How the Number of Required Gray Levels Depends on the Gamma of the Display

Stanley A. Klein, Thom Carney
University of California at Berkeley, Berkeley, CA

Abstract
In order to adequately present slowly varying shades of gray without visible steps, a monitor must have very fine luminance steps. We discuss how the monitor's gamma is surprisingly well matched to the increment threshold of human vision. This allows the digital to analog converter to have 12 bits whereas 15 bits are needed for encoding the image.

1. Introduction

Last year at this meeting we calculated the number of bits/min² that are needed for lossless perception by the human visual system (Klein & Carney, 1990). We claimed that 100 bits/min² were needed. After correcting an error in our calculations (to be discussed) this number should have been 100 bits/min², which is still surprisingly large. This year we will reexamine the calculation on the experiment that we recently did to measure the number of discriminable grey levels. We will also go into detail about how the gamma of the monitor affects the calculation. We hope to clarify these issues by using a nonsquare lookup table (LUT) in which the number of addresses is different from the range of numbers stored in the addresses. We will show that the number of addresses of the LUT is based on the sensitivity of the human visual system (HVS) at low luminances, and the range required of the DAC is based on the HVS sensitivity at high luminances.

The number of bits/min² needed for a perceptually lossless display is equal to the product of the number of bits/pixel times the number of pixels/min². The latter quantity is determined by the resolution limit of human vision. We have found that two lines that are separated by .33 mm can be distinguished from a single line, even if contrast is jittered (Levi & Klein, 1990; Carney, Klein & Levi, 1988). Thus 9 pixels/mm² are needed by the perfect display, a number that is higher than commonly realized. The number of bits per pixel is more complicated to calculate and is the topic of the present paper.

2. Last year's calculation of the number of bits/pixel.

In order to calculate the number of bits/pixel we need to know the span of luminances needing to be covered and we need to know the smallest luminance that can be discriminated. Last year we made the assumption that a .01 cd/m² luminance increment could be discriminated at a luminance of 2 cd/m² (a Weber's contrast of .5%). We made the assumption that the monitor's luminance is linearly related to input voltage ("gamma" = 1) so that in order to go from 0 to 200 cd/m², the number of luminance levels is 200/0.01 = 20,000. Last year this identical calculation plus a misplaced decimal point led to 200,000 levels. The number of levels is converted to bits by taking the logarithm to the base 2. Last year we calculated 17.6 bits whereas it should have been 14.3 bits of grey levels which leads to a total of 14.3 x 9 x 130 bits/mm². This year we will calculate the number of bits/pixel more accurately. Our discussion of nonsquare lookup tables should help clear up an ambiguity about whether the number of grey levels refers to the number of levels of the stimulator or of the monitor.

3. The gamma of the monitor.

This year we would like to refine our calculation by taking into account the gamma of the scope similar to what was done by Marsh & Weiman (1990) in the paper following ours at last year's meeting. Our approach differs from theirs. We will use an analytic expression for both the gamma characteristic of the scope and for the Weber's Law characteristic of the human visual system. Our use of analytic expressions give us greater understanding of how the monitor's gamma affects the calculation of the number of grey levels that are needed to avoid noticeable unintended jumps in luminance.

The luminance (cd/m²) of a display is typically a power function of the voltage, v, that is output from the computer's digital to analog converter (DAC). We assume that the relationship between luminance and voltage is given by:

L = L₀ + L₁vγ

where L₀ is the minimum luminance, L₁ is produced by the room lighting, γ is the gamma of the display, L₀ is the range of luminances. The voltage, v, produces the highest luminance which is L₀ + L₁. We have defined the origin of the voltage scale to be the voltage that produces the lowest luminance. The exponent, γ, is called the gamma of the display. The gamma is typically between 2.0 and 3.0 (Netravali, 1986) with γ = 2.2 being the most commonly found value. The relationship in Eq. 1 results from the nonlinear connection between beam current and applied voltage.

