Gamma 6

In physics, *intensity* is defined as radiant power per unit solid angle; it has units of watts per steradian ($W \cdot sr^{-1}$). Grayscale image data is normally based upon *relative luminance*, which is intensity per unit area, weighted by the spectral sensitivity of human vision, and normalized to a reference white. This chapter concerns the nonlinear mapping of relative luminance. In this chapter, I use the term *intensity* to emphasize the linear-light nature of the associated quantity. In the following chapter, I will detail luminance.

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. This chapter explains 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 chapter focuses on electronic reproduction of images, using video and computer graphics techniques and equipment. I deal mainly with the reproduction of intensity, or, as a photographer would say, *tone scale*. This is one important step to achieving good color reproduction; more detailed information about color can be found in *Color science for video*, on page 153.

A *cathode-ray tube* (CRT) is inherently nonlinear: The intensity of light reproduced at the screen of a CRT monitor is a nonlinear function of 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.

As explained in *Luminance and lightness*, on page 81, the human perceptual response to intensity is distinctly

nonuniform: The *lightness* sensation of vision is roughly a power function of intensity. This characteristic needs to be considered 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.

Combining these two concepts – one from physics, the other from perception – 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. If gamma correction were not already necessary for physical reasons at the CRT, we would have to invent it for perceptual reasons.

Photography also involves nonlinear intensity reproduction. 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 concerns the appearance of pictures rather than the accurate reproduction of intensity values. The appearance aspects of *gamma* in film also apply to television and computer displays.

Finally, I will describe how video draws aspects of its handling of *gamma* from 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.

Gamma in physics

The physics of the electron gun of a CRT imposes 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 $5/2$. Intensity is roughly between the square and cube of the voltage. The numerical value of the exponent of

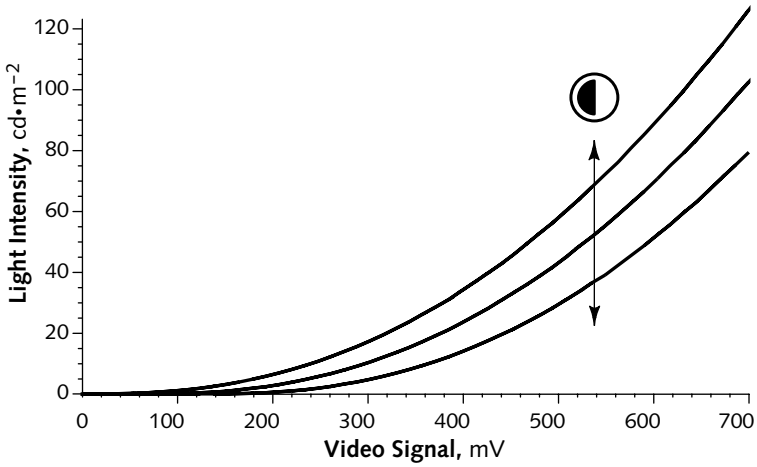

Figure 6.1 **CRT transfer function** involves a nonlinear relationship between video signal and light intensity, here graphed for an actual CRT at three different settings of the *Picture* control. Intensity is approximately proportional to input signal voltage raised to the 2.5 power. The *gamma* of a display system – or more specifically, a CRT – is the numerical value of the exponent of the power function.

the power function is represented by the Greek letter γ (*gamma*). CRT monitors have voltage inputs that reflect this power function. In practice, most CRTs have a numerical value of *gamma* very close to 2.5.

Figure 6.1 above is a sketch of the power function that applies to the single electron gun of a grayscale CRT, or to each of the 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 construction of the electron gun; the CRT's phosphor has no significant effect.

Gamma correction involves a power function, which has the form $y = x^a$ (where *a* is constant). It is sometimes incorrectly claimed to be an exponential function, which has the form $y = a^x$ (where *a* is constant).

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 framebuffer's lookup table.

Alan Roberts, "Measurement of display transfer characteristic (gamma, γ)," *EBU Technical Review* 257 (Autumn 1993), 32–40.

The actual value of *gamma* for a particular CRT may range from about 2.3 to 2.6. Practitioners of computer graphics often claim numerical values of *gamma* quite different from 2.5. But the largest source of variation in the nonlinearity of a monitor is caused by careless setting of the *Black Level* (or *Brightness*) control of your monitor. Make sure that this control is adjusted so that black elements in the picture are reproduced correctly before you devote any effort to determining or setting *gamma*.

Getting the physics right is an important first step toward proper treatment of gamma, but it isn't the whole story, as you will see.

The amazing coincidence!

In *Luminance and lightness*, on page 81, I described the nonlinear relationship between luminance and perceived lightness. The previous section described how the nonlinear transfer function of a CRT relates a voltage signal to intensity. Here's the surprising coincidence: The CRT voltage-to-intensity function is very nearly the *inverse* of the luminance-to-lightness relationship of vision. Representing lightness information as a voltage, to be transformed into luminance by a CRT's power function, is very nearly the optimal coding to minimize the perceptibility of noise. CRT voltage is remarkably perceptually uniform.

