Frequently Asked Questions about Gamma

Charles A. Poynton

www.inforamp.net/~poynton
poynton@inforamp.net

tel +1 416 486 3271
fax +1 416 486 3657

In video, computer graphics and image processing the *gamma* symbol γ represents a numerical parameter that describes the nonlinearity of intensity reproduction. 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 FAQ is intended to clarify aspects of nonlinear image coding in computer graphics, image processing, video, and the transfer of digital images to print.

This document is available on the Internet from Toronto at:

<ftp://ftp.inforamp.net/pub/users/poynton/doc/colour/>

It is mirrored to space provided by Fraunhofer Computer Graphics in Rhode Island, U.S.A. at <ftp://elaine.crcg.edu/pub/doc/colour/>, and in Darmstadt, Germany at <ftp://ftp.igd.fhg.de/pub/doc/colour/>.

Table of Contents

1 What is intensity?

Intensity is a measure over some interval of the electromagnetic spectrum of the flow of power that is radiated from, or incident on, a surface. Intensity is what I call a *linear-light* measure, expressed in units such as watts per square meter.

The voltages presented to a CRT monitor control the intensities of the colour components, but in a nonlinear manner. CRT voltages are not proportional to intensity.

Image data stored in a file (TIFF, JFIF, PPM, etc.) may or may not represent intensity, even if it is so described. The *I* component of a color described as *HSI* (hue, saturation, intensity) does not accurately represent intensity if *HSI* is computed according to any of the usual formulae.

2 What is luminance?

Brightness is defined by the Commission Internationale de L'Éclairage (CIE) as *the attribute of a visual sensation according to which an area appears to emit more or less light*. Because brightness perception is very complex, the CIE defined a more tractable quantity *luminance*, denoted *Y*, which is radiant power weighted by a spectral sensitivity function that is characteristic of vision. To learn about the relationship between physical spectra and perceived brightness, and other color issues, refer to the companion *Frequently Asked Questions about Colour*.

The magnitude of luminance is proportional to physical power. In that sense it is like intensity. But the spectral composition of luminance is related to the brightness sensitivity of human vision.

3 What is lightness?

Human vision has a nonlinear perceptual response to brightness: a source having a luminance only 18% of a reference luminance appears about half as bright. The perceptual response to luminance is called *Lightness* and is defined by the CIE [1] as a modified cube root of luminance:

$$L^* = 116\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16; \qquad 0.008856 < \frac{Y}{Y_n}$$

Y_n is the luminance of the white reference. If you normalize luminance to reference white then you need not compute the quotient. The CIE definition applies a linear segment with a slope of 903.3 near black, for (Y/ Y_n) < 0.008856. The linear segment is unimportant for practical purposes but if you don't use it, make sure that you limit L^* at zero. L^* has a range of 0 to 100, and a "delta L-star" of unity is taken to be roughly the threshold of visibility.

Stated differently, lightness perception is roughly logarithmic. You can detect an intensity difference between two patches when the ratio of their intensities differs by more than about one percent.

Video systems approximate the lightness response of vision using RGB signals that are each subject to a 0.45 power function. This is comparable to the 1⁄3 power function defined by L^*.

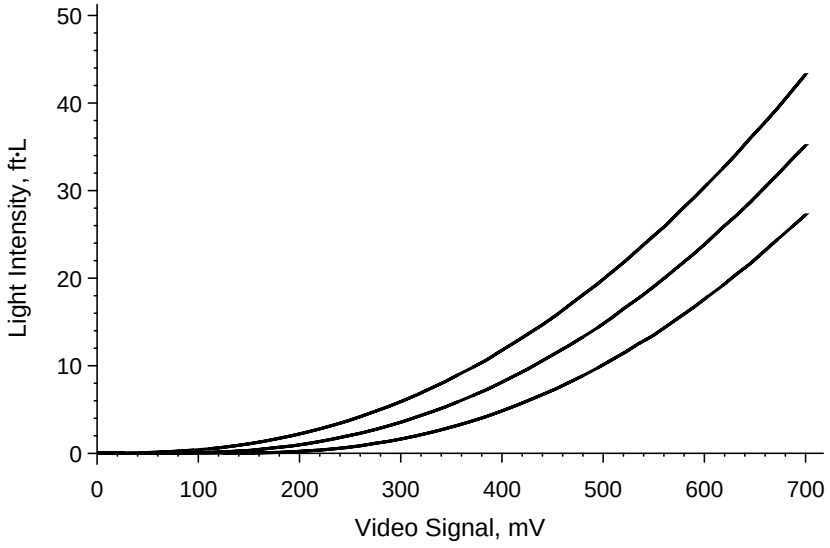
The *L* component of a color described as *HLS* (hue, lightness, saturation) does not accurately represent lightness if *HLS* is computed according to any of the usual formulae. See *Frequently Asked Questions about Colour*.

