DIGITAL HALFTONING METHODS
FOR SELECTIVELY PARTITIONING ERROR
INTO ACHROMATIC AND CHROMATIC CHANNELS

Jeffrey B. Mulligan
NASA Ames Research Center
Mail Stop 239-3, Moffett Field, CA., 94035-1000

ABSTRACT

A method is described for reducing the visibility of artifacts arising in the display of quantized color images on CRT displays. The method is based on the differential spatial sensitivity of the human visual system to chromatic and achromatic modulations. Because the visual system has the highest spatial and temporal acuity for the luminance component of an image, we seek a technique which will reduce luminance artifacts at the expense of introducing high-frequency chromatic errors. In this paper we explore a method based on controlling the correlations between the quantization errors in the individual phosphor images. The luminance component is greatest when the phosphor errors are positively correlated, and is minimized when the phosphor errors are negatively correlated. The greatest effect of the correlation is obtained when the intensity quantization step sizes of the individual phosphors have equal luminances. For the ordered dither algorithm, a version of the method can be implemented by simply inverting the matrix of thresholds for one of the color components.

1. INTRODUCTION

Human color vision is made possible by the initial encoding of the visual stimulus by the three photosensitive cone pigments in the retina. These signals are transformed and processed by subsequent neural stages in ways that are still not completely understood; empirical data does suggest, however, that certain special transformations are likely to occur. These inferences are based in large part upon the fact that the spatial and temporal response of the visual system varies with the color composition of the stimulus.

The differences in the visual system’s responses to chromatic and achromatic modulations are clearly illustrated by differences in the contrast sensitivity function, as was measured by Kelly[1]. Kelly found that sensitivity to stationary luminance modulation had a band-pass character, with a broad peak of sensitivity occurring at a spatial frequency of around 2 cycles per degree. Sensitivity to equiluminant chromatic modulation on the other hand was more low-pass in nature, with the peak occurring at approximately 0.5 cycles per degree, followed by a sharp decline in spatial sensitivity for higher spatial frequencies. There are also comparable differences in temporal sensitivity.

Results concerning spatial sensitivity are only a part of the body of evidence indicating that the linear combination of the cone signals known as "luminance" is treated specially by the visual system. The evidence concerning the particular combinations which encode the remaining color information is more equivocal; it is widely believed, however, that the first step is to form two color opponent signals: a red-green signal based on the difference between the excitations of the red and green cones, and a blue-yellow signal based of the difference of the blue cone signal and the luminance signal. Because of the low density of blue cones in the retina[2], the blue-yellow system has particularly poor spatial resolution.

In this paper we examine the question of whether these differences in sensitivity based on color can be exploited to improve the fidelity of quantized images which are reproduced on a CRT display. In particular, we are interested in reducing the luminance error, even at the expense of introducing substantial additional chromatic error, since it is the luminance signal to which the visual system is most sensitive at the high frequencies where most quantization procedures attempt to concentrate the noise.

2. THEORY

2.1. Applying ordered dither to color images

We will examine the problem of minimizing the luminance component of quantization errors in the context of the halftoning method known as *ordered dither*[3,4]. Although this method has some drawbacks, it has the advantages of simplicity and speed; the operations are performed independently at each pixel in the image, so the algorithm may be implemented in parallel. We consider the case of binary quantization, i.e., where each pixel in a monochrome image is represented by either black or white.

The algorithm is implemented as follows: the image is first divided up into subregions, or "cells." The cell size is a parameter which may be tuned to achieve a tradeoff between spatial and intensity resolution. Within each cell, a matrix of values is added to the corresponding matrix of pixels, which are then quantized using a simple fixed threshold. Different design rules are used to construct the dithering matrix depending on the final output device; different types of matrices are usually used for printers and CRT's.

When we apply this method to the dithering of color images, the simplest approach is to treat the red, green, and blue component phosphor images as independent monochrome images, dither each one, and finally combine the three quantized component images to form a quantized composite image. With a palette of eight colors, (three bits per pixel), the allocation of one bit to each phosphor is the only palette choice which allows the entire monitor gamut to be reproduced. When a larger palette is used, it is possible to achieve substantial improvements by analyzing the actual content of the target image; this involves significant preprocessing, however, so in applications where speed is important, such as interactive viewing of a large number of different images, it is convenient to work with a fixed palette which allows the phosphor component images to be processed independently. In following discussion, we will only consider palettes of this type.