Suppose you have a luminance value that you wish to communicate to a distant observer through a channel having only 8 bits. Consider a linear light representation, 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 of intensity values between adjacent codes is less than 1 percent; and for codes below 100, the ratio of intensity values between adjacent code values is greater than 1 percent.

For luminance values below 100, as the code value decreases toward black, the difference of luminance

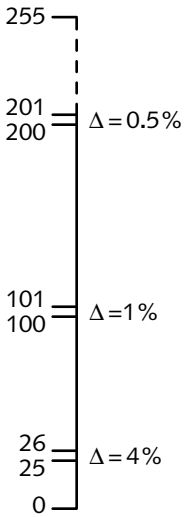

Figure 6.2 **Fixed-point linear-light coding.**

values between adjacent codes becomes increasingly visible: At code 25, the ratio between adjacent codes is 4 percent, which is objectionable to most observers. These errors are especially noticeable in pictures having large areas of smoothly varying shades, where they are known as *contouring* or *banding*.

Luminance codes above 100 suffer no artifacts due to visibility of the jumps between codes. However, as the code value increases toward white, the codes have decreasing perceptual utility. For example, at code 200 the ratio between adjacent codes is 0.5 percent, 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, sketched in Figure 6.2 in the margin, shows that a linear-luminance representation is a bad choice for an 8-bit channel.

In an image coding system, it is sufficient, for perceptual purposes, to maintain a ratio of luminance values between adjacent codes of about a 1 percent. 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 contrast sensitivity function, full perceptual use is made of every code.

As 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 a power of approximately the one-third: 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.

S. S. Stevens, *Psychophysics*. New York: John Wiley & Sons, 1975.

Incidentally, other senses behave according to power functions:

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

Gamma in film

This section describes gamma in photographic film. I give some background on the photographic process, then explain why physically accurate reproduction of luminance values gives subjectively poor 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 film is exposed, light imaged from the scene onto the film causes a chemical change to the emulsion of the film, and forms a *latent image*. Subsequent development causes conversion of the latent image into small grains of metallic silver. This process intrinsically creates a negative image: Where light causes silver to be developed, the developed film 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 to the transmittance of the developed film. When film is exposed in a camera, the exposure value at any point on the film is proportional to the luminance of the corresponding point in the scene, multiplied by the exposure time.

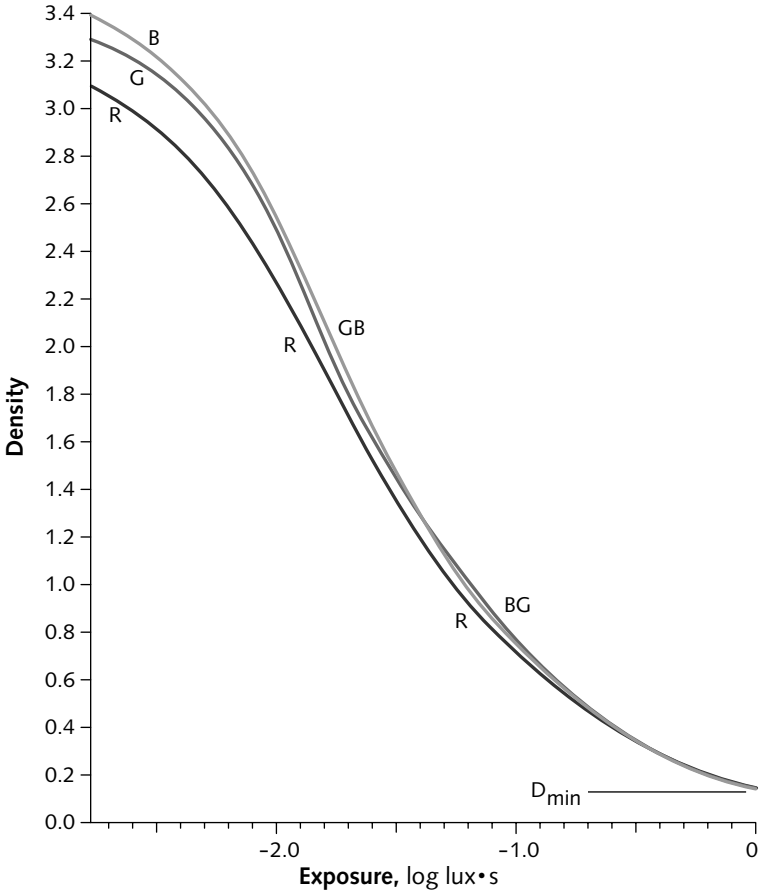

Figure 6.3 **Tone response of color reversal film.** This graph is redrawn, with permission, from Kodak Publication H-1. It shows the S-shaped exposure characteristic of typical color-reversal photographic film. Over the *straight-line* portion of the log-log curve, the density of the developed film is a power function of exposure intensity.

