"Gamma" and Its Disguises: The Nonlinear Mappings of Intensity in Perception, CRTs, Film, and Video

By Charles A. Poynton

In photography, video and computer graphics, the gamma symbol γ represents a numerical parameter that describes the nonlinearity of intensity reproduction. Gamma is a mysterious and confusing subject, because it involves concepts from four disciplines: physics, perception, photography, and video. In this note I will explain how gamma is related to each of these disciplines. Having a good understanding of the theory and practice of gamma will enable you to get good results when you create, process, and display pictures. This note focuses on electronic reproduction of images, using video and computer graphics techniques and equipment. However, the discussions of perception and photography may be of general interest as well. This note deals mainly with the reproduction of intensity, or as a photographer would say, tone scale. This is one important step to achieving good color reproduction, but issues specific to color are covered elsewhere.

The nonuniform perception of intensity is the subject of the first section of this note. The human perceptual response to intensity is distinctly nonuniform: the *lightness* sensation of vision is roughly a power function of intensity. This characteristic of vision needs to be accommodated if an image is to be coded so as to minimize the visibility of noise and make effective perceptual use of a limited number of bits per pixel.

The origin of *gamma* in the physics of a cathode ray tube (CRT) is covered in the second section of this note. The CRTs that are ubiquitous in workstation displays and in television sets are inherently nonlinear devices: the intensity of light reproduced at the screen of a CRT monitor is not proportional to its voltage input. From a strictly physical point of view, *gamma correction* can be thought of as the process of compensating for this nonlinearity in order to achieve correct reproduction of intensity.

The third section of this note discusses how combining these two concepts — one from perception, the other from physics — reveals an amazing coincidence: the nonlinearity of a CRT is remarkably similar to the *inverse* of the lightness sensitivity of human vision. Coding intensity into a gamma-corrected signal makes maximum perceptual use of the channel and simultaneously corrects for intensity nonlinearity at the CRT display. If gamma correction were not already necessary for physical reasons at the CRT, we would have to invent it for perceptual reasons.

The next section of this note switches to a completely different topic: photography. Photography also involves nonlinearity of intensity reproduction. The nonlinearity of film is characterized by a parameter gamma. As you might suspect, electronics inherited the term from photography! The effect of gamma in film primarily concerns the appearance of pictures rather than the accurate reproduction of intensity values. Some of the appearance aspects of gamma in film also apply to television and computer displays.

In the fifth section of this note I will describe how video draws aspects of its handling of gamma from of all of these areas: knowledge of the CRT from physics, knowledge of the nonuniformity of vision from perception, and knowledge of viewing conditions from photography. I will also discuss additional details of the CRT transfer function that you will need to know if you wish to calibrate a CRT or determine its nonlinearity.

In the final section of this note I will make some comments about gamma correction in computer graphics.

Luminance

The subject of colorimetry concerns the relationship between the sensation of color and spectral power at different wavelengths in the visible band. Detailed discussion of colorimetry is beyond the scope of this note, except for one fact that is important here. Human vision gives special treatment to luminance, or roughly speaking, brightness. Luminance is a weighted mixture of spectral energy, where the weights are determined by the characteristics of human vision. Luminance comprises roughly 11% power from blue regions of the spectrum, 59% from green, and 30% from red. The coefficients in this weighted sum are derived from the sensitivity of human vision to the corresponding wavelengths; a saturated blue color is quite dark, having a brightness of about 0.11 relative to white. A saturated red is considerably lighter, and saturated green is lighter still.

The Commission Internationale de L'Éclairage (CIE, or International Commission on Illumination) is the international body responsible for standards in the area of color perception. The CIE has standardized a weighting function, defined numerically, that relates spectral power to luminance. The symbol for luminance is Y, sometimes emphasized as being standard by the prefix *CIE*.

Unfortunately, in video practice, the term luminance has come to mean the *video signal representative of luminance*, even though that video signal has been subjected to a nonlinear transfer function. In the early days of video, the nonlinear signal was denoted Y', where the prime symbol indicated the

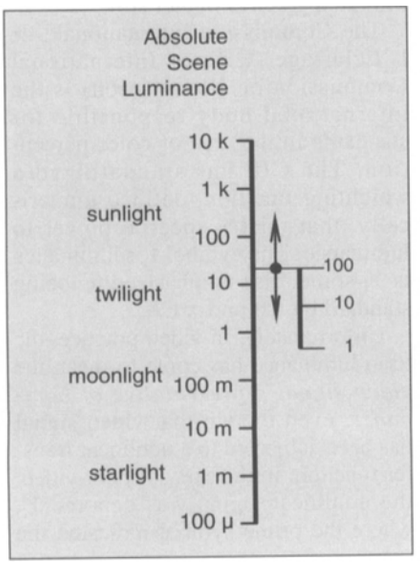
Charles A. Poynton is with Sun Microsystems Corp., Mountain View, CA 94043.

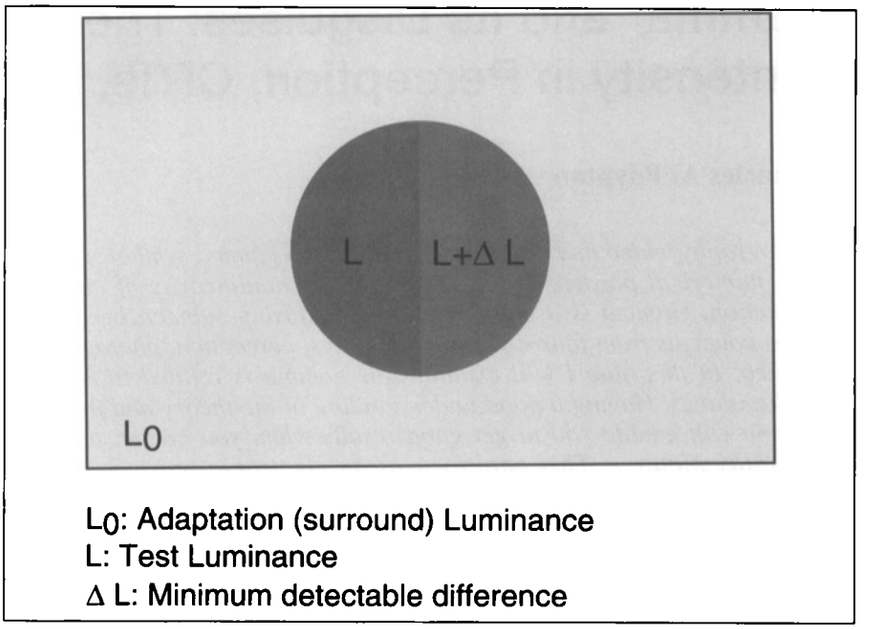
nonlinear treatment. But over the last 40 years the prime has been elided in video usage and now both the term *luminance* and the symbol *Y* collide with the CIE, making both ambiguous! This has led to great confusion, such as the incorrect statement commonly found in computer graphics and color textbooks that in the *YIQ* or *YUV* color spaces, the *Y* component is CIE luminance. I use the term *luminance* according to its standardized *CIE* definition and use the term *luma* to refer to the video signal. But my convention is not yet widespread, and in the meantime you must be careful to determine whether a linear or nonlinear interpretation is being applied to the word and the symbol.