4 What is gamma?

The intensity of light generated by a physical device is not usually a linear function of the applied signal. A conventional CRT has a power-law response to voltage: intensity produced at the face of the display is approximately the applied voltage, raised to the 2.5 power. The numerical value of the exponent of this power function is colloquially known as *gamma*. This nonlinearity must be compensated in order to achieve correct reproduction of intensity.

As mentioned above (*What is lightness?*), human vision has a nonuniform perceptual response to intensity. If intensity is to be coded into a small number of steps, say 256, then in order for the most effective perceptual use to be made of the available codes, the codes must be assigned to intensities according to the properties of perception.

Here is a graph of an actual CRT's transfer function, at three different CONTRAST settings:



This graph indicates a video signal having a voltage from zero to 700 mV. In a typical eight-bit digital-to-analog converter on a framebuffer card, black is at code zero and white is at code 255.

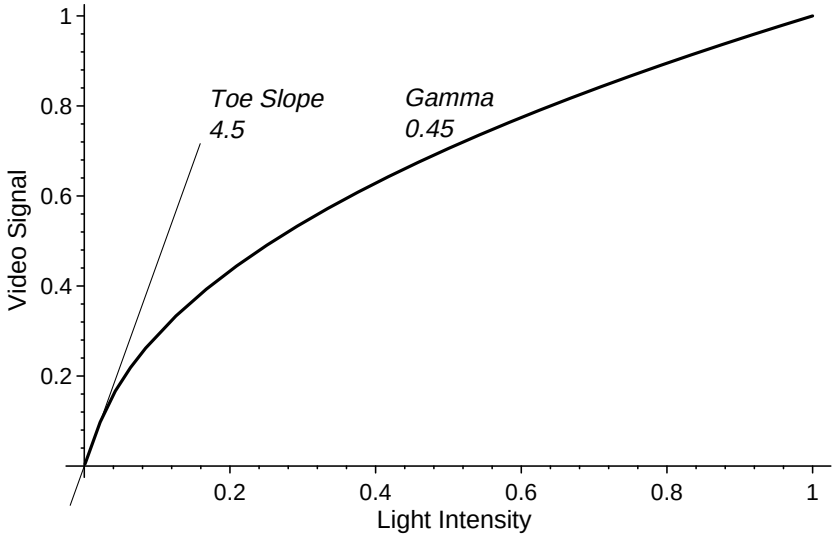
Through an amazing coincidence, vision's response to intensity is effectively the inverse of a CRT's nonlinearity. If you apply a transfer function to code a signal to take advantage of the properties of lightness perception – a function similar to the L^* function – the coding will be inverted by a CRT.

5 What is gamma correction?

In a video system, linear-light intensity is transformed to a nonlinear video signal by *gamma correction*, which is universally done at the camera. The Rec. 709 transfer function [2] takes linear-light intensity (here R) to a nonlinear component (here R'), for example, voltage in a video system:

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

The linear segment near black minimizes the effect of sensor noise in practical cameras and scanners. Here is a graph of the Rec. 709 transfer function, for a signal range from zero to unity:



An idealized monitor inverts the transform:

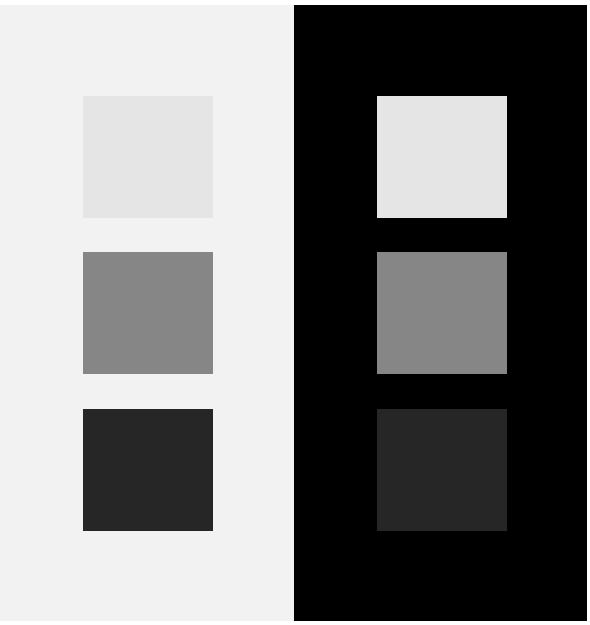
$$R = \begin{cases} \dfrac{R'_{709}}{4.5}, & R'_{709} \le 0.081 \\ \left(\dfrac{R'_{709} + 0.099}{1.099} \right)^{\frac{1}{0.45}}, & 0.081 < R'_{709} \end{cases}$$