EASTMAN Professional Motion Picture Films, Kodak Publication H-1, Fourth Edition. Rochester, NY: Eastman Kodak Company, 1992. Figure 26.

$$D = \log_{10}\left(\frac{P_0}{P_T}\right)$$

D: Optical density

P_0: Incident power

P_T: Transmitted power

See Table 7.5, *Density examples*, on page 153.

Transmittance is defined as the fraction of light incident on the developed film to light absorbed. *Density* is the logarithm of incident power divided by transmitted power. The characteristic of a film is 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.

A typical film plotted in this way is shown in the plot in Figure 6.3 above. The plot shows an S-shaped curve that compresses blacks, compresses whites, and has 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 function: In terms of physical quantities, transmittance of a typical film is a power function of exposure. The slope of the linear segment, in the log-log domain, is the exponent of the power function; in the straight-line region of the film's response curve its numerical value 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 from a negative on film. In the *reversal* film used in 35 mm 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.

This cascaded process is repeated twice in the processing of motion picture film. It is important that the individual power functions at each stage are kept under tight control, both in the design and the 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: the *surround effect*.

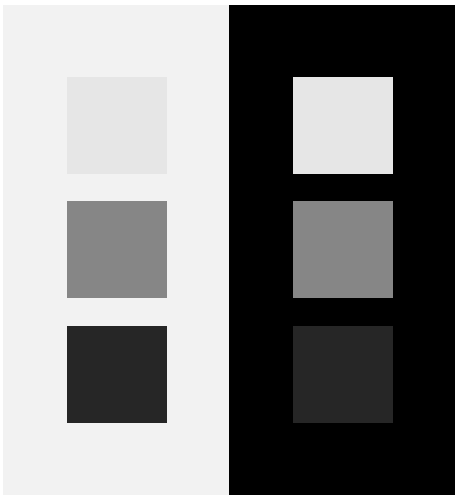

Figure 6.4 **Surround effect.** The three gray squares surrounded by white are identical to the three gray squares surrounded by black, but the contrast of the black-surround series appears lower than that of the white-surround series.

Surround effect

As explained in *Adaptation*, on page 85, human vision adapts to an extremely wide range of viewing conditions. One of the mechanisms involved in adaptation increases our 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 vision system compensates for this effect 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 Figure 6.4 above, from DeMarsh and Giorgianni.

LeRoy E. DeMarsh and Edward J. Giorgianni, "Color Science for Imaging Systems," in *Physics Today*, September 1989, 44–52.

The surround effect has implications for the display of images in dark areas, such as projection of movies in a cinema, projection of 35 mm 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 with correct physical intensity, the image will appear lacking in contrast.

Film systems are designed to compensate for viewing surround effects. Transparencies (slide) film is intended for viewing in a dark surround. Slide film is designed to have a gamma considerably greater than unity – about 1.5 – so that the contrast range of the scene is expanded upon display. Video signals are coded in a similar manner, taking into account viewing in a dim surround, as I will describe in a moment.

The important conclusion to take from this section is that image coding for the reproduction of pictures for human viewers is not simply concerned with mathematics, physics, chemistry, and electronics. Perceptual considerations play an essential role in successful image systems.

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. Figure 6.5 opposite summarizes the image reproduction situation for video. Gamma correction is applied at the camera, at the left; the display, at the right, imposes the inverse power function.

Coding into a perceptual domain was important in the early days of television because of the need to minimize the noise introduced by over-the-air transmission. However, the same considerations of noise visibility apply to analog videotape recording, and also to 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, it is universal to convey video signals in gamma-corrected form.

As explained in *Gamma in film*, on page 96, it is important for perceptual reasons to "stretch" the contrast ratio of a reproduced image when viewed in a dim surround. The dim surround condition is characteristic