The appearance of a single monochrome image which has been quantized by ordered dither will not be greatly affected if we rotate or flip the dither matrix prior to quantization. Such transformations have a large effect only on the high frequency structure of the quantization noise, but do not substantially affect the reproduction of spatial frequencies having periods of many cell sizes. Given the arbitrariness of the choice of the dither matrix for a single monochrome image, we can ask the question of whether we should choose identical or different matrices for quantizing the component phosphor images that make up a color image.

2.2. Multi-level quantization

It is easy to generalize the method to multi-level quantization as follows: if we desire a final image having N uniformly spaced levels, we first quantize the image to N-1 levels by simply thresholding each pixel, rounding down to insure that no pixels are assigned to the highest level. We then form an error image by subtracting the quantized image from the original. Because we have rounded down to quantize, the error is strictly positive, ranging from 0 to the quantization step size. We can then perform ordinary binary ordered dither on the error image to provide an additional bit of pixel information. This one bit dithered image is then used to promote pixels in the quantized image to the next highest level.

The main drawback of this technique is that the use of a uniformly quantized grey scale makes the dither pattern selectively more visible in dark regions of the image. It has the advantage, however, of simplicity and speed.

2.3. Analysis of an idealized two phosphor system

To gain some insight into the general problem we will first analyze a simple idealized case consisting of a two phosphor system. We will ignore the two-dimensional discrete nature of the dither matrices, and idealize the quantization interval as a one-dimensional continuum. We will also assume that the desired color is uniform over the quantization interval. Thus we will treat the problem as analogous to pulse-width modulation, where the pulse rate is much higher that the highest input frequency. Having made these simplifying assumptions, we now define the following symbols:

r desired fractional intensity of the red phosphor
g desired fractional intensity of the green phosphor

B fraction of the quantization interval over which neither phosphor is on (black)
R fraction of the quantization interval over which only the red phosphor is on (red only)
G fraction of the quantization interval over which only the green phosphor is on (green only)
Y fraction of the quantization interval over which BOTH phosphors are on (yellow)

It should be noted that **B**+**R**+**G**+**Y**=1 by definition.

The quantization is performed in our idealized case by comparing the desired intensity to a thresholding function which varies over the quantizing interval, such as a linear ramp. A linear ramp insures that we obtain a linear representation after integrating over the quantization interval, but any function constructed by chopping the ramp up into subintervals which are then permuted will have the same effect as far as the low frequencies are concerned. We will assume that the red signal is quantized using a positive-going ramp for the threshold function; if we use the same function to quantize the green signal, then we can easily write expressions for the fractional areas **B**,**R**,**G** and **Y**:

$$\mathbf{Y} = \min(\mathbf{r},\mathbf{g}), \tag{1}$$

$$\mathbf{B} = 1-\max(\mathbf{r},\mathbf{g}), \tag{2}$$

$$\mathbf{R} = \max(0,\mathbf{r}-\mathbf{g}), \tag{3}$$

$$\mathbf{G} = \max(0,\mathbf{g}-\mathbf{r}). \tag{4}$$

We shall refer to this case as the positively correlated condition. If instead we choose a negative-going ramp for the green threshold function, we have the negatively correlated condition, with the following quantities for the fractional areas:

$$\mathbf{Y} = \max(0,(\mathbf{r}+\mathbf{g})-1) \tag{5}$$

$$\mathbf{B} = \max(0,1-(\mathbf{r}+\mathbf{g})) \tag{6}$$

$$\mathbf{R} = \min(\mathbf{r},1-\mathbf{g}) \tag{7}$$

$$\mathbf{G} = \min(\mathbf{g},1-\mathbf{r}). \tag{8}$$

A final case of interest is one in which the two threshold functions are uncorrelated. This case is easier to understand if the quantization interval is again considered to have two spatial dimensions; the two threshold functions are still one-dimensional ramps over this interval, but have orthogonal tilts. (This type of correlation is also what we would expect from independent noise dithering, or error diffusion.) We obtain

$$\mathbf{Y} = \mathbf{r}\ \mathbf{g} \tag{9}$$

$$\mathbf{B} = (1-\mathbf{r})\ (1-\mathbf{g}) \tag{10}$$

$$\mathbf{R} = \mathbf{r}\ (1-\mathbf{g}) \tag{11}$$

$$\mathbf{G} = (1-\mathbf{r})\ \mathbf{g} \tag{12}$$

Now let us consider the error that results from quantization. We desire a certain level of red ($\mathbf{r}$), but we produce instead two spatial intervals, one having a value of 1 with an error of (1-$\mathbf{r}$), and another having a value of 0 with a resulting error of (-$\mathbf{r}$). The average red error, $\mathbf{e_r}$, can be expressed as

$$\mathbf{e_r} = (1-\mathbf{r})\ (\mathbf{Y}+\mathbf{R}) + (-\mathbf{r})\ (\mathbf{B}+\mathbf{G})\ . \tag{13}$$