Real monitors are not as exact as this equation suggests, and have no linear segment, but the precise definition is necessary for accurate intermediate processing in the linear-light domain. In a colour system, an identical transfer function is applied to each of the three *tristimulus* (linear-light) *RGB* components. See *Frequently Asked Questions about Colour*.

By the way, the nonlinearity of a CRT is a function of the electrostatics of the cathode and the grid of an electron gun; it has nothing to do with the phosphor. Also, the nonlinearity is a power function (which has the form $f(x) = x^a$), not an exponential function (which has the form $f(x) = a^x$). For more detail, read Poynton's article [3].

6 Does NTSC use a gamma of 2.2?

Television is usually viewed in a dim environment. If an images's correct physical intensity is reproduced in a *dim surround*, a subjective effect called *simultaneous contrast* causes the reproduced image to appear lacking in contrast. The effect can be overcome by applying an end-to-end power function whose exponent is about 1.1 or 1.2. Rather than having each receiver provide this correction, the assumed 2.5-power at the CRT is under-corrected at the camera by using an exponent of about $1/2.2$ instead of $1/2.5$. The assumption of a dim viewing environment is built into video coding.



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. – *LeRoy DeMarsh*

7 Does PAL use a gamma of 2.8?

Standards for 625/50 systems mention an exponent of 2.8 at the decoder, however this value is unrealistically high to be used in practice. If an exponent different from 0.45 is chosen for a power function with a linear segment near black like Rec. 709, the other parameters need to be changed to maintain function and tangent continuity.

8 I pulled an image off the net and it looks murky.

If an image originates in linear-light form, gamma correction needs to be applied exactly once. If gamma correction is not applied and linear-light image data is applied to a CRT, the midtones will be reproduced too dark. If gamma correction is applied twice, the midtones will be too light.

9 I pulled an image off the net and it looks a little too contrasty.

Viewing environments typical of computing are quite bright. When an image is coded according to video standards it implicitly carries the assumption of a dim surround. If it is displayed without correction in a bright ambient, it will appear contrasty. In this circumstance you should apply a power function with an exponent of about $1/1.1$ or $1/1.2$ to correct for your bright surround.

Ambient lighting is rarely taken into account in the exchange of computer images. If an image is created in a dark environment and transmitted to a viewer in a bright environment, the recipient will find it to have excessive contrast.

If an image originated in a bright environment and viewed in a bright environment, it will need no modification no matter what coding is applied. But then it will carry an assumption of a bright surround. Video standards are widespread and well optimized for vision, so it makes sense to code with a power function of 0.45 and retain a single standard for the assumed viewing environment.

In the long term, for everyone to get the best results in image interchange among applications, an image originator should remove the effect of his ambient environment when he transmits an image. The recipient of an image should insert a transfer function appropriate for his viewing environment. In the short term, you should include with your image data tags

that specify the parameters that you used to encode. TIFF 6.0 has provisions for this data. You can correct for your own viewing environment as appropriate, but until image interchange standards incorporate viewing conditions, you will also have to compensate for the originator's viewing conditions.

10 What is luma?

In video it is standard to represent brightness information not as a nonlinear function of true CIE luminance, but as a weighted sum of nonlinear R'G'B' components called *luma*. For more information, consult the companion document *Frequently Asked Questions about Colour*.

11 What is contrast ratio?

Contrast ratio is the ratio of intensity between the brightest white and the darkest black of a particular device or a particular environment. Projected cinema film – or a photographic reflection print – has a contrast ratio of about 80:1. Television assumes a contrast ratio – in your living room – of about 30:1. Typical office viewing conditions restrict the contrast ratio of a CRT display to about 5:1.

12 How many bits do I need to smoothly shade from black to white?

At a particular level of adaptation, human vision responds to about a hundred-to-one contrast ratio of intensity from white to black. Call these intensities 100 and 1. Within this range, vision can detect that two intensities are different if the ratio between them exceeds about 1.01, corresponding to a *contrast sensitivity* of one percent.