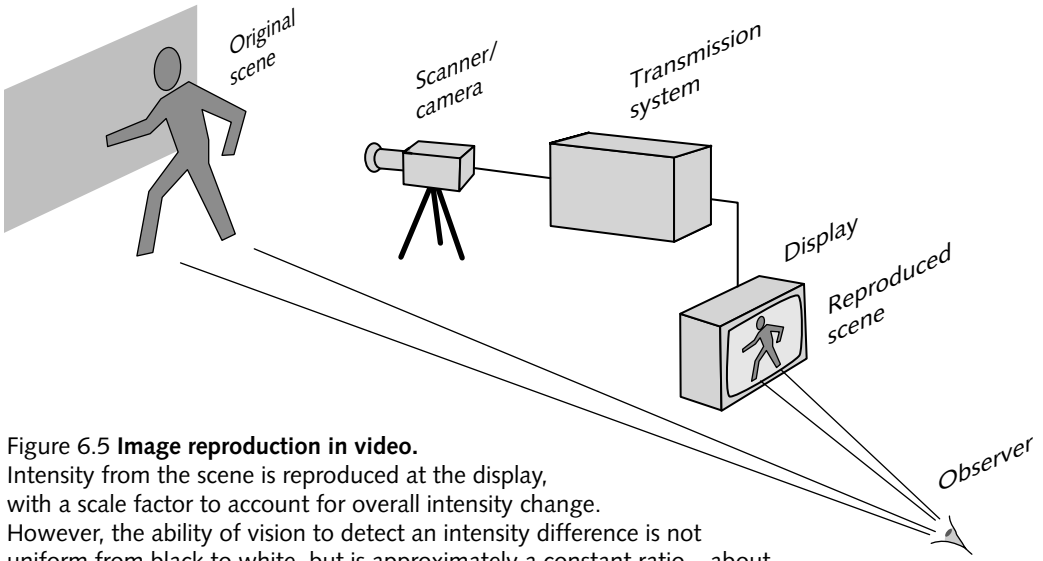

Figure 6.5 **Image reproduction in video.**
Intensity from the scene is reproduced at the display, with a scale factor to account for overall intensity change. However, the ability of vision to detect an intensity difference is not uniform from black to white, but is approximately a constant ratio – about 1 percent – of the intensity. In video, intensity from the scene is transformed by a function similar to a square root into a nonlinear, perceptually uniform signal that is transmitted. The camera is designed to mimic the human visual system, in order to "see" lightness in the scene the same way that a human observer would; noise introduced by the transmission system then has minimum perceptual impact. The nonlinear signal is transformed back to linear intensity at the display, using the 2.5-power function that is intrinsic to the CRT.

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 achieves pictures that are more subjectively pleasing than would be produced by a mathematically correct linear system.

$$0.45 = \frac{1}{2.222\dot{}}$$

$$\frac{1}{2.2} = 0.45\dot{4}\dot{5}$$

$$0.45 \times 2.5 \approx 1.13$$

Rec. 709 specifies a power function exponent of 0.45. The product of the 0.45 exponent at the camera and the 2.5 exponent at the display produces the desired end-to-end exponent of about 1.13. An exponent of 0.45 is a good match for both CRTs and for perception. Some video standards have specified an exponent of $1/2.2$.

Emerging display devices such as liquid crystal displays (LCDs) have nonlinearity different from that of a CRT. But it remains important to use image coding that is

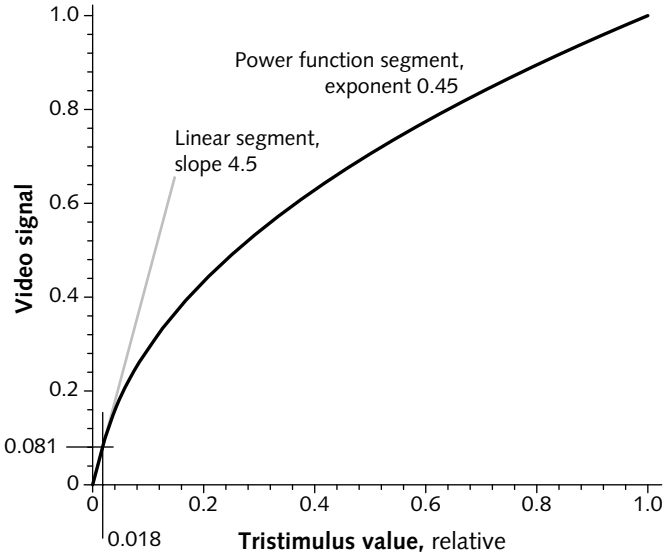

Figure 6.6 **Rec. 709 transfer function.**

well matched to perception. Furthermore, image interchange standards using the 0.45 value are very well established. The economic importance of equipment that is already built to these standards will deter any attempt to establish new standards just because they are better matched to particular devices. We can expect new display devices to incorporate local correction, to adapt between their intrinsic transfer functions and the transfer function that has been standardized for image interchange.

Rec. 709 transfer function

Figure 6.6 above illustrates the transfer function defined by the international Rec. 709 standard for high-definition television (HDTV). It is basically a power function with an exponent of 0.45. Theoretically a pure power function suffices for gamma correction; however, the slope of a pure power function is infinite at zero. In a practical system such as a television camera, in order to minimize noise in the dark regions of the picture it is necessary to limit the slope (gain) of the function near black. Rec. 709 specifies a slope of 4.5 below a tristimulus value of +0.018, and stretches the remainder of the curve to maintain func-

ITU-R, *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange*, Rec. BT.709 (Geneva: ITU).

tion and tangent continuity at the breakpoint. In this equation the red tristimulus (linear light) component is denoted R, and the resulting gamma-corrected video signal is denoted with a prime symbol, R'_{709}. The computation is identical for the other two components:

Eq 6.1

$$E'_{709} = \begin{cases} 4.5L, & L \le 0.018 \\ 1.099L^{0.45} - 0.099, & 0.018 < L \end{cases}$$

Standards for conventional 525/59.94 video have historically been very poorly specified. The original NTSC standard called for precorrection assuming a display power function of 2.2. Modern 525/59.94 standards have adopted the Rec. 709 function.

