Digital Color Reproduction

OSA Handbook: ***The Science of Color, 2nd. Edition*** **1.0**

Brian Wandell
Psychology Department
Building 420, Jordan Hall
Stanford University
Stanford, California
e-mail: wandell@stanford.edu

Louis D. Silverstein
VCD Sciences, Inc.
Scottsdale, Arizona
e-mail: lou-s@vcdsciences.com

Key words: color, displays, image capture, digital cameras, printing, scanning, LCD

Acknowledgements: We thank P. Catrysse, A. El Gamal, J. Farrell and G. Starkweather. Supported in part by the Programmable Digital Camera project at Stanford University.

Number of Figures: 24

Introduction and Overview

In this chapter, we describe how principles of human vision are used to design image capture and display devices. The chapter is divided into four sections. First, we provide an overview of two properties of human vision that are essential in designing color imaging technologies. The next three sections describe the application of these and related principles along with the specific technologies. The second section reviews digital cameras and scanners. The third section reviews displays with a particular emphasis on cathode ray tube (CRT) and liquid crystal display (LCD) technologies. The fourth section describes aspects of color printing.

A number of topics in color technologies are not covered in this chapter. We do not include implementation details or discussions of any specific technology. This is a fascinating and rapidly developing area, but the advances are so rapid that our discussion would be out of date by the time an archival chapter is published or read. Also, we do not discuss image processing methods, such as compression standards or graphics rendering techniques, even though the color vision principles described here are fundamental to these methods. We have excluded this topic because this chapter is a compromise between breadth of coverage and existence.

Our focus is on the fundamental principles of color imaging technology that must be addressed in the design of capture and display technology. Quantitative methods useful for certain specific devices are described, and we expect that these methods will be useful for future generations of display and capture technologies as well. It is in this sense that we hope this chapter will serve as a practical reference for the general principles of color imaging technologies.

Imaging as a Communications Channel

In this review, we emphasize the aspects of imaging devices that are important in characterizing their role within a communications channel. An overview of how imaging devices form a communications channel is shown in Figure 1. The input signal is the original scene. This scene is captured and communicated over a transmission channel. This transmission usually includes various computational operations that facilitate inter-device communication and efficient transmission and storage. The transmitted image is then converted to a form where it can be rendered by a display device. Finally, the displayed image is acquired by the human visual system. When the image communications channel works well, the visual experience of seeing the original image matches the visual experience of seeing the reproduction. Hence, channel metrics must be based on how well the system performs with respect to the human visual system.

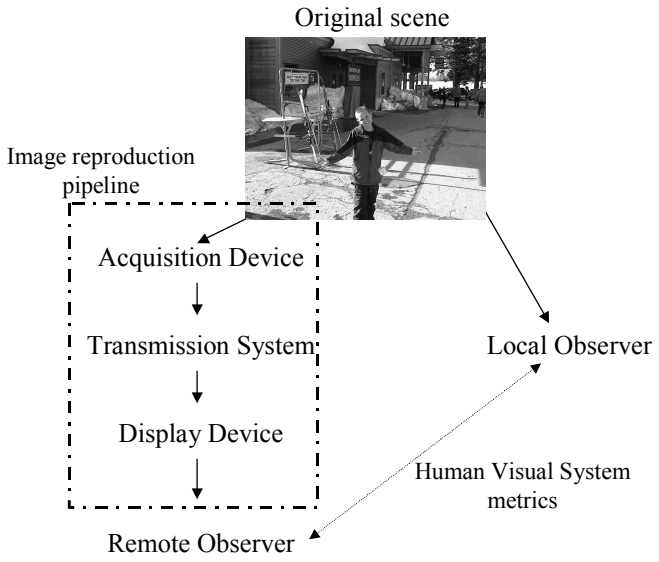

Figure 1. The image reproduction pipeline shares some properties with a general communications channel. The quality of the reproduction pipeline, the channel metrics, should be based on a comparison of the appearance of the original image with the appearance of the original scene. Hence, the visual significance of image features are an essential component in defining the quality of the channel.

From examining the imaging channel description, several requirements of the devices on the communications channel are evident. Capture devices must measure the original image over a range that matches the signals captured by the human visual system. Display devices must be able to deliver accurately controlled signals to the human visual system. Measures evaluating the quality of the communications channel must include comparisons of the *visual appearance*, a psychological quantity, associated with the original scene and the image delivered by the reproduction.

Two properties of human vision are central in the design of color imaging technologies. The first is trichromacy, a principle that has already been introduced in this book from the point of view of the behaviorist (Smith & Pokorny, In press) and from the point of view of the physiologist (Lennie, This volume). Here, we will introduce the principle from the point of view of the technologist. The second is the spatial resolution of the eye, and in particular spatial resolution limits for various types of colored stimuli. We briefly touch on each of these topics in the introduction. In the course of the chapter, we will return to explain how both aspects of human vision are important in the design of various technologies.

Trichromacy

The color-matching experiment coupled with the physiological and anatomical measurements of the three cone types (trichromacy) forms a beautiful story that relates brain and behavior. From the technologist's point of view, abstracting the story into mathematical terms, the color-matching experiment can be summarized by a very brief mathematical expression using simple linear algebra. Suppose the spectral power distribution of a light is $E(\lambda)$. Trichomacy tells us that the visual system makes a linear, three-dimensional measurement of this function. The three measurements can be expressed as the inner product of the cone photopigment absorption functions with the input spectral power distribution. For the L, M and S cones the values are $\langle L(\lambda), E(\lambda)\rangle$, $\langle M(\lambda), E(\lambda)\rangle$ and $\langle S(\lambda), E(\lambda)\rangle$. It is efficient to use matrix notation to express these three inner products. Create a matrix, **A**, whose columns are the three cone absorption functions. The photopigments measure the three values $\mathbf{A}^t\mathbf{E}$. The photopigments do not change their absorption rates to any input signal in the null space of the matrix $\mathbf{A}^t$.

Seen from the technologist's viewpoint, the major goal of the image communications channel can be expressed by a *color-reproduction equation*. At a point in the original scene, the eye encodes three values, $\mathbf{A}^t\mathbf{E}$. When the ambient viewing conditions at the time of capture are the same as the ambient viewing conditions at the time of redisplay, the color-reproduction equation defines how to obtain a perfect color match: the transmission system must capture the original image and display a new image, with spectral composition, $E'(\lambda)$, such that $\mathbf{A}^t\mathbf{E} = \mathbf{A}^t\mathbf{E}'$. This simple equation, is fundamental to the engineering of all color devices. Color engineers must analyze how design decisions influence the ability to satisfy the match in this equation.

Imaging systems never make a perfect match in terms with respect to the color-reproduction equation. Consequently, color metrics (e.g., CIELAB) are an essential tool for analyzing how well the imaging pipeline succeeds. A few moments of thought suggest that certain types of errors are far worse than others. For example, if the original, $\mathbf{A}^t\mathbf{E}$, differs from the reproduction, $\mathbf{A}^t\mathbf{E}'$, only by a common scale factor across the entire image, the two scenes will look quite similar. In that case, the scenes will look rather like one another because it is as if we are looking at the original through dark glasses. If the original and reproduction differ by an additive offset, however, the color appearance in many color regions will be changed and the reproduction will not be satisfactory.

The color-reproduction equation is only accurate when the original and reproduction are viewed in the same general viewing conditions, including size and ambient lighting. If the reproduction covers a very large portion of the visual field, the reproduction context may not be important. On the other hand, if the reproduction covers only a small part of the visual field the context must be taken into account when considering the color-reproduction errors. Attempts to generalize the color-reproduction equation when the viewing conditions at time of image capture and redisplay differ are an important open problem in color engineering.

Spatial Resolution and Color

The spatial and temporal resolutions of human vision are also of great importance in the design of capture and reproduction devices. One reason for their importance is that there will be no improvement in image quality if the reproduction exceeds the spatial or temporal resolution of human vision. Hence, manufacturing cost is sensitive to these limits. There is a second subtler but equally important reason. The ability to control the acquisition and reproduction of spectral

information is quite limited. Often, capture and display devices trade spatial and temporal information for color information. For example, color prints are often made by printing dots of colored inks adjacent to one another on the page (halftoning). When the dots are finely spaced, they blur together and are not individually resolved. Color is adjusted by varying the relative area covered by dots, effectively trading spatial resolution for color control. The spatial and temporal resolution limits of the human eye, and how these depend on color, are a key factor in designing this and other color imaging technologies.

The main properties of human spatial and temporal resolution are described in several reference sources (e.g.,(DeValois & DeValois, 1988; Wandell, 1995)). An important feature of human vision is the poor spatial resolution for certain types of colors. The largest effect arises in the short-wavelength region of the spectrum. In this region, chromatic aberration of the human cornea and lens limits spatial resolution to 6 cycles per degree (cpd) [Chapter Packer and Williams; (Wandell, 1995, Chapter 2)). But, there are other effects, too. Perceptual experiments show that certain patterns seen by the L and M cones can be difficult to detect as well. For example, if the sum of the L and M cone absorptions is constant across the image (L+M = constant), so that the pattern is defined only by a change in the difference (L-M) of the absorptions, spatial resolution is reduced to below 20 cpd (Anderson, Mullen, & Hess, 1991; Mullen, 1985; Sekiguchi, Williams, & Brainard, 1993a, 1993b). An intensity variation, however, in which the value of L+M varies, can be seen at spatial frequencies of 50 cpd or more.

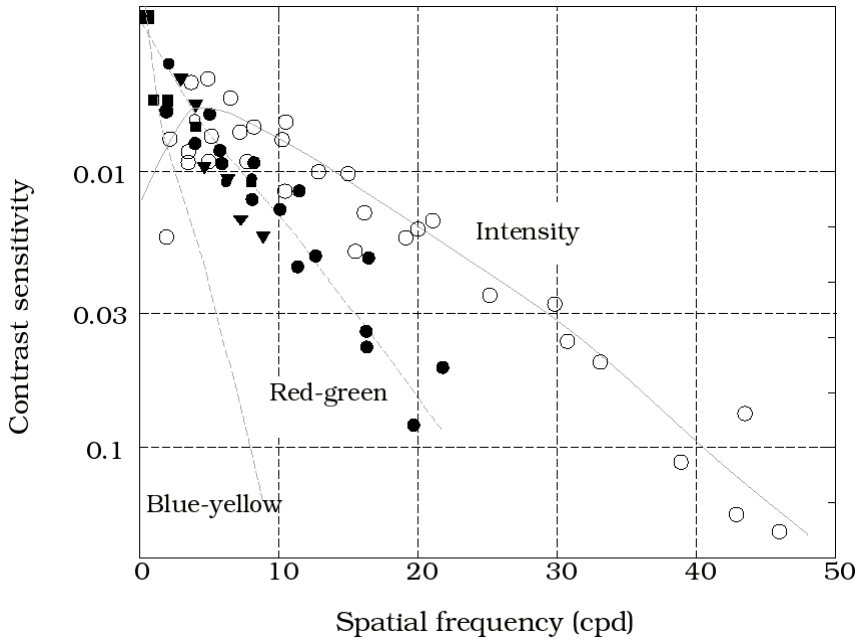

Figure 2. Human spatial contrast sensitivity depends on the color of the pattern. The lightly shaded curves show the general trends for stimuli containing mainly a luminance signal, a red-green signal, or a blue-yellow signal. The symbols are data reported in different papers from several groups (Anderson et al., 1991; Sekiguchi et al., 1993b; Poirson & Wandell, 1993). The figure is adapted from (Wandell, 1999) where further details are provided.

Figure 2 compares human spatial resolution to several types of colored targets. The curves and data show the contrast sensitivity necessary to perceive harmonic patterns at different spatial frequencies. Measurements from several labs are plotted to describe the luminance and red-green spatial sensitivity. The luminance contrast sensitivity function shows a much higher spatial frequency limit and also a pronounced low-frequency decline. The spatial resolution to red-green stimuli falls off at higher spatial frequencies and has no low frequency fall off. The lowest resolution, limited to less than 8 cycles per degree, is for the blue-yellow stimuli. These values show that image capture, image coding, and image display devices require more spatial detail for luminance stimuli than red-green stimuli; very little spatial information about S-cone (blue-yellow) image data is required.

Image capture

Overview

In this section we review general principles of color image acquisition and how these principles are applied to the design of color cameras and scanners. We consider only image capture intended for subsequent display to a human observer, excluding devices designed for computer vision or other physical experiments. Our emphasis is on the capture of wavelength information, though we will consider how this interacts with spatial variables as well.

The general goal of wavelength capture for scanners and cameras is to acquire enough information about the input material to enable creation of a reproduction that will look similar and pleasing to the human eye. Because of the limited sensitivity of the human eye to variations in the wavelength composition, a complete spectral measurement of the image is unnecessary. The very existence of inexpensive color capture devices is possible only because of the savings that are possible because of human trichromacy: image capture devices achieve enormous efficiencies in representing the wavelength composition of the original source by measuring only those portions of the signal that human observers perceive. Capturing more wavelength information is wasteful of resources, needlessly driving up the cost of the device; capturing less will cause perceptually significant differences in the reproduction.

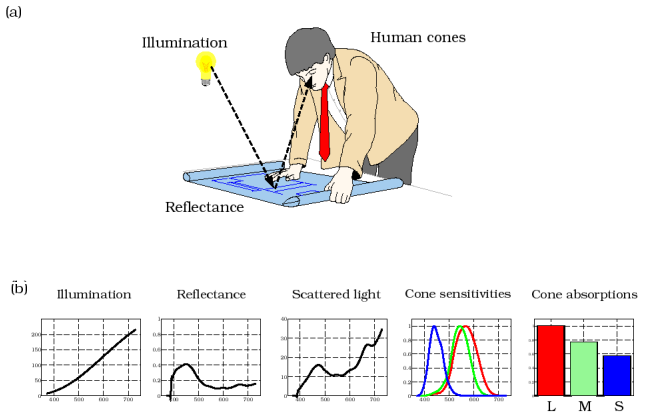

Figure 3. The physical factors governing sensor absorptions. The wavelength composition of the light sent to the eye (the color signal) depends on the ambient illumination and the surface reflectance. The number of photons absorbed in each of the eye's receptor types depends on the relative spectral absorption of the media of the eye and the photopigments within the receptors.

Figure 3 shows the physical factors that determine the wavelength composition of the image and thus the sensor absorptions. These factors are illustrated for capture by the human visual system, but the general formulation applies to other capture devices as well.

Suppose the illumination is diffuse and has radiance $E(\lambda)$ (measured in watts per square meter per steradian per nanometer). Given the particular viewing geometry, the illuminant radiance produces an irradiance at the surface that is specified in terms of watts per square meter per nanometer. The surface absorbs a portion of the irradiance and scatters a proportion back to the eye. The angular distribution of the scattered light depends on the imaging geometry and the properties of the surface. The angular distribution can be measured using goniophotometers (ASTM, 1991) or, more recently, conoscopic measurement systems (Fritsch & Mlynski, 1991; Leroux & Rossignol, 1995; Saleh, 1996). Fully predicting this geometry requires extensive theory and modeling of the surface properties. Because our emphasis is only the wavelength, and not the angular distribution of light, we restrict our calculations to Lambertian surfaces, that is surfaces that scatter uniformly in all directions. As a first approximation, the light emitted from CRTs and many printed material follow Lambert's law. A somewhat better approximation, useful in some applications of illuminant estimation, is the dichromatic reflectance model (Lee, March 18, 1985; Shafer, 1985; Tominaga & Wandell, 1989)

Using a Lambertian model, the effect of the surface on the scattered wavelengths is described by the *surface reflectance function*, $S(\lambda)$, a dimensionless quantity. The scattered light is again

defined by a radiance measurement, and it is given by the product $C(\lambda) = E(\lambda)S(\lambda)$ and (units of watts per steradian per meter squared per nanometer).