Until now I have used the familiar term *intensity* — including in the title of this note! — but from now on I will use the technical term *luminance* to refer to the brightness response of human vision. I will continue to use *intensity* for linear-light red, green, and blue quantities. Luminance is a linear-light quantity, although its spectral composition is defined according to human vision. The response of human visual system to luminance is the subject of the next section.

Perception

Human vision adapts over a remarkably wide range of intensity levels — about seven decades of dynamic range in total. This is illustrated in the sketch below. For about two decades at the bottom end of the intensity range, the retinal photoreceptor cells called *rods* are employed. Since there is only one type of rod cell, what is loosely called night vision cannot discern colors. About one decade of adaptation is effected by the iris; the remainder of the adaptation is due to a photochemical process that involves the visual *pigment* substance contained in photoreceptor cells.





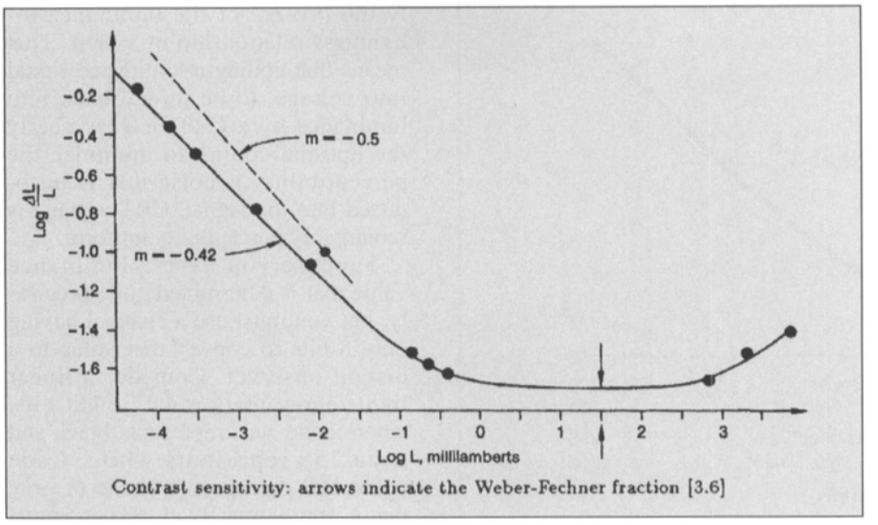
The adaptation is controlled by total retinal illumination. *Dark adaptation* to a lower intensity is slow: it can take many minutes to adapt from a bright sunlit day to the dark ambient light of a cinema. *Light adaptation* towards higher intensity is more rapid but can be painful, as you may have experienced when walking out of the cinema back into daylight. Since adaptation is controlled by total retinal illumination, your adaptation state is closely related to the intensity of "white" in your field of view.

At a particular state of adaptation, human vision can distinguish different luminance levels down to about 1% of peak white. In other words, our ability to distinguish luminance differences extends over a luminance range of about 100:1. Loosely speaking, intensities less than about one percent of what is sometimes called peak white appear simply "black": different luminance values below this luminance level cannot be distinguished.

Contrast ratio is defined as the ratio of luminance between the lightest and darkest elements of a scene. Contrast ratio is a major determinant of perceived picture quality, so much so that an image reproduced with a high contrast ratio may be judged sharper than another image that has higher measured spatial resolution. In a practical imaging system, many factors conspire to increase the luminance of blacks, and thereby lessen the contrast ratio and impair perceived picture quality.

Within the two-decade range of luminance over which vision can distinguish luminance levels, human vision has a certain discrimination threshold. For a reason that will become clear in a moment, it is convenient to express the discrimination capability in terms of *contrast sensitivity*, which is the ratio of luminances between two adjacent patches of similar luminance.

The diagram above shows the pattern presented to an observer in an experiment to determine the contrast sensitivity of human vision. Most of the observer's field of vision is filled by a *surround* luminance level, L_0, in order to fix the observer's state of adaptation. In the central area of the field of vision is placed two adjacent patches having slightly different



Contrast sensitivity: arrows indicate the Weber-Fechner fraction [3.6]

luminance levels L and $L+\Delta L$. The experimenter presents stimuli having a wide range of test values with respect to the surround, that is, a wide range of L/L_0 values. At each test luminance, the experimenter presents to the observer a range of luminance increments with respect to the test stimulus, that is, a range of $L+\Delta L/L$ values.

The result of conducting the experiment is shown in the graph above, Fig. 3.4 of W.F. Schreiber's *Fundamentals of Electronic Imaging Systems*. Plotting log $\Delta L/L$ as a function of *log* L reveals an interval of more than two decades of luminance, between about zero and +2.5 log millilamberts, over which the discrimination capability of vision is about 1% of the test luminance level. This leads to the conclusion that — for *threshold* discrimination of two adjacent patches of nearly identical luminance — the discrimination capability is very nearly logarithmic.

The contrast sensitivity function begins to answer the question, what is the minimum number of discrete codes required to represent luminance over a particular range? In other words, what intensity codes can be thrown away without the observer noticing? If codes are placed at exactly 1% intervals over a 100:1 range, the number of codes required is *log* 100/*log*1.01, or 460. About 460 codes are required to cover a dynamic range of 100:1, at a particular adaptation state, at a discrimination threshold of 1%.

The logarithmic relationship is based on measurements of contrast sensitivity at threshold. That is, we are measuring the ability of the visual system to discriminate between two nearly identical luminance values. Over a wider range of luminance, strict adherence to logarithmic coding is not necessary for perceptual reasons. Also, the discrimination capability of vision degrades for very dark shades of gray, below a few percent of peak white.