Formal standards for 625/50 video call for precorrection for an assumed power function exponent of 2.8 at the display. This is unrealistically high. In practice the Rec. 709 transfer function works well.

SMPTE 240M transfer function

SMPTE Standard 240M for 1125/60 HDTV was adopted several years before international agreement was achieved on Rec. 709. Virtually all HDTV equipment that has been deployed as I write this uses SMPTE 240M parameters. The 240M parameters are slightly different from those of Rec. 709:

Eq 6.2

$$E'_{240} = \begin{cases} 4.0L, & L \le 0.0228 \\ 1.1115L^{0.45} - 0.1115, & 0.0228 < L \end{cases}$$

The difference between the SMPTE 240M and Rec. 709 transfer functions is negligible for real images. It is a shame that international agreement could not have been reached on the SMPTE 240M parameters that were widely implemented at the time the CCIR (now ITU-R) discussions were taking place.

The Rec. 709 values are closely representative of current studio practice, and should be used for all but very unusual conditions.

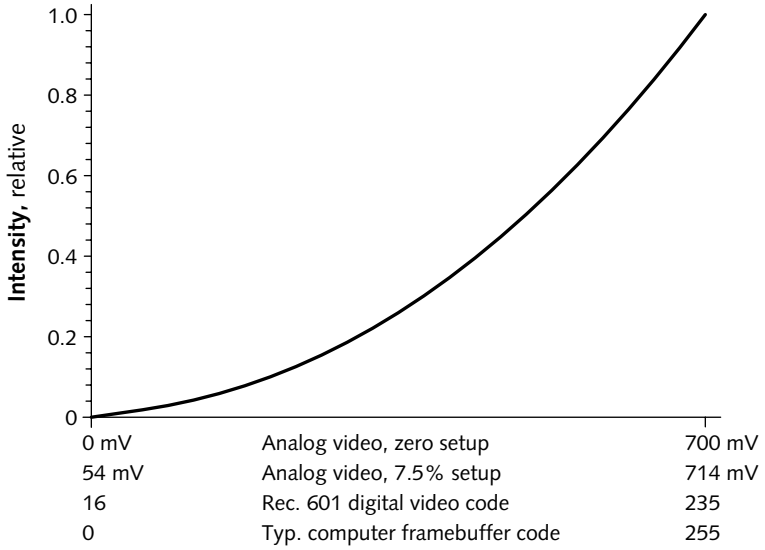

Figure 6.7 **CRT signal levels and intensity.** A video signal may be represented as analog voltage, with zero setup or with 7.5-percent setup. Alternatively, the signal may be represented digitally using coding from 0 to 255 (for computer graphics), or Rec. 601 coding from 16 to 235 (for studio video).

CRT transfer function details

This section provides technical information concerning the nonlinearity of a CRT. This section is important if you wish to determine the transfer function of your CRT, to calibrate your monitor, or to understand the electrical voltage interface between a computer framebuffer and a monitor.

Figure 6.7 above illustrates the function that relates signal input to a CRT monitor to the light intensity produced at the face of the screen. The graph characterizes a grayscale monitor, or each of the red, green, and blue components 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 analog voltage according to one of several standards. 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-percent setup*, including composite 525/59.94 systems such as NTSC, and computer video systems that conform to the *levels* of the archaic *EIA RS-343-A* standard. Computer framebuffer digital-to-analog converters often have 7.5-percent setup; these almost universally have very loose tolerance of about ±5 percent of full scale on the analog voltage associated with reference black. 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 framebuffer black-level errors by adjusting the *Black Level* (or *Brightness*) control on your monitor. This act effectively marries the monitor to the framebuffer.

The accuracy of black level reproduction is greatly improved in newer analog video standards that have *zero setup*. The voltage range 0 to 700 mV is used in 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 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-percent setup or zero setup. Coding of imagery with an extended color gamut may place the black and white codes even further inside the coding range, for reasons having to do with color reproduction that are outside the scope of this chapter.

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 eight-tenths of peak white at the onset of saturation.

Knowing that a CRT is intrinsically nonlinear, and that its response is based on a power function, many users have attempted to summarize the nonlinearity of a CRT display in a single numerical parameter using this relationship, where E is code or voltage and L is luminance:

Eq 6.3

$$L = E^{\gamma}$$

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 upward can be "fit" only by choosing a gamma value that is much smaller than 2.5. Black-level errors that displace the curve downward – saturating at zero over some portion of low voltages – can get a good "fit" only 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 a single parameter to accommodate black-level error:

Eq 6.4

$$L = (E + \varepsilon)^{2.5}$$

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

William B. Cowan, "An Inexpensive Scheme for Calibration of a Colour Monitor in terms of CIE Standard Coordinates," in *Computer Graphics*, vol. 17, no. 3 (July 1983), 315–321.

