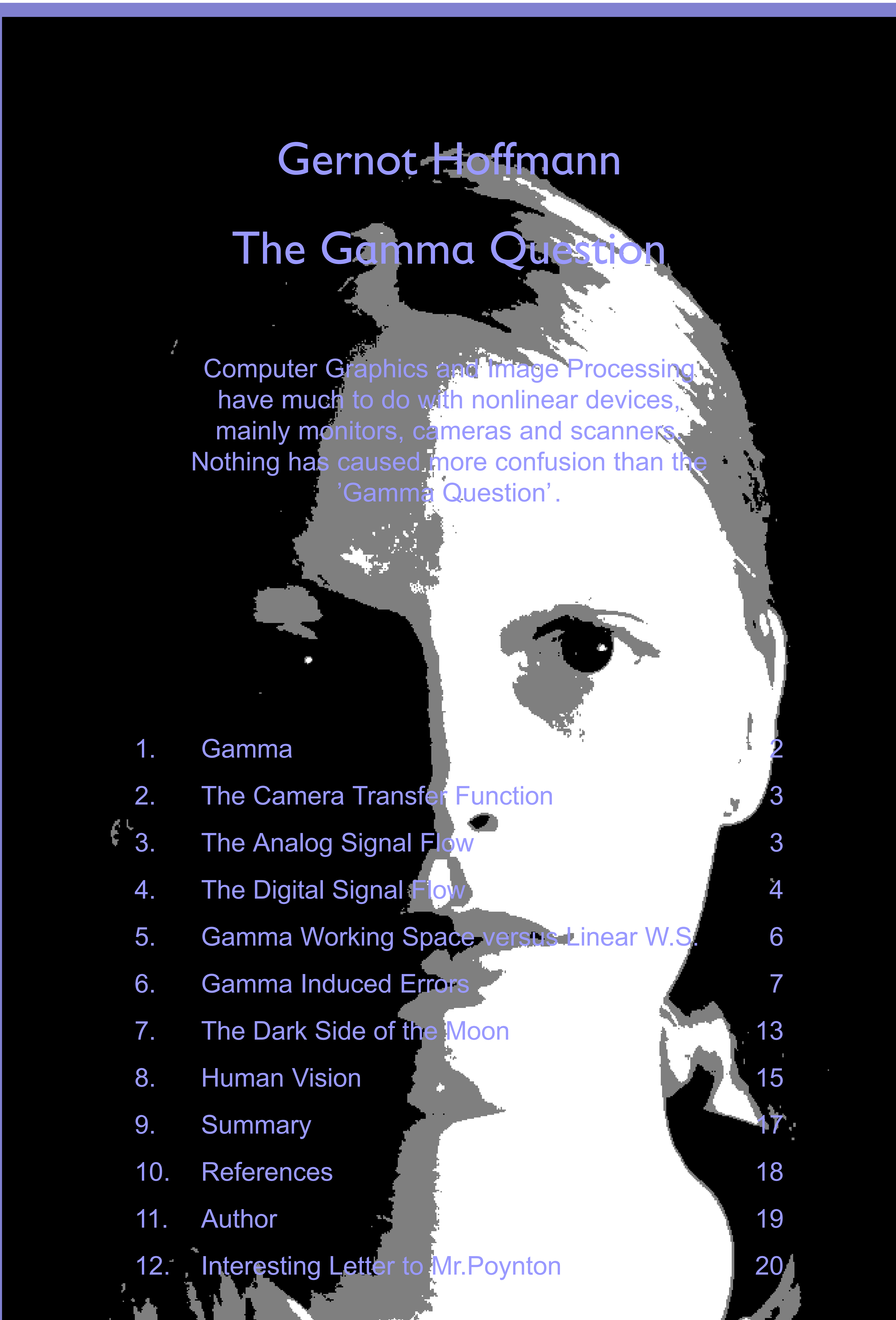


Gernot Hoffmann

The Gamma Question

Computer Graphics and Image Processing
have much to do with nonlinear devices,
mainly monitors, cameras and scanners.
Nothing has caused more confusion than the
'Gamma Question'.

1. Gamma

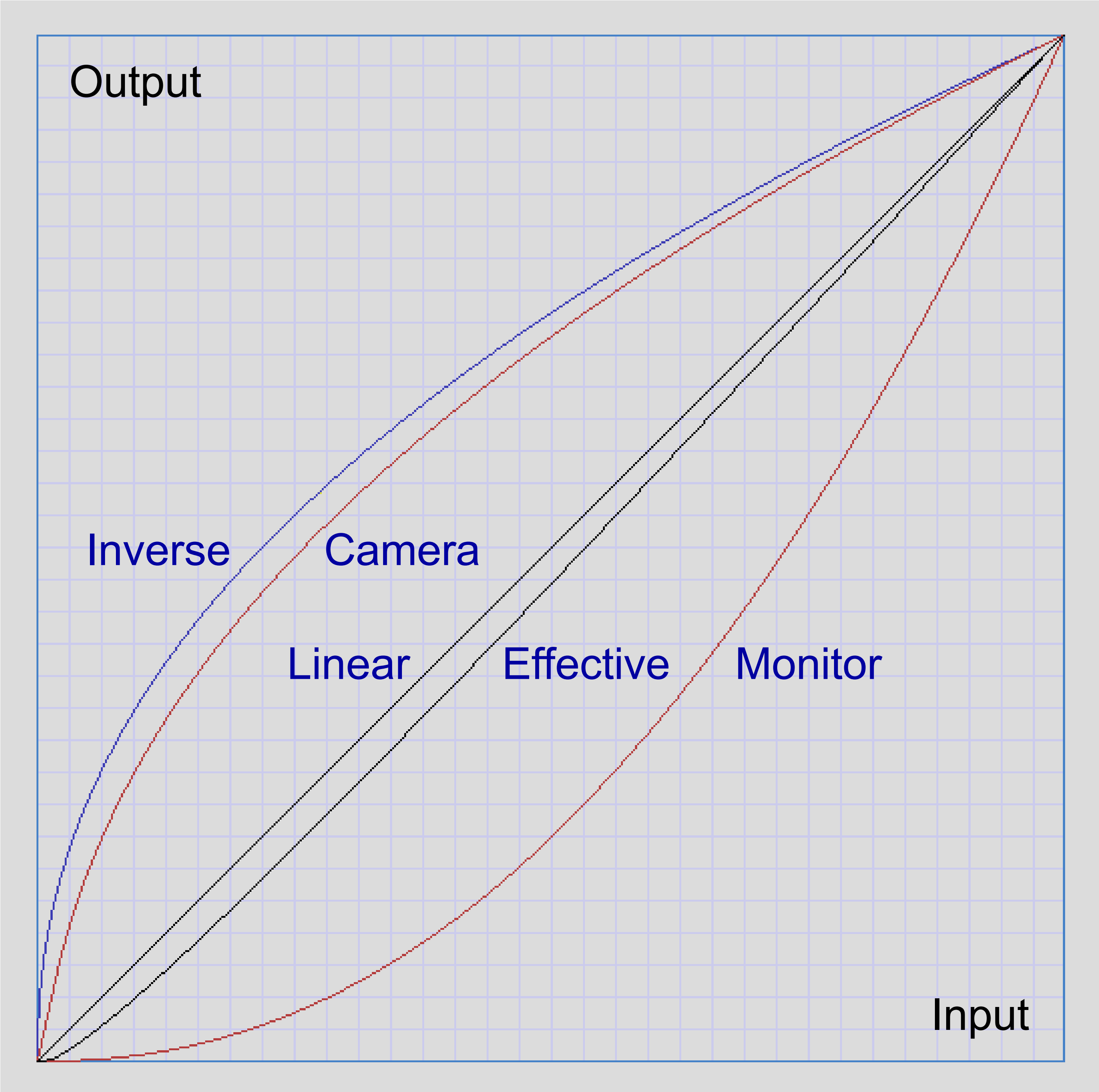


Figure 1
Source Image
and
Video Signal
Analog Coding

L_m	Monitor luminance
Y	Control signal
$L_m = Y^G$	
G = 2.5	Natural gamma
G = 2.2	Calibrated monitor
L_s	Scene luminance
X	Camera output
$X = L_s^{1/G}$	
$L_m = L_s$	Linear transfer funct.
$L_m (L_s)$	Eff. transfer function

The Monitor transfer function for a CRT screen shows that the luminance L_m depends on the control signal by $L_m = Y^G$. The exponent G is called Gamma.
The natural value is about G=2.5, but for computer applications the transfer function is usually calibrated by Lookup-Tables (LUTs) on the graphics card for G=2.2, as shown in the Monitor graph.
A television camera should have the inverse transfer function between scene luminance L_s and output signal X as in the graph Inverse, $X = L_s^{1/G}$.
According to ITU-R BT.709 [2], the function has a linear slope for low luminances, as shown in Camera. The Effective transfer function between scene luminance L_s and monitor luminance L_m is slightly curved.

2. The Camera Transfer Function

The Camera transfer function, as used in television broadcast systems, is defined as below. The exponent is 0.45 = 1/2.2222 instead of 1/2.20 .