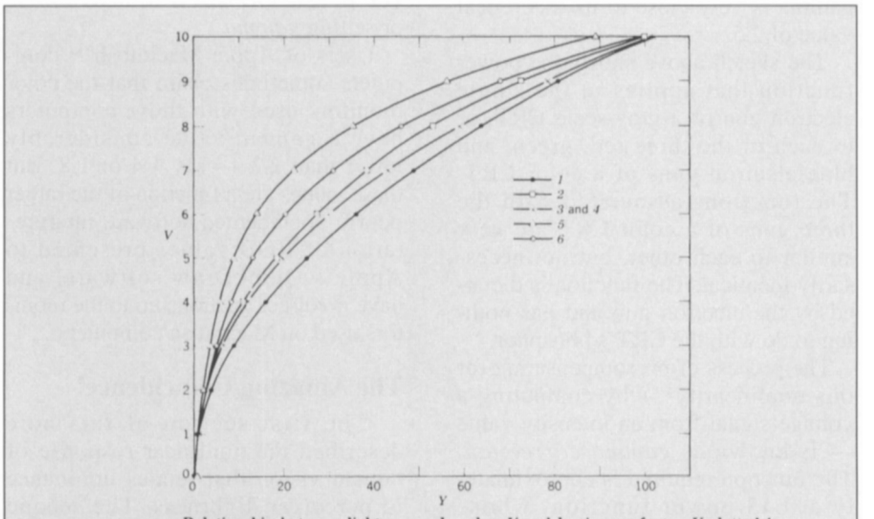
The graph below is Fig. 2 (6.3) from Wyszecki and Stiles' *Color Science*. It shows several lightness scales that have been used by different workers in the field. Some are polynomials, others are power functions, still others are logarithmic. There is no agreement on a single curve for all applications, but all of the curves have the same basic form. In 1976, the CIE standardized the L^* function, which is a power function of luminance, modified slightly by the introduction of a linear segment near black. In the following equations, Y is CIE luminance (proportional to intensity), and Y_n is the luminance of reference white, usually normalized to either 1.0 or 100:

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16, \quad \frac{Y}{Y_n} > 0.008856$$

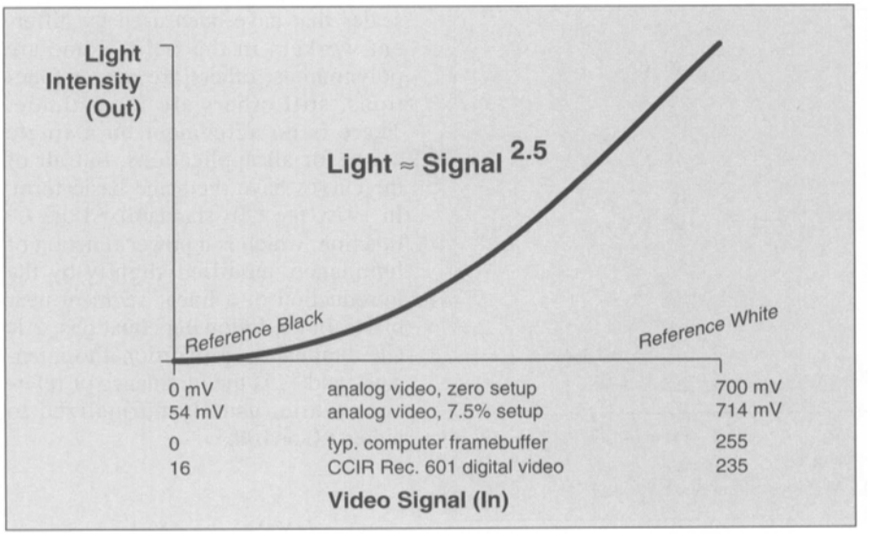
$$L^* = 903.3\left(\frac{Y}{Y_n}\right), \quad \frac{Y}{Y_n} \le 0.008856$$

This is the standard function that relates physical luminance to perceived lightness. The CIE refers to L^* as the lightness component of a *uniform color space. The term perceptually linear* is not appropriate: since we cannot directly measure the quantity in question, we cannot ascribe to it the properties of mathematical linearity.

Roughly speaking, perceived lightness is the cube root of luminance.



Relationship between lightness-scale value V and luminance factor Y plotted in accordance with different formulae [see Table 1(6.3)]. 1: Priest, Gibson, and McNicholas (1920); 2: Munsell, Sloan, and Godlove (1933); Godlove (1933); 3: Newhall, Nickerson and Judd (1943) (also known as the *Munsell Renotation-Value Scale*); 4: CIE (1976) lightness scale (plotting $V = L^*/10$); 5: Foss (1944) (gray scale of the *Color Harmony Manual*); 6: Richter (1953) (gray scale of the German *DIN Color Chart*; plotting $V = 10 - D$).



Gamma in Physics

The physics of the electron gun of a CRT dictates a relationship between voltage input and light output that a physicist calls a *five-halves power law:* the intensity of light produced at the face of the screen is proportional to the voltage input raised to the power 5/2. In other words, intensity is roughly between the square and cube of the voltage. The numerical value of the exponent of the power function is represented by the Greek letter γ (gamma). CRT monitors have voltage inputs that reflect this power function, and in practice the numerical value of gamma is very close to its theoretical value of 2.5.

The sketch above shows the power function that applies to the single electron gun of a gray-scale CRT, or to each of the three red, green, and blue electron guns of a color CRT. The functions associated with the three guns of a color CRT are very similar to each other, but not necessarily identical. The function is dictated by the electron gun and has nothing to do with the CRT's phosphor.

The process of precompensating for this nonlinearity — by computing a voltage signal from an intensity value — is known as *gamma correction.* The function required is approximately a 0.45-power function, whose graph is similar to that of a square root function. In video, gamma correction is accomplished by analog circuits at the camera. In computer graphics, gamma correction is usually accomplished by incorporating the function into a frame buffer's lookup table.

The actual value of *gamma* for a particular CRT may range from about 2.3 to 2.6. Practitioners of computer graphics frequently claim that the numerical value of *gamma* can vary widely from 2.5. Actually, by far the largest source of variation in the nonlinearity of a monitor is caused by the *black level* or *brightness* adjustment of the monitor. Make sure that your display's brightness control is adjusted so that black elements in the picture are reproduced correctly before you devote any effort to determining or setting *gamma.*

Users of Apple Macintosh™ computers sometimes claim that the color monitors used with those computers have a *gamma* value considerably lower than 2.2 — say 1.4 or 1.8. But these values are a function of the rather poorly documented software interpretation of RGB values presented to Apple's QuickDraw software, and have no direct relationship to the monitors used on Macintosh computers.

The Amazing Coincidence!

The first section of this note described the nonlinear response of human vision that relates luminance to perceived lightness. The second section described how the nonlinear transfer function of a CRT relates a voltage signal to intensity. Here's the amazing coincidence: the CRT voltage-to-intensity function is very nearly the *inverse* of the luminance-to-lightness relationship of vision. That means that coding a luminance signal into voltage, to be turned back into luminance by a CRT, is very nearly the optimal coding to minimize the perceptibility of noise that is introduced into the signal. CRT voltage is remarkably perceptually uniform.