Eq. 1 can be inverted (swapping and absissa) to produce a lookup table that specifies what voltage is needed to obtain any given luminance:

v = L₀vγ

The top panel in Fig. 1 shows the function in Eq. 2 for γ = 1.2, and 3.0. L₀ = 2.0, L₁=100, and v₀ = 3.0. Notice that for γ = 2 and 3, at the lowest luminances a sizeable change in voltage is needed to produce a small change in luminance. This plot of voltage vs. luminance is closely related to a lookup table (LUT). The lookup tables that we will discuss will be optimized to make full use of the range of luminances covered by the monitor.

4. The Weber fraction of vision

The visibility of a step luminance increment determines the fineness of the quantization steps that the monitor must be able to generate. The detectable step size increases with luminance and according to a modified Weber's Law:

ΔL/L = W (L/L₀)γ

where ΔL/L is called the Weber fraction and W is the Weber fraction for the "Weber" luminance, L₀. The exponent α controls the rate at which the Weber fraction increases with luminance.

Figure 1. The "gamma" relationship between voltage and luminance. The voltage is plotted on the ordinate rather than the abscissa to emphasize the use of this plot as a lookup table where one starts with an intended luminance from the image and one must then determine what voltage is needed to generate that luminance. The three curves are for γ = 1.2, 2.0, and 3.0 in Eq. 2.

5. Conclusion

The Weber fraction of vision is a critical parameter for video systems. The Weber fraction increases with luminance and can be used to determine the number of bits/pixel needed for lossless perception. The gamma of the display is also important for accurately calculating the number of bits/pixel needed, especially at low luminances. Further research is needed to refine the calculation and optimize the lookup tables for video systems.
which the Weber fraction gets smaller as luminance increases. The values of \( x = 0 \) and \( x = -5 \) correspond to Weber's Law and the De Vries Law respectively. We will choose \( x \) to be at the geometric mean of the luminance range so that changes in \( x \) will have minimal effect on the Weber fraction over a broad luminance range.

Before we obtain an analytic solution for the total number of bits that are needed by the DAC to be able to produce all discriminable luminance steps it is useful to give an example. The example will, in addition, show the need for nonsquare lookup tables.

5. Lookup tables

5.1 An example using a square lookup table

The example shown in Table 1 assumes \( x = 2.2, x = -5, L_0 = 2 \text{ cd/m}^2 \), and \( L_r = 198 \text{ cd/m}^2 \) so that the maximum luminance is 200 \text{ cd/m}^2. The Weber fraction has been chosen to be .6% at \( L_w = (2 \times 200)^{0.5} = 20 \text{ cd/m}^2 \), the geometric mean of the luminance range. We will discuss this choice of Weber fraction later. The first column is the intended luminance, \( L_i \), that goes from 2 to 200 \text{ cd/m}^2. In this example, we are assuming a 10-bit DAC so the luminance step size is 198 \text{ cd/m}^2/1024 = .193 \text{ cd/m}^2. The second column is the LUT address, \( A \), given by:

\[
A = 1023(L_i - L_0)/L_r.
\]

<table>
<thead>
<tr>
<th>( L_i )</th>
<th>Address value</th>
<th>( L_U )</th>
<th>( L_r )</th>
<th>( \text{LHS} )</th>
<th>( \text{DHV} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.10</td>
<td>1.00</td>
<td>43.85</td>
<td>44</td>
<td>2.19</td>
<td>0.195</td>
</tr>
<tr>
<td>2.20</td>
<td>2.00</td>
<td>87.70</td>
<td>44</td>
<td>2.29</td>
<td>0.191</td>
</tr>
<tr>
<td>2.30</td>
<td>3.00</td>
<td>131.55</td>
<td>44</td>
<td>2.39</td>
<td>0.186</td>
</tr>
<tr>
<td>2.40</td>
<td>4.00</td>
<td>175.40</td>
<td>44</td>
<td>2.49</td>
<td>0.181</td>
</tr>
<tr>
<td>2.50</td>
<td>5.00</td>
<td>219.25</td>
<td>44</td>
<td>2.59</td>
<td>0.176</td>
</tr>
<tr>
<td>2.60</td>
<td>6.00</td>
<td>263.10</td>
<td>44</td>
<td>2.69</td>
<td>0.170</td>
</tr>
<tr>
<td>2.70</td>
<td>7.00</td>
<td>306.95</td>
<td>44</td>
<td>2.79</td>
<td>0.165</td>
</tr>
<tr>
<td>2.80</td>
<td>8.00</td>
<td>350.80</td>
<td>44</td>
<td>2.89</td>
<td>0.159</td>
</tr>
<tr>
<td>2.90</td>
<td>9.00</td>
<td>394.65</td>
<td>44</td>
<td>2.99</td>
<td>0.154</td>
</tr>
<tr>
<td>3.00</td>
<td>10.00</td>
<td>438.50</td>
<td>44</td>
<td>3.09</td>
<td>0.148</td>
</tr>
</tbody>
</table>

