Model-Based Halftoning of Color Images

Thrasyvoulos N. Pappas, *Senior Member, IEEE*

***Abstract*—We present a new class of models for color printers. They form the basis for model-based techniques that exploit the characteristics of the printer and the human visual system to maximize the quality of the printed images. We present two model-based techniques, the modified error diffusion (MED) algorithm and the least-squares model-based (LSMB) algorithm. Both techniques are extensions of the gray-scale model-based techniques and produce images with high spatial resolution and visually pleasant textures. We also examine the use of printer models for designing blue-noise screens. The printer models can account for a variety of printer characteristics. We propose a specific printer model that accounts for overlap between neighboring dots of ink and the spectral absorption properties of the inks. We show that when we assume a simple "one-minus-RGB" relationship between the red, green, and blue image specification and the corresponding cyan, magenta, and yellow inks, the algorithms are separable. Otherwise, the algorithms are not separable and the modified error diffusion may be unstable. The experimental results consider the separable algorithms that produce high-quality images for applications where the exact colorimetric reproduction of color is not necessary. They are computationally simple and robust to errors in color registration, but the colors are device dependent.**

I. Introduction

DIGITAL halftoning is the process of generating a pattern of pixels with a limited number of colors that creates the illusion of a continuous-tone image. Digital halftoning is necessary to display continuous-tone images in media in which the direct rendition of the tones is impossible. The most common example of such media is paper. Another example is flat panel displays, which use both spatial and temporal halftoning.

Model-based halftoning techniques exploit the properties of the display device and the human visual system to maximize the quality of the displayed images. Thus, as illustrated in Fig. 1, they rely on accurate models of the display and human perception. In [1]–[3], Pappas and Neuhoff presented model-based techniques for binary displays and specifically for black and white (B&W) laser printers. This paper extends these models and techniques to color displays and specifically to color printers.

In recent years, several different new color printers have become available in the market. A variety of technologies exist: electrographic, ink jet, thermal, etc. Most of these color printers produce halftone images. They cover a wide range, from low-resolution personal printers for home and office use to high-resolution phototypesetters used for high-quality and high-volume printing.

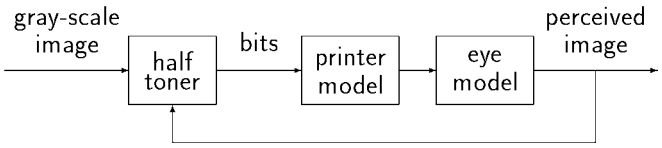


Fig. 1. Model-based halftoning.

Color printers exhibit many of the characteristics of B&W printers. For example, they tend to produce dots of ink that are roughly circular and overlap adjacent dots or spaces that are intended to remain white. We refer to this phenomenon as *dot overlap*. We use the word *dot* to refer to the smallest spot of ink that the printer is capable of producing, usually on a Cartesian grid whose spacing is determined by the printer resolution. However, some color printer characteristics are unique to color. Color printers use ink of at least three different colors, typically cyan, magenta, and yellow, while B&W printers use only one type of ink. Thus, we can have overlap between dots of the same or different colors. Moreover, the spectral absorption characteristics of the inks can vary significantly from printer to printer. This further complicates the process of accurate color reproduction. Finally, inexact registration of the different colors can create a number of distortions of the printed images (Moire patterns, color shifts, etc.).[1]

Conventional halftoning techniques resist printer distortions by printing colored dots in clusters or *macrodots*. One such technique is classical screening [4]. The image intensity is represented by the size of the macrodots, while their spacing is fixed. This mimics traditional analog halftoning techniques used in printing. The main advantage of the classical screen in B&W printing is its robustness to printer distortions; however, this is achieved at the expense of spatial and gray-scale resolution. In color printing, screens of three or four different colors are superimposed. This creates a number of problems including Moire patterns, color shifts, and loss of spatial resolution [5], [6]. The Moire patterns can be minimized by rotating the screens relative to each other, but this further constrains the spatial resolution and increases the visibility of the macrodots. In contrast, the model-based techniques can exploit both dot overlap and the spectral absorption characteristics of the inks to produce printed images with high spatial and color-scale resolution and visually pleasant

Manuscript received April 17, 1996; revised January 28, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark D. Fairchild. Portions of this paper were presented at IS&T's 8th International Congress on Advanced Non-Impact Printing Technology, Williamsburg, VA, October 1992; and at *ICASSP'93*, Minneapolis, MN, April 1993.