Suppose you have a luminance value that is determined quite precisely, but you must use a channel having only 8 bits to convey that value to a distant observer. Consider a linear light representation with eight bits, where code zero represents black and code 255 represents white. Code value 100 represents a shade of gray that is approximately at the perceptual threshold: for codes above 100, the ratio between intensity values of adjacent codes is less than 1%, and for codes below 100, the ratio between intensity values of adjacent code values is greater than 1%.

Luminance codes below 100 suffer increasing artifacts as the code value decreases towards black, due to the visibility of the luminance difference between adjacent codes: at code 50, the ratio between adjacent codes is 2%, which is noticeable to most observers. These artifacts are especially objectionable in pictures having large areas of smoothly varying shades.

Luminance codes above 100 suffer no artifacts due to visibility of the jumps between codes. However, as the code value increases towards white, the codes have decreasing perceptual utility. For example, at code 200, the ratio between adjacent intensity steps is 1/2%, well below the threshold of visibility. Codes 200 and 201 are visually indistinguishable: code 201 is perceptually useless, and could be discarded without being noticed. This example shows that a linear-luminance representation is a poor choice for an 8-bit channel.

The *Perception* section of this note drew the conclusion that it is sufficient for perceptual purposes to maintain about a 1% luminance ratio between adjacent codes. This can be achieved by coding the signal nonlinearly, as roughly the logarithm of luminance. To the extent that the log function is an accurate model of the

Power Laws in Perception

Percept	Physical quantity	Power
Loudness	Sound pressure level	0.67
Saltiness	Sodium chloride concentration	1.4
Smell	Concentration of aromatic molecules	0.6
Heaviness	Mass	1.45

contrast sensitivity function, full perceptual use is made of every code.

As I mentioned in the previous section, logarithmic coding rests on the assumption that the threshold function can be extended to large luminance ratios. Experiments have shown that this assumption does not hold very well, and coding according to a power law is found to be a better approximation to lightness response than a logarithmic function.

The lightness sensation can be computed as intensity raised to approximately the third power: coding a luminance signal to a signal by the use of a power law with an exponent of between 1/3 and 0.45 has excellent perceptual performance. Incidentally, other senses behave according to power laws, according to the table above.

Gamma in Film

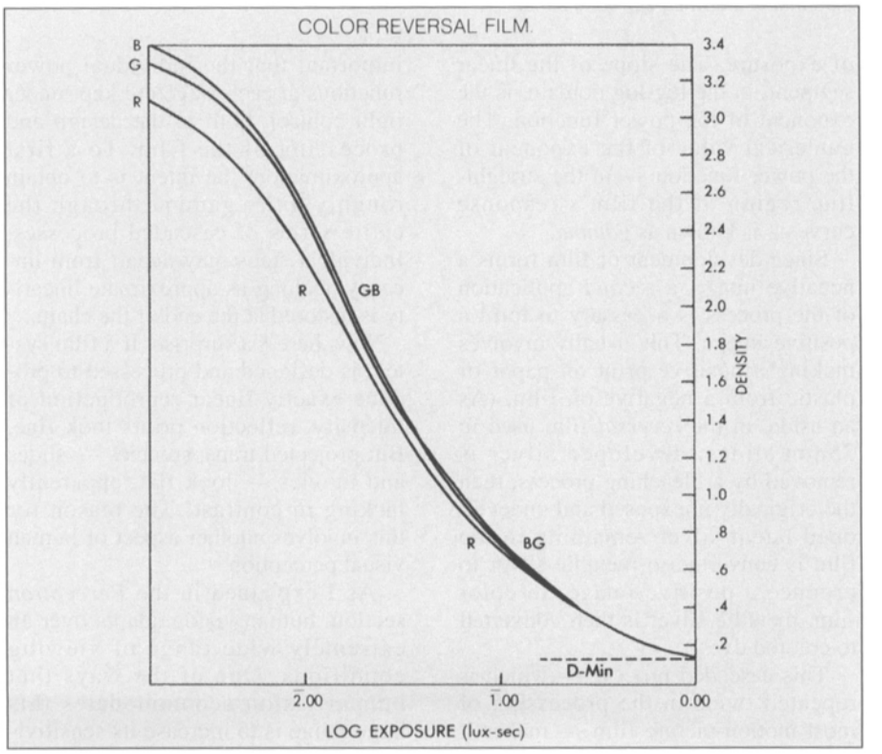
This section describes gamma in photographic film. I will give some background on the photographic process, then explain why physically accurate reproduction of luminance values does not give subjectively good results. Video systems exploit this gem of wisdom from photography; subjectively better images can be obtained if proper account is taken of viewing conditions.

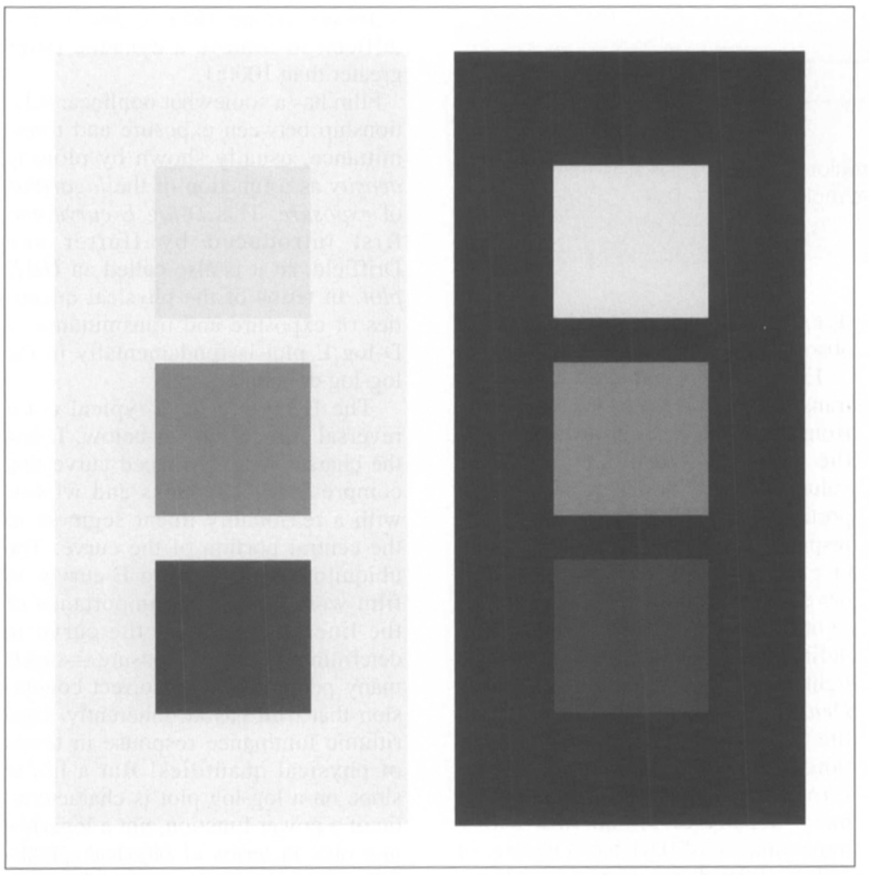
When photographic film is exposed to light and then developed, light imaged from the scene onto the film causes the development process to produce small grains of metallic silver. This process intrinsically creates a negative image: where light causes silver to be developed, the developed silver absorbs light and appears dark. Color film comprises three layers of emulsion sensitized to different wavelength bands, roughly red, green, and blue. The development process converts silver in these three layers into dyes that act as colored filters to absorb red, green, and blue light.