If you want to determine the nonlinearity of your monitor, consult the article by Cowan. In addition to describing how to measure the nonlinearity, he 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 antialiasing using intensity values that model the physical mixing of light. Intensity values stored in the framebuffer 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.

The voltage E between 0 and 1 required to display a red, green, or blue luminance L between 0 and 1 is this:

Eq 6.5

$$E = L^{\left(\frac{1}{\gamma}\right)}$$

In the C language this can be represented as follows:

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

In the absence of data regarding the actual gamma value of your monitor, or to encode 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 suitable for computer graphics applications, like this:

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

Loading this table into the hardware lookup table at the output side of a framebuffer will cause *RGB* luminance 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_tristim[r];
green_signal = sig_from_tristim[g];
blue_signal = sig_from_tristim[b];
```

A lookup table at the output of the framebuffer enables signal representations other than linear-light. If gamma-corrected video signals are loaded into the framebuffer, then a unity ramp is appropriate at the lookup table. This arrangement will maximize perceptual performance.

The availability of a lookup table at the framebuffer makes it possible for software to perform tricks, such as inverting all of the lookup table entries momentarily to flash the screen without modifying any data in the framebuffer. Direct access to framebuffer 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.

Gamma in video, computer graphics, SGI, and Macintosh

Transfer functions in video, computer graphics, Silicon Graphics, and Macintosh are sketched in Figure 6.8 opposite. Video is shown in the top row. Gamma correction is applied at the camera, and signals are maintained in a perceptual domain throughout the system until conversion back to intensity at the CRT.

Computer graphics systems generally store intensity values in the framebuffer, and gamma-correct on the fly through hardware lookup tables on the way to the display, as illustrated in the second row.

Silicon Graphics computers, by default, use a lookup table with a 1.7-power function; this is shown in the third row.

What are loosely called *JPEG files* use the *JPEG File Interchange Format* (JFIF). Version 1.02 of that specification states that linear-light coding (gamma 1.0) is used. That is seldom the case in practice. Instead, power laws of 0.45, $^{1}/_{1.8} \approx 0.55$, or $^{1.7}/_{2.5} \approx 0.68$ are used.

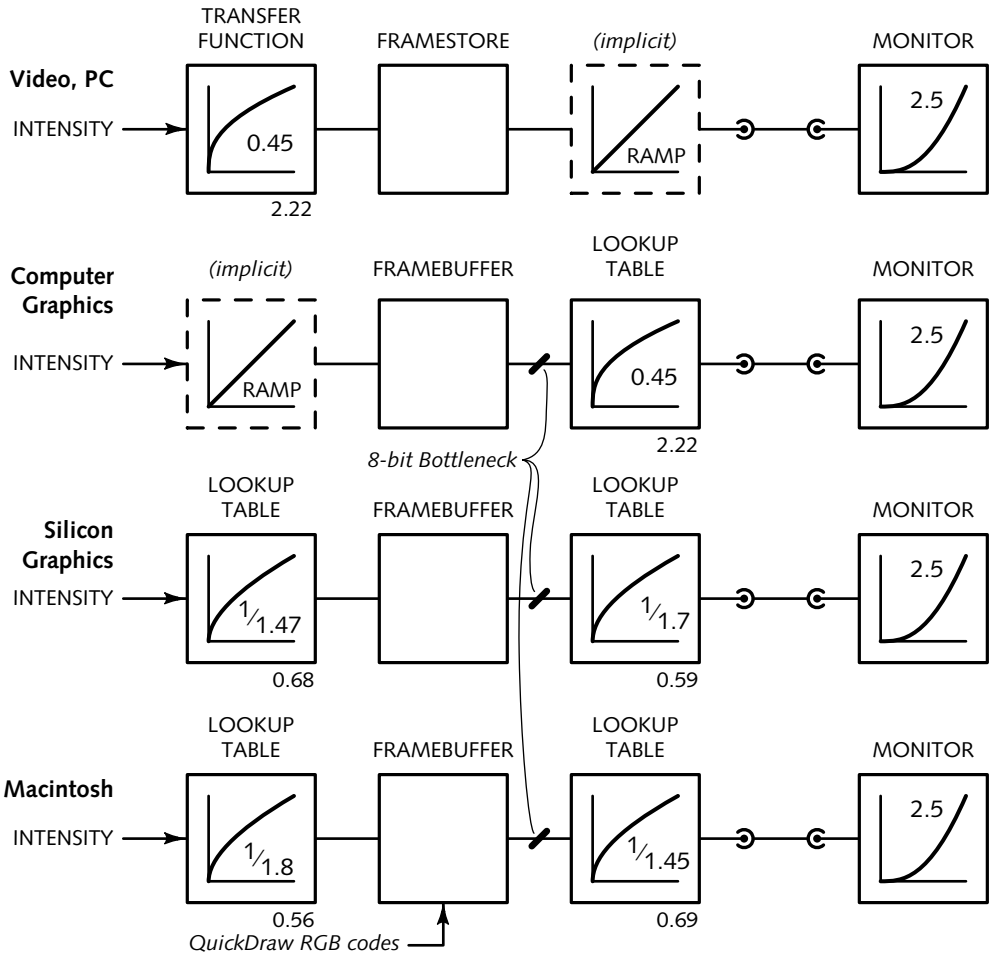

Figure 6.8 **Gamma in video, computer graphics, SGI, and Macintosh.** In a video system, shown in the top row, a transfer function in accordance with vision is applied at the camera. The middle row illustrates computer graphics: Calculations are performed in the linear light domain and gamma correction is applied in a lookup table at the output of the framebuffer. Silicon Graphics computers take a hybrid approach: Part of the correction is accomplished in software, and a $1/1.7$ power function is loaded into the lookup table. The approach used by Macintosh computer sketched in the bottom row.

JFIF files originated on Macintosh ordinarily encode *R*, *G*, and *B* tristimulus (intensity) values raised to the $1/1.8$ power.

Macintosh computers use the approach shown in the bottom row. Part of gamma correction is effected by application software prior to presentation of *RGB* values to the QuickDraw graphics subsystem; the remainder is accomplished in the lookup tables. The dominance of Macintosh computers in graphic arts and prepress has made "gamma 1.8" a *de facto* standard.

Pseudocolor