The first term in the sum, which is always positive, represents the excess red light from the regions where the desired value was rounded up to 1; the second term represents the error resulting from the regions where the desired value was rounded down to 0. Similarly, the green error may be expressed

$$\mathbf{e_g} = (1-\mathbf{g})\ (\mathbf{Y}+\mathbf{G}) + (-\mathbf{g})\ (\mathbf{B}+\mathbf{R})\ . \tag{14}$$

It can be verified that these quantities have a value of zero when the values for the fractional areas from any of the three cases listed above are substituted into the equations. (When spatial quantization is introduced this will not remain true in general.) A more interesting quantity is the error variance, which is the square of the local error integrated over the quantization interval:

$$\sigma^2{}_{\mathbf{r}} = (1-\mathbf{r})^2\ (\mathbf{Y}+\mathbf{R}) + (-\mathbf{r})^2\ (\mathbf{B}+\mathbf{G})\ , \tag{15}$$

$$\sigma^2{}_{\mathbf{g}} = (1-\mathbf{g})^2\ (\mathbf{Y}+\mathbf{G}) + (-\mathbf{g})^2\ (\mathbf{B}+\mathbf{R})\ . \tag{16}$$

The values of these quantities depends on the particular levels which we are trying to reproduce; they attain a maximum value of 0.25 when the desired level is 0.5, and the local error has a constant magnitude of 0.5 over the entire interval.

It should be noted that the error variance of the red phosphor depends only on the fraction of the area over which the red phosphor is turned on (**R**+**Y**), and is independent of how we quantize the green signal. As was suggested in the introduction, however, the red phosphor is unlikely to correspond to the visual system's internal representation of color. In order to predict the visual salience of the errors, therefore, it may be useful to recode the error into assumed perceptual dimensions prior to computing the variance.

We can write simple expressions for the achromatic and chromatic components of the signal as follows:

$$l = l_{\mathbf{r}}\mathbf{r} + l_{\mathbf{g}}\mathbf{g} \tag{17}$$

$$c = l_{\mathbf{r}}\mathbf{r} - l_{\mathbf{g}}\mathbf{g} \tag{18}$$

We can similarly write expressions for the luminance and chrominance of each of our quantization colors:

$$l_{\mathbf{Y}} = l_{\mathbf{r}} + l_{\mathbf{g}}\ , \qquad c_{\mathbf{Y}} = l_{\mathbf{r}} - l_{\mathbf{g}}\ , \tag{19}$$

$$l_{\mathbf{R}} = l_{\mathbf{r}}\ , \qquad c_{\mathbf{R}} = l_{\mathbf{r}}\ , \tag{20}$$

$$l_{\mathbf{G}} = l_{\mathbf{g}}\ , \qquad c_{\mathbf{G}} = -l_{\mathbf{g}}\ , \tag{21}$$

$$l_{\mathbf{B}} = 0\ , \qquad c_{\mathbf{B}} = 0\ . \tag{22}$$

We can express the luminance error integrated over the quantization interval:

$$\mathbf{e}_l = (l_{\mathbf{Y}}-l)\mathbf{Y} + (l_{\mathbf{R}}-l)\mathbf{R} + (l_{\mathbf{G}}-l)\mathbf{G} + (l_{\mathbf{B}}-l)\mathbf{B}\ , \tag{23}$$

Similarly, we can express the integrated chrominance error as sum of the local chrominance errors weighted by the corresponding areas:

$$\mathbf{e}_c=(c_{\mathbf{Y}}-c)\mathbf{Y} + (c_{\mathbf{B}}-c)\mathbf{B} + (c_{\mathbf{R}}-c)\mathbf{R} + (c_{\mathbf{G}}-c)\mathbf{G}. \tag{24}$$