Table 1: Constructing and testing a square LUT

It is seen that \( A \) ranges from 0 to 1023. The third column is the voltage that is needed to produce the intended luminance. This voltage is given by Eq. 2, where \( x \), the voltage range, is 1023. The fourth column is the third column rounded off. The value in this column is an integer since it is the value that is output to the DAC. Columns 2 and 4 constitute the lookup table. When the value of the 4th column is converted to a voltage by the DAC and then the voltage is applied to the monitor, the luminance that results is specified by Eq. 1, and is shown in column 5 as \( L_r \). Notice that column 5 is very close to column 1, the only difference coming from the round off in column 4. Columns 6 and 7 are most important. Column 6 shows the delta luminance jumps (DL monitor) of column 5. For example, in going from an address of 0 to an address of 1, the LUT value goes from 0 to 44 and the luminance changes from 2.000 to 2.195 \text{ cd/m}^2. The entry in column 6 is .195 \text{ cd/m}^2. Column 7 is the luminance change that can be seen by the human visual system (DLHVS) at that luminance. At a luminance of 2.19 \text{ cd/m}^2, according to our assumptions and Eq. 3, a luminance change of .0598 should be visible. Our goal in creating a perceptually lossless display is to have the values in column 6 be smaller than what the HVS can discriminate as shown in column 7. Unfortunately, in this example column 6 is larger than column 7, so it is shown as a bold number to indicate that the step size would be visible. Column 6 is seen to contain several bold values. This implies that a larger LUT or a larger DAC (a DAC with more output voltage levels) is needed.

5.2 A nonsquare lookup table

The situation in Table 1 is not really that bad. By comparing columns 7 and 4, it is seen that in most of the places where there is a bold number, the value in column 4 jumped by more than 1 step. Thus rather than getting a larger DAC we can simply expand the lookup table. This is done in Table 2. All the conditions are the same as for Table 1 except that now in column 1 the luminance increments are in steps of 198/4092 rather than 198/1023. The number 4092 is 4 times 1023. Thus what we have done is a 4-fold finer sampling of the luminance. We are using 12 bits of addresses and still only 10 bits of LUT values (column 4). We call this situation where the number of address bits differs from the number of value bits, a nonsquare LUT. It can be seen that the luminance increments of the monitor (column 6) and of the human visual system (column 7) are now much closer together. The discrepancies (the bold entries in column 7) that were present at the middle luminance levels in Table 1 are now gone in Table 2. The discrepancies at the very lowest luminances are associated with jumps in column 4 that are bigger than 1. Thus if we had gone to 13 bits of addresses (an 8-fold interpolation) then we would have generated a perfectly perceptually lossless display except for remaining discrepancies at the highest luminances which we now consider.