After passage through the optics of the eye, an image is formed at the retina. This can be expressed as irradiance at the retina (Rodieck, 1998). The sensor absorptions by the photoreceptors (or camera sensors) are calculated by an inner product between the image irradiance at the retina and the absorption function of the photoreceptor photopigment. For the i^{th} receptor class this value is

$$a_i = \int_{370}^{730} A_i(\lambda)E(\lambda)S(\lambda)d\lambda$$

where $A_i(\lambda)$ is the spectral absorption of the relevant sensor class.

For practical calculations, the wavelength functions are sampled and the integral is replaced by a summation. A matrix can then be used to find the predicted responses as follows. Place the three device spectral absorption functions, $A_i(\lambda)$, in the columns of an *absorption matrix*, **A**. To convert the continuous functions into discrete vectors, the CIE recommends using sampling intervals of 5 nm steps ranging from 380 to 780 nm. Most sensor absorption functions are smooth with respect to wavelength, so that the proper wavelength-sampling rate is limited by the expected variation in the irradiance signals, $C(\lambda)$. Expressing the image irradiance as a vector with the same sampling interval, **C**, the three response values are predicted by the matrix product $\mathbf{A}^t\,\mathbf{C}$.

Visible and hidden portions of the signal

Most cameras and scanners have three sensors. The three wavelength measurements, a_i, represent only a coarse sampling of the wavelength function $C(\lambda)$. Consequently, many different spectral power distributions can cause the same triplet of responses. A pair of lights, (**C, C'**), that cause the same responses in the capture device but that have different spectral power distributions are called *metamers*.

Once the sensor wavelength response functions of a device are known, it is straightforward to specify its metamers. Two lights **C** and **C**' are metamers if $\mathbf{A}^t\,\mathbf{C} = \mathbf{A}^t\,\mathbf{C}'$, or equivalently if $\mathbf{A}^t\,(\mathbf{C}-\mathbf{C}') = \mathbf{0}$. That is, two lights are metamers if and only if their difference falls in the null space of $\mathbf{A}^t$.

Again using conventional linear algebra, the signal measured by any image capture device can be divided into two parts. One part of the signal influences the sensor response. We say this part is *visible* to the device. It can be expressed as a weighted sum of the columns of the sensor matrix, **A**. The part that is *hidden* from the device is orthogonal to the columns of **A**. Metamers differ only in their "hidden" part.

Because image capture devices serve as a substitute for the visual system, it is desirable they encode precisely the same part of the input signal as the visual system. An ideal image capture device must encode **only** those parts of the wavelength signal as the human visual system. Responding to portions of the signal to which humans are blind (e.g., infrared), or failing to respond to portions the human visual system sees, usually will introduce errors into the image reproduction pipeline.

As a practical matter, the sensors in consumer devices do not align precisely, in the sense described above, with human vision. Much of the engineering of capture devices involves compensating for this basic difference in acquisition. These methods will be discussed after describing some of the basic features of practical capture devices.

Scanners for Reflective Media

Figure 4 shows two designs of scanners used to capture signals from printed material. The scanners illuminate the page with an internal lamp. In the one-pass designs shown here, three sensors encode light scattered from the print surface. Most modern scanners use a one-pass design, though original designs were often based on three separate measurements acquired using one sensor and three different colored light sources.

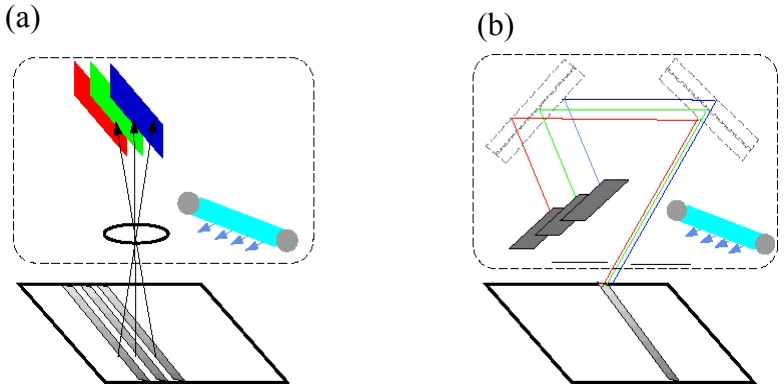

Figure 4. One-pass color scanner designs. In panel (a) the light from a line is measured three times as the three linear arrays pass. Because of this pipeline architecture, the total time to scan in three sensors is only slightly longer than the time to scan with a single (monochrome) sensor. In panel (b) light from a line is separated by means of dichroic mirrors into three wavelength regimes. Each of these separate wavelength regimes is imaged on a different sensor to produce the color image. See text for details.

Figure 4 shows an overview of the scanning elements in two patented designs . In the Canon design a small region in the image is focused onto an array of three parallel sensors (Tamura, 1983). In most modern implementations, the sensors are linear arrays of *charged-coupled devices* (CCDs) whose spectral sensitivities may be altered by the superposition of small colored filters. In this design, as the imaging element of the scanner moves across the document each line is focused, in turn, on one of the three different types of CCD arrays. By the time the entire document has been scanned all three arrays have scanned the entire page. By registering the signals acquired at different times, color images are obtained.

Hewlett-Packard has patented a design in which the capture device acquires signals through a set of dichroic mirrors (Vincent & Neuman, 1989). These mirrors reflect all wavelengths less than a cutoff wavelength and transmit all wavelengths above that cutoff. By arranging two sets of stacked mirrors, light in different wavebands is separated onto three identical linear CCD arrays. Using this method, all of the light analyzed at a single moment in time comes from the same source. Also, almost every photon in the visible range is acquired by one of the sensors. In this design the three sensor arrays are the same; the different spectral tuning of the sensors arises because of the properties of the dichroic mirrors along the light path.

The design of the Hewlett-Packard scanner forces the sensor wavelength responsivities to be essentially block functions, unlike the sensors in the human eye. Consequently, it is impossible to use this design to measure the wavelength spectrum in the same way as the human eye. Even though it is impossible to guarantee that the color of the reproduction and original match, the simplicity and elegance of this engineering design has many practical advantages so that the design is still used in scanners and cameras. We will discuss how problems introduced by the mismatch between the acquisition device and the human eye can be minimized later in this section.

Finally, we conclude with a few of the properties of the capture environment that make the design of scanners relatively simple. First, scanners work in a closed environment: The illuminant is known, unlike the operating environment for cameras or the human eye. Knowledge of the illuminant simplifies color estimation and eliminates problems caused by the need to manage exposure duration and color balancing. Second, scanners mainly acquire data about a limited set of inputs: flat, printed material. It is possible to make a best guest, or even have the user specify, the type of printed material in the scanner. Knowledge about the source of the input can be a significant advantage for color processing. When the properties of the input material are known, better inferences about the input can be made. We will describe this principle at greater length after introducing color acquisition with digital cameras.

Digital Cameras

There are two basic digital cameras designs. In one design, three or four color sensors are interleaved in mosaics within a single sensor array. Figure 5a shows a popular sensor in which four sensors are combined into three (R,G,B) signals. This is accomplished by forming weighted sums of the outputs in various combinations. Figure 5b illustrates the most commonly used mosaic for image acquisition, the *Bayer* pattern [(Bayer, 1973)]. In this design (R,G,B) sensors

are used and the middle-wavelength (G) sensor is present at twice the spatial sampling rate as the red and blue sensors. This design is effective because when the camera data are converted to a digital image, data from the green sensor are critical in defining the luminance representation. The human visual system is more sensitive to the luminance spatial component than the chromatic variations. The increased density of the green sensor improves the spatial sampling of the luminance signal and thus provides information that is better matched to the spatial resolution of the eye.

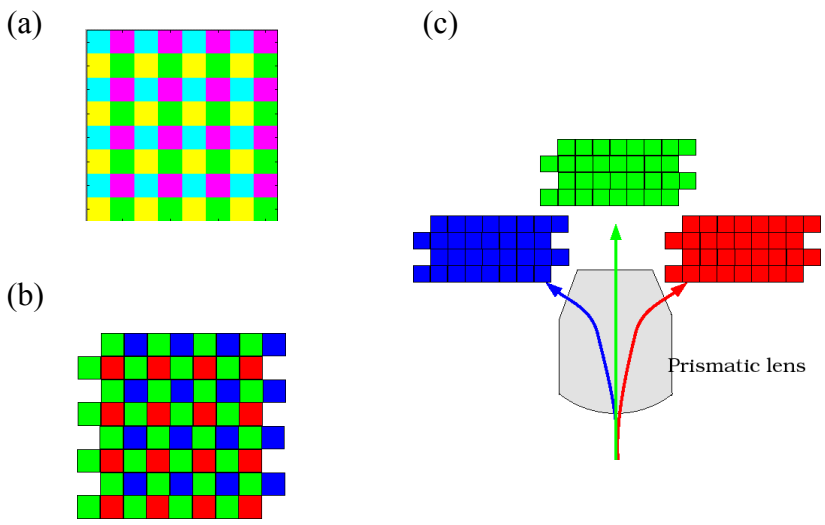

Figure 5. Arrangements of the color sensors in digital cameras. (a) Four color CMYG color filter array. (b) Three color Bayer pattern color filter array (c) Prismatic optics used by Minolta.

A design using prismatic optics is shown in Figure 5c. This design is analogous to the dichroic mirrors used in the Hewlett-Packard scanner. The prismatic optics forms three images of the scene, separated by wavelength bands. These images are each captured by three independent sensor arrays. As in the dichroic mirror design, the three images represent non-overlapping portions of the spectrum so that, again, matching the human wavelength responsivity is not possible.

The sampling mosaic design is usually built with a single monochrome sensor with a superimposed color filter array (CFA). (For a novel development in which the sensor wavelength responsivity is changed electronically see (Silicon Vision, 2000)). In this design camera spatial resolution is traded for color information. To render images captured with this design, the data from the three types of color sensors must be interpolated to form an image with (R,G,B) values at every location. This interpolation process is called *demosaicing*, and a variety of demosaicing algorithms have been proposed (see e.g., (Adams, Parulski, & Spaulding, 1998)).

Demosaicing algorithms are a very important component of the digital camera system design.

Some of the artifacts that can be introduced by using poor demosaicing algorithms are illustrated in Figure 6. The original image is shown in panel (a). This image was sampled to simulate an image acquired by a Bayer color filter array. The reconstructed image from a linear interpolation of the missing values is shown in panel (b). The reconstructed image formed by replicating pixel values is shown in (c). Neither method is acceptable and a variety of linear and nonlinear methods have been proposed and used in products (Adams et al., 1998; Brainard & Sherman, 1995).

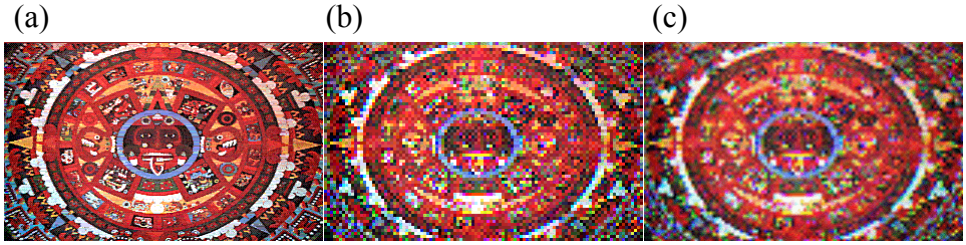

Figure 6. Spatial artifacts caused by demosaicing algorithms. The original image is shown in (a). Interpolation errors when using (b) linear interpolation and (c) pixel replication are shown.

There are three advantages of the prismatic optics approach. First, for the same spatial resolution as the mosaic design, the sensor arrays can be smaller, and it is less expensive to build three smaller sensor arrays than one larger one. Second nearly every photon is captured, producing a very efficient capture device. The mosaic design intentionally permits photons to fall on sensors that will not respond to them. Efficient photon capture is an important element of final image quality, giving the prismatic optics a design advantage. Finally, prismatic optics eliminates the need for demosaicing. The main disadvantage, of course, is the complexity of the prismatic optics, including the packaging and additional electronics needed to accommodate the additional sensor arrays.

Calibration and Characterization

In modern image capture applications, color characterization means finding a method to convert the measured (R, G, B) values into a description based on the CIE tristimulus values (or equivalently the human cone absorptions). For most image capture devices the conversion

process will depend on the specific state of the device; calibration refers to the process of adjusting the device parameters so that the device is in a known state where the characterization is accurate. Because the (R, G, B) responses in a scanner or camera are unique to that device, the measured values are called *device-dependent*. Because the CIE values are not tied to the device, but rather to human vision, these are called *device-independent.*

The characterization process is usually divided into two parts. First, measurements are made of the relationship between light input intensity and scanner or camera output. The function relating these quantities is called the transduction function, also called the gamma function. In most systems, the output follows the same function of intensity no matter what the spectral composition of the input source. The sensors themselves respond linearly to the input signal, and any nonlinearities arise from processing after the initial capture. A simple model for this type of system is given by the formula for a static nonlinearity

$$d = F(\sum s(\lambda)r(\lambda)d\lambda)$$

where d is the digital value from the system, $s(\lambda)$ is the input signal spectral power distribution, $r(\lambda)$ is the sensor spectral responsivity, and F() is a monotonic function. Because F() is a fixed, monotonic, nonlinearity, it is possible to estimate the function and remove its effect. After correcting for F(), the sensor wavelength responsivity can be estimated using standard linear methods. In the following sections, we describe some of the basic features of the nonlinear function used in most cameras. Then, we describe estimation methods for the sensor spectral responsivity.

Dynamic Range and Quantization

The dynamic range of a capture system is the ratio of the light level that produces a response just below system saturation and the light level needed to produce a response just above the dark noise. The device quantization describes how many intensity levels are classified in the digital output. For example, an 8-bit device classifies the input intensities into 256 levels. Each of these factors plays a significant role in determining the camera image quality.

The dynamic range and quantization properties are determined by different parts of the camera system. The dynamic range is an input-referred measurement; that is, its value is the ratio of two input light levels. Signal quantization is a description of the number of signal output levels and does not depend on the input signal at all. Despite this huge difference, one often hears the dynamic range of a device described in terms of the number of bits it codes. This is incorrect. A system that quantizes the output signal to 12 bits can have the same dynamic range as a system that quantizes the output to 8 bits. Two 8-bit systems can have very different dynamic ranges. To link the two measures, one must make a set of assumptions about how the camera designer chose the quantization levels, the properties of the sensor, and other system features. There is no guarantee that these assumptions will be met.

The *dynamic range* of commonly used CCD sensors is on the order of a factor of 500-1000 (60 dB), though devices with much higher dynamic range exist. Operationally, call one standard deviation of the sensor noise variability 1. Then, if the maximum response that we can read out

prior to sensor saturation is 100, the dynamic range is 100. Photomultiplier tubes, an older but still important technology, have a dynamic range in excess of 1000. Dynamic range is commonly described in log units or decibels (20 log units). Hence, it is often said that CCD sensors have a dynamic range of 2-3 log units (40-60 dB) and photomultiplier tubes have a dynamic range of 3-4 log units (60-80 dB) [(dpreview.com, 2000; Janesick, 1997). It is difficult to compare the dynamic range of these devices with that of the human eye; while the responses of these devices is roughly linear with input intensity, the visual system encodes light intensity using a nonlinear (compressive) transduction function (Cornsweet, 1970; Wandell, 1995).