To shade smoothly over this range, so as to produce no perceptible steps, at the black end of the scale it is necessary to have coding that represents different intensity levels 1.00, 1.01, 1.02 and so on. If linear light coding is used, the "delta" of 0.01 must be maintained all the way up the scale to white. This requires about 9,900 codes, or about fourteen bits per component.

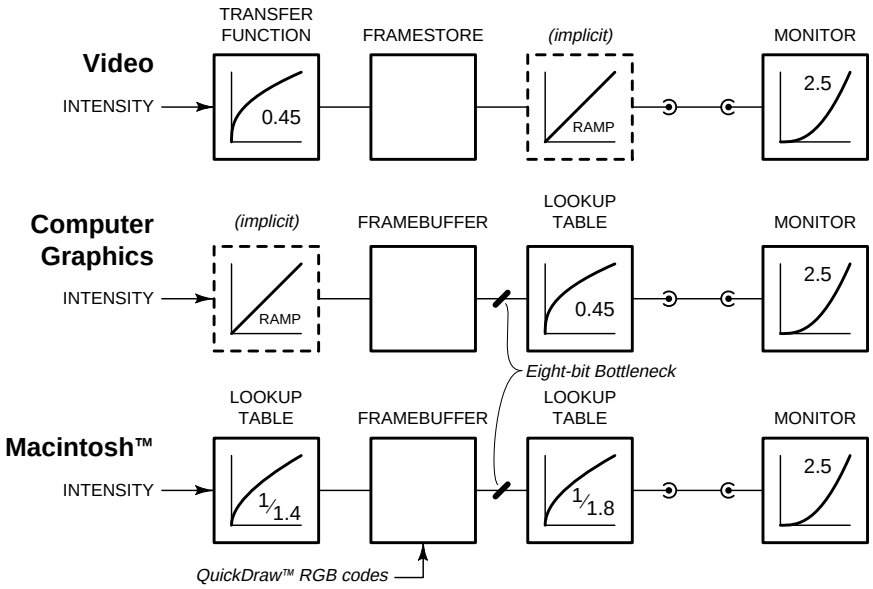
If you use nonlinear coding, then the 1.01 "delta" required at the black end of the scale applies as a ratio, not an absolute increment, and progresses like compound interest up to white. This results in about 460 codes, or about nine bits per component. Eight bits, nonlinearly coded according to Rec. 709, is sufficient for broadcast-quality digital television at a contrast ratio of about 50:1.

If poor viewing conditions or poor display quality restrict the contrast ratio of the display, then fewer bits can be employed.

If a linear light system is quantized to a small number of bits, with black at code zero, then the ability of human vision to discern a 1.01 ratio between adjacent intensity levels takes effect below code 100. If a linear light system has only eight bits, then the top end of the scale is only 255, and contouring in dark areas will be perceptible even in very poor viewing conditions.

13 How is gamma handled in video, computer graphics and desktop computing?

As outlined above, gamma correction in video effectively codes into a perceptually uniform domain. In video, a 0.45-power function is applied at the camera, as shown in the top row of this diagram:



Synthetic computer graphics calculates the interaction of light and objects. These interactions are in the physical domain, and must be calculated in linear-light values. It is conventional in computer graphics to store linear-light values in the framebuffer, and introduce gamma correction at the lookup table at the output of the framebuffer. This is illustrated in the middle row above.

If linear-light is represented in just eight bits, near black the steps between codes will be perceptible as banding in smoothly-shaded images. This is the *eight-bit bottleneck* in the sketch.

Desktop computers are optimized neither for image synthesis nor for video. They have programmable “gamma” and either poor standards or no standards. Consequently, image interchange among desktop computers is fraught with difficulty.

14 What is the gamma of a Macintosh?

Apple offers no definition of the nonlinearity – or loosely speaking, *gamma* – that is intrinsic in QuickDraw. But the combination of a default QuickDraw lookup table and a standard monitor causes intensity to represent the 1.8-power of the R, G and B values presented to QuickDraw. It is wrongly believed that Macintosh computers use monitors whose transfer function is different from the rest of the industry. The unconventional QuickDraw handling of nonlinearity is the root of this misconception. Macintosh coding is shown in the bottom row of the diagram. More detail is available [4].

