Error Diffusion on CMY space

Wilkin Chau and William B. Cowan
Department of Computer Science
University of Waterloo
Ontario, Canada

Abstract

With recent advances in the resolution of colour printers, halftoning is increasingly used in digital colour printing. It takes advantage of the limited spatial resolution of human vision to simulate continuous-tone imagery by the judicious distribution of small ink dots of discrete sizes. Several halftoning techniques have been developed. Among them, error diffusion is very commonly used becuase its implementation is simple and its output quality is good. Originally designed for monochrome images, error diffusion is problematic to extend to colour printing. For example, in which colour space should the errors be propagated? How should errors be distributed in the multidimensional colour space ?

To understand error diffusion in colour printing, we are studying dithered images produced using a dye-sublimation printer. Without the dot gain and dot overlay problems, images produced by propagating error in CMY space are brighter, less satuarated and lack detail. In this paper, the implications of using CMY colour space for dithering are discussed.

Introduction

With the increasing availability of inexpensive bi-level colour printers, colour halftoning becomes more and more common. It is usually provided by dithering, using algorithms[1,2,3] that have been extensively optimized for monochrome imagery. These algorithms are applied to each colour channel (or primary) of the image, dithering the primaries independently of one another[4,5]. Unfortunately, this simple extension of monochrome algorithms, while easy to implement, only sometimes produces good results. The characteristics of the output devices play an important role on how an algorithm should be extended to produce colour dithered images. This paper describes the interaction between device technology, and dithering algorithm, explaining why algorithms that are effective for one device fail with others.

Colour halftoning is normally done in two stages. First possible colour outputs are considered. They use the primaries in all possible combinations; these are the only colours that can appear in the halftoned image. Second, the image is dithered into a representation that uses only those colours. Light provided by the image is spatially mixed in the eye of the observer to produce the impression of colours that are intermediate between the colour combinations. The second stage is the same for all devices that use halftone reproduction, but the first stage is strongly affected by device characteristics, because different devices produce colour combinations from primaries in different ways. The important distinction is between device that use additive and subtractive colour mixing. Since these two classes of devices generate colours using completely different methods, their first colour generation stages of colour halftoning are quite different. To understand how colour halftoning is affected by the output devices, we apply a standard error diffusion algorithm to additive RGB display surfaces (such as CRTs) and to subtractive CMY display surfaces (such as printers).

In this study, each colour plane of an image is treated as a monochrome image and the halftoning algorithm is applied accordingly. The separate primary images are then combined to produce the final dithered image. The algorithm generates good results for additive RGB display surfaces and poor results for CMY display surfaces. As discussed in the later sections of this paper, the discrepancy of the results are caused by nonlinearities in the colour combinations of the subtractive device. They cannot be removed because the primaries are imperfect, a concept that is formalized below. It is derived from the properties of ideal inks, the non-linearities of which can be corrected, allowing the production of good halftone production on subtractive devices.

The Basic Concept

Like all of the halftoning algorithms, error diffusion uses the limited spatial resolution of human vision to simulate continuous-tone imagery. As the ration of viewing distance to image pixel size increases the observers ability of resolve spatial detail in the image decreases. Thus, larger numbers of pixels contribute to the perceived colour at each point in the image. This process is known to be additive in the standard colour spaces of colorimetry[6] (specifically the space of CIE tristimulus values). Specifically, a

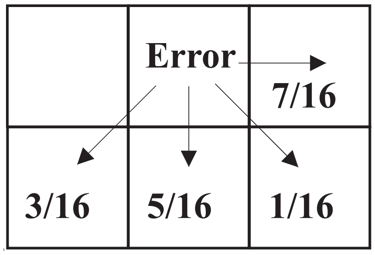

Figure 1: Error propogation weights

monochrome image with its pixels alternatively off and on, produces a sensation of grey, identical to the sensation produced by a continuous tone display surface displaying all pixels with tristimulus values that are exactly half of the full-on values. Effective halftone algorithms must therefore be able to maintain the average luminance of local regions of the original image. In other words, the low frequency components of the original image should be maintained. This is possible only if the device output can be linearized in the space of tristimulus values. (For reflective display surfaces, like printer output seen under constant illumination, this is equivalent to linearizing the device in a space that describes the reflectance of each pixel.) Error diffusion does this effectively. In addition it is simple to implement and generates a pleasingly structureless distribution of pixels. Consequently, it is the most popular halftoning algorithm.