The author is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: pappas@bell-labs.com).

Publisher Item Identifier S 1057-7149(97)04722-2.

[1] Usually, the different colors are printed on separate passes through the page.

textures. As in the gray-scale case, the basis of model-based techniques is an accurate printer model.

In this paper, we present a new class of models for color printers. The task of developing color printer models is considerably more complicated than that of developing B&W models. In principle, the number of parameters to be estimated increases exponentially with the number of inks. For this reason, we propose a simple model that captures the most important printer distortions. As in the B&W case [1], [2], the color model assumes that the printed dots are approximately circular and possibly larger than the minimal size required to completely cover the page. The model accounts for overlap between neighboring dots of the same and different colors. In addition, the proposed printer model accounts for the spectral absorption characteristics of the inks.

We present two model-based techniques for color halftoning, the *modified error diffusion (MED)* algorithm and the *least-squares model-based (LSMB)* algorithm. These techniques were initially presented in [7] and [8]. Both are extensions of the gray-scale model-based techniques, and produce images with high spatial resolution and visually pleasant textures. The B&W MED algorithm [1], [2] uses the printer model to account for printer distortions. Experiments show that it provides high quality reproductions with reasonable complexity. The B&W LSMB algorithm [3] uses models for visual perception and printing to produce an optimal halftone reproduction, in the following sense: It minimizes the squared error between the perceived intensity of the continuous-tone original image and the perceived intensity of the printed halftone image. Similar least-squares algorithms, for different display media, have been considered in [9] and [10]. The least-squares algorithm is iterative and requires significantly more computation than error diffusion.

We show that when the original images are specified in the red–green–blue (RGB) domain, and we assume a simple "one-minus-RGB" relationship between the cyan, magenta, and yellow inks and the RGB image values [11, pp. 152–153], then the printer models, and hence the algorithms, are *separable*. That is, there is no coupling between the three color components, and the algorithms can be applied to each color component independently. When this relationship does not hold, the algorithms are not separable and we show that the modified error diffusion may be unstable, that is, the quantization error may not be bounded.

Finally, we also examine the use of printer models in color screening. In particular, we examine blue-noise screening techniques, which attempt to approximate the performance of error diffusion with minimal computation [12], [13]. However, their performance cannot match that of the model-based techniques because of the constraints of the screening procedure. The design of blue-noise screens can also make use of models of visual perception and printing [14]. However, color screening can only be applied to each color independently.

Model-based techniques for color halftoning were presented by several authors. Miller and Sullivan [15] used error diffusion with a visual model to exploit the different sensitivity of the human visual system to the luminance and the chromatic dimensions. Mulligan and Ahumada in [16] and Flohr *et al.* in [17] proposed least-squares techniques that also exploit the different sensitivity of the human visual system to the luminance and the chromatic dimensions. Kolpatzik and Bouman [18] used visual models to optimize error diffusion filters for monochrome and color image displays. Kim *et al.* [19] adopted our color printer model (which was presented in [7] and [8]) to implement a variation of error diffusion whereby the quantization occurs in a perceptually uniform color space. Haneishi *et al*. [20] proposed an error diffusion algorithm that compensates for the ink characteristics but assumed no dot overlap. Finally, Rodriguez [21], [22] considered the application of various digital halftoning techniques to high-resolution printers used in the commercial graphic arts industry.