How much is enough dynamic range? If we consider only the surface reflectances of objects, a range of two log units is quite large. This spans reflectances from that of white reflective paper (100 percent) to very black ink (1 percent). The dynamic range of a CCD sensor is adequate to encode the dynamic range of printed material. Slides can represent a somewhat larger range of densities, exceeding two log units, so that in these applications either specially cooled CCDs or photomultiplier tubes may be appropriate. Natural scenes may have even larger dynamic ranges due to (a) geometric relationship between the light source, surface, and viewer, and (b) shadows. Images containing a portion in direct sunlight and a second portion in dark shadow, or a shadow within a shade can span 4 log units or more.

The analog-to-digital converters (ADCs) in the image capture system determine the signal quantization. In many early designs, uniform quantization steps were used, and the most frequently asked question was: How many bits of output are needed to capture the intensity differences seen by the human eye? The main principles of the answer are well understood: To match the intensity discrimination abilities of the human eye, the quantizer must classify intensities present at the finest discriminability. The finest human intensity resolution occurs at intensity levels somewhat lower than the mean image intensity. This demanding intensity region, then, determines the number of classification steps needed by a uniform quantizer,, and a uniform quantizer must classify the image intensities into more than 1024 bins (more than 10 bits). Using this scheme, the quantization step at very high or low intensities is spaced more finely than the visual system can discriminate.

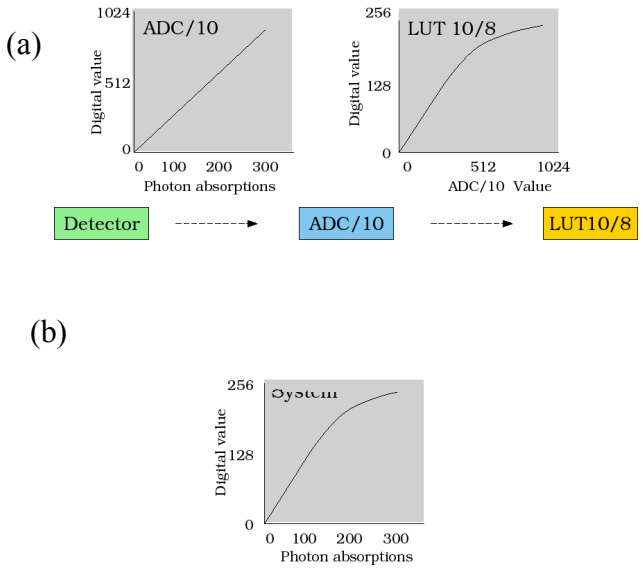

Figure 7. Quantization architecture of a digital camera. (a) Quantization is often performed in two steps. A linear, 10-bit ADC step converts the analog sensor to digital form. Certain types of linear processing, including demosaicing and certain color transformations, are performed in this linear representation. The signal is then reduced to 8 bits by a nonlinear lookup table that compresses the high intensity quantization bins. (b) The overall camera transduction function is compressive much as the visual sensitivity of the eye is a compressive function. Small differences in the dark regions are preserved, but these differences are not preserved in the light regions.

Although the inherent transduction of sensors used for digital imaging is linear, manufacturers often insert a nonlinear post-processing stage as shown in Figure 7a. The two-step process produces non-uniform quantization levels that approximate the discrimination capabilities of the human eye. In the method shown in panel (a), the image is first converted to 10 bits of quantization using a uniform conversion step. Then, a lookup table that produces a final result at 8 bits merges the quantization steps corresponding to high intensity levels. The design requires an extra look-up table beyond the ADC, but this results in an output that is only 8 bits and whose intensity classifications match the human eye more accurately. Reducing the number of bits to represent the image also has beneficial effects on signal storage and transmission Finally, as we shall see later, this quantization scheme is useful when the camera data are combined with a CRT display.

Wavelength

Once the transduction function is known, the sensor responsivity, r(), can be estimated from measurements with a variety of light sources, s(). The corrected digital value is related to the signal and responsivity by the linear equation

$$F^{-1}(d) = \sum s(\lambda) r(\lambda) d\lambda$$

One can only guarantee that the sensor response measures the same spectral image as the human cones if the sensor responsivities are linearly related to the human cones, or equivalently to the CIE Standard Observer functions $\bar{x}$, $\bar{y}$, and $\bar{z}$. That is, suppose there are three sensor responsivities, $r_i(\lambda)$. Then the image capture system will be guaranteed to see the same portion of the visible spectrum as the human eye if and only if there are weights, w_{ij}, such that

$$\bar{x} = w_{11} r_1 + w_{12} r_2 + w_{13} r_3$$

Two similar equations must hold for $\bar{y}$ and $\bar{z}$. When such a linear relationship exists, the sensors are *colorimetric*, and it is possible to guarantee that the sensor (R,G,B) responses can be converted to CIE tristimulus coordinates. The conversion step requires multiplying the (R,G,B) values by a 3x3 linear transformation comprised of the weights. It is possible to determine these weights from only a few measurements. Suppose that we know the (X,Y,Z) values of three color patches, and we know the linearized sensor values, $F^{-1}(R,G,B)$. Then one can determine the linear transformation that maps the (X,Y,Z) values into the linear sensor values.

In general, limitations on the cost of manufacturing make it impractical for the spectral sensitivity of these sensors to match the spectral sensitivity of the cones or the tristimulus functions. It is straightforward to show that when the sensors are not within a linear transformation of the tristimulus functions, there will be pairs of surfaces such that: (a) the sensor responses to the two surfaces are identical, but (b) the tristimulus coordinates of the surfaces differ. For such a pair of surfaces, it is impossible to guarantee a correct estimate of the tristimulus coordinates from the measured responses.

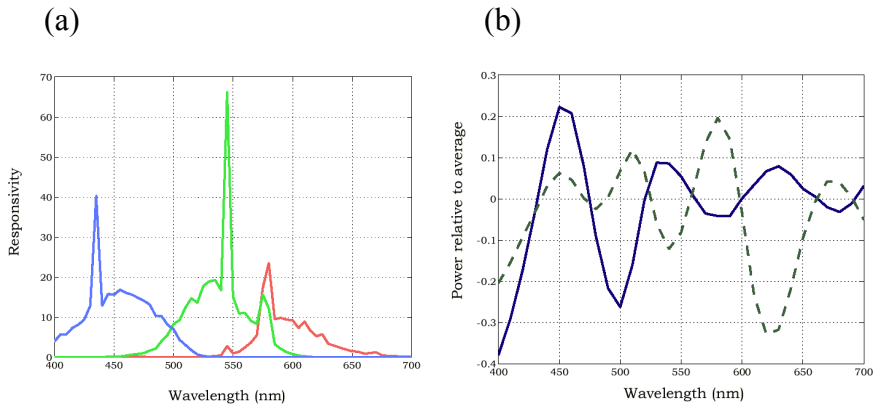

Figure 8. The spectral responsivity of cameras and scanners rarely match that of the human cones. Consequently, differences in spectral variations visible to the human eye may not be visible to the device. (a) Spectral responsivities of the MARC scanner. (b) Examples of modulations of the spectral power distribution that differ to a human observer but result in the same response of the MARC scanner.

Figure 8a shows the sensor spectral responsivity of the MARC system, an elegant device used to digitize paintings (Cupitt, Martinez, & Saunders, 1996; Farrell, Saunders, Cuppitt, & Wandell, 1999; Martinez, Cupitt, & Saunders, 1993). These sensors are not colorimetric, that is they are not within a linear transformation of the human cones. Consequently, there are variations in the spectral power distribution that are visible to the human visual system, but not to the MARC system. Two such variations are shown in Figure 8b. Unless such stimuli can be eliminated from the set of input signals or inferred by other means, it is impossible to guarantee that the sensor values can be accurately transformed into tristimulus values.

Characterization of non-colorimetric sensors

When the color sensors wavelength responsivity do not match the human cones, characterization means making a best estimate of the tristimulus (X,Y,Z) values from the sensor responses, (R,G,B). There are two basic techniques that are used for making this best estimate.

First, characterization can involve making measurements of multiple stimuli to find a functional form that relates the measured (R,G,B) to the (X,Y,Z) values. The functional forms that are commonly used include simple global linear transformations (3x3 matrix), linear transformations that vary smoothly with the input data (interpolation), nonlinear polynomial functions, and methods based on simple neural networks.

Tetrahedral interpolation is an elegant computational method that can be reduced to very efficient hardware computation. This method applies a linear transformation to the measured data, but the linear transform coefficients vary as a function of the measured (R,G,B) values. The method is called tetrahedral because the input space is partitioned into a set of non-overlapping tetrahedra using a characterization data set. The linear transformation applied to any (R,G,B) value depends on the measured characterization values at the vertices of the tetrahedra (Gill, 1999) (Hardeberg & Schmitt, 1997)]. The tetrahedral shape is preferred to cubes because tetrahedral are preserved across transformations from RGB to XYZ representations, so that transformations in either direction can be based on the same control points. Other geometric shapes, such as squares, can be transformed into curved shapes that are problematic when partitioning the response space. A patent has been obtained on the use of tetrahedral interpolation for color characterization (Sakamoto & Itooka, 1981).

The second technique that is helpful for characterization purposes is to specify the properties of the input signals. This can be a very powerful technique, particularly if the input signals fall in a sufficiently restricted set. For example, it is possible to use non-colorimetric camera sensors to estimate the tristimulus coordinates of a color display system that has only three independent primary lights (Horn, 1984; Wandell, 1986).

In many practical applications, for example when mainly processing a particular type of film print, the input material is restricted. Calibrating specifically for this print should lead to a relatively precise system compared to calibrating for arbitrary inputs. Hence, a target designed to span the color range of the print medium is helpful. Such a target, the ANSI (American National Standards Institute) IT8.7, has been standardized and is now provided by various vendors. These targets include examples of particular printed outputs and the manufacturer provides the tristimulus values of these prints. Hence, they form a good basis for calibrating a scanner or camera system that will be used regularly with one of a small set of targets. These targets may be purchased from a number of vendors.

Color Rendering of Acquired Images

Finally, we conclude this section with an observation about the role of camera characterization in the image systems pipeline. Often it is the case that an image is captured under one illumination condition and then rendered for an observer viewing the image under a different illumination. When this occurs, rendering the image with the same tristimulus coordinates as the original will not match the appearance of the original.

To understand the problem, consider that a black surface on a sunny beach may have a luminance of 200 cd/m^2. In a typical windowless office, a white surface will reflect on the order of 100 cd/m^2. Hence, to represent the color black on a display, one would not want to match the original scene tristimulus coordinates. The same principle holds for color variations as for luminance variations.

This issue is not important for scanners, which work in a fixed environment. However, digital cameras are used to acquire images under many different illuminants. One approach to solving this illuminant mis-match problem is to use algorithms that estimate the illuminant at the time of

the image capture. If the illumination is known, then it is possible to make a rough guess of new tristimulus coordinates that will match the original in appearance. This process is called *color balancing*. Algorithms for color balancing are an important part of digital camera design, though a review of the issues is beyond the scope of this chapter. A second approach is to build a model of color appearance and to render the image so that the appearances of the original and rendered images match. The CIE has recently standardized one model in what will probably be a series of color appearance models (Brainard, in this volume; Fairchild, 1997; Luo & Hunt, 1998; TC1-34, 1998).

Electronic Image Displays

Overview

Image rendering technologies can be divided into two major categories: electronic displays and printers. Electronic displays can be further sub-divided into emissive and non-emissive types. Emissive displays are those in which the image-forming element also serves as the source of light, while non-emissive displays modulate some aspect of an extrinsic illumination source. There are currently a large number of display technologies for rendering an electronic image, but two types dominate the market: the cathode ray tube (CRT) is the dominant emissive technology while the liquid crystal display (LCD) is the pervasive non-emissive technology. Printing is a non-emissive rendering technology.

We have separated the discussion of image displays into emissive and non-emissive technologies because the methods used to control light intrinsic to the device and those used to control transmitted or reflected light from an external source differ significantly. In this section we describe the basic principles of CRTs and LCDs. While there are many ways to utilize liquid crystals to modulate light and create a display device, we focus our attention on the ubiquitous transmissive, twisted-nematic (TN) color LCD that contains a separate but integrated illumination source. The basic principles of color synthesis and color control for these LCDs and CRT devices are similar and will play a role in most, if not all, of the display technologies that are envisioned over the next decade. We show how these color synthesis principles are used to satisfy the color-reproduction equation, described in the introduction to this chapter. We also review the general methods and computational tools that are used to characterize such electronic display devices.

CRT Devices

The venerable CRT has dominated the display market for the past 45 years, despite repeated claims of its imminent demise. The principal technology for generating color in direct-view CRTs is the shadow-mask CRT, illustrated in Figure 9.

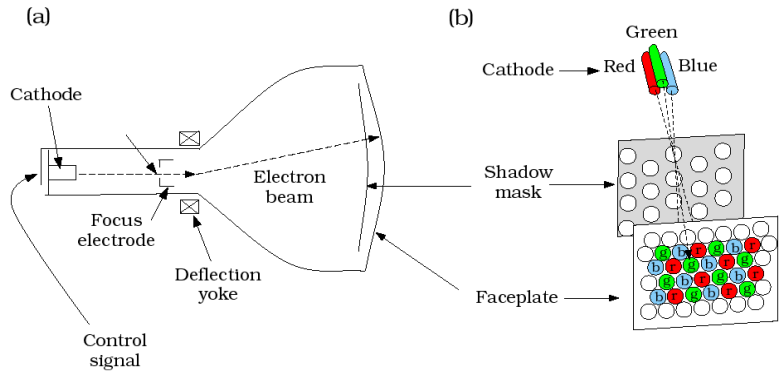

Figure 9. The shadow mask CRT. The basic architecture of a shadow-mask color CRT display is shown in (a), and the geometric relations between the cathodes, shadow-mask and phosphor-coated faceplate in determining color selection are shown in (b).

In this design, the three electron guns (one for each primary color phosphor) house a thermionic cathode that serves as a source of electrons. Video input voltages are applied to each electron gun assembly, which includes control grids for modulating the beam current flowing from the cathodes as well as electrodes to accelerate, shape and focus the electron beams on the phosphor-coated faceplate. The applied video signals cause the intensity of the electron beams to vary synchronously as the beams sweep out a raster path. The electrons that pass through the shadow-mask apertures excite the R, G and B phosphors. The geometry of the apertures is coordinated with the phosphor pattern on the CRT faceplate. Electron absorptions cause the phosphors to emit light in a process called cathodoluminescence. As illustrated in Figure 9b, color selection and registration are determined by the position of the electron guns and their geometric relations to the shadow-mask and phosphor-coated faceplate.

Although there are several places in the CRT image pathway where sampling artifacts are introduced, sampling artifacts are minimized because the electron beam cross-section is approximately Gaussian and spans several groupings or triads of color phosphor dots. This shape imparts a low-pass spatial filter to the signal path so that the sampling rate does not introduce any appreciable spatial aliasing (Lyons & Farrell, 1989).

In designing CRTs, the positions of the electron guns, shadow-mask apertures, and phosphors must all be taken into account and many configurations are currently available. In recent years there has been a trend toward the use of in-line configurations of electron guns, in which the electron beams are arrayed along a line rather than in a triangular configuration, due to their simpler alignment and deflection considerations (Sherr, 1993; Silverstein & Merrifield, 1985). In

addition, slotted-mask and strip-mask (e.g., the familiar Sony Trinitron tube) color CRTs which use continuous vertical RGB phosphor stripes on the CRT faceplate have become popular. Current technology has enabled mask-pitch and associated phosphor component pitch (i.e., the center-to-center distance between RGB phosphor groupings or between like-color phosphor components) to be reduced to the range of 0.25 to 0.31 mm. (Lehrer, 1985; Sherr, 1993; Silverstein & Merrifield, 1985).