$X = 1.099 \cdot L_s^{0.45} - 0.099$ for $0.018 \le L_s \le 1.0$

$X = 4.50 \cdot L_s$ for $0.0 < L_s < 0.018$

Best approximation by a single power function:

$X = L_s^{0.518}$

These transfer functions are now called TRCs, Tone Reproduction Curves.

3. The Analog Signal Flow

Scene luminance L_s is measured by a CCD camera, which is more or less linear. The signal is converted by the Camera transfer function into the output voltage X. The Transmission Line is linear and delivers the voltage Y. The monitor creates the luminance L_m by the Monitor transfer function.

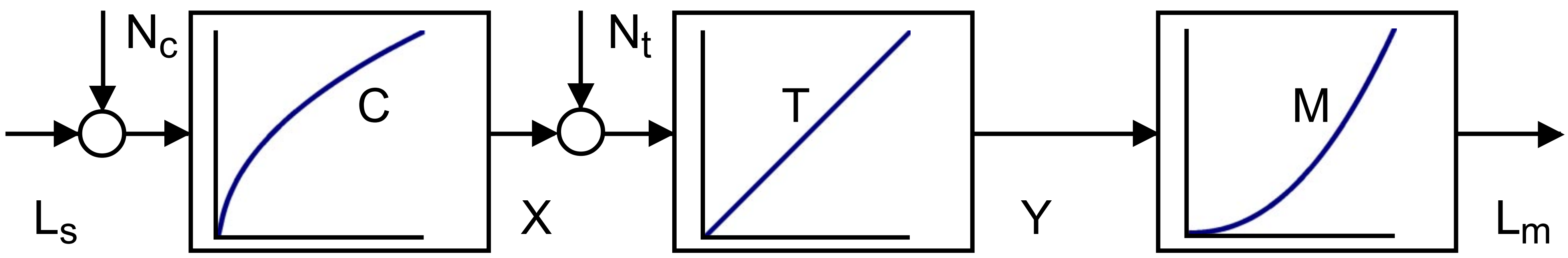


Figure 2 Analog Signal Flow

The Effective transfer function is valid for the relation between L_m and L_s. This is nearly linear (the minor deviation from the Linear transfer function improves the perceptual quality, it´s a flare compensation).
This means: the whole system design is based on the assumption, that human vision perceives an image on a monitor very similar to a real scene.

Sensor noise N_c, caused by the CCD electronics, is transmitted without any considerable attentuation to monitor luminance. The linear slope in the Camera transfer function helps a little for noise suppression.

Transmission Line noise N_t contributes much less to the luminance in the dark area because of the monitor Gamma function, but more in the light.

Obviously the nonlinearity of monitors, which is a historical fact, had influenced the design very much. Noise suppression could have been done by other methods as well. Monitor Gamma is a fact, it´s too late to build linear monitors.

4. The Digital Signal Flow

Here we see the signal flow as it is widely used in Image Processing.
Scene luminance L_s is measured by a CCD camera, which is more or less linear. Or an image is scanned by a scanner, which is also linear.
The signal is converted by the Camera transfer function or by the Inverse transfer function into the output signal X_d, where the index d indicates the digital coding. Both functions are usually not specified by the manufacturers.
Image Processing is linear, if nothing is modified. The output is still digital and then converted to the analog video signal Y. LUTs may be used, but this is not shown here.

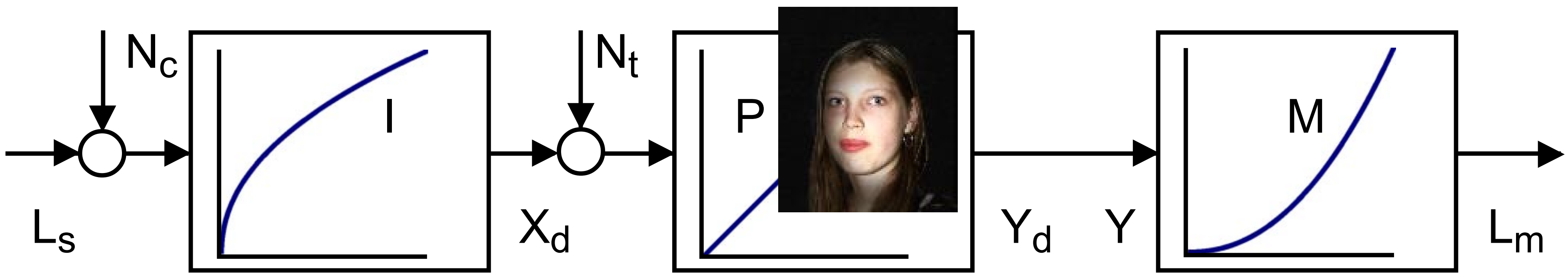


Figure 3 Digital Signal Flow

The Effective transfer function is valid for the relation between L_m and L_s.
It is exactly linear linear, if the Inverse transfer function was used.

This means again: the whole system design is based on the assumption, that human vision perceives an image on a monitor very similar to a real scene.

Sensor noise N_c, caused by the CCD electronics, is transmitted without any considerable attentuation to monitor luminance.
Transmission Line noise N_t is now zero because of digital coding. Noise on the video cables is still suppressed by the monitor Gamma function for dark areas, but not for light areas.
We have to state a very important fact:
Any deviation of the straight line Output code = Input code in a transfer function causes a loss of codes in the output. This is obvious for nonlinear transfer functions, but it is also valid for straight lines with attenuations of more or less than one, including clipping.

On the next page we can see the disappointing result of *two* sequential non-linear codings.

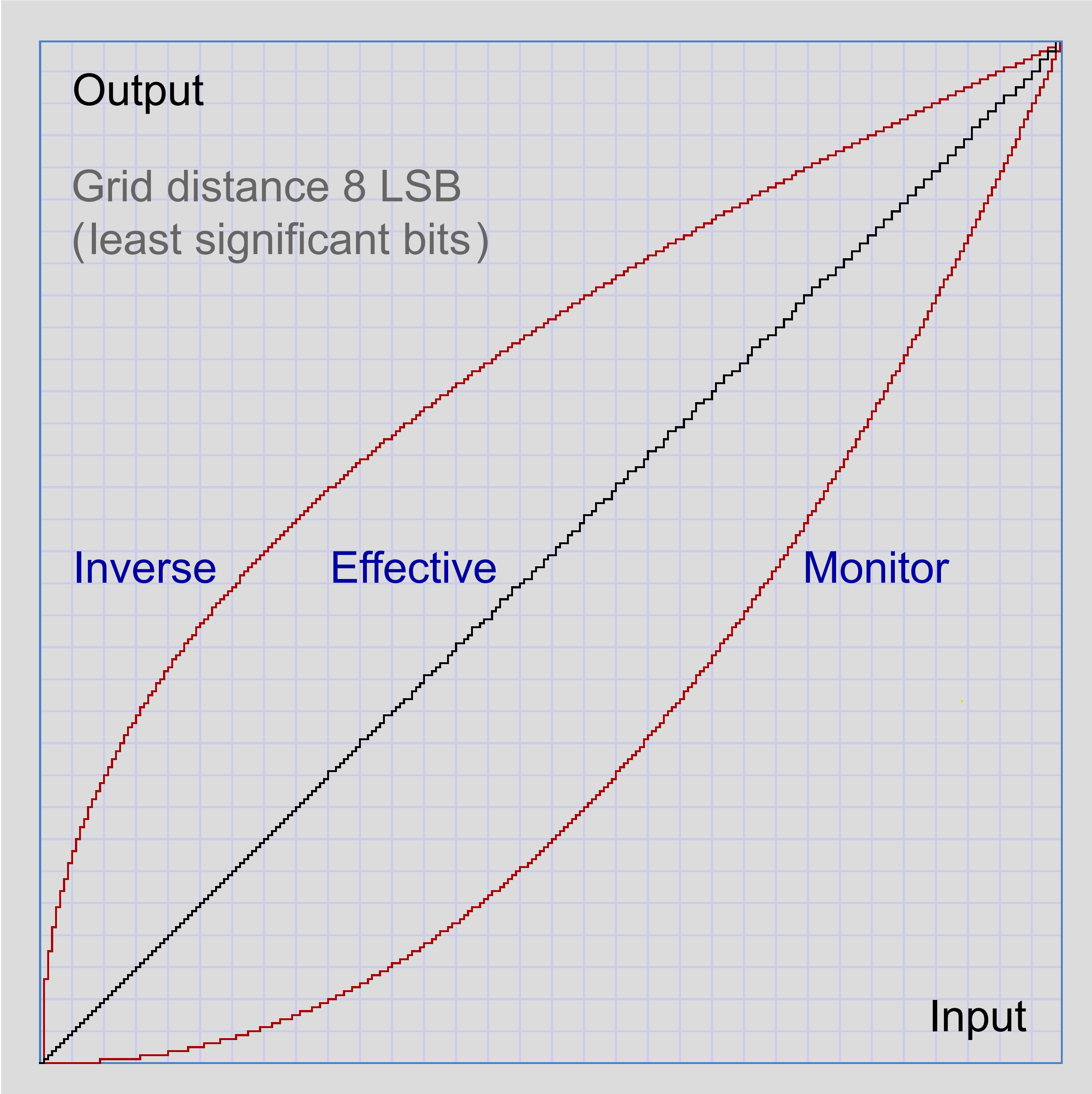


Figure 4
Source Image
and
Video Signal
8 Bit Coding

The quantized transfer functions use inputs and outputs in the range 0...255.

$$L_m = \text{Round}\left[\, 255\cdot\left(\left(\text{Round}\left[255\cdot(L_s/255)^{1/G}\right]\right)/255\right)^{G}\right]$$

Even if no Image Processing is applied - the quality loss is clearly visible in the Effective binary transfer function (which is additionally different to the transfer function in Figure 1, because now the Camera is replaced by the hypothetical Inverse).