In *Raster images in computing*, on page 1, I described how pseudocolor systems have lookup tables whose outputs are directly mapped to voltage at the display. It is conventional for a pseudocolor application program to provide, to a graphics system, *RGB* color values that are already gamma-corrected for a typical monitor. A pseudocolor image stored in a file is accompanied by a *colormap* whose *RGB* values incorporate gamma correction. If these values are loaded into a 24-bit framebuffer whose lookup table is arranged to gamma-correct intensity values, the pseudocolor values will be gamma-corrected a second time, resulting in poor image quality.

If you want to recover intensity from gamma-corrected *RGB* values, for example to "back-out" the gamma correction that is implicit in the *RGB* colormap values associated with an 8-bit colormapped image, construct an inverse-gamma table. You can employ a lookup technique as above, building an inverse table TRISTIM_FROM_SIG using code similar to the SIG_FROM_TRISTIM code on page 107, but with an exponent of $^{1}/_{0.45}$ instead of 0.45. Be aware that the perceptual uniformity of the gamma-corrected image will be compromised by mapping into the 8-bit intensity domain: *Contouring* – which I will discuss on page page 113 – will be introduced into the darker shades.

Halftoning

Figure 6.9

Continuous-tone (grayscale or color) image data can be reproduced using a process, such as color photography or thermal dye transfer printing, where a continuously variable amount of color material can be deposited at each point in the image. But some reproduction processes – offset lithographic printing and laser-printers, for example – place a fixed density of color at each point in the reproduced image. Grayscale and color images are *halftoned* – or *screened* – in order to be displayed on these devices. Halftoning produces apparent continuous tone by varying the area of small dots in a regular array. Viewed at a sufficient distance,

an array of small dots produces the perception of light gray, and an array of large dots produces dark gray.

Halftone dots are usually placed on a regular grid, although *stochastic screening* has recently been introduced, which modulates the spacing of the dots rather than their size. The screening is less visible because the pattern is not spatially correlated.

In order for halftoning to produce a reasonably good impression of continuous tone, the individual dots must not be too evident. If it is known that a reproduction will be viewed at a certain distance, then the number of screen lines per inch can be determined. This measurement can be expressed in terms of the angle subtended by the screen line pitch at the intended viewing distance: If the screen line pitch subtends any more than about one minute of arc at the viewer's retina, screen lines are likely to be visible.

Robert Ulichney, *Digital Halftoning*. Boston: MIT Press, 1988.

Peter Fink, *PostScript Screening: Adobe Accurate Screens*. Mountain View, CA: Adobe Press, 1992.

The standard reference to halftoning algorithms is Ulichney, but he does not detail the nonlinearities found in practical printing systems. For details about screening for color reproduction, consult Fink.

Printing

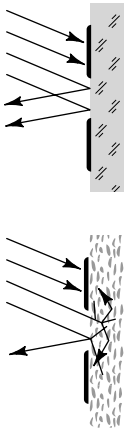

Figure 6.10 **Dot gain mechanism.**

An image destined for halftone printing conventionally specifies each pixel as *dot percentage in film*. An imagesetter's halftoning machinery generates dots whose areas are proportional to the requested coverage. In principle, *dot percentage in film* is inversely proportional to linear-light reflectance.

Two phenomena distort the requested dot coverage values. First, printing involves a mechanical smearing of the ink that causes dots to enlarge. Second, optical effects within the bulk of the paper cause more light to be absorbed than would be expected from the surface coverage of the dot alone. These phenomena are collected under the term *dot gain*, which is the percentage by which the light absorption of the printed dots exceeds the requested dot coverage.

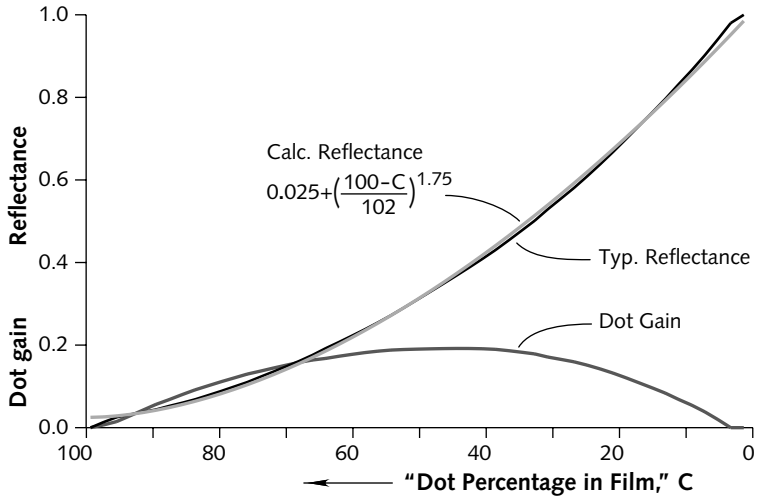