Error diffusion is easy to understand. Let $C(x)$ be the colour of continuous-tone original at pixel x, normalized so that $0 <= I(x) <= 1.0$. Let $H(x)$ be the colour of the halftone reproduced image at x, which is either 0 or 1. To start the error diffusion process, the value of the first pixel $C(1,1)$ is compared with the threshold, normally set to 0.5 for binary output. If it is greater than the threshold, the halftoned pixel $H(1,1)$ is set to 1; otherwise, it is set to 0. The error, which is the difference between the continuous tone value $C(1,1)$ and the halftone value $H(1,1)$, is added to (or subtracted from) neighbouring pixels according to predetermined scaling factors (Figure 2). At the next pixel, the continuous tone value $C(1,2)$ summed with the propagated errors from the neighbouring processed pixels is used to compare the threshold and the output value is determined based on the same rule stated above. The operation is repeated for each pixel from left to right and top to bottom until the whole image has been processed. Careful coordination between the scaling factors and the order in which pixels are processed ensures that every pixel is processed only once, ensuring that the algorithm is efficient.

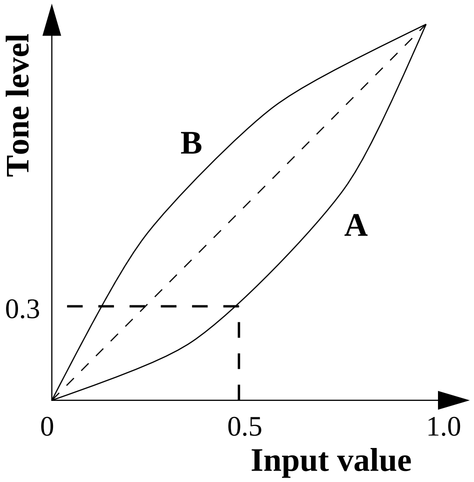

Figure 2: Device response functions

Spatial error diffusion depends on spatial averaging in the human visual system. Thus error calculations should be performed using a measurement scale that is linearly related to spatial averaging that occurs in vision. Otherwise, the resulting halftone image will be brighter or darker than the original. For example, when a display having the nonlinear response shown as curve A in Figure 2 shows the pixels with value of 0.5, the image has 30% of white luminance. Contrarily, if a dithered image of the same image is displayed, it is 50% of white luminance since the algorithm produces image with 50% black and 50% white pixels. The dithered image appears much brighter than the original. If, instead, the curve B in Figure 2 is the response curve for the display, then the dithered image is darker than the original. Images are normally linearized prior to dithering using calibration characteristics of the output display surface[7]. (Lookup tables are an efficient method for linearization.) Error diffusion using device linearization performs well on monochrome displays. However, for colour output devices, the issues of defining the error measurement metric and choosing the colour space for error propagation have to be resolved to provide a satisfactory error diffusion algorithm.

Error Diffusion and Colour Space

Extending error diffusion to cover a wide range of colour devices is very difficult. The characteristics of each individual device must be taken into account in determining how the algorithm should be modified.[8]

As mentioned above, two distinct colour mixing stages occur in colour halftoning. The first stage, called device colour mixture, is the generation of all the available colours from the primary colours of the output devices. For example, bi-level three colour printers can generate red, blue, green, black and white colours by appropriately mixing cyan, magenta, and yellow inks. The colorimetry of the resulting colours depends on the chemistry and physics of the device. Self-luminous displays, on the other hand, use a mixture method that is directly based on linear averaging.

The second colour mixing stage depends on spatial averaging in the human visual system, and is governed by the Grassman's law. The lights coming out from local regions of an image are mixing additively. As long as local spatial averages of light from the dithered image have the same tristimulus values as the continuous-tone original the two images are as perceptually identical.