A difference of one bit in two facing color patches, e.g. red, green or blue, cannot be distinguished. Two bits are mostly distinguishable.
This means: the double quantization causes dramatical round off errors, but for real photos, the quantization in the transfer function is probably not so obvious.
We have also to consider the noise in the analog video signal and this may be helpful to disguise the deterioration.

5. Gamma Working Space versus Linear W. S.

So far, we can the call this Image Processing in a Gamma Working Space, because any manipulation is done with inverse gamma compensated data.

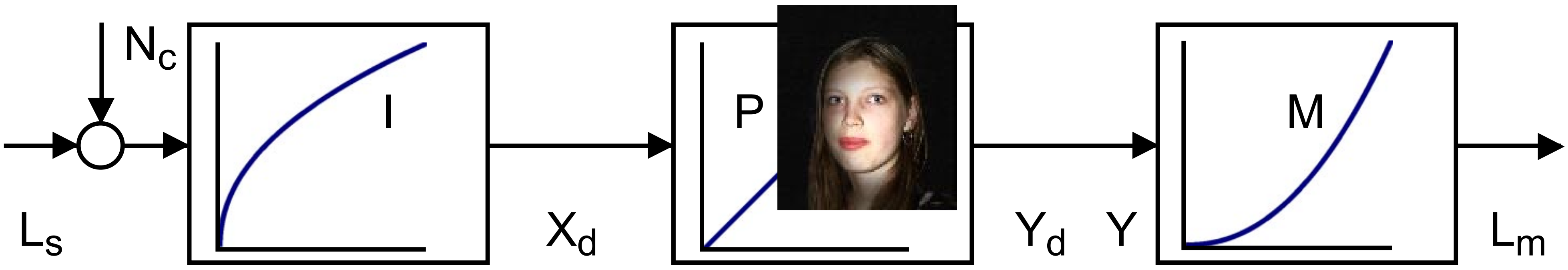


Figure 5 Gamma Working Space

The alternative is Image Processing in a Linear Working Space.

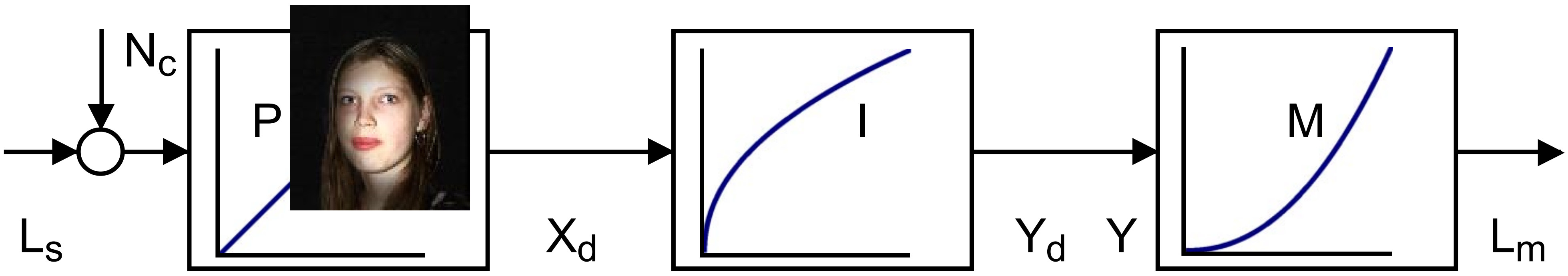


Figure 6 Linear Working Space

The data are correctly handled in a Linear Working Space, which doesn´t affect the features of physical light. Physical light adds linearly in reality.
The final results are compensated by an Inverse transfer function for the monitor characteristics.
This transfer function is established either by software LUTs or by so called User LUTs on the graphics card.
User LUTs can be expected in future, at present they are rare.
The software LUTs have to be established in present programs.
This means: all outgoing image data pass the software LUTs, but other data like menue graphics, don´t use them, because these colors are optimized for a proper visualization on calibrated monitors without any further correction.

The only objection to the Linear Working Space concept is indeed the necessary installation of software LUTs - not easy in already finished systems.

6. Gamma Induced Errors

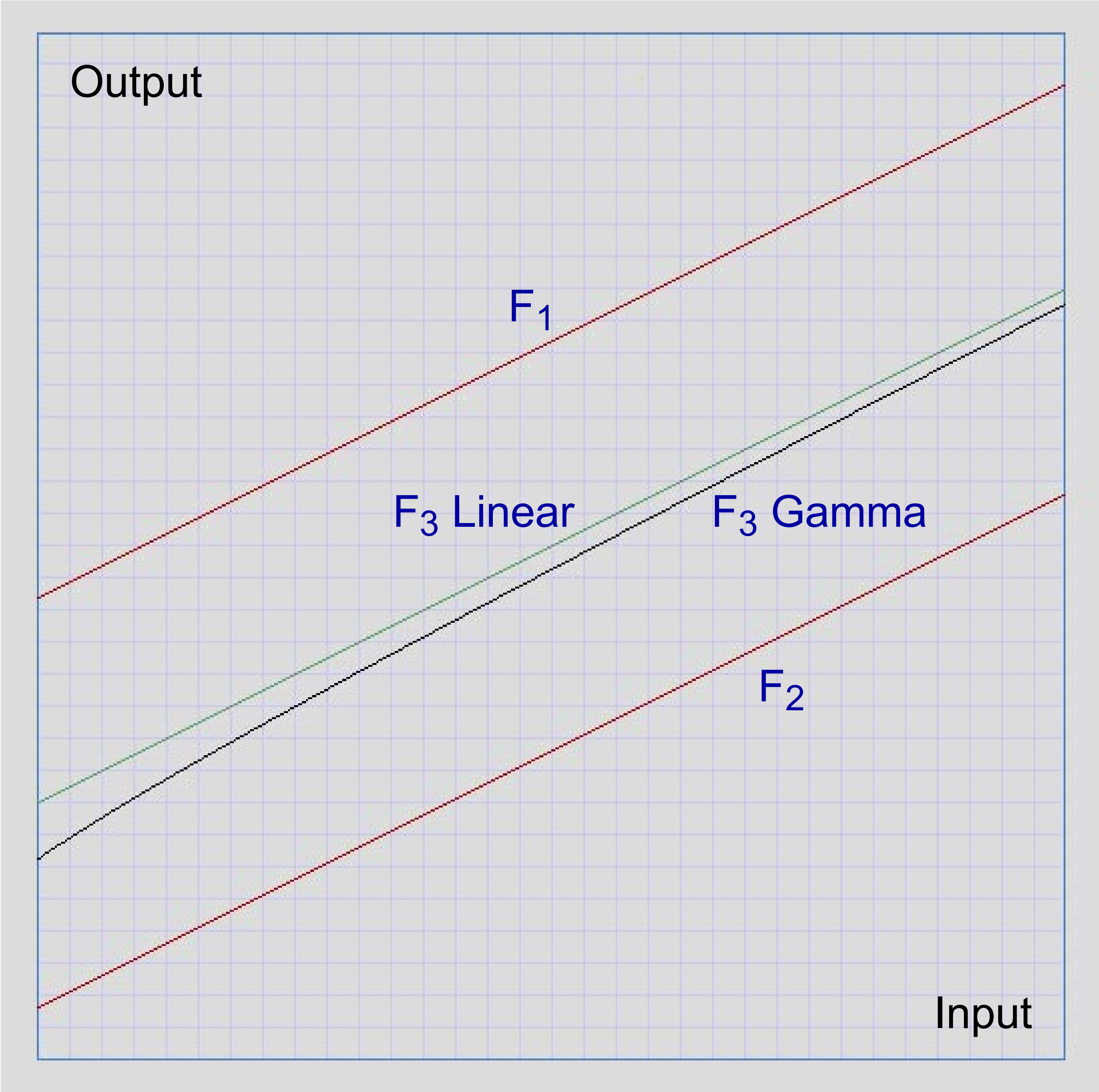


Figure 7

Average of F_1,F_2
Linear W. S.
and
Gamma W. S.

This example shows two functions F_1 and F_2 . They represent grayscales. Input is the coordinate of the grayscale, output is the gray value.
Both are shown in the real light space or in the Linear Working Space.
The third function F_3 Linear is the average $F_3 = 0.5 \cdot (F_1+F_2)$.
The graph F_3 Linear is obviously correct.

