Film Gamma Versus Video Gamma

CFS-242

February 1, 2004

David Dunthorn
www.c-f-systems.com

This document is a complete rewrite of a document we earlier placed on this web site. While nothing was specifically wrong with the earlier document, we believe this one is more detailed and easier to follow. This version arose as we worked on CFS-243 *Maintaining Color Integrity in Digital Photography*, a formal document explaining what we mean when we say that current widely accepted methods in digital photography severely compromise color integrity.

This document forms part of a series of three documents from early 2004 investigating some aspects of digital photography. In addition to CFS-243, the third companion document CFS-244 *Negative to Positive* explores the digital treatment of photographic negatives in the digital world. Refer to those documents, available on our web site, for the further explanation of the symbols, equations, and concepts used here.

This analysis is original with us and is not likely to be found elsewhere. You are welcome to use this analysis in articles, courses, or books, but please credit C F Systems and www.c-f-systems.com when you do.

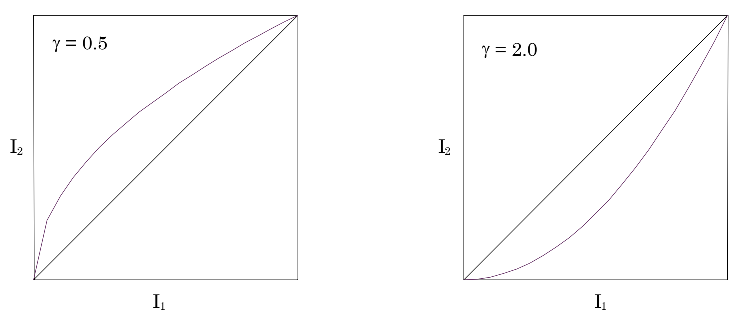
Video Gamma

Video gamma adjustments, as they are pervasively used in digital imaging, depend upon the fact that digital images are almost always expressed as intensity ratios. (More typically as *gamma-adjusted* intensity ratios, but we will deal with the implications of that later.) Images are usually stored as collections of integer values and the stored value is understood to be ratioed with the maximum possible value for that integer. For instance, 8-bit integers run in value from 0 to a maximum of 255, so a value of 130 is understood to represent the ratio 130/255 = 0.510. We can interpret 0/255 as the minimum possible intensity, or black, and 255/255 as the maximum possible intensity, or white. Such ratios can never be less than 0 nor greater than 1 and may also be used to represent individual primary colors, running from 0/255 black to 255/255 fully saturated red, for example.

Considering images that are stored as such ratios, $I / I_{\max}$, the basic gamma transformation is stated as:

$$\left(\frac{I_2}{I_{\max}}\right) = \left(\frac{I_1}{I_{\max}}\right)^{\gamma}$$

as illustrated in the following two example plots. These plots are identical to what Photoshop would show with the Curves tool and represent the curve necessary to produce the equivalent gamma adjustment.



Video Gamma Adjustment Plots

Note that no matter the value of γ, pure black remains pure black and pure white remains pure white. The equation is *normalized* so that all values must be between 0 and 1 inclusive. Normalization is the most important property of video gamma adjustment, making it lossless, in the sense that tonal values cannot be driven into saturation (pure white) nor can they be driven to pure black. In contrast to this, consider a simple, unnormalized brightness adjustment of an image file, $I_2 = I_1 + a$. Since pixels can only have values between 0 and $I_{\max}$, no pixel in the transformed image will have an intensity less than a, so there will be no true blacks and no very dark grays.

Any pixel with an intensity greater than $I_{\max} - a$ in the original image will come out with an intensity greater than $I_{\max}$. Image programs set these all to full white, $I_{\max}$, and any highlight detail they may have represented is lost. Since the video gamma adjustments are normalized, such losses do not occur (although other losses can occur due to precision problems).

Film photographers understand film "gamma" to be synonymous with contrast. Higher gamma means higher contrast, lower gamma means lower contrast. The plots above show that is not true with video gamma. To be increased contrast, the slope of the curve must be steeper than the diagonal line; that is, greater than 1. So, although higher video gamma produces more contrast in the lighter tones, it actually produces *lower* contrast in the darker tones. Lower video gamma does exactly the opposite, producing less contrast in the lighter tones but higher contrast in the darker tones.