<table>
<thead>
<tr>
<th>( L_i )</th>
<th>Address value</th>
<th>( L_U )</th>
<th>( L_r )</th>
<th>( \text{LHS} )</th>
<th>( \text{DHV} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.05</td>
<td>1.00</td>
<td>43.85</td>
<td>44</td>
<td>2.19</td>
<td>0.195</td>
</tr>
<tr>
<td>2.10</td>
<td>2.00</td>
<td>87.70</td>
<td>44</td>
<td>2.29</td>
<td>0.191</td>
</tr>
<tr>
<td>2.15</td>
<td>3.00</td>
<td>131.55</td>
<td>44</td>
<td>2.39</td>
<td>0.186</td>
</tr>
<tr>
<td>2.20</td>
<td>4.00</td>
<td>175.40</td>
<td>44</td>
<td>2.49</td>
<td>0.181</td>
</tr>
<tr>
<td>2.25</td>
<td>5.00</td>
<td>219.25</td>
<td>44</td>
<td>2.59</td>
<td>0.176</td>
</tr>
<tr>
<td>2.30</td>
<td>6.00</td>
<td>263.10</td>
<td>44</td>
<td>2.69</td>
<td>0.170</td>
</tr>
<tr>
<td>2.35</td>
<td>7.00</td>
<td>306.95</td>
<td>44</td>
<td>2.79</td>
<td>0.165</td>
</tr>
<tr>
<td>2.40</td>
<td>8.00</td>
<td>350.80</td>
<td>44</td>
<td>2.89</td>
<td>0.159</td>
</tr>
<tr>
<td>2.45</td>
<td>9.00</td>
<td>394.65</td>
<td>44</td>
<td>2.99</td>
<td>0.154</td>
</tr>
<tr>
<td>2.50</td>
<td>10.00</td>
<td>438.50</td>
<td>44</td>
<td>3.09</td>
<td>0.148</td>
</tr>
<tr>
<td>2.55</td>
<td>11.00</td>
<td>482.35</td>
<td>44</td>
<td>3.19</td>
<td>0.142</td>
</tr>
</tbody>
</table>

Table 2: A non-square lookup table
At a luminance of 200 cd/m² the luminance step-size of the 10 bit DAC would be .4256, but the Weber fraction is .39. The 10 bit DAC is therefore not quite sufficient, and an 11 bit DAC is needed to generate fine enough luminances such that the step can not be detected by the HVS. If the Weber fraction, W, had been slightly higher, then 10 bits would have been sufficient.

For the example given in Tables 1 and 3, it is seen that the number of bits of address is controlled by the Weber fraction of the human visual system at the lowest luminances, and the number of bits of the DAC is controlled by the Weber fraction at the highest luminances. As long as the gamma of the monitor is greater than unity, it will always be true that the number of address bits is controlled by the lowest luminances. The luminance that controls the number of bits of the DAC, however, is a delicate balance between the monitor's gamma and the HVS alpha as will be discussed in Section 6.

There are three methods to generate nonsquare lookup tables: a software LUT, a software transformation, and hardware. The software LUT is straightforward, but just before the image is sent to the DAC's square LUT, it is sent to a software LUT of the form given in columns 2 and 4 of Table 2. The problem with this method is that it is slow, and speed is needed for video presentations. The software transformation method is that when the image is first created it should undergo a nonlinear transformation such as given by Eq. 2. The problem with this method is that the nonlinear transformation makes it difficult to perform linear operations on the image such as adding two images together or filtering the image. The hardware method would involve a DAC with a built-in nonsquare LUT. This is the ideal method. The only problem with it is that to our knowledge such DACs are not yet available. The presently available hardware implementation requires the addition of memory chips with the larger address space that would feed into the DAC. The problem with this method is that it will decrease the bandwidth of the video signal (the decrease would be measured in nanoseconds, rather than microseconds in the software LUT). It would be better to build the larger memory into the DAC itself and avoid the speed loss. However, this may produce a heat problem.

6. An analytic solution to the number of needed bits.
6.1 The number of value bits. We would now like to develop an analytic expression for the critical luminance that controls the number of bits needed by the DAC. For the values of γ and α in the above example, it turned out that the critical luminance was at the highest luminance. This need not be always true.

We begin by getting an expression for the change in voltage LUT value that corresponds to a just noticeable change in luminance.

\[ \Delta V = \delta V / \Delta L \]  

The derivative can be obtained from Eq. 2, and \( \Delta L \) from Eq. 3:

\[ \Delta V / \Delta L = W L_{\text{avg}} / (L - L_0) / (1 - 1/\gamma) / (L_{\text{avg}} - L_0) \]  

The reciprocal of \( \Delta V / \Delta L \) gives the total number of voltage steps that are needed to cover the full voltage range if the step size is determined by the just noticeable change in voltage at the luminance 1. In Table 1, W was equal to 1023 and \( \Delta L \) was unity. That is, we want to have the lookup table large enough so that when the value changes by unity anywhere in the table, the change is just at threshold. The number of bits required of the voltage DAC to cover that range is thus given by:

\[ N = \log_2 (V / \Delta V) \log_2 (2) \]  