In the Gamma Working Space the calculation is done like this:

$$A = 0.5 \cdot (F_1^{1/G} + F_2^{1/G})$$

$$F_3 = A^G$$

The result F_3 Gamma is wrong. The level is considerably shifted to lower values and the average is now nonlinear (much more for lower levels).

Commercial programs cause always Gamma Induced Errors.
Mostly, the user adjusts images by appearance. Then these errors are not visible. They have to be discussed for formalistic conversions, like sharpening filters, contrast variation, accurate blending and any arithmetic operations.

Gamma Induced Errors in Calculations

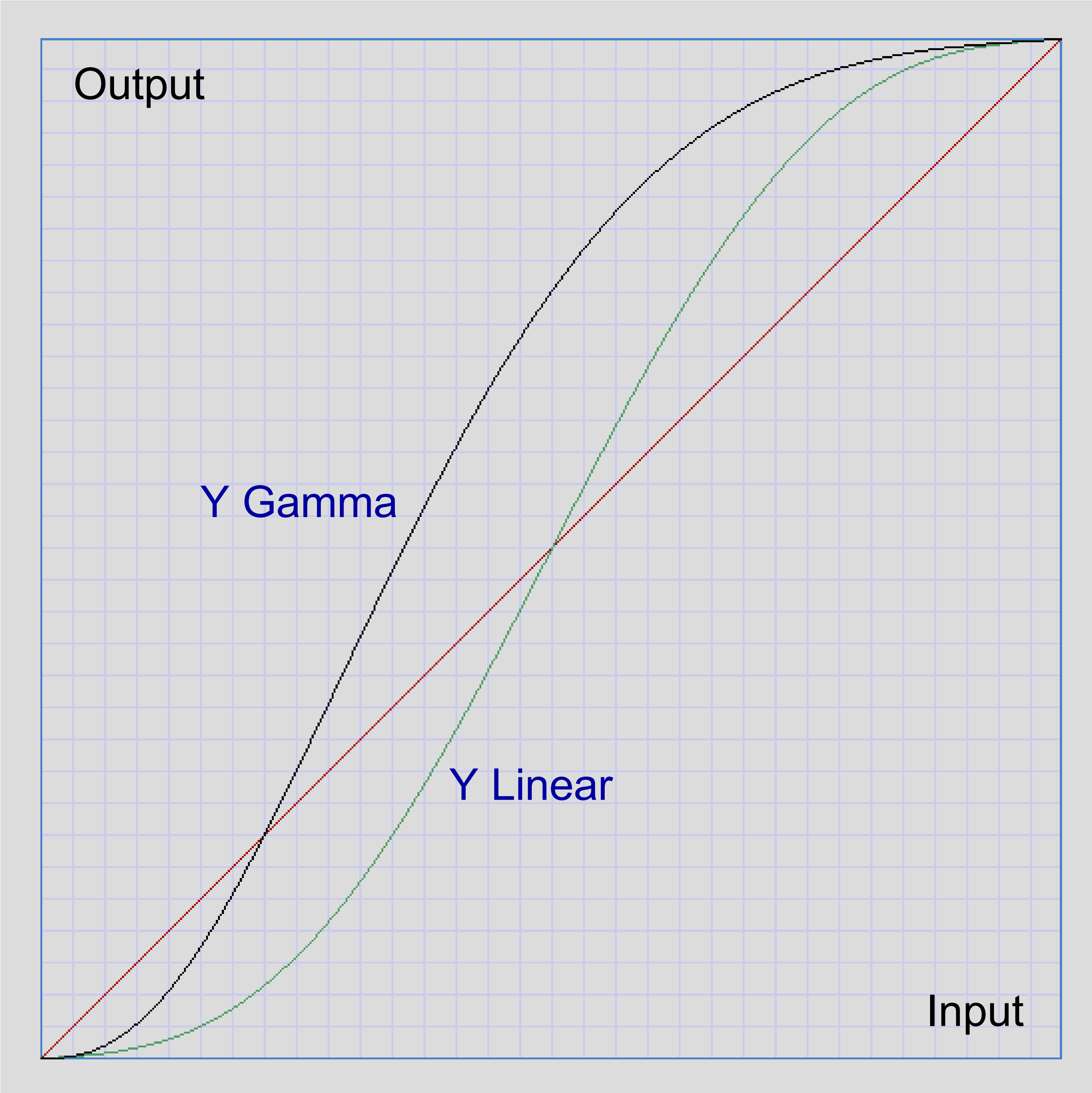


Figure 8

Nonlin. Function
Linear W.S.
and
Gamma W.S.

This is another example for Gamma Induced errors. It simulates an enlarged contrast by a gradation function (so called Curve).
Again, we show the result in the linear light space.

In the Linear Working Space we have Y Linear :

$$Y = X - 0.15 \cdot \sin(2 \cdot \pi \cdot X)$$

In the Gamma Working Space this is executed as Y Gamma :

$$Y = X^{1/G}$$
$$Y = Y - 0.15 \cdot \sin(2 \cdot \pi \cdot Y)$$
$$Y = Y^{G}$$

Y Gamma is probably not the desired result, but commercial programs work mostly like this.

Gamma Induced Errors
Example 1

Figure 9a (right)
Small part of original image

Figure 9b (bottom left)
Sharpening filter in Gamma W.S.

Figure 9c (bottom right)
Sharpening filter in Linear W.S

For Figure 9b, a strong sharpening filter was applied directly to the original image.
For Figure 9c, the image was transformed into the Linear Working Space by $Z = X^{2.2}$ for $X = R, G, B$. Then the filter was applied. Finally the image was transformed back into the Gamma Working Space by $Y = Z^{1/2.2}$.

Where are the differences ?
The text „Viking“ in Figure 9c looks probably better, compared to Figure 9b, because it has no halo.

Other experiments showed, that Gamma Induced Errors can be hardly detected in real images (the above image is a carefully chosen sample).

Tests with gradation functions (so called Curves, for increased contrast) didn´t show any improvement in the Linear Working Space.The manifold of perceptual and esthetical effects overrides the formalistic correctness.

Gamma Induced Errors
Example 2

Figure 10a
Interpolation for
saturated Colors
Gamma Working Space

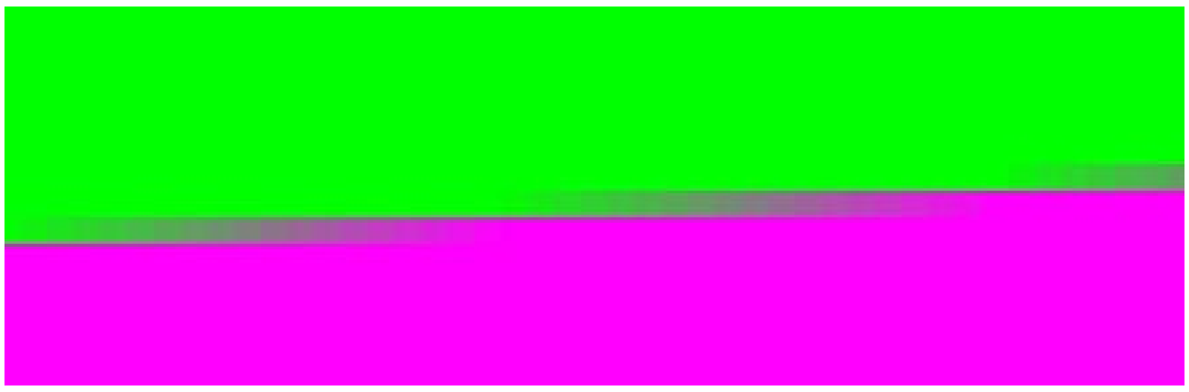


Figure 10b
Interpolation for
saturated Colors
Linear Working Space

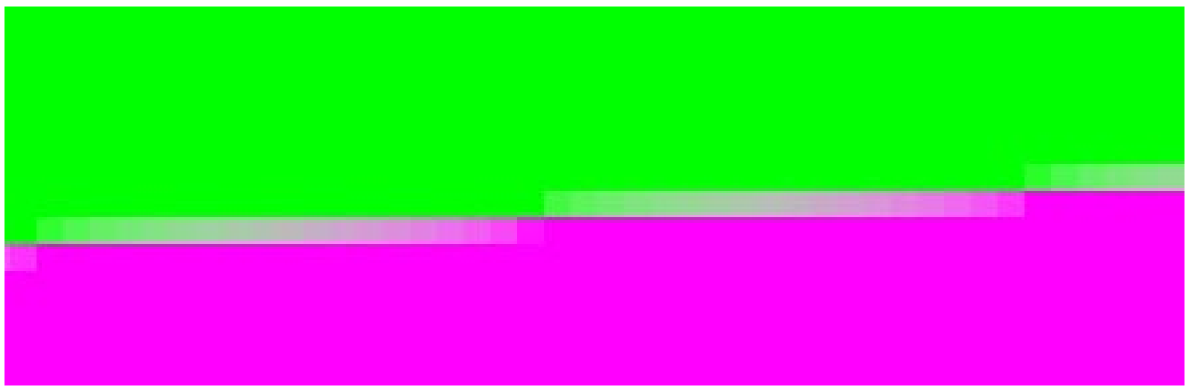


Figure 10c
Interpolation for
not saturated Colors
Gamma Working Space

These images are based on an example by Dersch [9].
They show the interpolation between the complementary colors Green and Magenta. It´s a zoom view for a slanted edge with anti-aliasing.