Error Diffusion for Colour Images

It is very important to distinguish the two colour mixture stages when using any halftoning algorithm. Since all the halftoning algorithms, either for monochrome or colour images, are designed based on the concept of linearity of light mixture, care must be taken to solve some problems if the device colour mixture stage is incompatible with the concept. In this section, we will explain why the simple modification of error diffusion that works for the RGB images fails when it is used to dither CMY images.

Extension of Monochrome Error Diffusion to Colour Images

The simplest way to use error diffusion for colour images is to treat each primary image as a single monochrome image and process the primary images independently. This algorithm produces good results for additive display surfaces, as shown in the following analysis.

Let C be the average colour of a local region $\Re$ of an image, that is too small to be resolved, and let C_i be the colour of pixel i. The perceived colour C is :

$$C = \frac{1}{|\Re|} \sum_{i \in \Re} C_i \left(R_i, G_i, B_i \right),$$

Where the sum is understood to be additive colour mixture, which amounts to ordinary addition of tristimulus values. Because the device mixes colour additively, the colour at pixel i is the sum of the colours of the individual primaries at the pixel, $C_i = \sum_p C_{pi}$, where p identifies the primary. Then, the averaged colour is

$$\begin{aligned} C &= \frac{1}{|\Re|} \sum_{i \in \Re} C_{Ri} + \frac{1}{|\Re|} \sum_{i \in \Re} C_{Gi} + \frac{1}{|\Re|} \sum_{i \in \Re} C_{Bi} \\ &= C_R + C_G + C_B. \end{aligned}$$

When the halftoning algorithm is applied to each primary the result is a different set of pixel values, but the region averages remain the same. Thus, if C'_p is the colour of the halftoned image of primary p,

$$C'_p = \frac{1}{|\Re|} \sum_{i \in \Re} C'_{pi} = C_p,$$

which is true if the colours C_{pi} are chosen in accord with the principles of monochrome dithering. Using this result

$$C' = C'_R + C'_G + C'_B = C.$$

A final step relates the pixel colours C_{pi} to device coordinates. For additive devices the colour of primary p at pixel i depends only on the device coordinate D_{pi}. That is the colour of the red primary of a CRT image depends only on voltages produced in the red channel of the CRT input. The dependence is monotonic, and can therefore be both inverted and linearized. Thus, once the device coordinates are linearized error diffusion and colour mixture commute, explaining why error diffusion is satisfactory for linearized additive devices.

The same process error diffusion algorithm is not satisfactory when applied to subtractive display surfaces. To show this we applied error diffusion algorithm to each individual CMY colour plane of printed images. The result was printed on a dye-sublimation printer with non-overlapping output dots. This avoids darkening from dot overlap. To minimize dot gain, each pixel of the dithered images is printed as a 4x4 block. Several problems can be identified immediately. The images are brighter, less saturated and lack details. As will be explained in the sections followed, the problems are the results of the nonlinear of device response.

Nonlinear Response

The device coordinate $D_{pi}, p = C, M, Y$ of a three colour printer has a non-linear relationship to reflectance similar to curve A of Figure 2. For printers that support continuous tone, the D_{pi} control the amount of dye transferred at pixel i. Colour is produced as ink absorbs part of the incoming light and reflects the rest of it across the visible spectrum. To simplify the discussion, consider a printing media follows the Beer-Bouguer Law[9] and its paper white background has uniform spectral power distribution. The relationship between the reflectance and the absorption of ink can be expressed as:

$$R_\lambda = \exp\left(-\alpha_\lambda\right)$$

where α_λ is the absorption. According to Beer's law, the absorption of a mixture of dyes is the sum of the absorption of the individual dyes, so that

$$R_\lambda = \exp\left(k_C \alpha_{C\lambda} + k_M \alpha_{M\lambda} + k_Y \alpha_{Y\lambda}\right),$$

where k_p is the normalized concentration of primary p, and $\alpha_{p\lambda}$ is the corresponding normalized absorption. The concentrations, k_p, are controlled by the corresponding device coordinates, D_p, by transfer functions that are presumably monotonic. The printer overall printer response is nonlinear; the colour images generated by error diffusion are