The transfer of image data in computing involves various transfer functions: at coding, in the framebuffer, at the lookup table, and at the monitor. Strictly speaking the term *gamma* applies to the exponent of the power function at the monitor. If you use the term loosely, in the case of a Mac you could call the gamma 1.4, 1.8 or 2.5 depending which part of the system you were discussing.

I recommend using the Rec. 709 transfer function, with its 0.45-power law, for best perceptual performance and maximum ease of interchange with digital video. If you need Mac compatibility you will have to code intensity with a $1/1.8$-power law, anticipating QuickDraw's $1/1.4$-power in the lookup table. This coding has adequate performance in the bright viewing environments typical of desktop applications, but suffers in darker viewing conditions that have high contrast ratio.

15 Does the gamma of CRTs vary wildly?

Gamma of a properly adjusted conventional CRT varies anywhere between about 2.35 and 2.55.

CRTs have acquired a reputation for wild variation for two reasons. First, if the model $intensity = voltage^{gamma}$ is naively fitted to a display with black-level error, the exponent deduced will be as much a function of the black error as the true exponent. Second, input devices, graphics libraries and application programs all have the potential to introduce their own transfer functions. Nonlinearities from these sources are often categorized as gamma and attributed to the display.

16 How should I adjust my monitor's BRIGHTNESS and CONTRAST controls?

On a CRT monitor, the control labelled *CONTRAST* controls overall intensity, and the control labelled *BRIGHTNESS* controls offset (black level). Display a picture that is predominantly black. Adjust *BRIGHTNESS* so that the monitor reproduces true black on the screen, just at the threshold where it is not so far down as to "swallow" codes greater than the black code, but not so high that the picture sits on a "pedestal" of dark grey. When the critical point is reached, put a piece of tape over the *BRIGHTNESS* control. Then set *CONTRAST* to suit your preference for display intensity.

17 Should I do image processing operations on linear or nonlinear image data?

If you wish to simulate the physical world, linear-light coding is necessary. For example, if you want to produce a numerical simulation of a lens performing a Fourier transform, you should use linear coding. If you want to compare your model with the transformed image captured from a real lens by a video camera, you will have to "remove" the nonlinear gamma correction that was imposed by the camera, to convert the image data back into its linear-light representation.

On the other hand, if your computation involves human perception, a nonlinear representation may be required. For example, if you perform a discrete cosine transform on image data as the first step in image compression, as in JPEG, then you ought to use nonlinear coding that exhibits perceptual uniformity, because you wish to minimize the perceptibility of the errors that will be introduced during quantization.

The image processing literature rarely discriminates between linear and nonlinear coding. In the JPEG and MPEG standards there is no mention of transfer function, but nonlinear (video-like) coding is implicit: unacceptable results are obtained when JPEG or MPEG are applied to linear-light data. In computer graphic standards such as PHIGS and CGM there is no mention of transfer function, but linear-light coding is implicit. These discrepancies make it very difficult to exchange image data between systems.

When you ask a video engineer if his system is linear, he will say "Of course!" referring to linear voltage. If you ask an optical engineer if her system is linear, she will say "Of course!" referring to linear intensity. But

when a nonlinear transform lies between the two systems, as in video, a linear transformation performed in one domain is not linear in the other.

18 What's the transfer function of offset printing?

A image destined for halftone printing conventionally specifies each pixel in terms of *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.

Standard offset printing involves a dot gain at 50% of about 24%: when 50% absorption is requested, 74% 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. This is lower than the optimum for perception, but works well for the low contrast ratio of offset printing.

The Macintosh has a power function that is close enough to printing practice that raw QuickDraw codes sent to an imagesetter produce acceptable results. High-end publishing software allows the user to specify the parameters of dot gain compensation.

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

19 References

[1] Publication CIE Nº 15.2, *Colorimetry, Second Edition* (1986), Central Bureau of the Commission Internationale de L'Éclairage, Vienna, Austria.

[2] ITU-R Recommendation BT.709, *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange* (1990), [formerly CCIR Rec. 709], ITU, 1211 Geneva 20, Switzerland.

[3] Charles A. Poynton, "Gamma and Its Disguises" in *Journal of the Society of Motion Picture and Television Engineers*, Vol. **102**, No. 12 (December 1993), 1099–1108, available on the Internet as <ftp://ftp.inforamp.net/pub/users/poynton/doc/Article_Reprints/SMPTE93_Gamma/>.

[4] Charles A. Poynton, "Gamma on the Apple Macintosh", <ftp://ftp.inforamp.net/pub/users/poynton/doc/Mac/>.