In Figure 10a we have Green=0,255,0 and Magenta=255,0,255.
The values themselves are not affected by any Gamma distortion.
The edge is obviously too dark. The line between Green and Magenta passes in the RGB color cube the Gray axis at Gray=128,128,128.
This is a relative dark gray, because a medium gray is at Gray=186,186,186 for Gamma=2.2 .

The Linear Working Space result is simulated in Figure 10b. The interpolation looks much better, but for sharp eyes the transition now is *too light*. It is overcompensated.

In Figure 10c we have Green=128,255,128 and Magenta=255,128,255.
The interpolation looks reasonable in the Gamma Working Space. Corrections for less saturated colors are obviously not necessary.

Gamma Induced Errors
Example 3

Floyd-Steinberg
Bilevel Dithering

The data are corrected
Source Pixels
$S = S^{1.6}$
$S = (L_s{}^{1/2.2})^{1.6} = L_s{}^{0.73}$
$S = R,G,B = 0 \ldots 255$
Destination pixels
$D = R,G,B = 0 / 255$

Correct view for Acrobat Zoom = 100% on calibrated screen G=2.2

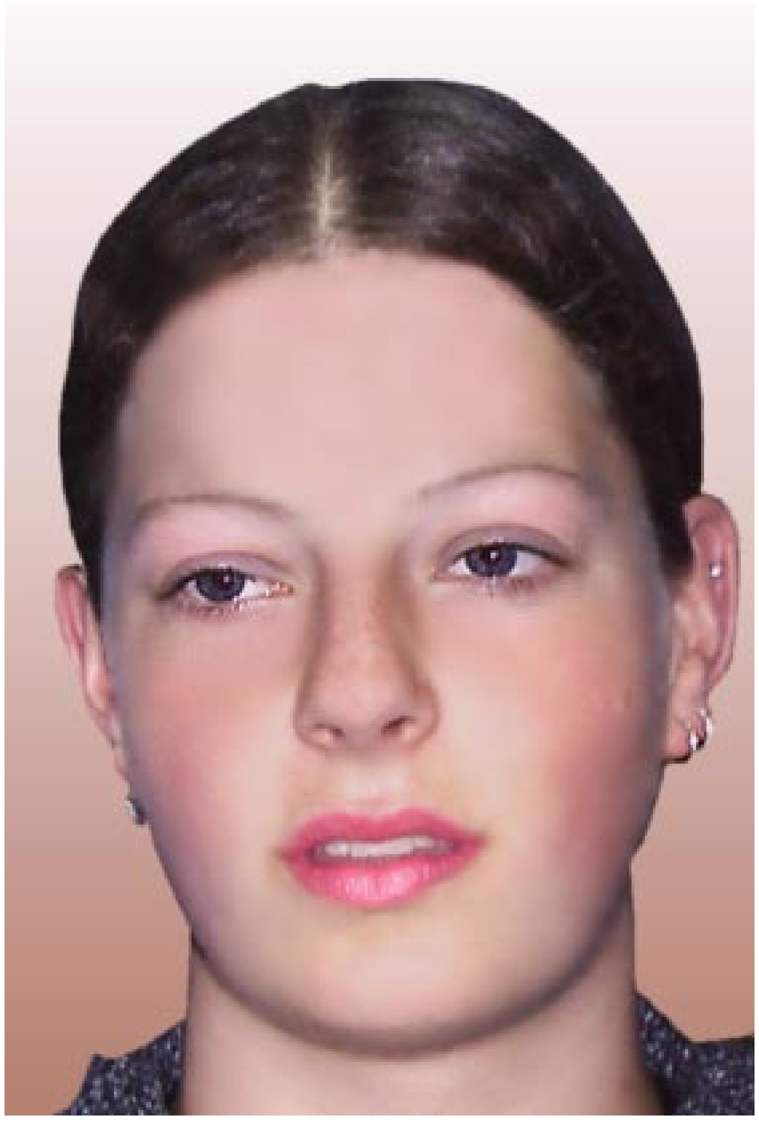
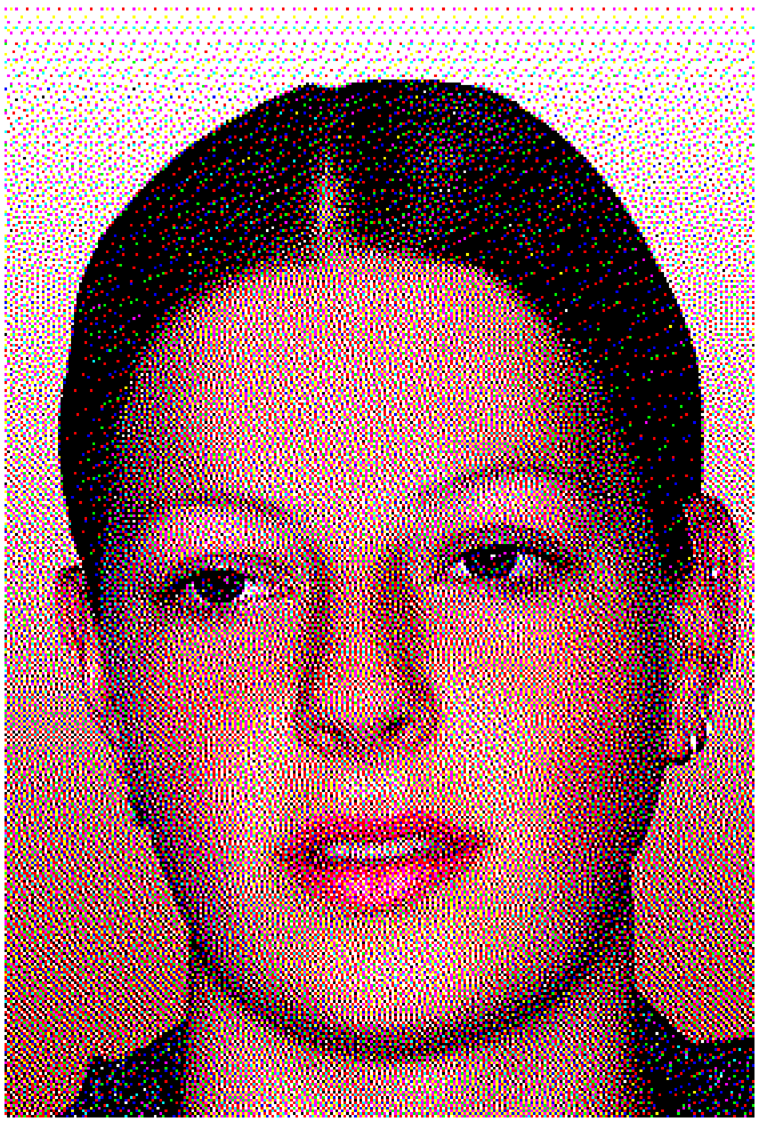


Figure 11a
Original image
Gamma Working Space

Figure 11b
FS Dithering
Gamma Working Space
Too light!

Figure 11c
FS Dithering
Linearized Working Space
G=1.6 correction, better!



