. 13,1945.

[4] J. C. Maxwell, XXI. On the Theory of Compound Colours, and the Relations of the Colours of the Spectrum, in The Scientific papers of James Cleck Maxwell, W. D. Niven ed., Dover, New York, Vol. 1, pp. 410-444, 1965.

[E]. H. Land, Smitty Stevens' test of Retinex Theory, in Sensation and Measurement, papers in honor of S. S. Stevens, H.R. Moskowitz, B. Scharf, and J. C. Stevens, D. Reidel Publishing Company, Boston, 1974, reprinted in Edwin H. Land's Essays, IS&T, Springfield, VA, Vol. 3, pp. 113-117, 1993.

[6] C. E. K. Mees, Brit. J. Photog., 55, p 41, 1908. and C. E. K. Mees and Pledge, Phot. J., 50, p.197, 1910.

[7]W. A. Stiehl, J. J. McCann and R. L. Savoy , "Influence of intraocular scattered light on lightness scaling experiments" J Opt. Soc. Am. **73**, pp. 1143-1148,1983.

[8] V. C. Smith and J. Pokorney, "Spectral Sensitivity of the Foveal Cone Photopigments between 400 and 500 nm," Vision Res., **15**, p. 161, 1975.

[9] E. Hering, "Outline of A Theory of the Light Sense," trans. by L. M. Hurvich and D. H. Jameson, Harvard University Press, Cambridge, MA, 1964.

[10]J. M. Eder, *History of photography*, E. Epsteam, trans. Dover, New York, (1978),453.

[11]H. deRidder, F. J. J. Blommaert, E. E. Fedorovskaya, *Naturalness and image quality: chroma and hue variation in color images and natural scenes*, SPIE Proc. 2411, (1995).

[12] Dowling, J., 'Neural and photochemical mechanism of visual adaptation in the rat". *J. Gen. Physiol.* **46**, 459-474, 1963