We are interested in finding the minimum value of \( \Delta V \) (or maximum value of N) so that we can be assured that the step size is adequate for the entire luminance range. To find the minimum N, we set its derivative with respect to L equal to zero:

\[ \frac{dN}{dL} = \frac{(1/\gamma - 1)(L - L_0) + (a + 1) / L}{(a + 1) / L} \]  

Eq. 8 can be solved for L. The solution will be called \( L_\alpha \), the critical luminance at which the number of bits is maximal:

\[ L_\alpha = L_0 (1 - 1/\gamma) / (a + 1) \]  

Special cases of this formula are:

\[ L_\alpha = L_0 \]  

for \( \gamma = 1 \)

\[ L_\alpha = \log_2 (W) \]  

for \( \alpha = 0 \)

For \( \gamma = 2.2 \) and \( \alpha = 0, -1, -2, -3, -4 \), and -5, \( L_\alpha \) in Eq. 8 becomes 5.4, 5.1, 6.3, 9.1, 22 and 200 respectively. The last value (where \( L_\alpha \) hits a ceiling) deserves special mention. If we make the approximation that \( L_\alpha \gg L_0 \) then \( L_\alpha \) is at the maximum luminance (200 in our case) when \( a \leq -1/\gamma \). For a monitor \( \gamma = 2.2 \) this occurs at \( a \leq -1.22 \approx -45 \). Thus if the luminance and obeys the De Vries-Rose law (see Section 7 for our evidence that \( a = -5 \)) the \( \alpha \leq -1/\gamma \) condition is met and we would expect that the highest luminance (200 cd/m² in our example) places the constraint on the number of bits required of the DAC.

By plugging Fig. 9 back into Eq. 7 and 6, we obtain the total number of bits that are needed to guarantee that all perceptually discriminable luminance changes can be generated by the DAC.

\[ N = \log_2 [(W - L_0) / L_\alpha (1 - 1/\gamma)] / \log_2 (2) \]

The numbers of bits corresponding to \( \gamma = 1 \) (solid curve) and \( \gamma = 2.2 \) (dashed curve) are shown in Fig. 2. The ordinate scale on the left corresponds to a Weber fraction, W = .006 as in Tables 1 and 2. The right-hand ordinate is for W = .0014 as will be discussed in Section 7 in connection with our experimental results. The lower curve (\( \gamma = 2.2 \)) specifies the number of bits needed by the DAC. The other curve, specifying the number of needed address bits will be discussed in Section 6.2. For the conditions of Tables 1 and 2 (\( \alpha = -2 \) and \( \gamma = 2.2 \)) the number of needed DAC bits is 10.16. Tables 1 and 2 lead to the same value since the discrepancy between columns 6 and 7 at the highest luminance predicts that \( \log_2 (L_{16} - 380) = 16.5 \) and extra bits are needed above the 10 bits with which we started, in agreement with Fig. 2.

Figure 2. The number of bits needed to adequately specify a lookup table. The upper curve is the number of address bits given by Eq. 14 or by Fig. 11 with \( \gamma = 1 \). The lower curve is the number of output (DAC) bits with from Eq. 11 with \( \gamma = 2.2 \). The left and right ordinates correspond to Weber fractions of W = .006 and .0014.
6.2 The number of address bits. The calculation of the number of bits that are needed for the addresses is a bit simpler. The address step size is determined by the lowest luminance level. The total number of these steps is $L_0/\Delta L_0$ so the total number of address bits is:

$$N_A = \log_2\left(\frac{L_0}{\Delta L_0}\right)$$  

where $\Delta L_0$ is the luminance and at the lowest luminance level. It can be calculated from Eq. 3:

$$\Delta L_0 = W L_0^{\alpha+1} L_0^{-\alpha}$$

so the number of address bits becomes:

$$N_A = \log_2\left(\frac{L_0 W}{\alpha+1}\right)$$

For the conditions of Tables 1 and 2, $L_0/L_0 = 100$ and $L_0/L_0 = 10$, $\alpha = 5$ and $W = 0.006$. Plugging these values into Eq. 13 gives the number of address bits to be 12.35, in agreement with our earlier statement that 12 bits was not quite sufficient for Table 2. The extra .36 bits is approximately the number of bits needed to account for the discrepancy between the .0469 cd/m² step of the monitor and the .0384 cd/m² detectable step of the human visual system.