We can determine exactly where this transition between raised contrast and lowered contrast happens by taking the derivative (slope) of the curve and finding where it is equal to 1:

$$I_2 = I_{\max}\left(\frac{I_1}{I_{\max}}\right)^{\gamma} = I_{\max}^{1-\gamma} I_1^{\gamma}$$

$$\frac{dI_2}{dI_1} = \gamma\, I_{\max}^{1-\gamma} I_1^{\gamma-1} = \gamma\left(\frac{I_1}{I_{\max}}\right)^{\gamma-1}$$

Setting

$$\gamma\left(\frac{I_1}{I_{\max}}\right)^{\gamma-1} = 1$$

and solving for $I_1/I_{\max}$ gives

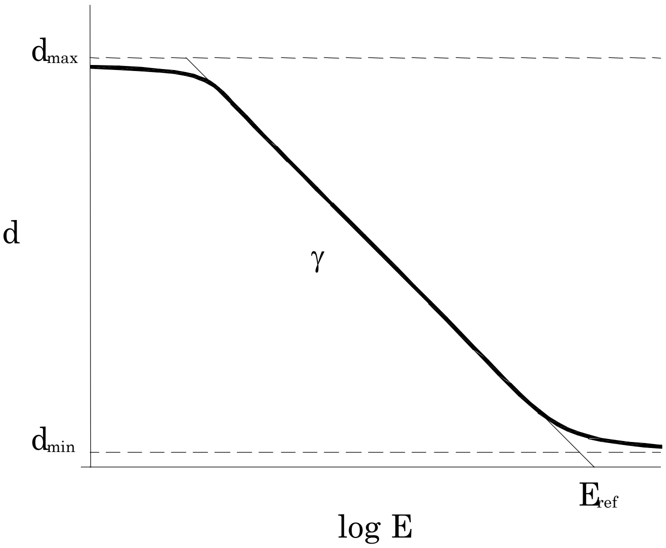
$$\text{Video Gamma Transition Point} = \left(\frac{I_1}{I_{\max}}\right) = \gamma^{\frac{1}{1-\gamma}}$$

Film Gamma

Now we examine the corresponding mathematics of film gamma. The basic gamma equation for a *positive* film is: (see the end of this document for the derivation of this from the more familiar form for a *negative* film)

$$d = -\gamma \log_{10}\left(\frac{E}{E_{ref}}\right)$$

and the equation represents the straight-line portion of the following curve:



Here d is the photographic density, with $d = 0$ meaning transparent film (or fully white) and as d becomes larger, transparency decreases (the film becomes darker). E is the light exposure the film receives, measured in candela-sec/m^2 or similar units. (–)γ is the slope of the straight-line portion of the S-curve. E_{ref} is the exposure for which $d = 0$ when the straight-line portion is extrapolated through the log E axis.

Photographic density is the base10 logarithm of an intensity ratio. In this case, it is the ratio between the intensity of the light impinging on a film transparency to the light passing through a transparency having density d. That is:

$$d = \log_{10}\left(\frac{I_{\max}}{I}\right)$$

and so

$$\log_{10}\left(\frac{I_{\max}}{I}\right) = -\gamma \log_{10}\left(\frac{E}{E_{ref}}\right)$$

giving

$$\left(\frac{I}{I_{\max}}\right) = \left(\frac{E}{E_{ref}}\right)^{\gamma}$$

which begins to look familiar. Now consider exposure, E, as applied to forming an image of a scene. E produced by a point in a scene is essentially the intensity of light at that point as attenuated through an *f*-aperture for a specific length of exposure time. This means that for any point in a scene, $E = K\,I_s$, where I_s is the intensity of light from that point in the scene and K is a constant factor that groups the shutter speed-*f*-aperture effects. We can also define I_{sref} using $E_{ref} = K\,I_{sref}$. Strictly speaking, we have described forming a direct positive image here. A negative-positive process is more complicated and flexible than a direct positive but the overall process can be expressed in this same way using a single *K*. Thus:

$$\left(\frac{I}{I_{\max}}\right) = \left(\frac{I_s}{I_{sref}}\right)^{\gamma}$$

In this form the film gamma equation looks both normalized and identical to the video gamma equation. We find that I_{sref} must correspond to some effective maximum intensity in the scene. Despite this misleading appearance of identity, there is one very critical difference when comparing gamma adjustments in digital images and in film images. To make a video gamma adjustment, which is done in Photoshop by making a Levels tool gray point adjustment; one adjusts γ and that is that. (Of course Photoshop makes you enter the value of $1/\gamma$ instead of γ for the middle gray adjustment. Who knows why?) For film, however, the value of E_{ref} is an experimentally determined result and *is different for every value of* γ, even if the change in γ is only the result of changing the film development time. Moreover, the standard approach when changing contrast in traditional photography is to *adjust the exposure* so that some particular tone in the changed contrast version will match the same tone in the original. Typically, a middle gray is matched between the two versions. Thus when a film gamma adjustment is made I_{sref} will *change*. For example, the exposure time and/or the *f*-aperture may be changed so that the same scene intensity I_s^* will produce the same density d^* for different film gammas. For purposes of comparison with video gamma, let us make two images of the same scene, one using γ_1 and the other using γ_2. The image exposures will be adjusted so that the scene illumination I_s^* will produce density d^* for both images.

$$d^* = -\gamma_1 \log_{10}\left(\frac{I_s^*}{I_{sref1}}\right) = -\gamma_2 \log_{10}\left(\frac{I_s^*}{I_{sref2}}\right)$$

$$\frac{\gamma_1}{\gamma_2}\log_{10}\left(\frac{I_s^*}{I_{sref1}}\right) = \log_{10}\left(\frac{I_s^*}{I_{sref2}}\right)$$

$$\left(\frac{I_s^*}{I_{sref1}}\right)^{\frac{\gamma_1}{\gamma_2}} = \left(\frac{I_s^*}{I_{sref2}}\right)$$

$$I_{sref2} = I_s^*\left(\frac{I_{sref1}}{I_s^*}\right)^{\frac{\gamma_1}{\gamma_2}}$$

The first image will then be:

$$\log_{10}\left(\frac{I_{\max}}{I_1}\right) = -\gamma_1 \log_{10}\left(\frac{I_s}{I_{sref1}}\right)$$

and the second:

$$\log_{10}\left(\frac{I_{\max}}{I_2}\right) = -\gamma_2 \log_{10}\left(\frac{I_s}{I_{sref2}}\right) = -\gamma_2 \log_{10}\left(\frac{I_s}{I_{sref1}} \frac{\left(I_s^*\right)^{\frac{\gamma_1}{\gamma_2}-1}}{\left(I_{sref1}\right)^{\frac{\gamma_1}{\gamma_2}-1}}\right)$$

$$\log_{10}\left(\frac{I_{\max}}{I_2}\right) = -\gamma_2 \log_{10}\left(\frac{I_s}{I_{sref1}}\right) + \left(\gamma_2 - \gamma_1\right)\log_{10}\left(\frac{I_s^*}{I_{sref1}}\right)$$

This can be rewritten in terms of I_1:

$$\log_{10}\left(\frac{I_{\max}}{I_2}\right) = \frac{\gamma_2}{\gamma_1}\log_{10}\left(\frac{I_{\max}}{I_1}\right) + \left(\gamma_2 - \gamma_1\right)\log_{10}\left(\frac{I_s^*}{I_{sref1}}\right)$$

$$\frac{I_2}{I_{\max}} = \left(\frac{I_1}{I_{\max}}\right)^{\frac{\gamma_2}{\gamma_1}} \left(\frac{I_s^*}{I_{sref1}}\right)^{(\gamma_1 - \gamma_2)}$$

So we can see that the equation actually is not normalized, the right side having the multiplying factor $(I_s/I_{sref1})^{(\gamma_1 - \gamma_2)}$. In the normalized video gamma adjustment, the range of basic scene intensities represented in the image is frozen. The unnormalized film gamma adjustment *changes the range of basic intensities* represented in the image. This can be confusing when doing plots or numerical comparisons. Some intensities I_s from the original scene that are less than I_{sref1} will exceed I_{sref2} (or vice versa). Thus unlike the normalized video gamma equation there will be intensities for which $I_1/I_{\max} < 1$ but $I_2/I_{\max} > 1$ (or vice versa). This is normal and expected - think of it as that the intensities will record on the lower contrast material but will block the highlights in the higher contrast material.