Gamma Induced Errors
Example 4

Figure 12a
Original Computer Graphic

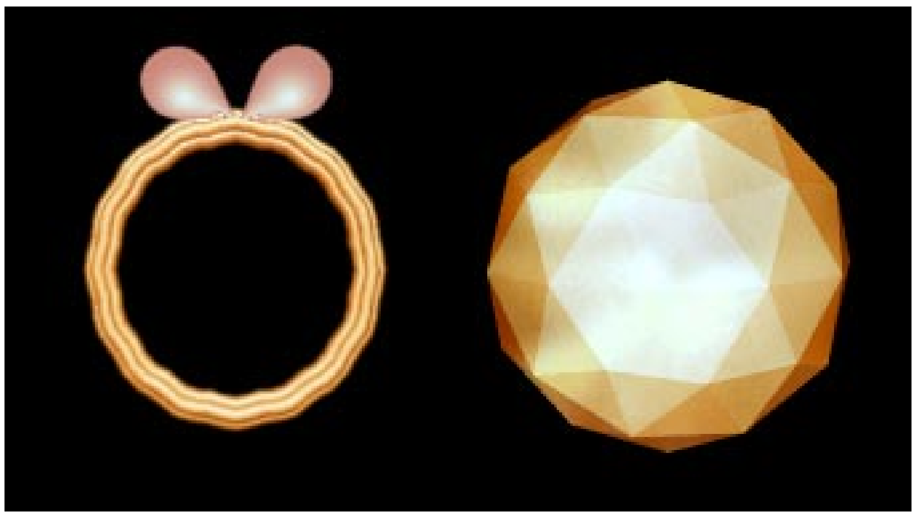


Figure 12b
Sharpening Filter
Gamma Working Space

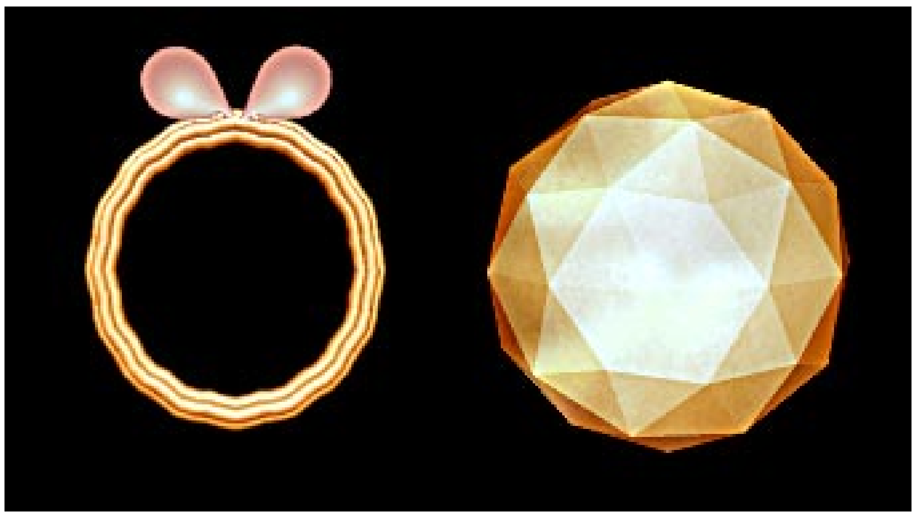


Figure 12c
Sharpening Filter
Linear Working Space

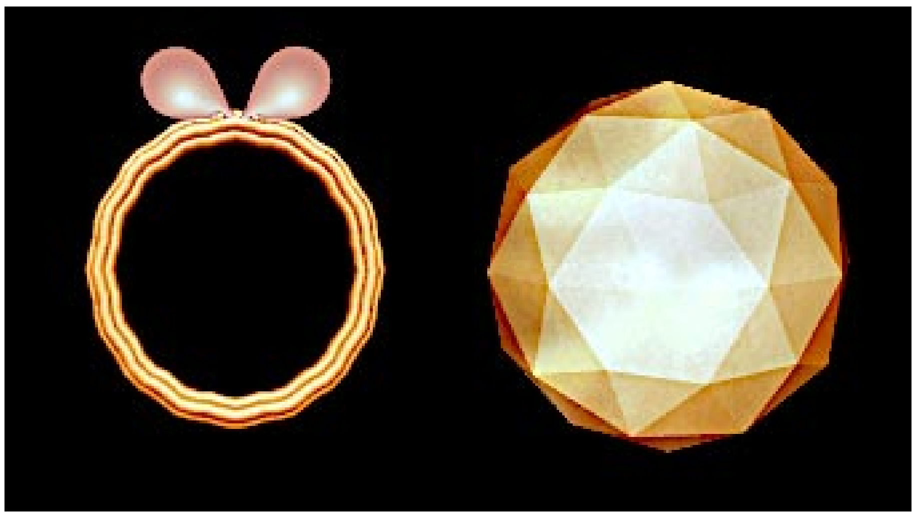


Figure 12a shows a computer graphic
In Figure 12b a strong sharpening filter was applied in the Gamma Working Space. Edges are unexpectedly enhanced.
In Figure 12c the image was transformed by $Z = X^{2.2}$ for $X = R, G, B$ into the Linear Working Space. Then the sharpening filter was applied. Finally the image was transformed back to the Gamma Working Space by $Y = Z^{1/2.2}$. The edges are sharp but not unusually enhanced.

Resumé:

The Gamma Induced Errors are not very relevant for filters in practical Image Processing for photos.
They are relevant for computer graphics, for correct blending, for general calculations - altogether for accurate Image Processing [4].

7. The Dark Side of the Moon

Very often we hear this argument: „The Gamma Working Space has more codes for dark signals. Here, the resolution of eye and brain is higher, therefore the code has to deliver more levels“.

Now let´s assume, as in all previous discussions, that human vision perceives the screen luminance as lightness. The television signal flow is based on this assumption (with a minor flare correction), though there are some doubts.

Here we see again Figure 4 and additonally as an example for Image Processing two functions $Y = X \pm 0.15 \cdot \sin(2 \cdot \pi \cdot X)$.

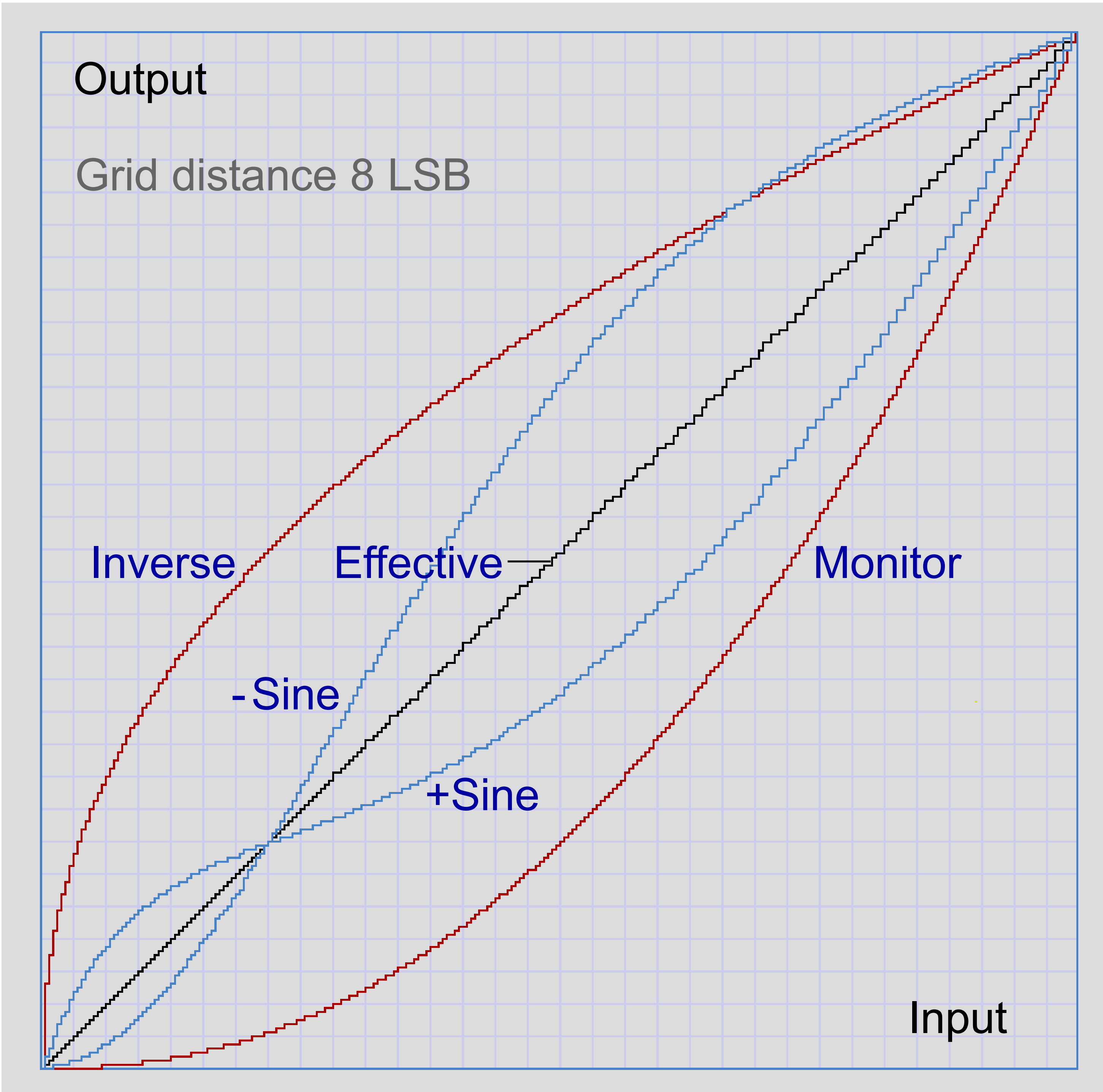


Figure 13

Source Image
and
Video Signal
8 Bit Coding

First, we discuss only the combination of Inverse input and Monitor output. The input signal luminance appears as output luminance linearly with some effects of quantization - it´s the Effective transfer function.

The loss of information at the dark end is not significant, as long as *no Image Processing* is applied.

But then, the quality will be affected, because the code sequence is rather sparse at the dark end. This is also shown in the table on the next page.

Table for the Inverse transfer function $X = L_S^{1/2.2}$

Dark Side ...