We express the variance as the square of the local error integrated over the quantization interval:

$$\sigma^2{}_l = (l_{\mathbf{Y}}-l)^2\mathbf{Y} + (l_{\mathbf{R}}-l)^2\mathbf{R} + (l_{\mathbf{G}}-l)^2\mathbf{G} + (l_{\mathbf{B}}-l)^2\mathbf{B} \tag{25}$$

$$\sigma^2{}_c = (c_{\mathbf{Y}}-c)^2\mathbf{Y} + (c_{\mathbf{R}}-c)^2\mathbf{R} + (c_{\mathbf{G}}-c)^2\mathbf{G} + (c_{\mathbf{B}}-c)^2\mathbf{B} \tag{26}$$

The goal of this exercise is to show that by selecting the method for obtaining **Y**,**B**,**R** and **G**, we can influence the relative magnitudes of $\sigma^2{}_l$ and $\sigma^2{}_c$.

Let us assume for the moment that the range of the red and green phosphors is such that 1 unit of red has the same luminance as 1 unit of green, i.e. $l_{\mathbf{r}}=l_{\mathbf{g}}=1$. Under this assumption we can easily calculate the quantities $\sigma^2{}_l$ and $\sigma^2{}_c$ under each of the correlation assumptions stated above in equations 1-4, 5-8, and 9-12. These quantities are plotted in figures 1-4. Figure 1 shows the luminance error variance as a function of gray level for a gray scale ramp. The three curves represent correlations of 1,0, and -1 between the red and green threshold functions. Note that the error variance is greatest for the positive correlation, intermediate for the uncorrelated case, and smallest for the negative correlation. The case of negative correlation exhibits a zero for a gray level of 0.5: at this level the quantization interval is half red and half green; since we have assumed that the red and green phosphor units have equal luminance, there is no luminance variation over the interval.

Figure 2 plots the variance of the chromatic error for the same gray level ramp. Note that the ordering of the curves is reversed. The case of positive correlation exhibits zero chromatic error variance at all gray levels; since the red and green signals overlap completely, the image consists of only yellow and black regions. Conversely, the case of negative correlation exhibits the maximum chromatic error variance, since at every gray level there are regions of both pure red and green.

In figures 3 and 4 we have plotted the same quantities for an equiluminant chromatic ramp. In figure 3 the luminance error variance is plotted against the chrominance. Note that the curves have the same form as figure 2 above, but that the ordering is reversed. Similarly in figure 4, where the chromatic error variance is plotted as a function of the target chrominance, the curves have the same form as in figure 1.

A shortcoming of this analysis is that the variance measure does not take into account the spatial profile of the errors, which greatly affects the visibility of artifacts. A single pixel with a large error can make the same contribution to the variance as a larger number of pixels with smaller errors, although the latter will be less visible both because the amplitude is smaller, and more importantly because the average spatial frequency is higher. For the case of ordered dither considered above, however, this deficiency will only overestimate the benefit to be derived from the method; the method described never makes the final image worse than that obtained with correlated dither matrices.

2.4. Effects of unequal phosphor luminances

In performing the above calculations, we have assumed that the red and green phosphor excitations produce the same luminance. In practice, this is rarely the case: typically, the green phosphor produces substantial excitation in both the long-wave sensitive and middle-wave-sensitive cones, and consequently has a higher luminance. The NTSC standard specifies luminance ratios between the red, green, and blue phosphors of approximately 3:6:1, but substantial variation can be found in between actual monitors.

This effect can be incorporated into our analysis by using different values for the constants $l_{\mathbf{r}}$ and $l_{\mathbf{g}}$ which were introduced above in equations (17) and (18). These constants represent the relative luminances of the red and green phosphors, which were assumed above to each have a value of 1.

To illustrate the effects of this parameter, we consider an extreme case where the phosphor luminances have a ratio of 4 to 1. In order to compare this case to the previous example, the total luminance was held constant at 2, with $l_{\mathbf{r}}$ having a value of 0.4 and $l_{\mathbf{g}}$ having a value of 1.6. The resulting variances of the luminance and chromatic errors are plotted in figures 5-8. The calculations performed were identical to those that generated figures 1-4 with the exception of the luminance parameter. Note that the ordering of the curves is the same as before, but that the vertical differences are reduced. Another feature which can be noted in figures 5-8 is that the middle curve, representing the case of uncorrelated dither images, is unaffected by the change in the relative phosphor luminances as long as the total luminance is held constant. The fact that the differences between the curves is reduced as the phosphor luminance ratio departs from 1 reflects the fact that when one phosphor has a luminance much greater than the others, its errors will dominate the luminance error regardless how the other phosphors are quantized.