The number of necessary address bits is the upper curve in Fig. 2. It is given by either Eq. 14 or by the prescription of Section 6.1 with $\gamma = 1$ since this choice of $\gamma$ forces the critical luminance to be at $L_0$. One interesting property of this curve is that the $\alpha$ dependence is a straight line as seen in Fig. 2 and in Eq. 14.

7. An experiment to measure the Weber fraction, $W$.

We now consider the question of how to choose the Weber fraction, $W$. In Tables 1 and 2, the Weber fraction was .6% at 20 cd/m². This implies a Weber fraction of about .2% at 200 cd/m² and 2% at 2 cd/m² assuming a De Vries-Rose square root improvement in detection contrast as a function of luminance. These are reasonable values for standard experimental conditions. One can, however, do quite a bit better.

We recently carried out a preliminary experiment to place very stringent constraints on the size of the just detectable luminance increment. We measured the contrast discrimination threshold of a sinusoidal grating drifting at 5 Hz. Two luminances were used: 1 cd/m² and 100 cd/m². The 1 cd/m² luminance was produced using a ND filter (a 100-fold luminance reduction).

The purpose for using drifting gratings is that (contrary to Murch & Weiman, 1986) we expected that the presence of drift should increase the visibility of the gratings about two-fold. The purpose for using contrast discrimination rather than contrast detection is that for a pedestal that is a few times threshold (at the bottom of the "dipper" function) the facilitation effect should reduce the visible test threshold by a factor of 2 - 3 (Stromeyer & Klein, 1974; Nachmias & Sansbury, 1974).

At the 1 cd/m² luminance a 2 c/deg sinusoidal grating was used. We could discriminate (d = 1) between gratings whose contrasts were 1.5% and 1.9%. Thus the discrimination threshold (detection of the contrast increment) was 4%. A square wave matched in visibility to the sine wave (one with the same strength fundamental) would have a contrast of $\pi^2/4 = 0.314$. This corresponds to a luminance Weber fraction of $0.314 \times 2 = 0.628$, or a luminance increment of 1 cd/m². 0.628% = 0.00628 cd/m². At a luminance of 100 cd/m² we found the discrimination threshold of a 4 c/deg drifting sinusoid to be .04. This remarkably low value needs further verification since it is getting close to the limits of the 12 bit DAC of our VENUS display generator.

Our finding that a 100-fold increase in luminance resulted in a 10-fold decrease in contrast threshold was the reason that we have used the value of $\alpha = 0.5$ (De Vries-Rose Law) in Tables 1 and 2. We assume that the De Vries-Rose Law will extend to the 2 to 200 cd/m² range that we have been using. At the geometric mean luminance of $L_0 = 20$ cd/m² the expected Weber fraction should be $W = 0.00628 \times 20^{-0.5} = 0.0014$. This value of $W$ based on our experiment is a bit more than 4 times lower than the value of .006 used in Tables 1 and 2 and in Fig. 2. This factor of 4 translates into 2 more bits being needed for both the address and DAC. The effect of the two extra bits can be seen by looking at the right-hand abscissa of Fig. 2. The upper curve generated from Eq. 14 or by setting $\gamma = 1$ in Eq. 11 corresponds to the number of address bits. The lower curve, from Eq. 14 with $\gamma = 2.2$, corresponds to the number of DAC bits. It is seen that about 12-13 bits are needed for the DAC and about 15-16 bits are needed for the address, for a wide range of $\alpha$ values.

In summary, by using an analytic expression for the nonlinear "gamma" relationship between voltaga and luminance we have shown how to calculate the number of bits that are needed by a "perfect" display to avoid unwanted visible luminance steps. The nonlinear gamma of display controls varies about these bits when a nonsquare lookup table is used. The scope had a unity gamma then the video DAC would have needed 15 - 16 bits of output levels, whereas with a gamma between 2 and 3 only 12 - 13 bits are needed.

References