Film can be characterized by the transfer function that relates *exposure* from the scene to the *transmittance* of the developed film. The exposure value at any point on the film is proportional to the luminance of the corresponding point in the scene. A higher exposure causes more silver to be developed, hence more absorption of light in the negative. *Transmittance* is defined as unity minus the fraction of light absorbed by the developed film. *Density* is defined as the negative of the base 10 logarithm of transmittance. Clear film has a density of zero; film with a transmittance of 0.1 has a density of 1, and film with a transmittance of 0.01 has a density of 2. It is difficult in practice to achieve a density greater than 3, that is, it is difficult to achieve a dynamic range greater than 1000:1.

Film has a somewhat nonlinear relationship between exposure and transmittance, usually shown by plotting *density* as a function of the *logarithm* of *exposure*. This *D-log E curve* was first introduced by Hurter and Driffield, so it is also called an *H&D plot*. In terms of the physical quantities of exposure and transmittance, a D-log E plot is fundamentally in the log-log domain.

The H&D plot of a typical color reversal film is shown below. It has the characteristic S-shaped curve that compresses both blacks and whites, with a reasonably linear segment in the central portion of the curve. The ubiquitous use of D-log E curves in film work — and the importance of the linear segment of the curve in determining correct exposure — leads many people to the incorrect conclusion that film has an inherently logarithmic luminance response in terms of physical quantities! But a linear slope on a log-log plot is characteristic of a power function, not a logarithmic one: in terms of physical quantities, transmittance is a power function





Surround Effect. The three gray squares surrounded by white are identical to the three gray squares surrounded by black, but the contrast of the black-surround series appears lower than that of the white-surround series.

of exposure. The slope of the linear segment, in the log-log domain, is the exponent of the power function. The numerical value of the exponent of the power function — in the straight-line region of the film's response curve — is known as *gamma*.

Since development of film forms a negative image, a second application of the process is necessary to form a positive image. This usually involves making a positive print on paper or plastic from a negative on film. (As an aside, in the *reversal* film used in 35mm slides, developed silver is removed by a bleaching process, then the originally unexposed and undeveloped latent silver remaining in the film is converted to metallic silver to produce a positive image. In color film, metallic silver is then converted to colored dyes.)

This cascaded process — which is repeated twice in the processing of most motion-picture film — makes it important that the individual power functions at each stage are kept under tight control, both in the design and processing of the film. To a first approximation, the intent is to obtain roughly *unity* gamma through the entire series of cascaded processes. Individual steps may depart from linearity, as long as approximate linearity is restored at the end of the chain.

Now, here's a surprise. If a film system is designed and processed to produce exactly linear reproduction of intensity, reflection prints look fine. But projected transparencies — slides and movies — look flat, apparently lacking in contrast! The reason for this involves another aspect of human visual perception.

As I explained in the *Perception* section, human vision adapts over an extremely wide range of viewing conditions. One of the ways that human vision accommodates this huge range is to increase its sensitivity to small brightness variations when the area of interest is surrounded by bright elements. Intuitively, light from a bright surround can be thought of as spilling or scattering into all areas of our vision, including the area of interest, reducing its apparent contrast. Loosely speaking, the effect is similar to "flare," and the visual system compensates for it by "stretching" its contrast range to increase the visibility of dark elements in the presence of a bright surround. Conversely, when the region of interest is surrounded by relative darkness, the contrast range of the vision system decreases: our ability to discern dark elements in the scene decreases.

The effect is demonstrated in the sketch at the left, taken from the article "Color Science for Imaging Systems," by LeRoy E. DeMarsh and Edward J. Giorgianni, in *Physics Today* (Sept. 1989, pp. 44-52). The three gray squares surrounded by white are identical to the corresponding gray squares surrounded by black, but the contrast of the black-surround series appears lower than that of the white-surround series.

This has implications for the display of images in dark areas, such as projection of movies in a cinema, projection of 35mm slides, or viewing of television in your living room. If an image is viewed in a *dark* or *dim surround*, and the intensity of the scene is reproduced physically correctly, the image will appear lacking in contrast.

Film systems are designed to compensate for viewing surround effects. Film intended for viewing with a dark surround is designed and processed to have a gamma considerably greater than unity — about 1.5, in fact — so that the contrast range of the scene is expanded upon display. Video signals are coded in a similar manner, taking into account a dim surround viewing environment as I will describe in a moment.

The important conclusion to take from this section is that image coding is not simply concerned with mathematics, physics, chemistry, and electronics; perceptual considerations play an essential role in successful image systems.