The CRT design reduces spatial resolution and photon efficiency in exchange for color. It is important that the spatial patterning of the red, green and blue phosphors be invisible under normal operation. At a nominal display viewing distance of 61.0 cm, this spacing translates into a range of subtended visual angles from approximately 1.41 to 1.75 arc min. Given the resolving capability of the chromatic channels of the human visual system (also see Chapters 2 and 6 of the present volume), this spacing is sufficient to ensure reliable spatial-additive color synthesis (Glenn, Glenn, & Bastian, 1985; Mullen, 1985; Schade, 1958; VanderHorst & Bouman, 1969).

Color CRTs are inefficient compared to monochrome displays because of the shadow mask. The presence of the mask reduces the percentage of electrons that result in a electron absorption and subsequent photon emission, and such masks are not needed in monochrome displays. The market has demonstrated that to most consumers the value of color information is the tradeoff.

LCD Devices

Direct-view color LCDs are commonplace in portable computer and miniature color television applications. They are beginning to penetrate the market for larger, high-resolution, high-performance color displays. Figure 10 shows the major optical components of an active-matrix addressed transmissive TN LCD. The color LCD is composed of a backlight illumination source, diffuser, rear linear polarizer, glass sheets with transparent thin-film indium-tin-oxide (ITO) electrodes and thin-film transistors (TFTs), optically active layer of birefringent LC material, absorbing thin-film color selection filters, and a front polarizer. The operation of the LCD depends mainly on the polarization properties of light. Light from the illumination source is plane polarized by the rear (entrance) polarizer. The light passes through the liquid crystal (LC) layer where its polarization state can be altered. Depending on the polarization state after passing through the LC, the light is either absorbed or transmitted by the front (analyzing) polarizer.

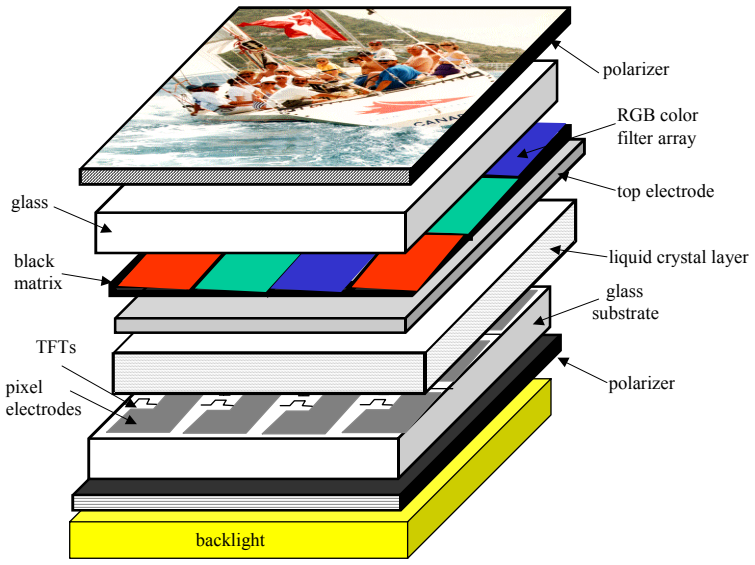

Figure 10. The architecture of a transmissive color LC display employing active-matrix addressing is shown. See text for details.

Three components have the principle effects on the colorimetric and photometric characteristics of the emitted light: the spectral power distribution (SPD) of the illumination source; the spectral transmission of the thin-film color selection filters; and the spectral transmission of the LC cell (Silverstein, 2000) The largely clear optical elements, such as the glass containing the ITO electrodes only modify the spectral composition of the light by a small amount. Along the imaging path, prior to reaching the human observer, each optical component must be characterized by its full emission or transmission spectrum.

The backlight illumination for most direct-view color LCDs is either a hot-cathode (HCF) or a cold-cathode (CCF) fluorescent lamp. Fluorescent lamps have the advantages of high luminous efficiency and the ability to tailor the SPD of the lamp via the selection and mixture of individual phosphor components and their proportional contributions to the total phosphor blend. Tri-band phosphor mixtures are typically employed to improve color performance for these lamps. The final emission spectra are the weighted sum of the three phosphor emissions plus energy at the mercury emission lines.

Direct-view color LCDs typically use thin-film color absorption filters to determine the spectral composition of the three primary lights. Only a limited variety of dyes and pigments compatible with LC materials and the LCD manufacturing process exist. Once the filter materials are selected varying the filter thickness and dye concentration can make some adjustments to their spectral transmission, though the value of these parameters must fall within the limits of the thin-film deposition processes. If the spectral transmission of a set of reference filter materials is known, and the dye or pigment in concentration is known to follow Beer's Law within the range of concentrations used, then the spectral transmission of the filter material at other dye

concentrations and film thickness may be estimated via the use of the Beer-Lambert Laws (Wyszecki & Stiles, 1982) (Silverstein & Fiske, 1993).

The most complex spectral component of the system is the LC layer. The spectral properties of the LC cell depend on a variety of material parameters and the geometry of the LC cell. In addition, the spectral transmission depends on the display voltage (i.e. luminance or gray level) and the direction of light propagation (Silverstein & Fiske, 1993)

Liquid crystals (LCs) are complex, anisomeric organic molecules that, under certain temperature conditions, exhibit the fluid characteristics of a liquid and the molecular orientational order characteristics of a solid (Collings, 1990). A consequence of the ordering of anisomeric molecules is that LCs exhibit mechanical, electric, magnetic and optical anisotropy (Penz, 1985; Scheffer & Nehring, 1992). Most LC materials are uniaxial and birefringent. Uniaxial materials possess one unique axis, the optic axis, which is parallel to the liquid crystal director (i.e., the long axis of the molecules). The anisotropic nature of LC materials gives them the optical property of birefringence, which refers to the phenomenon of light traveling with different velocities in crystalline materials depending on the propagation direction and the orientation of the light polarization relative to the crystalline axes (Collings, 1990). For a uniaxial LC, this implies different dielectric constants and refractive indices for the unique or "extraordinary" direction and for other "ordinary" directions in the LC material.

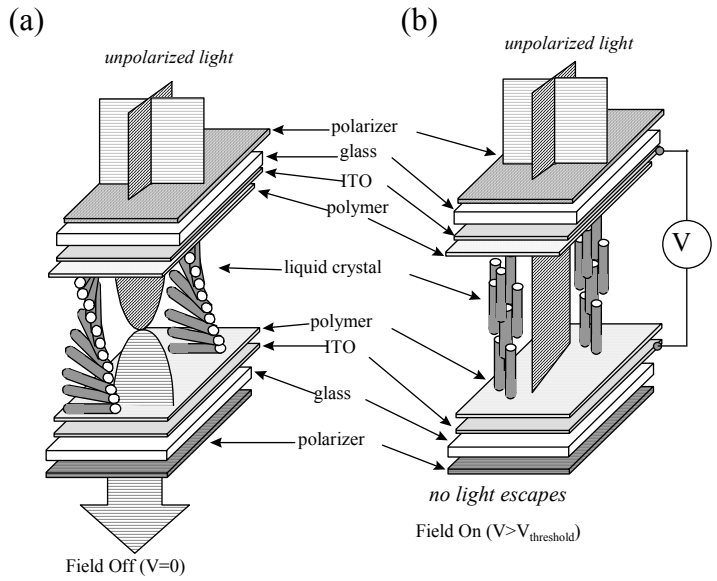

Figure 11. The composition of a TN LCD cell is shown. Applying voltage to the liquid crystal controls the transmission of light through the cell. These voltages alter the polarization of light passing through the cell. In (a), zero voltage is applied so that the twist structure is left undisturbed and rotates the polarization of the light 90^o where it passes through the exit polarizer. In (b) a supra-threshold voltage is applied such that the LC twist structure is destroyed, leaving the initial polarization of light intact as it passes through the LC layer where it is finally absorbed by the orthogonal exit polarizer.

As mentioned above, the predominant LC cell configuration for high-performance color LCDs is the TN cell, whose basic principles of operation are illustrated in Figure 11. An entrance polarizer linearly polarizes the source light. In the field-off state (panel a), with no voltage applied, the LC layer optically rotates the axis of polarization of the incoming light. The typical twist or rotation angle used for most TN LCDs is 90^o, although other twist angles may be used to achieve certain desired optical characteristics (Scheffer & Nehring, 1990, 1992). In the field-on state (panel b) the dielectric anisotropy of the LC material enables the applied electric field to deform the LC layer, destroying the twisted structure and eliminating the LC birefringence for normally incident incoming light. The LC layer does not rotate the axis of polarization of the incoming light. The difference in polarization state is the key variable for determining the display output.

After passage through the LC layer, the exit polarizer or "analyzer" analyzes the polarization state of light exiting the LC layer. Light polarized parallel to the analyzer polarization vector is transmitted, light polarized perpendicular to the analyzer polarization direction is extinguished, and light polarized at intermediate angles follows Malus' Law; $I' = I \cos^2 \theta$, where (I) is the intensity of polarized incident light from a first linear polarizer, (I') is the intensity of light output and (θ) is the relative angle between the orientations of the two linear polarizers (Collett, 1993). Two configurations of TN cell entrance and exit polarizers are used. LCDs that use crossed rear

and front polarizers operate in the normally-white (NW) mode. LCDs with parallel polarizers operate in normally-black (NB) mode. The TN cell of Figure 11 operates in the NW mode.

The precise polarization state of light exiting the LC cell depends on several liquid cell parameters, including the LC birefringence, LC layer thickness, twist angle and importantly for us, the wavelength of the light. As a consequence of this dependence, the transmitted spectrum (and thus color appearance) of the display can vary with viewing angle. This variation is an important consideration in the ultimate color performance of LCDs and models of the influence of these parameters are an important element in LCD design (Silverstein, 1991; Silverstein & Fiske, 1993). Various methods for compensating for this wavelength-dependence have been developed.

Other LCD Display Technologies

There is increasing use of LCs in color projection systems. A major advantage of LCD color projectors over CRT based systems is the ability to separate the image-forming elements and the illumination source. This permits the development of projectors with very high intensity output and thus extremely large projected images. Some of the key problems with LCD direct-view displays, such as viewing angle effects of the transmitted or reflected light, are eliminated in projection systems. The observer does not directly view the image, so LCD viewing angle effects are eliminated. Finally, the relatively small size of the LC image-forming elements permits a very compact optical design. Given the advantages of color projection systems based on LC technology, the market demand for these large-screen color displays continues to have substantial growth.

A second trend is the development of subtractive color displays. This approach offers the advantages of very high image resolution since each addressable pixel is capable of generating the full color gamut of the display, unlike typical color CRTs or LCDs which rely on additive spatial color synthesis of R, G and B pixels and thus sacrifice two-thirds or more of the available pixels to serve the color synthesis function. The development of subtractive color LCDs is an important technology initiative for full-color head-mounted displays, in which the image source must be small and very high pixel densities are required to support good color image resolution across a wide field of view.

Current embodiments of subtractive color LCDs use three LC layers, each controlling the spectral transmission of a portion of the visible spectrum from a broadband illuminant. Thus, each LC layers acts as an electronically controlled color filter that is analogous to ink (see next section). Three different approaches to subtractive color LCDs have been developed. In the first, dichroic dye molecules are suspended within the LC material in what is typically called a guest-host LC cell (Silverstein & Bernot, 1991). Subtractive primaries (cyan, magenta and yellow dyes) are used in the three respective LC cells. When the LC material is switched by the application of an applied electric field, the elongated dichroic dye molecules are reoriented along with the LC material, causing different degrees of spectral filtering in each cell as the LC director orientation is varied between alignment parallel and perpendicular to the cell surfaces. The second approach uses three TN LC cells with colored linear polarizers as the analyzers for each cell (Plummer, 1983). The cells are arranged such that each cell rotates the plane of

polarization of light entering the cell into the entrance plane of the next cell in the stack. The linear polarizers employed as the analyzers in this novel configuration utilize cyan, magenta and yellow dyes instead of the embedded iodine crystals found in standard, achromatic linear sheet polarizers. Each TN LC cell operates as a typical TN light valve, but instead of varying the transmission between an achromatic light and dark state the output of each cell varies from an achromatic state to the state produced by the spectral transmission of the respective dye. The stack of three such TN LC cells constitutes a full-color subtractive LCD. In a final approach, three LC cells configured as electrically controlled birefringence (ECB) cells are used to provide spectral shaping which approaches the subtractive color primaries (Conner, 1992).

Prototype subtractive LCDs yielding excellent color performance have been demonstrated for the guest-host configuration using appropriately selected dichroic dyes and for the stacked TN cell approach with high-quality color polarizers. The three-layer ECB cell configuration has been used in color projection panels for overhead projectors for a number of years, although good color saturation and precise color control have been difficult to achieve with ECB cells. Thus, while high-performance subtractive color LCDs are still in their early stages of development, their technical feasibility has been demonstrated. The potential advantages of subtractive color displays are compelling, and the technology will surely find a place in the future of electronic color imaging.

Color LCD technology is still relatively new and evolving at a rapid pace. Continuing advances in all key LCD technologies; LC materials, optical systems configurations, illumination sources, color filters, optical compensation techniques, driver chips and LC controllers, promise to raise the level of performance for each successive generation of color LCDs. Research into the spatial and temporal imaging characteristics of color matrix displays, including the effects of color mosaic patterns and methods of luminance quantization, remains a highly active area of investigation (Silverstein, Krantz, Gomer, Yei-Yu, & Monty, 1990). As the evolution of color LCD technology progresses, those concerned with electronic color imaging can look forward to brighter, higher contrast displays that exceed the color performance and image quality of today's color workstation standards.

Display Characterization

The purpose of display characterization is to specify the relationship between the values that control the input to the display and the light emitted by the display (Berns, Gorzynski, & Motta, 1993; Berns, Motta, & Gorzynski, 1993; Brainard, 1989). Hence, while the device physics of CRTs and transmissive color LCDs are completely different, the principles and methods of display characterization are quite similar (Silverstein, 2000). In this section we describe the principles of characterization of a specific display at a specific point in time, and we provide example measurements.

A digital frame buffer controls most displays. The intensities emitted by the three primaries comprising each pixel are specified by three digital values, (R, G, B). The potential scope of a complete characterization is enormous. The industry standard for color applications allocates 8 bits of intensity control for each display primary and a total of $2^{8\,(3)}$ or approximately 16.8 million combinations. Multiplied by roughly a million pixels on the screen, and taking into account interactions between pixels, makes it impossible to perform an exhaustive

characterization. Instead, characterizations are always based on simple models of the device that make powerful assumptions about the relationships between the display primaries and the spatial interactions between pixels.

With respect to color control, the most important modeling assumption is that the primary intensities can be controlled independently. Specifically, a control signal to the red primary will produce the same emission no matter what state the green or blue primaries. This assumption, *primary independence*, can and should be empirically verified during characterization. We recommend working only with display systems that satisfy primary independence. A second important assumption is that the SPD of the display primaries are invariant as their intensities are changed. If the SPD of the primaries change with intensity level, characterization becomes more complex. If these two assumptions hold, the characterization task is simplified and only a few dozen measurements need to be made.

There are many other issues that one might be concerned about in characterization. The spatial and temporal distribution of the signals may interact with the primary levels; the display may not be perfectly stable across time or with temperature; there can be variations across the surface of the display or with viewing angle. In general, complete characterization is not possible and some assumptions about these effects must be made and hopefully evaluated.

In many scientific studies, experiments are often based on only a small number of stimuli. In such cases, it is best to measure each of the stimuli individually. If too large a set of stimuli is used to measure them all, the first question to check is primary independence. To evaluate independence, measure the light emitted by the R primary alone, the G primary alone, and the sum of the R and G primaries (R+G). The sum of the R and G measurements alone should equal the measurement of R+G. Then, stimuli with spatial and temporal configurations similar to the ones that will be used in the experiments should be calibrated.