Figure 6.11 **Transfer function in offset printing.** *Dot gain* refers to light absorption in excess of that predicted by ink coverage, or *dot percentage in film*, alone. When expressed as intensity instead of absorption, and as total absorption instead of excess absorption, the standard dot gain characteristic of offset printing reveals a transfer function roughly similar to that of a CRT.

Standard offset printing produces a dot gain at 50 percent of about 22 percent: When 50 percent absorption is requested, 72 percent absorption is obtained. The midtones print darker than requested. This results in a transfer function from code to reflectance that closely resembles the voltage-to-light curve of a CRT. Correction of dot gain is conceptually similar to gamma correction in video: Physical correction of the "defect" in the reproduction process is very well matched to the lightness perception of human vision. Coding an image in terms of dot percentage in film involves coding into a roughly perceptually uniform space. The standard dot gain functions employed in North America and Europe correspond to intensity being reproduced as a power function of the digital code, where the numerical value of the exponent is about 1.75, compared to about 2.2 for video or 3 for CIE *L**. The value used in printing is lower than the optimum for perception, but works well for the rather low contrast ratio of offset printing.

The default power function used in the Macintosh produces a mapping from code to reflectance that is similar to that of printing: Raw QuickDraw *R′G′B′* codes subtracted from 255 and sent to an imagesetter produce reflectance that closely matches the intensity displayed on the Macintosh monitor.

I have described the linearity of conventional offset printing. Other halftoned devices have different characteristics, and require different corrections.

Limitations of 8-bit intensity

As mentioned in *Gamma in computer graphics*, on page 107, 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 framebuffer. Eight-bit intensity representations suffer contouring artifacts, due to the contrast sensitivity threshold of human vision discussed in *Lightness sensitivity*, on page 85. 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 approximately the threshold of visibility at 1 percent contrast sensitivity: Code 100 represents the darkest gray that can be reproduced without the increments between adjacent codes being perceptible. I call this value *best gray*. One of the determinants of the quality of an image is the ratio of intensities between *brightest white* and *best gray*. In an 8-bit linear-light system, this ratio is a mere 2.5:1. 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.

High-end systems for computer generated imagery (CGI) usually do not depend on hardware acceleration. They perform rendering calculations in the intensity domain, perform gamma correction in software, then write gamma-corrected values into the framebuffer. These systems produce rendered imagery without the quantization artifacts of 8-bit intensity coding.

The future of gamma correction

Work is underway to implement facilities in graphics systems to allow device-independent specification of color. Users and applications will be able to specify colors, based on the CIE standards, without concern for gamma correction. When this transition is complete, it will be much easier 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 your monitor's *Black Level* (or *Brightness*) control is set to correctly reproduce black elements on the screen.

- Use gamma-corrected *R′G′B′* representations whenever you can. An image coded with gamma correction has good perceptual uniformity, resulting in an image with much higher quality than one coded as 8-bit intensity values.

- When you exchange images either in truecolor or pseudocolor form, code *R′G′B′* color values using the Rec. 709 gamma value of $1/0.45$.

- In the absence of reliable information about your monitor, display pictures assuming a monitor gamma value of 2.5. If you view your monitor in a dim surround, use a lower value of about 2.2.