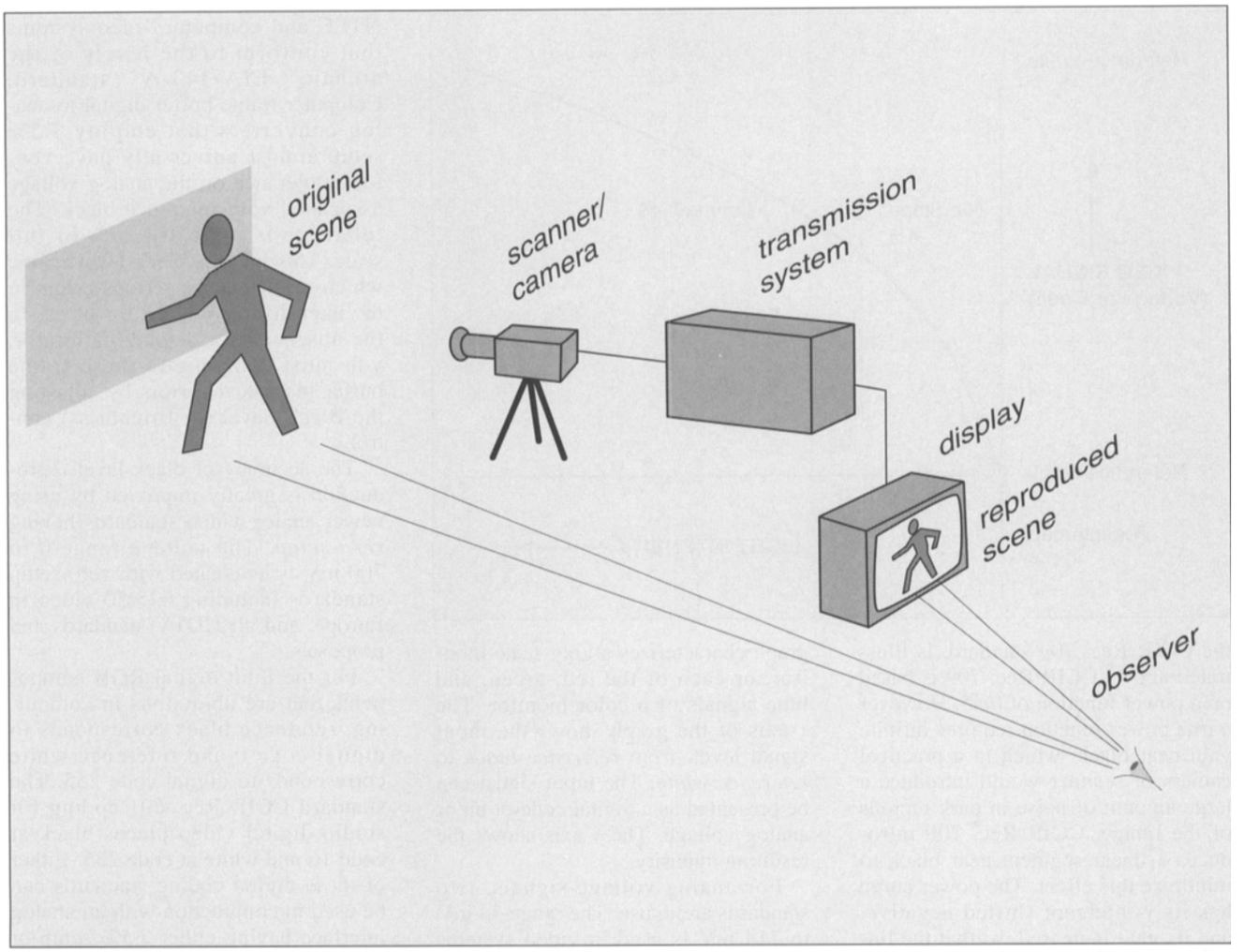


Image Reproduction in Video. To convey a convincing impression of the scene at the left to the observer at the right, intensity from the original scene must be reproduced at the display, possibly with a scale factor to account for an overall intensity change. However, the ability of human vision to detect an intensity difference is not uniform across the range from black to white, but is approximately a constant ratio — about 1% — of the intensity. In order for noise introduced by the transmission system to have minimum perceptual impact, intensity from the scene is transformed by a function similar to a square root into a nonlinear, perceptually-uniform signal that is transmitted. The nonlinear signal is transformed back to linear intensity at the display. Essentially, the camera is designed to have a signal response that mimics the human visual system, in order to "see" lightness in the scene the same way that a human observer would.

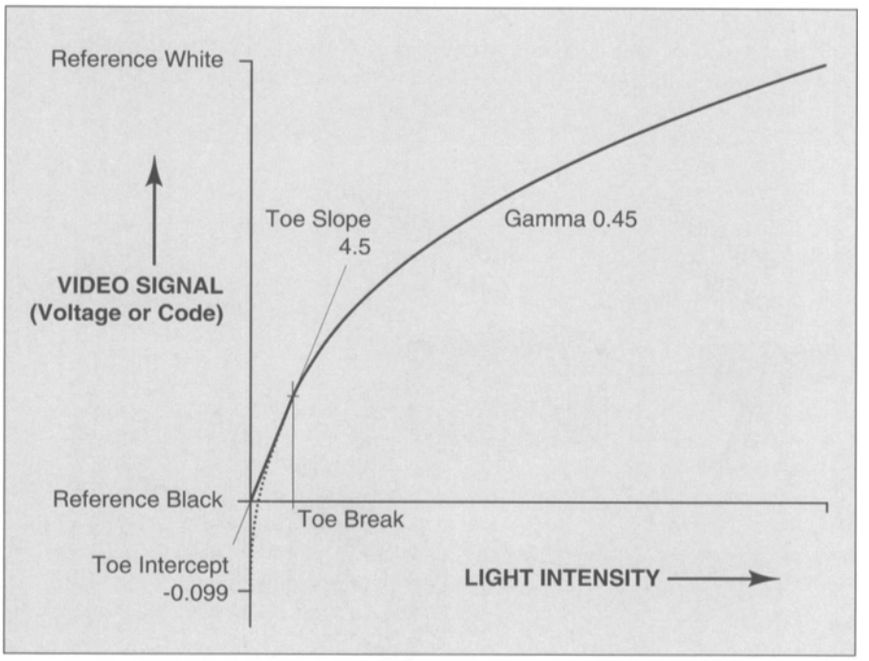
Gamma in Video

In a video system, gamma correction is applied at the camera for the dual purposes of coding into perceptually uniform space and precompensating the nonlinearity of the display's CRT. The first of these considerations was important in the early days of television, because of the need to minimize the noise introduced by VHF over-the-air transmission of broadcast television. However, the same considerations of noise visibility apply to analog videotape recording, and also to minimization of the quantization noise that is introduced at the front end of a digital system when a signal representing intensity is quantized to a limited number of bits. Consequently, video signals are almost always conveyed in gamma-corrected form.

As I explained on the facing page, when viewing an image on a screen in a dim surround it is important for perceptual reasons to "stretch" the contrast ratio of the reproduced image. The dim surround condition is characteristic of television viewing. In video, the "stretching" is accomplished at the camera, by slightly undercompensating the actual power function of the CRT to obtain an end-to-end power function with an exponent of 1.1 or 1.2. This technique produces pictures that are much more subjectively pleasing than those produced by a mathematically correct linear system.

Video standards specify a gamma value of about 2.2 for purposes of precorrection: the product of the 1/2.2 exponent at the camera and the 2.5 exponent at the display produces the desired end-to-end exponent of about 1.13.