To specify the characterization process, we must measure the relationship between the digital control signals (frame buffers), the light emitted by each of the primaries (primary spectra and transduction), and the effect this light will have on the human observer (tristimulus and chromaticity values). An excellent review of the principles and issues of display characterization may be found in (Brainard, 1989). A discussion can also be found in (Wandell, 1995) and publications from the Rochester Institute of Technology (Berns, Gorzynski et al., 1993; Berns, Motta et al., 1993), and CIE technical reports (CIE, 1996).

Frame buffers

The primary roles of the frame buffer are the storage, conditioning and output of the video signals that drive the display device. The industry standard for color applications allocates 8 bits of intensity control for each display primary or approximately 16.8 million discretely addressable colors. The match between the sampled values and human color sensitivity is imperfect, however. Consequently, not all displayed colors can be discriminated from one another, and many colors that differ by a single bit in their digital representation are significantly above the threshold discriminability of the eye. This results in various types of color artifacts, such as contouring artifacts on shaded graphics. High quality rendering and other demanding

applications, such as psychophysical measurements, can require finer (10 or 12 bits) control over the primary intensity level.

Many of the features of the frame buffer are determined by cost considerations, and the primary costs relate to the size and speed of the frame buffer memory. Consider a display system with 1280 x 1024 addressable pixel resolution and each pixel controlled by a 24-bit value. This system requires 4 million bytes of (fast) memory to represent the frame. An economical alternative is the look-up table (LUT) architecture. In this design, the intensity levels of the primary colors are controlled by a list of entries in a look-up table. Hence, a single number represents the three voltage levels that control the primary intensities, say between 0 and 255. At display time, the system retrieves the primary values from the LUT. In this way, each pixel is represented by one 8-bit quantity. While the precision of intensity control is established by the 8-bit resolution of the digital-to-analog converters (DACs), the number of entries in the LUT limits the total number of colors available for simultaneous display.

One benefit of a LUT design is to reduce image memory. There are other benefits as well, for certain types of display conditions. For example, LUTs provide an efficient means for implementing image operations that depend only on display value, and not on display position. To alter the image luminance or contrast one can re-write the 256 LUT entries rather than the entire image buffer. In this way various image processing operations, spanning the control of drifting and flickering patterns, can be implemented by controlling only the LUT.

Primary Spectra and transduction

Figure 12a shows the SPDs of the primary phosphor emissions in a modern, high-performance color CRT monitor. The phosphors in this particular display were from the P22 family. There are many different phosphors available to manufacturers. Notice that the red phosphor SPD has several discrete spikes. Such spikes are not commonly found in nature, and consequently the CRT emissions almost never match the spectral power distribution found in the original scene. The color match can only be arranged because of the analysis, based on the color-matching experiment, of the eye's inability to distinguish between different spectral power distributions (metamerism).

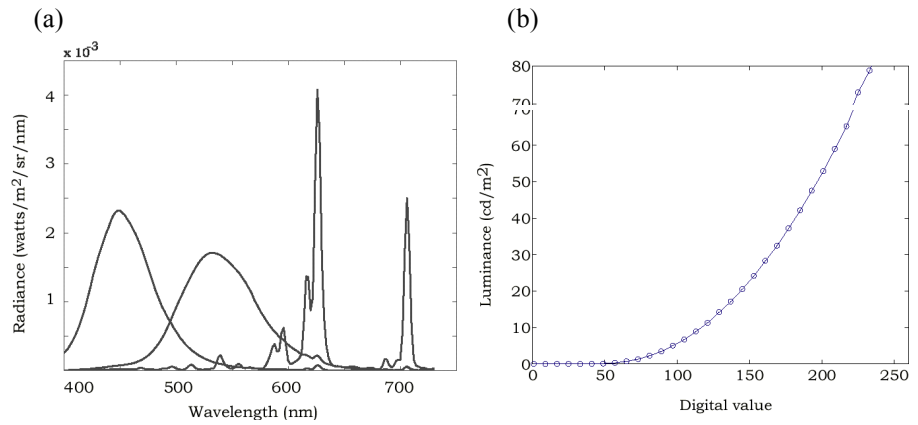

Figure 12. (a) The spectral power distributions of the three primary phosphor emissions in a CRT display. (b) The transduction function relating digital value of the frame buffer to screen luminance for the same display is shown.

Figure 12b shows a typical transduction function for a CRT. The digital count, shown on the horizontal axis, is related to the primary luminance by a nonlinear function. The nonlinear relationship is a characteristic of the CRT tube itself, not the surrounding circuitry. Typically, the relationship is close to a power function, Luminance = α(Digital Count)$^{\gamma}$ + β, or one of a set of other similar equations. Hence, because of the common use of gamma for the exponent, the transduction function is often called a "gamma" curve. The value of the exponent differs between displays, but is generally between 1.7 and 2.2 for CRT displays, and its value is not under user control. User-controlled knobs manipulate the values of the gain (α) and offset (β) parameters. Extensive discussions of this curve and its implications for imaging can be found in the literature (Poynton, 1996).

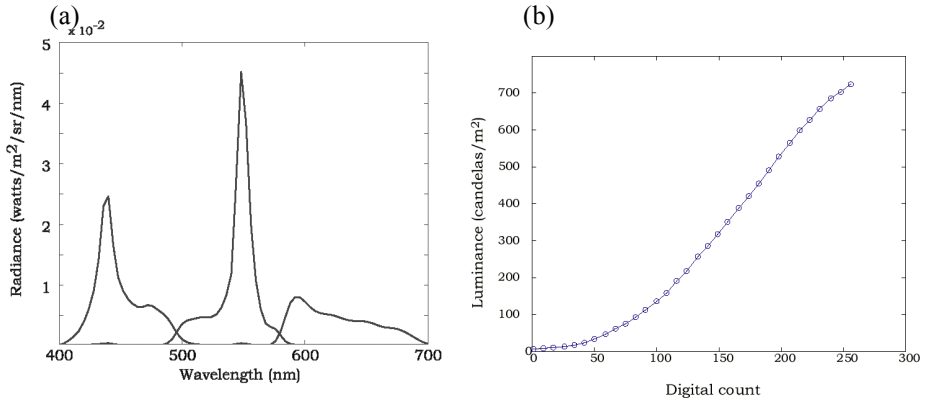

Figure 13. (a) The spectral power distributions of the three primaries in an LC display. (b) The transduction function of the same display.

Figure 13a shows the spectral power distributions of three primaries in a particular LCD display. (Displays vary considerably). The spikes in the distributions are due to materials placed in the fluorescent backlights. The peaks of the backlight emissions are designed to fall at the centers of the passbands of the thin-film color filters that are part of the LCD assembly. Notice that the shapes of the blue and green primary spectral power distributions are narrower than the corresponding distributions for the CRT. This results in a larger range of displayable colors, as will be described below.

Figure 13b shows the transduction function of an LCD. The relation between the digital frame buffer value and the light intensity is nonlinear, as for the CRT. The relationship is not a natural consequence of the LCD display physics, but rather it is arranged by the manufacturer to be close to that of the CRT. For many years image data have been adjusted to appear attractive on CRT displays. Were the LCD transduction different from the CRT function, these images would not appear attractive, and customers would be dissatisfied.

The difference between the CRT and LCD properties raises the following question: How can one be sure that image data will appear as intended without knowing the display? An industry consortium, the International Color Consortium (ICC), has worked on proposals to solve this problem. The ICC recommends a solution in which display device characteristics are represented in a *device profile* file. Software manufacturers are urged to include routines that read and interpret the device profile. Hardware manufacturers and third-party vendors are urged to provide enough characterization data to permit accurate color rendering. The device profile is based on the CIE system (e.g., tristimulus coordinates or CIELAB coordinates) and several basic operators. To learn more about these proposals, which continue to evolve, you may wish to

consult the Internet at <http://www.color.org>.

Tristimulus and chromaticity values

For most applications, it is not necessary to know the complete SPD of light emitted from a display. Rather, it is enough to know the effect of this light on the human cones. Usually, this is specified using CIE standard tristimulus coordinates. There are three major sets of color-matching functions (CMFs) used in display colorimetry (Wyszecki & Stiles, 1982). First, the CIE 1931 2-degree standard observer CMFs, which are appropriate for the calculation of tristimulus values when color fields spanning less than 4 degrees are used. Second, the CIE 1964 10-degree supplementary standard observer CMFs, which are appropriate for color fields > 4 degrees and reflect the shift toward increased short-wavelength sensitivity for large color fields. Finally, the Judd modification of the CIE 1931 2-degree CMFs correct for underestimates of the short-wavelength photopic sensitivity (i.e., < 460 nm) for the original 1931 CMFs. This last set of CMFs are important in basic color vision research and are the basis for a number of linear transformations to the cone absorption curves. They also serve as the basis for the CIE 1988 2-degree supplementary luminous efficiency function for photopic vision (CIE, 1990).

The tristimulus values of the three primary color phosphor emissions can be computed from the color rendering equation described in Section XXX. Suppose that the columns of the matrix **C** contain the color matching functions and the columns of the matrix **P** contain the spectral power distributions of the three primary lights at maximum intensity. The tristimulus coordinates of the three primaries are contained in the columns of the matrix product $\mathbf{C}^t\mathbf{P}$**.** To predict the tristimulus coordinates of a light emitted when the frame buffer values are **v**' = (r',g',b'), first correct for the nonlinear transduction function, Fi(). This produces three linear primary intensity values, **v** = (r,g,b) = (Fr(r), Fg(g), Fb (b)). The tristimulus coordinates, **c**, are $\mathbf{c} = \mathbf{C}^t\mathbf{P}\mathbf{v}$**.** To find the frame buffer values that will display a given set of tristimulus coordinates, **c,** invert the calculation to find $\mathbf{v} = (\mathbf{C}^t\mathbf{P})^{-1}\,\mathbf{c}$ and then apply the inverse of the transduction value to obtain **v'**. If the resulting values are negative or exceed the maximum intensity of one of the primaries, the desired color is called *out of gamut*.

The tristimulus calculations specify the part of the emitted light that is visible to the human observer. Two lights presented in the same context that have the same visible components will have the same color appearance. Even two lights with the same spectral power distribution may appear different when presented in different contexts (Smith & Pokorny, In press).

It is common to express the tristimulus coordinates in a form that captures separately the luminance and color of the signal. To do this, the values (X,Y,Z), are converted to the form (Y, x, y) = (Y, X/(X+Y+Z), Y/(X+Y+Z)). The Y value is luminance and the values, (x,y), are *chromaticity coordinates*. These coordinates are invariant with the intensity of the signal. Doubling the intensity of a light doubles its Y (luminance) value, but leaves the (x,y) chromaticity coordinates unchanged.

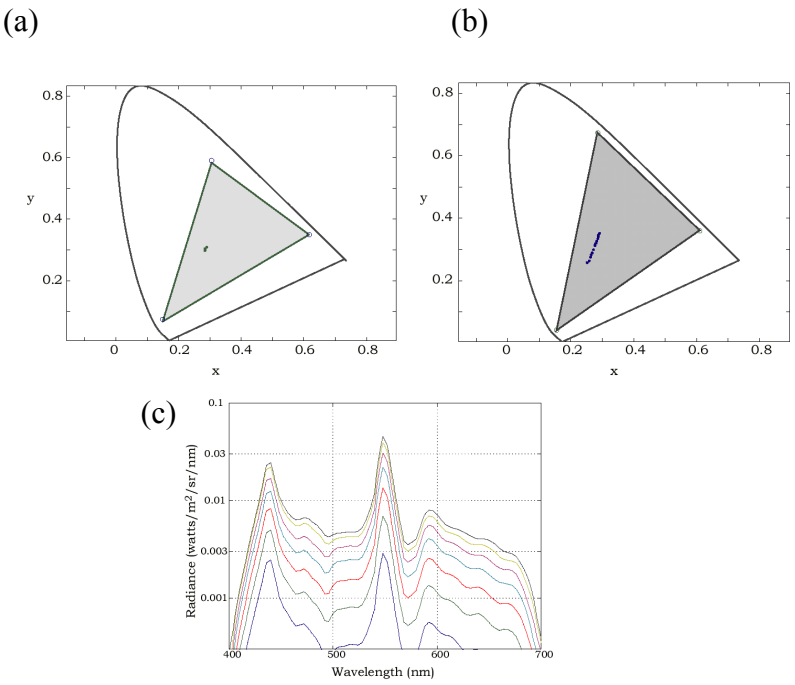

Figure 14. The color gamuts of (a) a CRT display and (b) an LC display plotted in the (x,y)-chromaticity diagram. The chromaticity coordinates of a series of grays are shown as the points in the middle of the graphs. (c) The spectral power distribution of the LCD gray series.

The three pairs of chromaticity coordinates (one pair for each primary) define the range of colors that can be produced by the display. Figure 14 shows the chromaticity coordinates of each of the three primaries in a CRT display (panel a) and an LCD display (panel b). The triangle that connects these three points defines the device *color gamut*. The gamut represents the range of colors that can be displayed by the device. The smooth curve on the graph denotes the chromaticity coordinates of spectral lights. This curve is called the *spectrum locus* and all lights must have chromaticity coordinates within this area.

The color gamut of LCDs can be made larger than that of CRTs. This is because the choice of backlights and thin film color filters in LCDs offer display designers additional degrees of freedom, including primaries with narrower spectral distributions that fall closer to the spectrum locus than the broadband phosphors in CRTs. The difference in color gamuts between devices poses a problem for color reproduction. Suppose that an image is designed on an LCD, but we wish to display it on a CRT. It may be impossible to reproduce a light with the same tristimulus coordinates. This is called the *gamut-mapping* problem. How to compensate for mismatches in the gamut between display devices or between displays and printers is an active area of research.

Although direct-view LCDs are generally brighter and can have larger gamuts than CRTs, they do have one significant problem. A very desirable feature of a display is that scaling the digital counts of the frame buffer should preserve the chromaticity coordinates. Perhaps the most important values to preserve (for customer satisfaction) are the gray series, comprised of the digital values, say (10,10,10), (20,20,20) and so forth. Figure 14c shows the spectral power

distribution of an LCD at a series of gray values. The chromaticity values of a gray series are shown in the center of the panels (a) and (b) for the CRT and LCD. The chromaticity shifts are much larger for the LCD than the CRT. This is caused by a change in the SPD passed by the liquid crystal layer and the polarizers. Panel (c) shows the SPD of the gray series at several mean levels. Were the SPDs invariant with level, the curves would be shifted copies of one another on this log radiance axis. The curves are not shifted copies, and notice that there are significant differences in the spacing in the long- and short-wavelength regions compared to the middle-wavelength regions. These differences occur because the LC polarization is not precisely the same for all wavelengths and also as a result of spectral variations in polarizer extinction. For the viewer, this results in a shift in display chromaticity of the primaries when they are presented at different intensity levels. It is possible to compensate for these changes using algorithms described by Spiegle and Brainard (Speigle & Brainard, 1999).

Printing

Introduction

Reflection prints are a convenient and flexible means of viewing and exchanging images. To view a reflection print, one needs no energy apart from the ambient light. The print itself is light and easy to transport. Printing can be applied to many different substrates, making it convenient to mark all kinds of objects. Look around the room you are in. You will find printing on many of the objects near you, perhaps even your clothes. The printing industry is enormous, and managing the color appearance of printed copy is an important part of that industry.

Improvements in printing come from three basic sources: the ability to create papers and inks with improved ink absorption properties; the ability to control the placement of the inks on the page; and, the ability to predict the perceptual consequences of the first two processes. Over the last two decades there have been advances in all three areas, though perhaps the most impressive advances have been in the ability to control the placement of ink on the page. In this chapter we will be concerned mainly with methods of creating colored prints under the control of digital computers that is *digital printing*.