L_S	X	L_S	X	L_S	X	L_S	X
0	0	64	136	128	186	192	224
1	21	65	137	129	187	193	225
2	28	66	138	130	188	194	225
3	34	67	139	131	188	195	226
4	39	68	140	132	189	196	226
5	43	69	141	133	190	197	227
6	46	70	142	134	190	198	227
7	50	71	143	135	191	199	228
8	53	72	144	136	192	200	228
9	56	73	144	137	192	201	229
10	59	74	145	138	193	202	229
11	61	75	146	139	194	203	230
12	64	76	147	140	194	204	230
13	66	77	148	141	195	205	231
14	68	78	149	142	195	206	231
15	70	79	150	143	196	207	232
16	72	80	151	144	197	208	232
17	74	81	151	145	197	209	233
18	76	82	152	146	198	210	233
19	78	83	153	147	199	211	234
20	80	84	154	148	199	212	234
21	82	85	155	149	200	213	235
22	84	86	156	150	200	214	235
23	85	87	156	151	201	215	236
24	87	88	157	152	202	216	236
25	89	89	158	153	202	217	237
26	90	90	159	154	203	218	237
27	92	91	160	155	203	219	238
28	93	92	160	156	204	220	238
29	95	93	161	157	205	221	239
30	96	94	162	158	205	222	239
31	98	95	163	159	206	223	240
32	99	96	164	160	206	224	240
33	101	97	164	161	207	225	241
34	102	98	165	162	207	226	241
35	103	99	166	163	208	227	242
36	105	100	167	164	209	228	242
37	106	101	167	165	209	229	243
38	107	102	168	166	210	230	243
39	109	103	169	167	210	231	244
40	110	104	170	168	211	232	244
41	111	105	170	169	212	233	245
42	112	106	171	170	212	234	245
43	114	107	172	171	213	235	246
44	115	108	173	172	213	236	246
45	116	109	173	173	214	237	247
46	117	110	174	174	214	238	247
47	118	111	175	175	215	239	248
48	119	112	175	176	215	240	248
49	120	113	176	177	216	241	249
50	122	114	177	178	217	242	249
51	123	115	178	179	217	243	249
52	124	116	178	180	218	244	250
53	125	117	179	181	218	245	250
54	126	118	180	182	219	246	251
55	127	119	180	183	219	247	251
56	128	120	181	184	220	248	252
57	129	121	182	185	220	249	252
58	130	122	182	186	221	250	253
59	131	123	183	187	221	251	253
60	132	124	184	188	222	252	254
61	133	125	184	189	223	253	254
62	134	126	185	190	223	254	255
63	135	127	186	191	224	255	255

The resolution of the effective monitor luminance is especially at the light end affected. Partly we have only 3 or 4 least significant bits, LSBs.

Now we see how many different codes are left, linearly and for the Sine functions:

Maximum	256	100%
Inverse	184	72%
Monitor	184	72%
Effective	184	72%
+Sine	147	57%
- Sine	170	66%

We have the paradox situation, that the digital input has a low resolution at the dark end, but the Effective transfer function looks reasonably.

The above mentioned statement 'better resolution at the dark end' is correct, if the Inverse transformation is done by an analog module or a high resolution digital device.

It´s wrong if the transformation is applied *after* an 8-bit analog-digital conversion.

8. Human Vision

Luminance is a measurable physical quantity. Brightness is the correlate for perceived luminance. Lightness is relative brightness, related to the reference white by adaption of eye and brain [1].
The lightness is responsible for the impression of „darker“ or „lighter“.
Eye and brain adapt to a monitor image or a paper image of medium size once on an average level. The Weber Law says:
Two color or gray patches are just distinguishable, if they have a relative difference. One patch has the gray level C, the other C + dC.
The just distinguishable level is defined by the relative level r, not by dC.
The relative level is constant, e.g. $r = dC / C = 0.01$.
Therefore, two patches C and $C + r \cdot C = C \cdot (1+r)$ are distinguishable.
The absolute threshold dC is small for dark patches and larger for light patches.
All this cannot be applied to images, because the Weber Law is a result of variable adaption (sitting in a dark room and observing two large patches).
For images, the adaption is more or less fixed, the Weber Law is not valid, as demonstrated on the next page. Further investigations by the author [7] have shown some results for the human vision of grayscales.

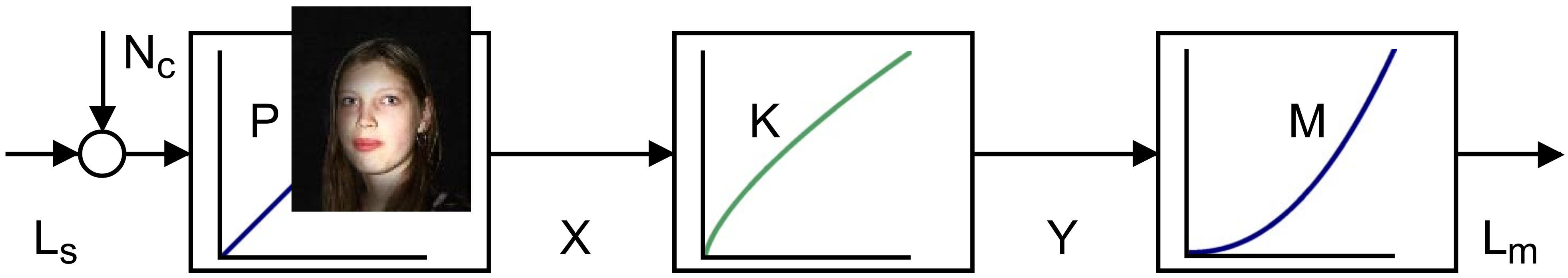


Figure 14 Perceptual Correction

This signal flows shows in the left block the Linear Working Space with no further Image Processing. The right block is the pure Monitor transfer function. The middle block is a Correction for perceptually optimized grayscales, which is used instead of the Inverse transfer function:

$$Y = X^{0.7}$$

The Effective transfer function is then

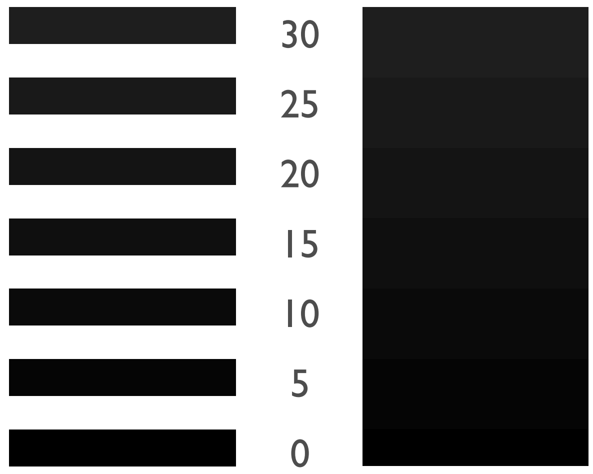
$$L_m = L_s^{0.7 \cdot 2.2} = L_s^{1.54}$$

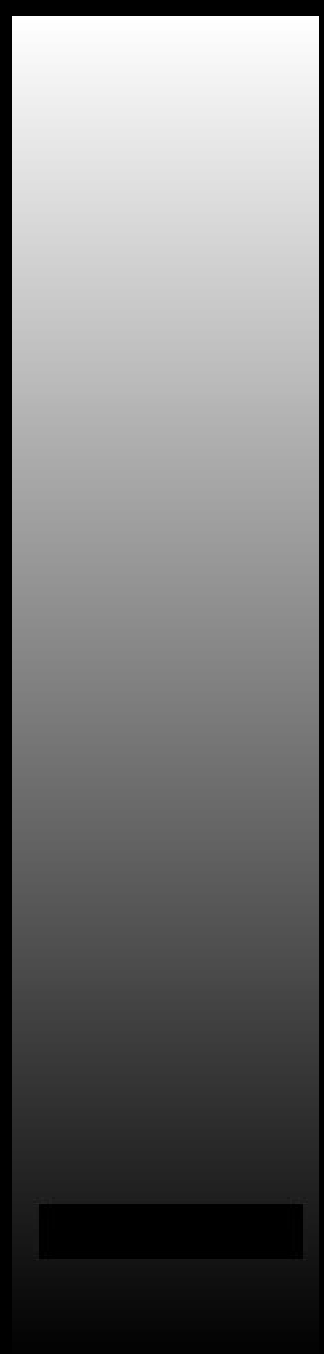
This can be compared to the effective transfer function for television systems. As mentioned in chapter 2, the camera can be represented by approximately $X = L_s^{0.518}$ and the standard uncorrected TV monitor may have $L_m = Y^{2.5}$.

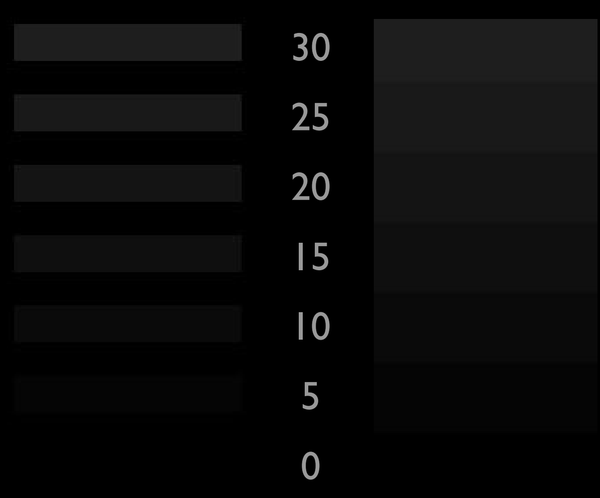
$$L_m = L_s^{1.295} .$$

The Weber Law is not valid for images.