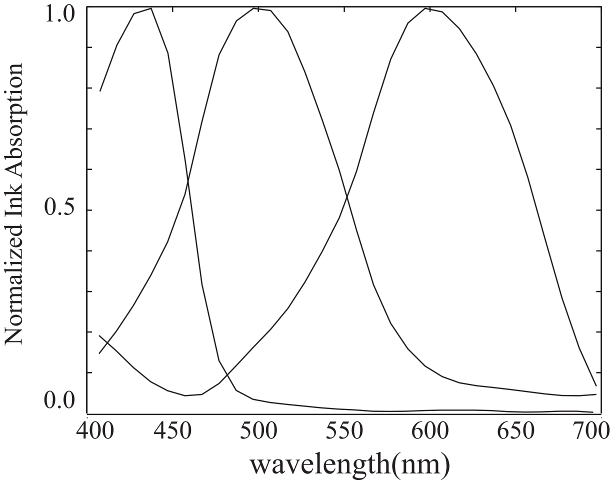

Figure 3: Normalized Ink Absorptions

always brighter than the original. This brightening effect is further complicated by the subtractive mixture of imperfect inks, which makes it impossible to correct the problem by linearizing the device coordinates.

To see why, consider a perfect set of subtractive primaries. Each primary has the property that its absorption is zero or a constant, α_p at every wavelength. Furthermore, no two elements of the set have non-zero absorption at the same wavelength. The set of wavelengths for which the absorption of primary p is non-zero is Λ_p. Ideal inks are a perfect set of subtractive primaries, with each ink absorbing in a contiguous set of wavelengths. For one of a set of perfect primaries the reflectance of the display surface, which is linearly related to the tristimulus values, can be written

$$\begin{aligned} R_{\lambda \in \Lambda_p} &= \exp(-k_p \alpha_p) \\ &= (\log \alpha_p)^{-k_p} \\ R_{\lambda \notin \Lambda_p} &= 1.0\,. \end{aligned}$$

Because the intersection of any pair of sets Λ_p is empty the reflectance depends on only a single primary at every wavelength. Thus, any linear function of reflectance, is a linear function of the values $(\log \alpha_p)^{-k_p}$. The device can be linearized by exponentiating the density of the ink, a long established practice. In such a case error diffusion in the exponentiated space would provide perfect results.

Unfortunately, the perfect primaries, like ideal inks, do not exist. The wavelength that should be completely absorbed are partially transmitted, and those should be completely transmitted are partially absorbed (Figure 3). These unwanted absorption and transmission have profound effects on printer response function. Without them subtractive devices can be linearized; with them such devices cannot be linearized.

Unwanted Absorptions

We now consider how imperfect primaries affect the error diffusion. The tristimulus values for a given mixture can be expressed as $Y_n = \int R_\lambda \Phi_\lambda \bar{y}_{n\lambda} \delta\lambda$, where Φ_λ is the illuminant and $\bar{y}_{n\lambda}$ is the colour matching function, for tristimulus value n. Any absorption can be represented as the sum of a perfect absorption, $\alpha^0_{i\lambda}$, and an unwanted absorption, $\alpha'_{i\lambda}$. Then the tristimulus values are

$$\begin{aligned} Y_n &= \int_\lambda \exp\left(-\sum_p k_p \left(\alpha^0_{p\lambda} + \alpha'_{p\lambda}\right)\right) \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda \\ &= \int_\lambda \exp\left(-\sum_p k_p \alpha^0_{p\lambda}\right) \exp\left(-\sum_p c_p \alpha'_{p\lambda}\right) \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda. \end{aligned}$$

As long as the imperfection is small the unwanted absorption can be written in the form

$$\exp\left(-\sum_p c_p \alpha'_{p\lambda}\right) = 1 - \sum_p c_p \alpha'_{p\lambda}$$