Our review of color printing is separated into two parts. First, we will introduce some of the concepts used by the color printing community. Because of their very separate historical developments, the printing and display communities color reproduction terminology differ even when the concepts are closely related. We will introduce the concepts and terms used by the printing community but with an additional emphasis on showing how the emissive and reflective display methods are designed to obey the same color reproduction equations.

Second, we describe the ideas behind two printing methods, *continuous tone* and *halftone* printing. In continuous tone printing the printed page is covered with a very fine array of ink drops. The droplet density, but not position, is controlled by the printing method. Hence, the control of continuous tone printing is conceptually similar to controlling the appearance of overlaid sheets of colored transparencies, a method called subtractive reproduction.

In halftone printing, the ink drops are larger and the printing process controls their position and size. The color appearance of the print is controlled by the position and size of the dots in the array and these dots do not often overlap. Thus, the dots from different inks form several spatial mosaics, and color reproduction is more akin to an additive process: the reflected light is the sum of light scattered from the several mosaics.

Inks and Subtractive Color Calculations

Conventional color printing relies on three different types of colored ink; cyan, magenta and yellow. A fourth ink, black, is also used in a special and important role that will be described later. In continuous tone printing, the amount and spectral composition of the light reflected from the page is controlled by superimposing these colored inks and controlling their density on the page.

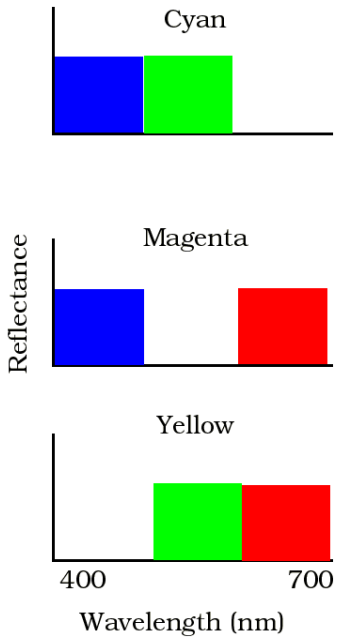

Figure 15. The reflectance functions of three (theoretical) block inks.

The way in which the reflected light is controlled is illustrated using very simple, theoretical inks whose reflectance spectra are shown in Figure 15. These are called *block inks* because they divide the wavelength spectrum into three bands corresponding, roughly, to a red, green, and blue. Each of the inks is transparent to light in two block regions, while absorbing light in a third block region. The inks take their names from the two regions in which they do not absorb. The cyan ink does not absorb in the blue and green bands, the magenta does not absorb in the red and blue, and the yellow does not absorb in the red and green.

The amount of light absorbed by ink is controlled by the amount of ink on the page. If a very thin layer of, say, cyan ink is placed down on white paper a small amount of the long wavelengths will be absorbed; the page will appear mainly white. If a great deal of cyan ink is placed on the page, a great deal of the long wavelength region will be absorbed and the ink will

take on its characteristic cyan appearance. Seen from this perspective, controlling the amount of the ink on the page is analogous to controlling the intensity of a display primary.

Density

To control any display technology, it is necessary to understand the relationship between the control variables and their action on the light arriving at the observer's eye. To understand the reflected light in the continuous tone process, we must predict the effect of altering the amount of ink on the page and the consequence of superimposing inks.

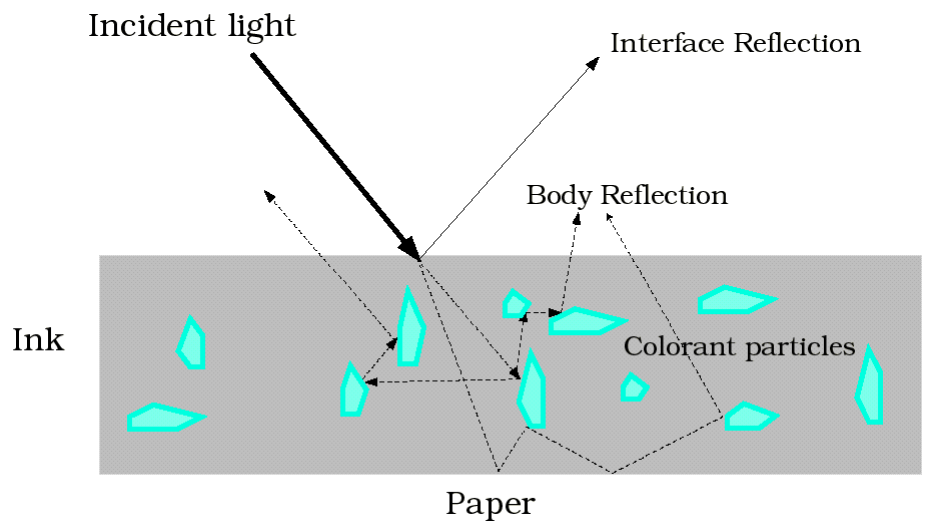

Figure 16. The scattered reflection from a layer of ink includes rays with different path lengths. Consequently, accurate prediction of the reflected light is a theoretically challenging problem.

Suppose we place a very thin layer of ink, of thickness δ, on a piece of paper. The probability that a monochromatic ray of light, at wavelength λ, will be absorbed by the ink is proportional to the thickness of the layer, a(λ) δ. Next, consider what will happen when the thickness of the ink is increased. Figure 16 shows some possible light paths as ambient light passes through the ink and is scattered towards an observer. Imagine that the average optical path length, including both passage towards the white page and passage back towards the eye, is D. Divide this path into N thin layers, each of thickness $\delta = D / N$. The absorption process follows the proportionality law within each thin layer. Consequently the chance that a ray will be reflected after traversing the entire optical path, D, is equal to the chance that it is not absorbed in any of the thin N layers, namely $(1-\delta a(\lambda))^N$. As the thickness is subdivided into more layers, N, the probability of absorption is expressed as *Beer's Law*

$$\lim_{n\to\infty}\left(1-a(\lambda)\tfrac{D}{N}\right)^{N}=e^{-Da(\lambda)}$$

Equation 1. Beers Law

The proportion of reflected light depends on the optical path length, D, which is controlled by the amount of ink placed on the page. As D increases, the fraction of light reflected becomes zero (unless $a(\lambda) = 0$). The constant of proportionality, $a(\lambda)$, is the *absorption function* of the ink.

Conventionally, the ink is described using by an *optical density* function $od\{\lambda\} = -\log_{10}(a(\lambda))$. From Equation 1 we find that optical density is proportional to the thickness,

$$od(\lambda)=2.3Da(\lambda)$$

Equation 2. Optical density defined.

Moreover, the optical density of two inks that are superimposed to form a single layer should add. Suppose the absorption functions of the inks are $a_i(\lambda)$. Then probability of absorption after traversing through the two layers is the product of the individual absorption functions, $\prod_{i=1,2}(1-a_i(\lambda))$. Since the absorption probabilities multiply, the optical densities add:

$$od(\lambda)=\log_{10}\left[\prod_{i=1,2}(1-a_i(\lambda))\right]=od_1(\lambda)+od_2(\lambda)$$

Equation 3. Density summation.

Equation 3 shows that when two inks are overlaid: There is a linear relationship between density and the control variable (density of colorant in the ink) and the optical density. This is the key reason why it is convenient to use optical density, rather than reflectance, in describing inks. In meeting the requirements of the color reproduction equations, however, the observer sees the reflected light and the observer's cone absorptions must be predicted. The nonlinear relationship between the control variable (ink density) and the light absorption by the cones is significantly more complex than the parallel nonlinear relationship between frame buffer intensity and cone absorptions that must be addressed in display technology. In displays, the spectral shape of the phosphors does not change a great deal with level. In the case of ink reflections, however, the spectral reflectance function changes considerably with density level. Thus predicting and controlling the light signal in printing is a much more difficult computational challenge.

Continuous Tone Printing

The physical principles described above are only a general overview of the reflection process; they do not capture many of the details necessary to make a precise prediction of the properties of color prints. In practice, a number of factors arise that make the predictions based on these very simple calculations inaccurate. Figure 16 shows one aspect of the reflection process that we have omitted: the light that is scattered to the observer's eye from a single point on the page may have traversed one of many different optical paths. The optical path will depend on the viewing geometry and on microscopic details of the paper and surrounding inks. Hence, computing the true optical path is very difficult and the calculations we have reviewed only serve as a first order approximation to the true reflection.

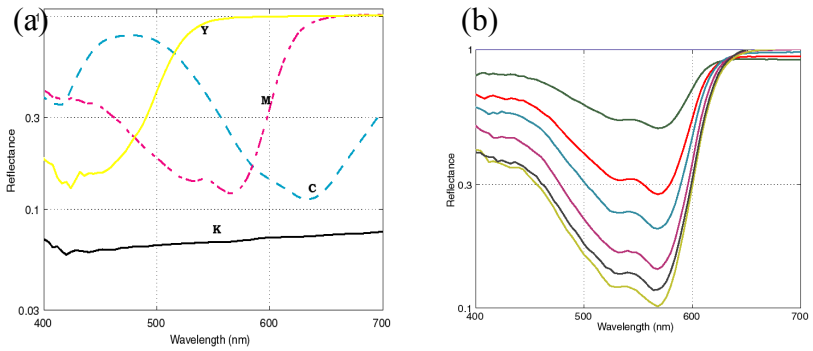

Figure 17. (a) Reflectance functions of the four primaries of an ink jet printer measured at high density. (b) The reflectance function of the magenta ink at various densities.

Finally, there are a great many limitations on the inks that can be used. As our colleague Gary Starkweather reminds us, many perfectly good inks are destroyed by exposure to the human environment and conversely perfectly good humans are destroyed by exposure to the ink's environment. Consequently, the supply of possible inks is limited and none are very close to the theoretical block inks. In addition, for all but the theoretical block inks, the shape of the reflectance function varies with density. Figure 17a shows the reflectance functions of four inks in a modern printer. The real inks overlap in their spectral reflection and absorption regions, unlike the perfect block inks. Panel (b) shows how the reflectance function of the magenta primary varies as a function of ink density. Notice that unlike the ideal block inks, the reflectance in both the short and middle wavelengths change as the magenta density varies.

The overlap in the absorption functions of the inks and the change in reflectance as a function of density make characterization calculations very difficult. We describe some of the basic methods later in this section. In addition, there are a great many ingenious efforts to understand and control such effects in order to make attractive prints. An excellent overview of these

technologies, and many of the technologies described in this chapter, can be found in R.W. Hunt's book (Hunt, 1987).

Finally, we conclude with a discussion of the very important role of the black ink in printing. To form a black or gray color using the cyan, magenta and yellow inks requires mixing all three together. These inks are generally very expensive, and even worse combining the three inks results in a very wet piece of paper. To reduce cost and to reduce bleeding within the paper, it is very effective to replace an equal mixture of the three colored inks with a single black ink in the appropriate density. The specific implementation of this will depend on the paper, colored inks, and black ink. The processes for substituting black ink for the colored inks are called *gray component removal* (GCR) or *undercolor removal* (UCR). Often, this process also improves the appearance of the print because the black ink has higher contrast than the combination of CMY.

Halftoning

Traditional halftoning

Halftoning is a printing method that simulates continuous tone dots of varying size or position. In traditional halftoning, illustrated in Figure 18, intensity is adjusted by varying the size of the printed dot. The image shows light from an original being imaged onto a fine mesh screen, often called the halftone *screen*. The screen converts the original image into a collection of point sources whose intensities depend on the intensity of the original. These point sources form images on a high contrast negative film. The fine screen mesh causes a set of pointspread images to be formed on the film.

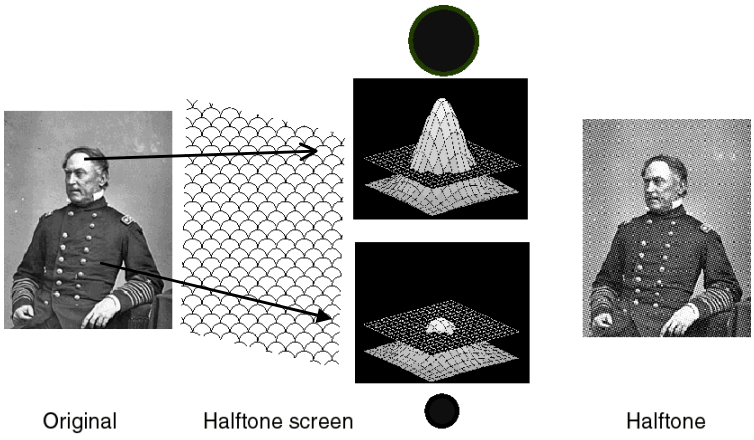

Figure 18. The steps involved in traditional screening are illustrated. See the text for

details.

The pointspread images differ as shown in the intermediate levels of the figure. When the dot is low intensity, only a small amount of the dot area exceeds the threshold of the high contrast film. When the dot is high intensity a large fraction of the dot area exceeds the threshold. Depending on the intensity of the original, a larger or smaller region of the high contrast film is exposed. Hence, the halftone screen process converts intensity into dot area.

The negative high contrast film is printed as a positive image. The dot sizes substitute for the original light levels, so that dark regions of the image have many large black dots and light regions have few. In a high quality print, the dots themselves are below visible resolution. Even when they are visible, the regular pattern of dots does not interfere strongly with the content of most images.

To create a color halftone, the process is repeated for the cyan, magenta, yellow and black components. In traditional halftoning, regular screens are used and this can cause an additional problem. When overlaying images from different halftone screens, the separate images produce unwanted interference known as Moiré patterns. To reduce such artifacts, the screens are rotated to different angles. Conventionally, the screen for black is oriented at 45 degrees (straight up is zero degrees). The screen angles for the other inks are cyan (105), magenta (70), and yellow (90).

Digital halftoning

Digital halftoning is a method for converting digital image files that represent the image intensity into an array of binary level values that represent dots in the printing process. The digital halftoning algorithms compute the positions where dots will be printed, and these are used to direct a digitally controlled printer. Both laser and ink jet technologies have evolved to a point where it is possible to control the dot placement with extraordinary accuracy and speed.

Three features of digital halftoning, that extend traditional halftoning, are of particular note. First, with digital techniques one does not need to place the dots in a perfectly regular array. Instead, it is possible to randomize the dot positions into disordered arrays. Using this method, it becomes harder for the eye to discern the screen pattern in monochrome printing and the problem of colored Moiré is also reduced. Methods using randomized screen positions are sometimes called *stochastic screening, frequency modulated (FM) halftoning,* or *blue noise* methods (Allebach & Lin, 1996; Mitsa & Parker, 1992; Mitsa, Ulichney, Parker, & Andre, 1991; Ulichney, 1988). While it is technically possible to achieve these results with traditional screening methods, it is very easy to achieve these results with digital halftoning.

Second, in addition to computational methods for randomizing the dot placement, it is now commonplace to control the dot placement at a very fine level using piezo-electric positioning devices on the print head. A printer that emits 600 dots per inch on a single line across the page may be able to place these dots at any of 1440 different positions.

Third, with digital control and computation it is possible to extend halftoning to a slightly more general process in which there is not just a single density level of, say, magenta ink printed on the page but one of two density levels. The multiple levels are achieved by including not just one source of magenta ink in the printer but also two sources, with different density. This process, often called *multi-level halftoning*, is in widespread use in digital printers such as the ink jet products.