The transfer function standardized for high-definition television, part of



the CCIR Rec. 709 standard, is illustrated above. CCIR Rec. 709 is based on a power function of 0.45. However a true power function requires infinite gain near black, which in a practical camera or scanner would introduce a large amount of noise in dark regions of the image. CCIR Rec. 709 introduces a linear segment near black to minimize this effect. The power curve has its *y*-intercept shifted negative, and its gain increased, so that the linear segment meets the curve where their values and slopes match. The mapping of unity is undisturbed. (Details of this transfer function will be included in a tutorial article, "Introduction to Component Video Coding," to appear in a forthcoming issue of the *SMPTE Journal.*)

CRT Transfer Function Details

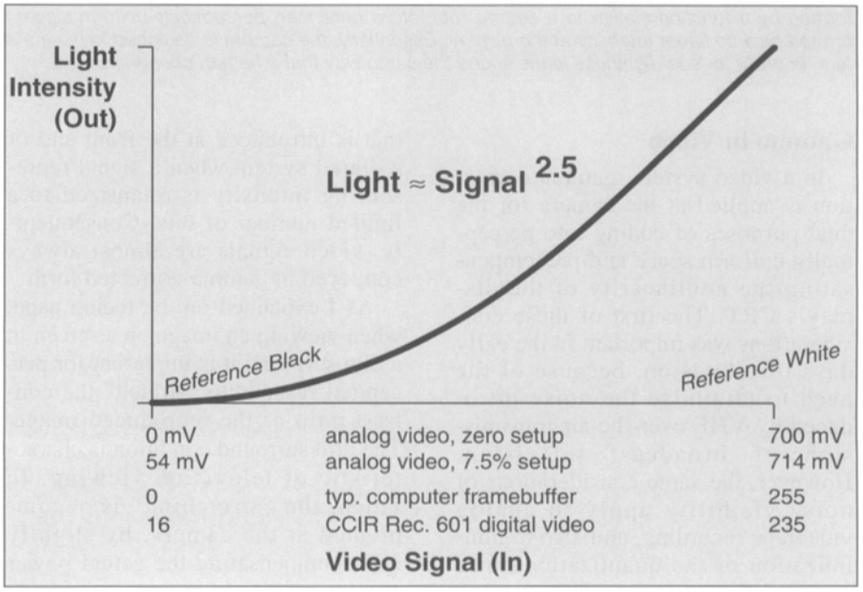
This section provides technical information concerning the nonlinearity of a CRT that is important if you wish to determine the transfer function of your CRT, or to calibrate your monitor, or to understand the electrical voltage interface between a computer frame buffer and a monitor. If you are not interested in these details, I encourage you to skip this section!

The graph to the right illustrates the relationship of the signal input to a monitor to the light luminance produced at the face of the screen. The graph characterizes a gray-scale monitor, or each of the red, green, and blue signals of a color monitor. The *x*-axis of the graph shows the input signal level, from *reference black* to *reference white*. The input signal can be presented as a digital code or an or analog voltage. The *y*-axis shows the resulting intensity.

For analog voltage signals, two standards are in use. The range 54 mV to 714 mV is used in video systems that have *7.5% setup*, including composite 525/59.94 systems such as NTSC and computer video systems that conform to the levels of the archaic EIA-343-A standard. Computer frame buffer digital-to-analog converters that employ 7.5% setup almost universally have very loose tolerance on the analog voltage associated with reference black. The tolerance is typically ±5% of full scale. This induces black-level errors, which in turn cause serious errors in the intensity reproduced for black. In the absence of a display calibrator, you must compensate these frame buffer black-level errors by adjusting the Black Level (or Brightness) control.

The accuracy of black-level reproduction is greatly improved by using newer analog video standards having *zero setup*. The voltage range 0 to 700 mV is associated with zero-setup standards including 625/50 video in Europe, and all HDTV standards and proposals.

For the 8-bit digital RGB components that are ubiquitous in computing, reference black corresponds to digital code 0 and reference white corresponds to digital code 255. The standard CCIR Rec. 601 coding for studio digital video places black at code 16 and white at code 235. Either of these digital coding standards can be used in conjunction with an analog interface having either 7.5% setup or zero setup. Coding of imagery with extended color gamut may place the



black and white codes even further inside the range 0 and 255, for reasons having to do with color reproduction that are outside the scope of this note.

The nonlinearity in the voltage-to-intensity function of a CRT originates with the electrostatic interaction between the cathode and the grid that controls the current of the electron beam. Contrary to popular opinion, the CRT phosphors themselves are quite linear, at least up to an intensity of about 8/10 of peak white where saturation begins to set in.

Knowing that CRT is intrinsically nonlinear, and that its response is based on a power function, many users attempt to summarize the nonlinearity of a CRT display in a single numerical parameter using the obvious relationship:

$$intensity = voltage^{\gamma}$$

This model shows wide variability in the value of gamma, mainly due to black-level errors that the model cannot accommodate due to its being "pegged" at zero: the model forces zero voltage to map to zero intensity for *any* value of gamma. Black-level errors that displace the transfer function upwards can only be "fit" by choosing a gamma value that is much smaller than 2.5. Black-level errors that displace the curve downwards — saturating at zero over some portion of low voltages — can only get a good "fit" by having a value of gamma that is much larger than 2.5. In effect, the only way the single gamma parameter can fit a black-level variation is to alter the curvature of the function. The apparent wide variability of gamma under this model has given *gamma* a bad reputation.

A much better model is obtained by fixing the exponent of the power function at 2.5, and using the single parameter to accommodate black-level error:

$$intensity = (voltage + \varepsilon)^{2.5}$$

This model fits the observed nonlinearity much better than the variable-gamma model.

If you want to determine the nonlinearity of your monitor, I recommend the article, "An Inexpensive Scheme for Calibration of a Colour Monitor in terms of CIE Standard Coordinates," by William B. Cowan (*Computer Graphics*, 17:315-321, July 1983). In addition to describing how to measure the nonlinearity, the article also describes how to determine other characteristics of your monitor — such as the chromaticity of its white point and its primaries — that are important for accurate color reproduction.

Gamma in Computer Graphics

Computer graphics software systems generally perform calculations for lighting, shading, depth-cueing and anti-aliasing using intensity values that model the physical mixing of light. Intensity values stored in the frame buffer are gamma-corrected by hardware lookup tables "on the fly" on their way to the display. The power function at the CRT acts on the gamma-corrected signal voltages to reproduce the correct intensity values at the face of the screen. Software systems usually provide a default gamma value and some method to change the default.

Since color monitors take red, green, and blue voltage inputs, most computer software uses the *RGB color model*. Each color *component* (or *channel*) — red, green, and blue — is typically represented in high-level software by a floating point number in the range 0 to 1. Graphics library software typically translates these floating point values to 8-bit integers in the range from 0 to 255 for use by the graphics hardware. In both the floating point and integer representations, the minimum value in the red component means "as black as you can get". The maximum value in the red component means "as red as you can get", without regard to whether the monitor displays its reds more red or more orange or more purple than another. The color characterization of monitors outside the scope of this note.

A *true color* system has separate red, green, and blue components for each pixel of an image or in a frame buffer. In most computer systems, each component is represented by a *byte* of 8 bits, so each pixel has 24 bits of color information. It is convenient to build a frame buffer or memory system with each complete pixel having a number of bits that is a power of 2. Consequently, a true color frame buffer usually has 32 bits/pixel, where the 8 additional bits are used for purposes other than representing color.