Then,

$$\begin{aligned} Y_n &= \int_\lambda \left(1 - \sum_p k_p \alpha'_{p\lambda}\right) \exp\left(-\sum_p k_p \alpha^0_{p\lambda}\right) \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda \\ &= \sum_p \int_{\Lambda_p} \left(1 - \sum_p k_p \alpha'_{p\lambda}\right) \exp\left(-\sum_p k_p \alpha^0_{p\lambda}\right) \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda \\ &= \sum_p \int_{\Lambda_p} \left(1 - \sum_p k_p \alpha'_{p\lambda}\right) \exp\left(-k_p \alpha^0_p\right) \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda \end{aligned}$$

Now let $M_{pn} = \int_{\Lambda_p} \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda$ be the tristimulus value of an uniform spectral in the absorption region p, and $N_{pqn} = \int_{\Lambda_q} \alpha'_{p\lambda} \Phi_\lambda \bar{y}_n(\lambda)\delta\lambda$ be the tristimulus value of the unwanted absorption of primary p in the absorption region of primary q. Then:

$$Y_n = \sum_p M_{pn} \exp\left(-k_p \alpha^0_p\right) - \sum_{pq} \exp\left(-k_p \alpha^0_p\right) N_{pqn} k_q.$$

If the primaries are perfect the second term is zero; the tristimulus values are a linear transformation of the exponentiated densities of the primaries. Given perfect primaries and a linearized display surface, tristimulus values are preserved in error diffusion. When the inks are close to perfect, the second term provides an estimate of the effect of unwanted absorption. If it is sufficiently small the perfect primary approximation is satisfactory. Otherwise, the tristimulus values are not preserved in error diffusion due to the nonlinearity, and the achromatic information are distorted.

Error Diffusion and Device Response Function

The above discussion can be given a fully general mathematical form. The ultimate goal of error propagation is to preserve spatial averages of colour coordinates that are averaged linearly by the human visual system. For printed images reflectance is the most useful, so suppose that errors in reflectance are propagated when creating the halftoned image. For a set of control values D_p, in terms of which the image to be halftoned is specified, there is a function g mapping the control values into the displayed reflectance R. In the dithering operation we set the device coordinates at pixel i to $\hat{D}_{ip}$. This produces a reflectance error equal to

$$\Delta R_i = g(D_{ip}) - g(\hat{D}_{ip}).$$

The amount of this difference that is moved to pixel j depends on the distribution matrix w_{ji}. Now let's look at pixel j. The reflectance requireded by the image specification is $R_j = g(D_{jp})$. To it should be added reflectance error distributed from other pixels, $\sum_i w_{ji} \Delta R_i$. To compute the appropriate device values it is necessary to calculate

$$\begin{aligned} D'_{jp} &= g^{-1}(g(D_{jp} + \sum_i w_{ji} \Delta R_i)) \\ &= g^{-1}(g(D_{jp} + \sum_i w_{ji}(g(D_{ip}) - g(\hat{D}_{ip}))\,))\,. \end{aligned}$$

When g is linear the above equation simplifies to

$$D'_{jp} = D_{jp} + \sum_i w_{ji}(D_{ip} - \hat{D}_{ip})\,.$$

This a general form of the result that spaces which are linear in reflectance allow error propagation in the space of device coordinates. The device coordinates of devices that produce colour additively, such as linearized CRTs, are examples. We showed above that subtractive display surfaces with perfect primaries can also be linearized.

For devices that do not produce colour additively g is the device characterization function. It is necessary both to apply and invert this characterization function for every pixel that is set. This operation is expensive, further emphasizing the benefit that is obtained by successful linearization of the device.

Discussion and Conclusion

Error diffusion can only be performed in a space that is linearly related to the colour space in which the spatial averages are performed in the observer. There is no simple error diffusion algorithm for non-linear colour spaces. To ensure the image details are preserved, the colour spaces that have linear relationship with the achromatic signal are suitable. With these criterion colour space such as linearized RGB or CIE tristimulus values, can be used to diffuse error, while the CIE L*a*b*, CIE L*u*v* should be avoided.