There are two computational methods for implementing digital halftoning. One method, called dithering, is illustrated in Figure 19. This method approximates traditional screening and can be calculated at very high speeds. In this approach, the user selects a small matrix to serve as a *dither pattern,* also called a *mask*, for the calculation. Suppose the image and the mask are both represented at the resolution of the printer. Further suppose the mask is NxN. Digital halftoning begins by comparing the mask values with the intensity levels in an NxN region of the image. The image intensities are compared with the entries in the mask. Each of these is shown in the small surface plots in the figure. Suppose d_{ij} is a dither matrix entry and p_{ij} is a image intensity, then if $d_{ij} > p_{ij}$ set the printed point white, and otherwise set the point black. This process is repeated for each NxN block in the original picture until the entire image is converted from multiple intensity levels to a binary output suitable for printing as a halftone.

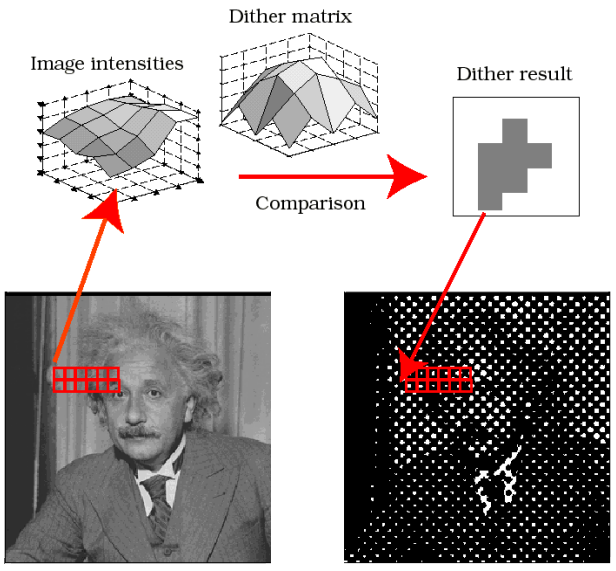

Figure 19. The steps involved in digital halftoning are illustrated. See the text for details.

The entries of the dither pattern partition the intensity levels in the original image. If the mask size is set to, say, NxN then the mask can partition the original image into $N^2 + 1$ intensity levels. Increasing the size of the dither pattern simulates more intensity levels, but reduces the effective spatial resolution of the printed picture. Several types of dither matrices are commonly used, and these are described below.

Digital halftoning algorithms are straightforward, but the terminology associated with digital halftoning and traditional screening has become intertwined and often very unclear. Printer dot density is usually specified in *dots per inch* (dpi). This describes the number of ink dots that can be placed on the page. Printer *addressability* refers to the number of positions where these dots can be placed. A printer may be able to place 300 dots per inch, but the center of these dots may fall at any of 1400 locations. Finally, the size of the dither pattern also influences the spatial resolution of the final print. A 3x3 mask reduces the spatial resolution of the print, and this is summarized by a quantity called the screen *lines*. A 300 dpi printer that uses a 3x3 mask is said to print at 100 screen lines per inch. A 300 dpi printer with a 6x6 mask is said to have 50 lines per inch.

Cluster Dot Dither

The *cluster dot* or *ordered dither* mask is designed to be similar to traditional screening. An example of a 5x5 dither pattern for a cluster dot is:

$$\begin{pmatrix} 1 & 9 & 16 & 8 & 7 \\ 10 & 17 & 21 & 20 & 15 \\ 2 & 22 & 25 & 24 & 6 \\ 11 & 18 & 23 & 19 & 14 \\ 3 & 12 & 4 & 13 & 5 \end{pmatrix} \left(\frac{255}{26} \right)$$

Consider the results of comparing this mask with uniform intensity patterns ranging from 0 to 255. The mask has twenty-five entries and can partition the input regions into 26 different levels. When the image intensity is very low (< 255/26), all of the points are set black. As the image intensity increases the size of the dot shrinks, so that at a value just less than 25*(255/26) only a single dot is left in the middle.

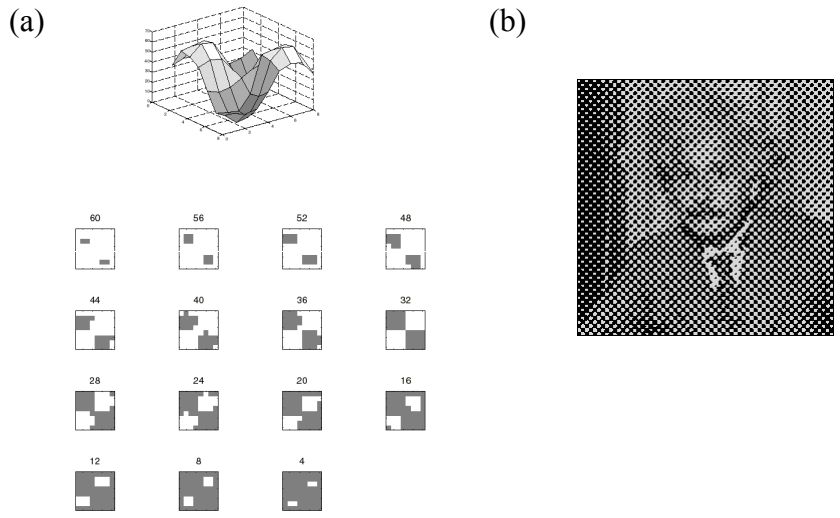

Figure 20. Cluster dot halftoning is illustrated. The dot masks are shown in (a). The result of applying these masks an image is shown in (b).

A variant of this mask is shown in Figure 20a. While basically a cluster dot, the thresholds in the mask are arranged so that the growth in dot size occurs on a diagonal pattern. This has the effect of arranging the dots to fall on the 45 deg line, as in traditional screening. The result of applying this mask to an image is illustrated in Figure 20b.

Bayer Dither and Void and Cluster dither

The Bayer (Bayer, 1973) dither pattern represented an early and important innovation that showed how digital halftoning might improve on traditional screening. The Bayer dither pattern was chosen so that the spatial structure of the printed dots would be less visible than the ordered dither dots. Figure 21a shows the result of applying an 8x8 Bayer dither mask to an image. The Bayer output compares favorably with the results of applying a traditional screening pattern, this time using an 8x8 digital dither pattern structured to give dots along the 45-degree diagonal (see Figure 21b).

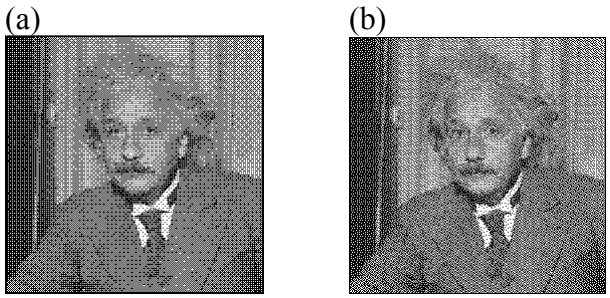

Figure 21. Bayer dither (a) and void and cluster dither (b) applied to an image.

For very low resolution printing, the Bayer dither mask results in images that are more appealing than ordered dither. Even so, the Bayer dither pattern contains a regular structure that is visible in the printed halftone. Ulichney (Ulichney, 1993) proposed a method of creating dither patterns with locally random structure that are even less visible because they have most of their spatial structure in the high spatial frequencies. Because the spatial frequency power is in the high frequencies, these are called *blue-noise* masks. One computational method for implementing blue-noise masks is called the v*oid and cluster* method. In this method, a pattern is selected so that there are no very large voids or very large clusters.

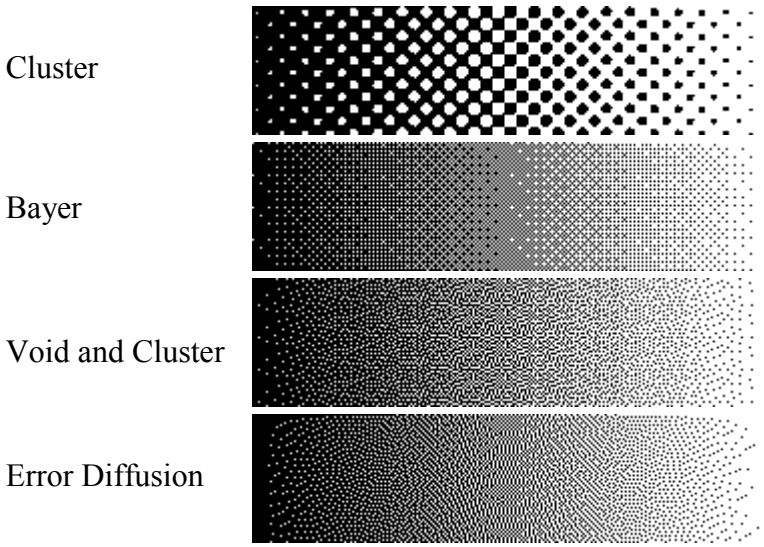

Figure 22. Cluster dot, Bayer, void and cluster and error diffusion applied to an intensity ramp.

Comparisons of the dither patterns from ordered dither, Bayer, void and cluster and error diffusion are shown applied to simple intensity ramps Figure 22. At low printer resolutions Bayer and the void and cluster method are preferred to ordered dither. They are no more computationally expensive. Notice the similarity between the void and cluster pattern and the error diffusion pattern. The error diffusion process is explained in the next section. The pattern of dots created by this method is similar to the pattern created by void and cluster.

There is an analogy between the different halftoning methods and signal transmission methods in a communication channel. The cluster dot method modulates the light intensity by controlling the size of a dot, much like controlling the amplitude of a signal. The Void and Cluster method modulates the light intensity by varying the spatial structure of a complex pattern, much like controlling the frequency of a signal. Hence, sometimes cluster dot is described as an amplitude modulation (AM) screening technique and while void and cluster is described as a frequency modulation (FM) screening technique.

Error Diffusion

At low print resolutions, the best halftoning results are obtained using an adaptive algorithm in which the halftoning depends upon the data in the image itself. Floyd and Steinberg (1976) introduced the basic principals of adaptive halftoning methods in a brief and fundamental paper. Their algorithm is called *error diffusion*. The idea is to initiate the halftoning process by selecting a binary output level closest to the original intensity. This binary level will differ substantially from the original,. The difference between the halftone output and the true image

(i.e., the error) is added to neighboring pixels that have not yet been processed. Then, the binary output decision is made on the next pixel whose value now includes both the original image intensity and the errors that have been added from previously processed pixels. Figure 23 shows a flow chart of the algorithm (panel a) and a typical result (panel b).

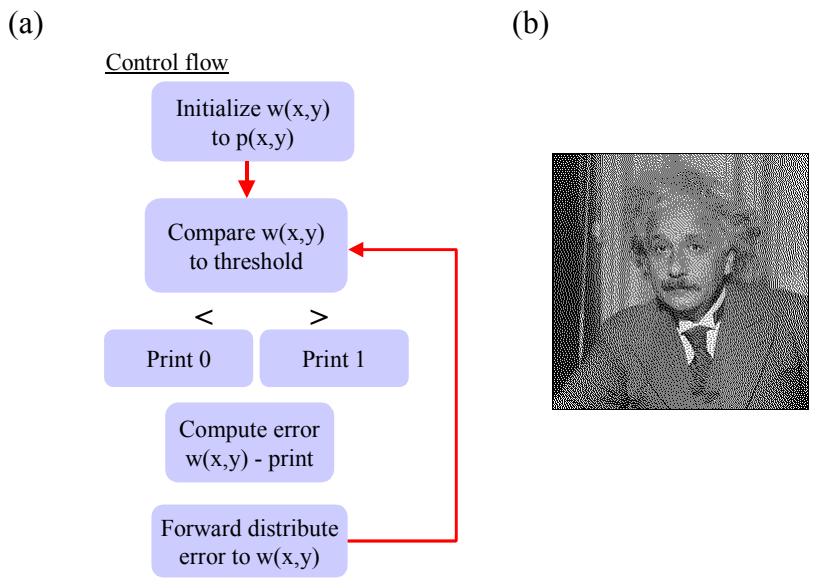

Figure 23. The steps involved in error diffusion algorithm (a) and the resulting image (b). See text for details.

The coefficients that distribute the error among neighboring pixels can be chosen depending on the source material and output device. Jarvis, Judice and Ninke (Jarvis, Judice, & Ninke, 1976) found the apportionment of error using the matrix

$$\begin{pmatrix} 0 & 0 & * & 7 & 5 \\ 3 & 5 & 7 & 5 & 3 \\ 1 & 3 & 5 & 3 & 1 \end{pmatrix} \left(\frac{1}{48}\right)$$

to be satisfactory, where * denotes the current image point being processed. Notice that the error is propagated forward to unprocessed pixels. Also, the algorithm works properly when applied to the linear intensity of the image. The algorithm should not be applied to images represented in a nonlinear space, such as the frame buffer values of a monitor. Instead, the image should be converted to a format that is linear with intensity prior to application of the algorithm.

For images printed at low spatial resolution, error-diffusion is considered the best method. Ulichney (Ulichney, 1987) analyzed the spatial error of the method and showed that the error was mainly in the high spatial frequency regime. The drawback of error diffusion is that it is very time-consuming compared to the simple threshold operations used in dither patterns. For images at moderate to high spatial resolution (600 dpi), blue-noise masks are visually as attractive as

error diffusion and much faster to compute. Depending on the nature of the paper, cluster dot can be preferred at high resolutions. The cluster dot algorithm separates the centroids of the ink so that there is less unwanted bleeding of the ink from cell to cell. In certain devices and at certain print resolutions, reducing the spread of the ink is more important than reducing the visibility of the mask.

Color digital halftoning

The principles of digital halftoning can be directly extended to making colored halftone prints. The most common extension to color for dither patterns is to convert the original image into a CMY representation, and then to apply the digital halftoning algorithm separately to each of the color planes. By using these angles the effect of Moire between the dots in the separations is minimized. This separable architecture is computationally efficient. The resulting output can be followed by a step of gray-component removal.

There are several technical issues that must be addressed when extending halftoning to color. First, the overlap between the dots comprising the colored planes can lead to undesirable spatial artifacts (moiré patterns). To minimize this spatial disturbance, the different color separations are printed with their dots arrays at different angles. Typically, the dots comprising the black separation (K) are printed at 45 deg, and the CMY dots at 105, 75 and 90 deg (vertical = 0). Second, depending on the type of printing process, the size of the dots may cause the mixture of halftones to overlap greatly or little. When the printing process produces sets of interleaved mosaics of dots, the characterization is similar to an additive system. When the printing process includes a great deal of overlap between the dots, the process is similar to a subtractive system characterized by the Neugebauer process described below. Hence, the color characterization needed to manage the different techniques will depend on the spatial structure of the printing process.

There are also issues that must be considered when extending to adaptive processing algorithms, such as error diffusion. In an adaptive process, the algorithm must decide which one of the usable dots should be printed next. Normally, the usable dots are one of the primaries alone (C,M,Y,K) or a mixture of the non-black primaries (CM, CY, MY, CMY). These algorithms can use the color error up to the current print position to decide which dot should be printed. The error can be a three-dimensional quantity, calculated jointly from all of the separation errors, which is far more complex than making a decision separately for each color separation.

Print Characterization

Transduction: The tone reproduction curve

The relationship between the device parameter that controls the ink density and the relative amount of reflected light is called the *tone reproduction curve* (TRC). This curve is the printer transduction function, analogous to the display transduction function. Figure 24a shows a tone reproduction curve measured for a monochrome laser printer. This printer achieves different gray levels by halftoning. Notice that the curve is similar to the curve measured on a CRT or

LCD device. Figure 24b shows the tone reproduction curve for the magenta primary of an ink jet printer. This curve has the same general form as the black ink, but it never reaches as low a luminance level.