The RGB components of each pixel in a 24-bit system can represent one of 16.7 million codes, but the number of colors that can be distinguished is considerably less than this. With 24-bit color, near-photographic quality images can be displayed and geometric objects can be rendered smoothly-shaded. A true-color frame buffer almost invariably has three lookup tables, one for each component.

The voltage signal between 0 and 1 required to display a red, green, or blue luminance between 0 and 1 can be computed as:

$$signal = intensity^{\left(\frac{1}{gamma}\right)}$$

In the C language this can be represented as follows:

```
signal = pow((double)intensity,
        (double)1.0/gamma));
```

In the absence of data regarding the actual gamma value of your monitor — or for an image intended for interchange in gamma-corrected form — the recommended value of gamma is 1/0.45 (or about 2.222).

You can construct a gamma-correction lookup table like this:

```
#define SIG_FROM_INTEN(i)
((int)( 255.0 *
pow((double)(i) / 255.0, 0.45)))
int sig_from_inten[256], i;
for (i=0; i<256; i++)
sig_from_inten[i] =
SIG_FROM_INTEN(i);
```

Loading this table into the hardware look-up table at the output side of a frame buffer will cause RGB intensity values with integer components between 0 and 255 to be gamma-corrected by the hardware as if by the following C code:

```
red_signal = sig_from_inten[r];
green_signal = sig_from_inten[g];
blue_signal = sig_from_inten[b];
```

The provision of a lookup table at the output of the frame buffer makes it possible to use frame-buffer signal representations other than linear light. For example, it is possible to load gamma-corrected video signals into the frame buffer and load a unity ramp into the lookup table.

The availability of a lookup table at the frame buffer also makes it possible for software to perform sleazy tricks,

such as inverting all of the lookup table entries momentarily to flash the screen without modifying any of the data in the frame buffer. Direct access to frame buffer lookup tables by applications makes it difficult or impossible for system software to avoid annoyances such as colormap flashing and to provide features such as accurate color reproduction. To allow the user to make use of these features, applications should access lookup tables in the structured ways that are provided by the graphics system.

Pseudocolor

A *pseudocolor* (or *indexed color*, or *color mapped*) system dedicates several bits — usually eight — to each pixel in an image or frame buffer. The content of each pixel is used as an index into a color lookup table (CLUT, or *colormap*) that retrieves red, green, and blue signal values upon display of the pixel. An *8-bit* frame buffer displays at most 256 different colors on the screen at once. A typical lookup table has 8-bit values for each of red, green, and blue, so each pixel can be chosen from set of 16.7 million possible codes. As in 24-bit color systems, the number of these colors from this set that can be distinguished is considerably less than 16 million.

Pseudocolor systems usually have lookup tables whose outputs produce voltage at the display. Consequently, it is conventional for a pseudocolor (or *"8-bit"*) application to provide, to a graphics system, RGB color values that are already gamma-corrected for a typical monitor. Consequently, a pseudocolor image stored in a TIFF, GIF, or SunRaster file is almost invariably accompanied by a *colormap* whose RGB values implicitly incorporate gamma correction. If these RGB values are loaded into a 24-bit frame buffer whose lookup table is arranged to gamma-correct intensity values, the pseudocolor values will be gamma corrected a second time, resulting in a severely "washed-out" appearance of the colors.

If you want to recover intensity from gamma-corrected RGB values, for example to "back-out" the gamma correction that is implicit in RGB colormap values associated with an 8-bit colormapped image, construct an inverse-gamma table. You can employ a lookup technique as above, but building an inverse table INTEN_FROM_SIG using the exponent (1.0/0.45) instead of 0.45. Be aware that the perceptual uniformity of the gamma-corrected image will be severely compromised by mapping into the 8-bit intensity domain: *contouring* will be introduced into the darker shades. This topic is covered in the next section.

The Limitations of 8-Bit Intensity

As I mentioned in the *Gamma in Computer Graphics* section earlier, computer graphics systems that render synthetic imagery usually perform computations in the linear-light (intensity) domain. Graphics accelerators usually perform Gouraud shading in the intensity domain, and store 8-bit intensity components in the frame buffer. Eight-bit intensity representations suffer *contouring* artifacts due to the poor perceptual performance of 8-bit intensities, due to the contrast sensitivity threshold of human vision that I discussed in the *Perception* section of this note. The visibility of contouring is enhanced by a perceptual effect called *Mach bands*, consequently the artifact is sometimes called *banding*.

In fixed-point intensity coding where black is code zero, code 100 is at the threshold of visibility at 1% contrast sensitivity: code 50 represents the darkest gray that can be reproduced without the increments between adjacent codes being perceptible. I call this value *best gray*. I consider one of the determinants of the quality of a computer graphics image to be the intensity ratio between *brightest white* and *best gray*. In an 8-bit linear-light system, this ratio is a mere 2.55:1. If an image is contained within this contrast ratio then it will not exhibit banding, but the low contrast ratio will cause the image to appear flat. If an image has a contrast ratio substantially larger than 2.5:1, then it is liable to show banding. In 12-bit linear-light coding the ratio improves to 40:1, which is adequate for the office but does not approach the quality of a photographic reproduction or the cinema.

High-end software systems that are not dependent on hardware acceleration usually perform rendering calculations in intensity domain, then map into the gamma-corrected domain through software and write gamma-corrected values into the frame buffer. These systems can produce rendered images that are free from the quantization artifacts of 8-bit intensity representations.

The Future of Gamma Correction

Color-management systems for PCs and workstations will soon allow device-independent specification of color. Users and applications will be able to specify colors in a device-independent form based on the CIE international standards for color, without any concern about gamma correction. This will make it easy to obtain color matching across different graphics libraries, and different hardware. In the meantime, you can take the following steps:

- Establish good viewing conditions. If you are using a CRT display, you will get better image quality if your overall ambient illumination is reduced.
- Ensure that the Black Level (or Brightness) of your CRT is set to correctly reproduce black elements on the screen.
- Use gamma-corrected representations of RGB values whenever you can. An image coded in gamma-corrected space has very good perceptual uniformity and will display much higher quality than if coded as 8-bit intensity values.
- When you exchange images either in true color or pseudocolor form, code RGB color values using the CCIR Rec. 709 transfer function, gamma corrected with a 0.45-power function.
- In the absence of reliable information about your monitor, assume a gamma value of 2.5 at the display and precorrect accordingly. If your monitor is viewed in a dim surround, precorrect for a gamma value of about 2.2.

Acknowledgments

I would like to thank LeRoy E. DeMarsh for many helpful discussions on these topics. I thank my employer, Sun Microsystems Computer Corp., — and particularly my manager, Dr. Paul Borrill — for permitting me to spend considerable time contributing to SMPTE activities. Linda Bohn and Dean Stanton of Sun also contributed to this note.