As with the video gamma equation, we can find where the transition between raised and lowered contrast exists, and if it exists, by finding where the slope is equal to 1.

$$\frac{dI_2}{dI_1} = \frac{\gamma_2}{\gamma_1} I_{\max}^{1-\frac{\gamma_2}{\gamma_1}} \left(\frac{I_s^*}{I_{sref1}}\right)^{(\gamma_1 - \gamma_2)} I_1^{\frac{\gamma_2}{\gamma_1}-1}$$

$$\frac{dI_2}{dI_1} = \frac{\gamma_2}{\gamma_1} \left(\frac{I_s^*}{I_{sref1}}\right)^{(\gamma_1 - \gamma_2)} \left(\frac{I_1}{I_{\max}}\right)^{\frac{\gamma_2}{\gamma_1}-1}$$

Setting this equal to 1:

$$\frac{\gamma_2}{\gamma_1} \left(\frac{I_s^*}{I_{sref1}}\right)^{(\gamma_1 - \gamma_2)} \left(\frac{I_1}{I_{\max}}\right)^{\frac{\gamma_2}{\gamma_1}-1} = 1$$

$$\frac{I_1}{I_{\max}} = \left(\frac{\gamma_1}{\gamma_2}\left(\frac{I_s^*}{I_{sref1}}\right)^{(\gamma_2 - \gamma_1)}\right)^{\frac{\gamma_1}{\gamma_2 - \gamma_1}} = \left(\frac{\gamma_2}{\gamma_1}\right)^{\frac{\gamma_1}{\gamma_1 - \gamma_2}} \left(\frac{I_s^*}{I_{sref1}}\right)^{\gamma_1}$$

Normally we will be making comparisons of non-unity gammas to unity gamma, and that also provides a direct comparison to a video gamma adjustment. We will set $\gamma_1 = 1$ and $\gamma_2 = \gamma$:

$$\text{Film Gamma Transition Point} = \frac{I_1}{I_{\max}} = \gamma^{\frac{1}{1-\gamma}}\left(\frac{I_s^*}{I_{sref1}}\right)$$

So, the transition point for photographic gamma is identical to the transition point for video gamma multiplied by the intensity ratio of the intensity used for the matching, to the maximum intensity of the $\gamma = 1$ image.

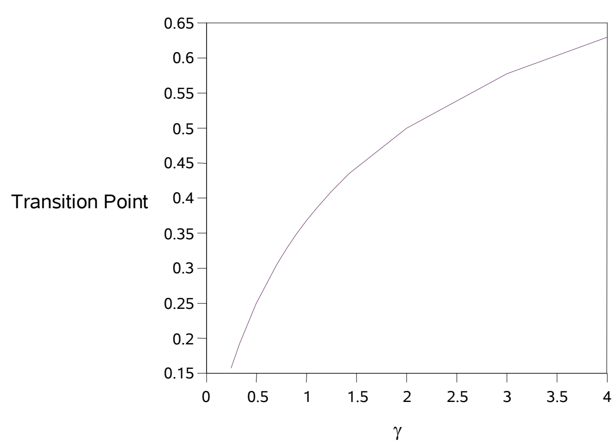
What does this mean?

What we have shown is that:

$$\text{Video Gamma Transition Point} = \left(\frac{I_1}{I_{\max}}\right) = \gamma^{\frac{1}{1-\gamma}}$$

$$\text{Film Gamma Transition Point} = \frac{I_1}{I_{\max}} = \gamma^{\frac{1}{1-\gamma}}\left(\frac{I_s^*}{I_{sref1}}\right)$$

Even if you have no problem following the mathematics, analyzing what this means in the practical sense is more complicated than it might seem. We can plot the Video Gamma Transition Point:



Video Gamma Transition Between Lowered and Raised Contrast

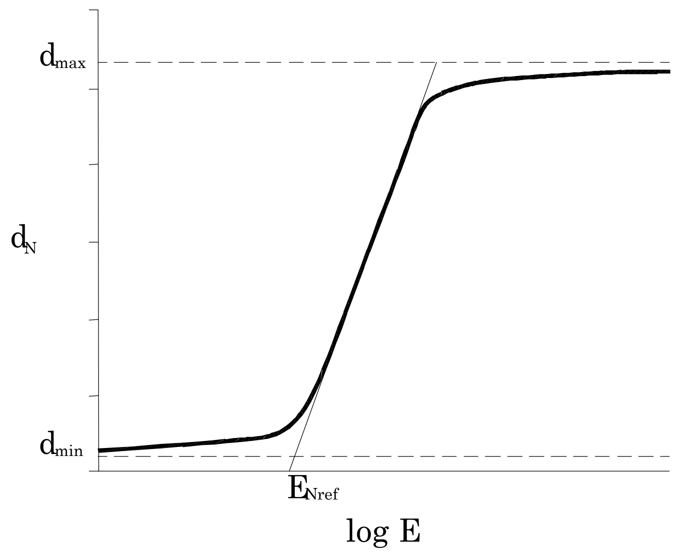
Note that the transition between raised and lowered contrast is well up into the middle intensities and *not* hidden down in the extreme blacks. Since I_{sref1} is effectively a maximum intensity for the reference case, then I_s* will be less than I_{sref1}. So, the curve for the film gamma transition point will be the same shape, but displaced downward from the video gamma transition. However, this is complicated in multiple ways by the fact that what they eye perceives as middle gray is really approximately 18% of maximum intensity. Here we accept this well-known experimental result, the reason for the standard 18% reflectance gray card, without further comment. This means that for video gamma middle gray is $I_1 / I_{max} = 0.18$, or very nearly at the bottom of the above graph. However, for film gamma, the match point is typically taken as middle gray, so that $\left(\frac{I_s^*}{I_{sref1}} \right) = 0.18$. Thus even for $\gamma = 4$, the highest gamma (film contrast) shown on the above graph, the transition point is $I_1 / I_{max} = 0.63 \times 0.18 = 0.11$. This places the transition well below middle gray and of course the transition point gets progressively lower as the amount of contrast shift is lessened.

Thus at this point it appears that for film work, the generally accepted concept that higher gamma means raised contrast and lower gamma means lowered contrast is basically correct. Except for very high gammas, in practical usage it works for tones well down into the darkest grays. However, the video gamma adjustment operating on basic intensity ratios is almost the opposite. The transition point is always above middle gray (0.18) in the above plot, so only at very high intensities is it true that higher gamma means raised contrast and lower gamma means lowered contrast. For most tonal values it is the other way around.

That is not the end of the story, however. Digital photography has one more twist to complicate the picture. By current convention, digital image files are almost always pre-adjusted for a gamma value of from 1.8 to 2.2. The conversion is such that the values stored for an image are *not* the actual basic intensity ratios, but the intensity ratio raised to the power of the inverse of the pre-adjusted gamma. Thus a middle gray of 0.18 basic intensity ratio will appear in a file as a value around 0.4. An image stored in this manner is considered the normal starting point, and any gamma adjustments are applied to that image. For most images processed in Photoshop and other image editing programs, the net result is that instead of middle gray being nearly at the bottom of the above Transition graph at 0.18, it is squarely in the middle at 0.4. This means that in typical digital video gamma adjustments - also known as the middle gray adjustment of the Photoshop Levels tool - the contrast typically behaves exactly the opposite for dark tones as it does for light tones, not at all the same as a gamma change in conventional photography. And this despite the fact that the equations for the two appear nearly identical.

Mathematical Proof of Photographic Gamma Equation for Positives

Traditional or silver-based images are usually described by a relationship that expresses the sensitivity of film to light. This relationship is called the S-curve and is physically measured for each type of film under specific processing:



S-Curve for a Negative Photographic Film

Where the straight line portion of the "S" has the equation:

$$d_N = \gamma_N \log_{10}(E / E_{Ref})$$

Here d_N is the density of a point in the negative image, γ_N is gamma, the slope of the straight line portion of the relationship, E is the exposure, which is normally the product of the light intensity of the image point on the film introduced from a scene, times the exposure duration (shutter speed). E_{Nref} is called the reference exposure but the name can be misleading as it really is just the intercept value chosen to make the equation lie on top of the straight line portion of the measured S-curve. This is the accepted form of the film gamma equation as found in standard references. As is traditional, we will deal only with this equation that represents just the straight-line portion and thus we will ignore the upper and lower tails of the S-curve.