L* a* b* / **XYZ**	WHITE	YELLOW	MAGENTA	RED	CYAN	GREEN	BLUE	BLACK
WHITE	190	0	0	0	0	0	0	0
YELLOW	249	410	0	3	0	0	0	0
MAGENTA	688	129	600	151	48	249	27	108
RED	44	100	171	699	0	21	0	45
CYAN	151	0	2	0	581	8	36	1
GREEN	36	1	147	35	127	927	0	822
BLUE	0	0	38	0	369	0	646	398
BLACK	0	0	1	10	0	8	21	964

*Table 1. Pixel colour assignment in the CIE XYZ vs. L*a*b* colour*

Unfortunately, this consideration runs counter to what is needed when thresholding pixels for pixel assignment. For this operation it is best to use a colour space that provides a good perceptual distance measurement. Uniform colour spaces like CIE L*a*b* and CIE L*u*v* are therefore best for pixel assignment because they minimize the local difference between the halftone pixel and the continuous-tone image, and thereby minimize the error to be propagated in perceptual terms. Unfortunately, the uniform colour spaces are always nonlinear. That means two colour spaces have to be used for the dithering process and the operation costs are subsequently increased.

The quality improvement of using the perceptual uniform space for pixel assignment depends on the content of the image. Some pixels are assigned to the same colour regardless of what colour space is used for pixel assignment. But the differences from one colour space to another are greater than might be expected. To make Table 1 we chose 9261 colours, uniformly spaced in CMY space derived from a printer model,[10] and did pixel assignment for each colour in two different colour spaces, L*a*b* and XYZ, the former presumably better, but more expensive. As Table 1 shows, only about half of these colours are given the same pixel assignment by the two colour spaces. Thus, it is essential that further study of colour halftoning be performed to determine when the image quality improvement is worth the increased cost.

In summary, we have studied the problems of applying error diffusion on CMY images. Performing error diffusion in a nonlinear colour space imperfectly reproduces both tone levels and achromatic information in our test images. The halftoned images are sometimes brighter, sometimes darker. In both cases achromatic detail is lost. Therefore, device colour spaces of devices that use subtractive colour mixing should be avoided for error diffusion because of unavoidable non-linearities from unwanted ink absorption. As shown in the last section, using a printer model and the complicated operations on ink density space can solve the problems. However, a linear additive colour space is a better choice for error diffusion.

References

1. R. Floyd and L. Steinberg, ”An adaptive algorithm for spatial grey scale,” Proc. Soc. Info. Display, **Vol. 17**, p.75 (1976).
2. J. F. Jarvis, C. N. Judice, and W. H. Ninke, ”A survey of techniques for the of continuous-tone pictures on bilevel displays,” *Comp. Graph. Im. Proc.*, **5**, p.13-40 (1976).
3. B. E. Bayer, ”An optimum method for two-level rendition of continuous-tone pictures,” *IEEE International Conference on Communications*, **Vol. 1** p.26-11 to 26-15 (1973).
4. R. Ulichney, *Digital Halftoning*, The MIT Press, Cambridge, MA. 1987.
5. G. Marcu, S. Abe, ”An error diffusion method for color reproduction in ink jet printering,” *Proc. IS&T’s Tenth International Congress on Advance in Non-Impact Printing Technologies*, Oct.30-Nov.4, New Orleans, Louisiana (1994).
6. G. Wiszecky, W. Stiles, *Color Science: Concepts, Methods, Quantitative Data and Formulae*, John Wiley & Sons, 1982.
7. C. J. Rosenberg, ”Measurement Based Verification of An Electrophotographics Printer Dot Model For Halftone Algorithm Tone Correction,” *Proc. IS&T’s Eighth International Congress on Advances in Non-Impact Printing Technologies*, Oct.25-30, 1992, Williamsburg, Virginia.
8. T. N. Pappas and D. L. Neuhoff, ”Model-Based Halftoning,” *SPIE/IS&T Symposium on Electronic Imaging Science and Technology*, San Jose, California, Feb. 1991. p.244-255.
9. F. Grum and C. J. Bartleson, Eds., *Optical Radiation Measurements*, Academic, New York, 1980, Vol. 2.
10. R. S. Berns, ”Spectral modeling of a dye diffusion thermal transfer printer,” *Journal of Electronic Imaging*, **2**: p.359-370 (1993).