This observation suggests a method for determining the optimal number of quantization levels for each phosphor when we have the luxury of more than one bit per phosphor. In order to minimize the luminance component of the error, we should choose the quantization step size for each phosphor in inverse proportion to its relative luminance. A common task in computer graphics is the display of a color image on a display having only 8 bits per pixel.

A common approach is to allocate 3 bits for green, 3 for red, and 2 for blue, yielding 8 levels each for green and red, and 4 for blue. The above analysis suggests that we should instead allocate 11 levels for green and 5 for red, thus making the step sizes have approximately equal luminances. Note that the number of levels is not constrained to be a power of two, merely that the product of the number of levels must be less than or equal to 256.

2.5. Other possible approaches

The technique described above was motivated by a desire to exploit the visual system's insensitivity to high frequency chromatic modulation in order to improve the visual appearance of quantized images. It should be emphasized that the technique described above is by no means the only way that this might be accomplished. In this section we present a few speculations on how this important property of vision might be exploited by other existing schemes.

Instead of processing the component images independently, it is possible to read the three component images in parallel, and at each pixel quantize the output value to one of the palette colors. This may be done for either a fixed palette, or one which has been optimized for the particular image. In any case, the quantization process requires the use of a metric for determining which color from the palette is the "closest" to the target color. Although using a perceptually uniform color space such as CIElab or CIEluv[5] is more sensible than simply using the distance in RGB space, these color spaces were designed to represent color discrimination data obtained with large uniform fields, and therefore probably give excessive weight to purely chromatic differences. Simply stretching the space along the luminance axis is not satisfactory, however, as that will continue to neglect the effects of chromatic errors even when they occur at low frequencies.

One of the better digital halftoning methods is the Floyd and Steinberg error diffusion algorithm[6]. If the palette is one which is separable into red, green, and blue quantization steps (such as was considered in the preceding sections), then the red, green and blue component images can be processed independently. Except in unusual cases where the component images are highly correlated (as in a grey scale ramp), the halftoning errors produced by this algorithm will tend to be uncorrelated, which we have seen produces an intermediate case in terms of how the error is partitioned into chromatic and achromatic components. One way in which the luminance error might be reduced would be to add the error signal from the first image processed (which should be the one which makes the largest contribution to the luminance error), and add it to the next component's image prior to dithering. The error signal is usually high-pass noise with few structured artifacts; since it contains little energy at low frequencies, it should not distort the image, but should simply cause the new error to be negatively correlated with that of the first dithered component.

3. CONCLUSION

The above analysis demonstrates that substantial tradeoffs between the chromatic and achromatic components of halftoning noise can be affected by manipulating the correlation between the errors in the individual phosphor images. The optimal correlation for a particular image will in general depend on the overall noise level and the viewing conditions. The assumption that the luminance errors are the most visible is likely to hold when the errors are near threshold and concentrated at high spatial frequencies, but may fail when the errors are suprathreshold and at reduced spatial frequencies due to a small viewing distance.

Nevertheless, when the viewing conditions are known, it should be possible to apply basic visibility data to determine the optimal combination rule.

Inverting one of the dither matrices in component ordered dither is a simple way of decorrelating the component errors. There is little additional computational cost associated with doing this, and while the degree of improvement depends on the input image and the quantization parameters, the errors are less than or equal to those obtained with correlated matrices.

4. REFERENCES

1. D.H. Kelly, "Spatiotemporal variation of chromatic and achromatic contrast thresholds," *J. Opt. Soc. Am.*, vol. **73**, no. 6, pp. 742-750, June 1983.

2. F.M. de Monasterio, E.P. McCrane, J.K. Newlander, and S.J. Schein, "Density profile of blue-sensitive cones along the horizontal meridian of macaque retina," *Invest. Ophth. and Vis. Sci.*, vol. **26**, no. 3, pp. 289-302, March 1985.

3. B.E. Bayer, "An optimum method for two level rendition of continuous-tone pictures," *Proc. IEEE Int. Conf. Commun., Conference Record,* pp. (26-11)-(26-15), 1973.

4. R. Ulichney, *Digital Halftoning,* ch. 6, MIT Press, Cambridge MA., 1987.

5. G. Wyszecki and W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae, 2nd ed.,* John Wiley and Sons, New York, 1982.

6. R.W. Floyd and L. Steinberg, "Adaptive algorithm for spatial grey scale," *SID Int. Sym. Digest of Tech. Papers,* pp. 36-37, 1975.