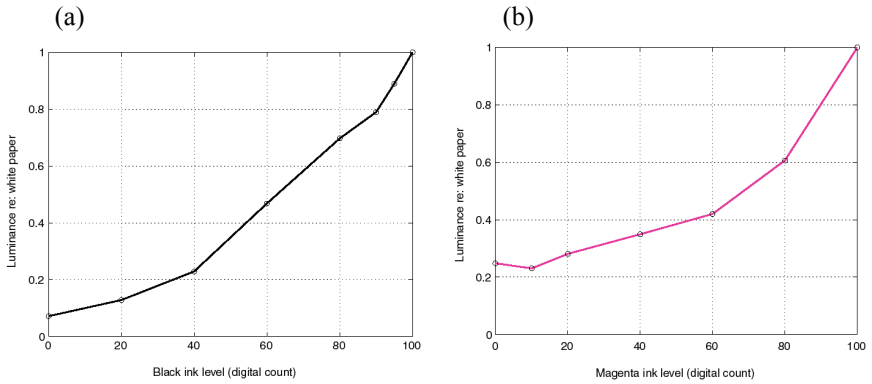

Figure 24. The printer tone reproduction curve. Curves are shown for (a) monochrome laser printer and (b) the magenta primary of an ink jet printer.

The TRC is a fundamental measurement of printer performance. For two reasons, however, the TRC cannot be combined easily with the ink primary reflectance functions to predict reflected light or solve the color-reproduction equation. First, the reflectance functions of the CMYK primaries do not combine additively. Overprinting inks and blending of the ink spots can cause light to follow optical paths whose spectral transmission is very complicated to predict. Second, even for pure ink, Beer's Law (Equation 1) shows that the ink reflectance function varies with density. Consequently, the tristimulus values of the reflected light will vary with density, making it impossible to use the simple principles that are applied to monitor characterization. Specifically, ink mixtures do not follow the rules of primary independence that are essential to the simple characterization of display devices.

An influential theory for predicting ink reflectance functions was provided by the brilliant color engineer, Neugebauer (Neugebauer, 1937). The idea introduced in the *Neugebauer equations* is based on the following physical intuition. Imagine partitioning the printed page into extremely small areas over which each region contains only an infinitesimal (binary) amount of ink. At this spatial scale, we will find only eight possible combinations of ink (C,M,Y,CM,CY,MY,K); other combinations are eliminated because K combined with anything is equivalent to K. Any small region will contain many of these infinitesimal dots, and the reflectance function of the small region will be a weighted sum of these eight basic terms. The weights will depend on the area allocated to the infinitesimal dots, and any small region is predicted to have a reflectance function that is the weighted sum of eight possible basis terms. Predicting these basis terms from

the ink primaries, and then predicting the ink reflectance functions is the main application of the theory. Practical difficulties in compensating for scatter into the paper and complicated light paths represent a continuing challenge to the theory, which continues to be an active area of investigation. For a modern update the reader may wish to consult the special symposium on this topic (Neugebauer, 1989).

Because of the difficulties in predicting the ink reflectance functions from first principles, printer characterization methods mainly rely on the use of extensive look-up tables. These tables are based on measurements of the tristimulus values measured from a variety of print samples on a variety of paper types. Tetrahedral look-up tables are often used for this purpose.

Conclusions

Color imaging technologies are central to many features of modern day life. The breadth and vitality of the industry that creates these technologies is extraordinary. In reviewing a small part of these technologies, we have tried to explain how knowledge of the human visual system plays an important role in many design decisions. The widespread use of tristimulus coordinates for characterization represents one major contribution of vision science to imaging technology. Understanding when spatial and wavelength measurements can be safely traded for one another is a second contribution.

Equally, the contributions of color technology have propelled forward experiments in vision science. Improvements in color characterization and color displays have made new experimental methods and precise control possible. The interaction between these fields, as represented by the inclusion of this chapter in this volume, enriches both.

References

Adams, J., Parulski, K., & Spaulding, K. (1998). Color Processing in Digital Cameras. *IEEE Micro, 18*(6).

Allebach, J., & Lin, Q. (1996). *FM screen design using DBS algorithm.* Paper presented at the 1996 IEEE International Conference on Image Processing, ICIP'96. Part 1 (of 3), Lausanne, Switz.

Anderson, S., Mullen, K., & Hess, R. (1991). Human Peripheral Spatial Resolution for Achromatic and Chromatic Stimuli: Limits Imposed by Optical and Retinal Factors. *J. Physiol., 442*, 47-64.

ASTM. (1991). Standard recommended practice for goniophotometry of objects and materials., *ASTM Standards on Color and Appearance Measurement* (3 ed.). Philadelphia: ASTM.

Bayer, B. E. (1973). *An optimum method for two-level rendition of continuous-tone pictures.* Paper presented at the Proc. IEEE Conf. Communications.

Berns, R. S., Gorzynski, M. E., & Motta, R. J. (1993). CRT colorimetry, Part I: Theory and practice. *Color Res. and Appl., 18*, 299-314.

Berns, R. S., Motta, R. J., & Gorzynski, M. E. (1993). CRT colorimetry, Part II: Metrology. *Color Res. and Appl., 18*, 315-325.

Brainard, D. (in this volume). Color Appearance? Something. In Shevell (Ed.), *Color?*: Optical Society of America.

Brainard, D. H. (1989). Calibration of a computer controlled color monitor. *Col. Res. Appl., 14*, 23-34.

Brainard, D. H., & Sherman, D. (1995). *Reconstructing images from trichromatic samples: from basic research to practical applications.* Paper presented at the Proceedings of 3rd Color Imaging Conference: Color Science, Systems and Applications, Scottsdale, AZ, USA.

CIE. (1990). *CIE 1988 2 deg Spectral Luminous Efficiency Function for Photopic Vision* (CIE 86): Commission Internationale de L'Eclairage (CIE).

CIE. (1996). *The Relationship Between Digital and Colorimetric Data for Computer-Controlled CRT Display* (technical report 122-1996): Commission Internationale de L'Eclairage (CIE).

Collett, E. (1993). *Polarized Light: Fundamentals and Applications*. New York: Marcel Dekker, Inc.

Collings, P. J. (1990). *Liquid Crystals: Nature's Delicate Phase of Matter*. Princeton: Princeton University Press.

Conner, A. R. (1992). The evolution of the stacked color LCD. *Society for Information Display Applications Notes*, 109-112.

Cornsweet, T. N. (1970). Visual Perception.

Cupitt, J., Martinez, K., & Saunders, D. (1996). A Methodology for art reproduction in colour: The MARC project. *Computers and the History of Art, 6*(2), 1-19.

DeValois, R. L., & DeValois, K. K. (1988). Spatial Vision.

dpreview.com. (2000). *Measuring Dynamic Range*. Available: http://www.dpreview.com/news/0011/00111608dynamicrange.aspThursday, 16 November].

Fairchild, M. (1997). *Color Appearance Models*. Reading, MA: Addison Wesley Longman.

Farrell, J., Saunders, D., Cuppitt, J., & Wandell, B. (1999). *Estimating spectral reflectance of art work.* Paper presented at the Chiba Conference on Multispectral Imaging, Chiba, Japan.

Floyd, R. W., & Steinberg, L. (1976). An adaptive algorithm for spatial greyscale. *Proceedings of the Society for Information Display, 17*, 75-77.

Fritsch, M. W., & Mlynski, D. A. (1991). Faster contrast measurement of LCDs with improved conoscopic methods. *Proceedings of the Society for Information Display, 32*, 207-211.

Gill, G. (1999). *ICC profile I/O library (icclib), README file*. Available: http://web.access.net.au/argyll/icc_readme.html99/11/29].

Glenn, W. E., Glenn, K. G., & Bastian, C. J. (1985). Imaging system design based on psychophysical data. *Proceedings of the Society for Information Display, 26*, 71-78.

Hardeberg, J. Y., & Schmitt, F. (1997). *Color Printer Characterization Using a Computational Geometry Approach.* Paper presented at the Proceedings of IS&T and SID's 5th Color Imaging Conference,, Scottsdale, Arizona.

Horn, B. K. P. (1984). Exact reproduction of colored images. *Computer Vision, Graphics and Image Processing, 26*, 135-167.

Hunt, R. W. G. (1987). The Reproduction of Colour.

Janesick, J. (1997, 10-11 Feb. 1997). *CCD Transfer method - standard for absolute performance of CCDs and digital CCD camera systems.* Paper presented at the Solid State Sensor Arrays: Development and Applications, San Jose.

Jarvis, J. F., Judice, C. N., & Ninke, W. H. (1976). A survey of techniques for the display of continuous tone pictures on bilevel displays,. *Computer graphics and image processing, 5*(1), 13-40.

Lee, H.-C. (March 18, 1985). Method for determining the color of a scene illuminant from a color image. US: Kodak.

Lehrer, N. H. (1985). The challenge of the cathode-ray tube. In J. L. E. Tannas (Ed.), *Flat-Panel Displays and CRTs* (pp. 138-176). New York: Van Nostrand Reinhold Company.

Lennie, P. (This volume). Chapter title. In Shevell (Ed.), *This book.*

Leroux, T., & Rossignol, C. (1995). Fast analysis of contrast and color coordinates vs. viewing angle. *Society for Information Display Digest of Technical Papers*, 739-742.

Luo, M. R., & Hunt, R. W. G. (1998). The structure of the CIE 1997 color appearance model. *Color Res. and Appl., 23*, 138-146.

Lyons, N. P., & Farrell, J. E. (1989). Linear systems analysis of CRT displays, *1989 SID International Symposium. Digest of Technical Papers* (pp. x+440). Playa del Ray, CA, USA: Soc. Inf. Display.

Martinez, K., Cupitt, J., & Saunders, D. (1993). *High resolution colorimetric imaging of paintings*.

Mitsa, T., & Parker, K. J. (1992). Digital halftoning technique using a blue-noise mask. *Journal of the Optical Society of America A (Optics and Image Science), 9*(11), 1920-1929.

Mitsa, T., Ulichney, R., Parker, K. J., & Andre, J. (1991). The construction and evaluation of halftone patterns with manipulated power spectra, *Raster Imaging and Digital Typography II* (pp. x+217). Cambridge, UK: Cambridge University Press.

Mullen, K. (1985). The constrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *J. Physiol., 359*, 381-400.

Neugebauer. (1989). *Neugebauer Memorial Seminar on Color Reproduction*.

Neugebauer, H. E. J. (1937). Die theoretischen Grundlagen des Meshrfarbendruckes. *Z. Wiss. Photogr., 36*, 73-89.

Penz, P. A. (1985). Nonemissive displays. In J. L. E. Tannas (Ed.), *Flat-Panel Displays and CRTs* (pp. 415-457). New York: Van Nostrand Reinhold Company.

Plummer, W. T. (1983). Color filter. USA.

Poirson, A. B., & Wandell, B. A. (1993). The appearance of colored patterns: pattern-color separability. *J. Opt. Soc. Am. A, 10*(12), 2458-2471.

Poynton, C. (1996). *A Technical Introduction to Digital Video*.: John Wiley & Sons.

Rodieck, R. W. (1998). *The First Steps in Seeing*. Sunderland, MA: Sinauer Press.

Sakamoto, T., & Itooka, A. (1981). Linear interpolator for color correction. USA.

Saleh, B. E. A. (1996). The Fourier scope: an optical instrument for measuring LCD viewing-angle characteristics. *Journal of the Society for Information Display, 4*(1), 33-39.

Schade, O. (1958). On the quality of color-television images and the perception of colour detail. *Journal of the Society of Motion Pictures and Television Engineers, 67*, 801-819.

Scheffer, T., & Nehring, J. (1990). Twisted nematic and supertwisted nematic mode LCDs, *Liquid Crystals: Applications and Uses, Volume I,* . New Jersey: World Scientific Publishing Company.

Scheffer, T., & Nehring, J. (1992). Twisted nematic (TN) and super-twisted nematic LCDs. *Society for Information Display Seminar Lecture Notes, 1*, M1/1?1/52.

Sekiguchi, N., Williams, D. R., & Brainard, D. H. (1993a). Aberration-free measurements of the visibility of isoluminant gratings. *J Opt Soc Am A, 10*(10), 2105-2117.

Sekiguchi, N., Williams, D. R., & Brainard, D. H. (1993b). Efficiency in detection of isoluminant and isochromatic interference fringes. *J Opt Soc Am A, 10*(10), 2118-2133.

Shafer, S. A. (1985). Using color to separate reflection components. *Col. Res. Appl., 10*, 210-218.

Sherr, S. (1993). *Electronic Displays* (2nd ed.). New York: John Wiley & Sons.

Silicon Vision, A. (2000). *TFA Color Image Sensor (COSIMA)*. Available: http://www.siliconvision.de/produkte/cosima-e.htm [2000.

Silverstein, L. D. (1991). *Description of an on-axis colorimetric/photometric model for twisted-nematic color liquid crystal displays* (Unpublished technical report for the NASA/ARPA Visual Display Engineering and Optimization System (ViDEOS) project): NASA/ARPA.

Silverstein, L. D. (2000). Color in Electronic Displays. *Society for Information Display Seminar Lecture Notes, 1*, M6/1-M6/88.

Silverstein, L. D., & Bernot, A. J. (1991). Apparatus and method for an electronically controlled color filter for use in information display applications. US.

Silverstein, L. D., & Fiske, T. G. (1993). Colorimetric and photometric modeling of liquid crystal displays. *Proceedings of the First IS&T/SID Color Imaging Conference: Transforms and Transportability of Color, 1*, 149-156.

Silverstein, L. D., Krantz, J. H., Gomer, F. E., Yei-Yu, Y., & Monty, R. W. (1990). Effects of spatial sampling and luminance quantization on the image quality of color matrix displays. *Journal of the Optical Society of America A (Optics and Image Science), 7*(10), 1955-1968.

Silverstein, L. D., & Merrifield, R. M. (1985). *The Development and Evaluation of Color Display Systems for Airborne Applications: Phase I - Fundamental Visual, Perceptual, and Display System Considerations* (Technical Report DOT/FAA/PM-85-19): FAA.

Smith, V., & Pokorny, J. (In press). This Volume. In Shevell (Ed.), *This book*.

Speigle, J. M., & Brainard, D. H. (1999). Predicting color from gray: the relationship between achromatic adjustment and asymmetric matching. *Journal of the Optical Society of America A (Optics, Image Science and Vision), 16*(10), 2370-2376.

Tamura, Y. (1983). Color original readout apparatus.: Canon Kabushiki Kaisha.

TC1-34, C. C. (1998). *Final CIE TC1-34 specification*, [Internet]. Available: http://www.cis.rit.edu/people/faculty/fairchild/PDFs/CIECAM97s_TC_Draft.pdf.

Tominaga, S., & Wandell, B. A. (1989). The standard surface reflectance model and illuminant estimation. *J. Opt. Soc. Am. A, 6*, 576-584.

Ulichney, R. (1987). *Digital Halftoning*.: MIT Press.

Ulichney, R. (1993). *The Void-and-Cluster Method for Generating Dither Arrays.* Paper presented at the Proc. SPIE, San Jose, CA.

Ulichney, R. A. (1988). Dithering with blue noise. *Proceedings of the IEEE, 76*(1), 56-79.

VanderHorst, G. J. C., & Bouman, M. A. (1969). Spatiotemporal chromaticity discrimination. *Journal of the Optical Society of America, 59*, 1482-1488.

Vincent, K., & Neuman, H. (1989). Color combiner and separator and implementations. USA: Hewlett-Packard Company.

Wandell, B. (1999). Computational Neuroimaging: Color representations and processing. In M. S. Gazzaniga (Ed.), *The New Cognitive Neurosciences* (2nd ed.). Cambridge, MA: MIT Press.

Wandell, B. A. (1986). Color rendering of camera data. *Col. Res. Appl., Supplement, 11*, S30-S33.

Wandell, B. A. (1995). *Foundations of Vision*. Sunderland, MA: Sinauer Press.

Wyszecki, G., & Stiles, W. S. (1982). *Color Science: concepts and methods, quantitative and formulae*. New York: Wiley.