In the digital world we tend to deal mostly in positive images rather than with negatives and we have previously stated that it is only necessary to insert a minus sign to have an equation that describes a positive image:

$$d_P = -\gamma_P \log_{10}(E/E_{Pref})$$

Intuitively one might expect this to be the case and it would, of course, be possible to plot experimental data for a positive image and, finding a reversed S-curve, simply apply the negative of the standard equation. However, given the form for negative images it is possible to prove that the sign reversal is the mathematically correct conversion for positive images. Published mathematical proofs of this are not easily found, so we offer one here.

In making a direct positive photographic image by the reversal process, a negative image is first developed, and as for any negative image:

$$d_N = \gamma \log_{10}(E/E_{Nref})$$

This step forms a silver image, selectively using up silver halide as it does so. At this point the developed silver may be bleached out when trying to produce a positive silver image or just left in place if the final object is a coupled dye image. In either case the film is exposed to light so that *all* the *remaining* silver halide will be converted to silver. Thus if d_{Nmax} is the maximum possible density that can be produced and d_N is the amount taken away by developing the negative image,

$$d_P = d_{Nmax} - d_N = d_{N\,max} - \gamma \log_{10}(E/E_{Nref})$$

recalling that E_{Nref} is just a constant with no physical importance, we can incorporate d_{Nmax} into it and call the combination E_{Pref}, which again has no physical significance and as with the negative serves only to position the straight line properly.

$$d_P = -\gamma \log_{10}(E/E_{Pref})$$

which shows the correct form for a direct positive image.

The negative/positive process of producing a positive image, where there is an actual negative and a positive print or transparency made from it, involves two separate light-sensitive media and is a little more complicated.

Here as before we start with a negative image, but as there will be separate exposures for the positive and negative images, we use E as the exposure to the scene for the negative and E_N as the exposure to the illuminated negative for the positive:

$$d_N = \gamma_N \log_{10}(E/E_{Nref})$$
$$d_P = \gamma_P \log_{10}(E_N/E_{Pref})$$

The positive media is exposed to light passing through the negative media. To find what exposure the positive will get from this negative, take the definition of density for the negative:

$$d_N = \log_{10}(I_{Nmax}/I_N) = -\log_{10}(I_N/I_{Nmax})$$

The exposure for the positive will be this intensity applied for an exposure time t_0 through a *f*-aperture so that $E_N = K\, I_N$ where K accounts for both the exposure time and *f*-aperture applied to the light passing through the negative:

$$d_N = -\log_{10}(I_N / I_{Nmax}) = -\log_{10}(K\, I_N/(K\, I_{Nmax})) = -\log_{10}(E_N/E_{Nmax})$$

so that $$d_N = -\log_{10}(E_N/E_{Nmax}) = -\log_{10}(E_N/E_{Pref}) - \log_{10}(E_{Pref}/E_{Nmax})$$

and $$\log_{10}(E_N/E_{Pref}) = -\, d_N - \log_{10}(E_{Pref}/E_{Nmax})$$

substituting the fundamental form for d_N

$$\log_{10}(E_N/E_{Pref}) = -\,\gamma_N \log_{10}(E/E_{Nref}) - \log_{10}(E_{Pref}/E_{Nmax})$$

The original d_P relationship: $$d_P = \gamma_P \log_{10}(E_N/E_{Pref})$$

Thus becomes $$d_P = -\,\gamma_P\, \gamma_N \log_{10}(E/E_{Nref}) - \gamma_P \log_{10}(E_{Pref}/E_{Nmax})$$

Again, the entire last term is a constant and can be combined with E_{Nref} into a new constant E'_{Pref}, while we can take an overall gamma as $\gamma' = \gamma_P\, \gamma_N$:

$$d_P = -\,\gamma' \log_{10}(E/E'_{Pref})$$

which is the required form.