The resolution for dark grays is not generally better than for light grays. If eye and brain can adapt to darkness, then the resolution is indeed better (bottom page).







9. Summary

The standard workflow in television systems and in Image Processing by programs on computers is strongly determined by the monitor charecteristics.

Therefore all image source data are usually distorted according to an inverse monitor transfer function.

In real life the light behaves linearly. Linear operations for distorted data result in nonlinear operations for the associated physical data.

The deteriorating effect is less visible in photos but more visible in synthetical graphics.

We can discern two main applications.

Blending operations for large ranges are wrong. Filter operations are more or less correct in photos, but in graphics the errors are obvious.

State of the art technical instruments don´t deliver linear data, though they measure in principle linear. The industrial standard has adopted the Gamma Working Space like an Eternal Law.
In rare cases the linear mode can be selected.

Programming in a Linear Working Space requires software LUTs.
It cannot be expected, that program manufacturers are willing to modify all the programs.

Altogether: the best technical solution would be to acquire data linearly, process them linearly and apply different nonlinearities for outputs to devices like monitors and printers.

For printers this is anyway established in qualified printing programs. These accept any Working Space profile.

Further complications arise from the fact, that monitor luminance is not perceived as eye + brain lightness. Human vision applies an additional transfer function which is not a simple power law and which depends much on the adaption to the average luminance level of the image.

10. References

[1] R.W.G.Hunt
Measuring Colour
Fountain Press
England, 1998

[2] E.J.Giorgianni + Th.E.Madden
Digital Color Measurement
Addison-Wesley
Reading Massachusetts ,..., 1998

[3] G.Wyszecki + W.S.Stiles
Color Science
John Wiley & Sons
New York ,..., 1982

[4] T.Autiokari
Accurate Image Processing
http://www.aim-dtp.net
2001

[5] Ch.Poynton
Frequently asked questions about Gamma
http://www.inforamp.net/~poynton/
1997

[6] M.Stokes + M.Anderson + S.Chandrasekar + R.Motta
A Standard Default Color Space for the Internet - sRGB
http://www.w3.org/graphic/color/srgb.html
1996

[7] G.Hoffmann
Corrections for Perceptually Optimized Grayscales
http://www.fho-emden.de/~hoffmann/optigray06102001.pdf
2001

[8] G.Hoffmann
Hardware Monitor Calibration
http://www.fho-emden.de/~hoffmann/caltutor270900.pdf
2001

[9] H.Dersch
http://www.fh-furtwangen.de/~dersch/gamma/gamma.html
1999

This document:
http://www.fho-emden.de/~hoffmann/gamquest18102001.pdf

11. Author

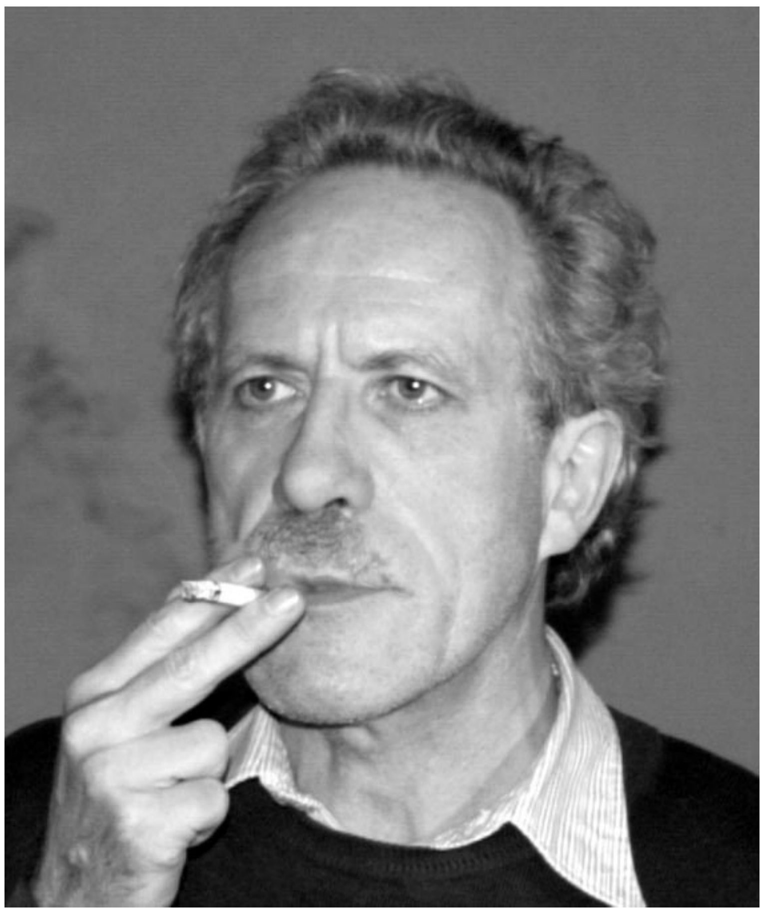


Image Processing:
ZEBRA

Computer Graphics:
ZEFIR

Document:
PageMaker

Compression:
ZIP
JPEG Medium
72/144 dpi

12. Interesting Letter to Mr. Poynton (1)

Found in Google by G.Hoffmann, dated June 1998.
An interesting contribution by M.Francois Esquirol
about Mr.Charles Poynton´s Gamma Mystifications.
Original version, only page layout and colors were modified.

Charles Poynton a écrit dans le message ...
>you [me] wrote:
>> If I have a camera delivering a video signal proportional to the quantity
>> of photons hiting the camera sensor, let me do my processing in a linear
>> manner.
>> [...] even if my eyes can’t *see* the differences !
>
>All of this is fine as long as you don’t display the image.
>
>"Intensity" has a special meaning to physicists, and scientists. Their
> meaning is not always respected by people in other fields. See
> <http://www.opt-sci.arizona.edu/summaries/James_Palmer/intenopn.html>

Right for both the sentences:
- when I use a linear image for processing, my goal is processing and not
  display, and don’t need to display images (except for debugging, but in this
  case I apply a gamma to the linear values and don’t mind the *banding*, I’m
  the sole viewer of this image)
- I prudently choose the term „quantity of photons", rather than „intensity"
  or „luminosity" or „luminance" or „lightness" or „brightness" because these
  terms are often confusing in many minds (including my own mind, and I must
  refere to definitions from the CIE or SI).

>[...]In video, we do operations like A+B all the
>time, but A and B are not usually proportional to intensities, they are
>typically proportional to roughly the square roots of intensities.

- saying „video", you must say „traditional TV video" or „gamma involved
  video"
- with a video signal proportional to intensity, I do video too. And
  calculating A+B, I get a value proportional to an intensity
- video just means *something related to vision*, no matter of gamma or not

>If you take linear-intensity image data, [...] gamma-correct by taking the
>0.4, 0.45, or 0.5 power, quantize or digitize to 8 bits, then send that
>data to a conventional CRT, no visible banding will be introduced under any
>reasonable conditions.

Right, but what is the origin of gamma ?
Is it *mainly* because of the human perception ? no !
Is it because of digitization ? no, obviously. Gamma exist since the birth
of television !
- in video (general term), the goal is to shoot a scene (with a sensor),
  transport the data to another place and render them on a screen
- a typical video system is a camera, a wire and a monitor
- even if the human perception is non-linear, a perfect video system can
  acquire the intensity linearly, transmit it without noise, and render it
  linearly: and the human eyes are satisfied
- but, in early TV, a camera (mainly made of a cathode ray tube) have a
  transfer function related to a power of the intensity of light of approx.
  0.45 (inherent to the sensor)
- two possibilities arrise:
  1) correct the signal before transmission, to achieve proportionality to
     the intensity
  2) transmit the signal, and correct it upon reception
- the second solution was retained, because:
  1) upon reception, the screen is a CRT, with a transfert function
     related to the voltage of the video signal with a power coefficient
     approx. 2.2, and that corrects the incident signal (at no expense ;
     for the first choice, signal must be transformed at the camera and at
     the monitor)
  2) and, *related to the human perception*, noise immunity is better
     achieved on a gamma-corrected video signal

12. Interesting Letter to Mr. Poynton (2)

```
Now, we have CCD sensors, delivering a video signal proportional to the
intensity of light.
We have LCD or plasma displays (with transfert function probably not the
same as a CRT)
We have digital transmission or storage systems (e.g. MPEG)
We not only take pictures to be transmited elsewhere, but to be processed
and give a result
But we must:
- in an analog world (traditional TV), be compatible with existing systems,
  and build cameras that deliver requested gamma-corrected video signal,
  and build display systems that properly render images with such a signal
- in a digital world:
  1) if the data are used for human seeing, take into account the human
     perception while quantizing, this attempt to limit artifacts (banding)
  2) to achieve image processing, use linear-video if you need
     proportionality to intensity of light

>Please read „Linear and nonlinear coding,“
> <http://www.inforamp.net/~poynton/notes/Timo/index.html>

I’ll do so, but there is a lot of stuff, and I need to regenerate my neurons
before digesting those writtings.

Francois Esquirol.
```