In the experimental results of this paper we assume that the "one-minus-RGB" relationship holds, and the algorithms are separable. Even though this assumption is not realistic, the separable model-based algorithms produce high-quality halftone images for applications where the exact colorimetric reproduction of color is not necessary. As in the B&W case, they produce images that are sharper, have better texture, and have richer and better color tones than both the classical and blue-noise screening techniques. In addition, due to the fact that we have eight different colors instead of two, the few artifacts and asymmetries that the B&W error diffusion algorithm is known to produce are a lot less visible. In fact, a 400 dpi electrographic printer produces images that look like continuous-tone images. The separable algorithms offer the advantages of computational simplicity and robustness to errors in color-plane registration. The main drawback of the separable algorithms is the fact that the colors of the resulting images are device dependent. The colors of each printed image look fine when viewed independently, but a comparison between images printed on different devices shows significant color differences.

The nonseparable algorithms offer the possibility of accurate color reproduction and maybe device independent color. They require an accurate device for estimating the ink parameters, and more importantly, accurate knowledge of the characteristics of the image acquisition device, in order to reproduce the colors of the original image. Also, one would expect them to will be more sensitive to color registration. A detailed study of the nonseparable algorithms is beyond the scope of this paper.

The new color printer models are presented Section II. Section III examines color screening techniques. Section IV presents the color modified error diffusion algorithm. Section V presents the color least-squares model-based algorithm.

II. Color Printer Models

In this section, we present models for color laser printers. The color printer models are extensions of the gray-scale printer models that were described in [1] and [2]. Our test vehicles included a CANON CLC300 printer and a Tektronix Phaser 220i printer. The first is an electrographic 400 dpi printer with four colors, and the second is a thermal 300 dpi printer with three colors. However, the models and the techniques that rely on them are intended to apply to any color printer, as well as other display devices.

To a first approximation, color printers are capable of producing colored dots on a piece of paper, at any and all sites of a Cartesian grid with horizontal and vertical spacing of T inches. The reciprocal of T is generally referred to as the *printer resolution* in dots per inch (dpi).[2] A typical color printer uses three different types of ink to produce colored dots: cyan (C), magenta (M), and yellow (Y). These colors form the basis of the subtractive systems for color reproduction [23]. We refer to them as the *primary* colors for printing. The relationship to the additive systems, which are based on the mixing of red (R), green (G), and blue (B) light, is very simple. Cyan ink absorbs red light, magenta absorbs green, and yellow absorbs blue. Dots of different types of ink can be printed on top of each other to produce dots of the *secondary* colors: red, green, blue, and black.

The spectral characteristics of the inks vary from printer to printer. In the ideal case, each ink absorbs 100% of the light in a different block of wavelengths (e.g. 400 to 490 nm for the yellow ink, 490 to 580 nm for magenta, and 580 to 700 nm for cyan) and transmits 100% of the light in the remaining portion of the visible spectrum [23, pp. 135–136]. These are often referred to as *block* inks. In practice, the inks are quite different from the ideal. They absorb some light outside the specified block (unwanted absorptions), they do not absorb 100% percent of the light inside the block, and the transitions are not sharp. Due to absorption deficiencies of the inks, many printers use a separate black ink (K) to produce better black dots. Without loss of generality, in the remainder of this section we will assume that the printer uses all four types of ink.

The printer is controlled by an $N_W \times N_H$ array of vectors with binary components

$$\boldsymbol{b}_{i,j} = (b_{i,j}^C, b_{i,j}^M, b_{i,j}^Y, b_{i,j}^K) \quad 1 \le i \le N_W, \quad 1 \le j \le N_H \tag{1}$$

where $b_{i,j}^C = 1$ indicates that a cyan dot is to be placed at site (i,j) and $b_{i,j}^C = 0$ indicates that no cyan dot is to be placed at the site. The magenta $b_{i,j}^M$, yellow $b_{i,j}^Y$, and black $b_{i,j}^K$ components are defined similarly. When all components are zero, the site is to remain white. When more than one component is equal to one, different inks are printed on top of each other to produce red, green, blue, or black dots. In principle, we can specify $2^4 = 16$ different colors for each dot, but nine of these colors are variations of black. Thus, each dot can take one of $2^3 = 8$ distinct colors. Sometimes, for various reasons, there are restrictions in the use of the black ink. For example, it could be used only in combination with the other three inks to produce solid black dots, or it could be used instead of the other three inks in order to reduce ink consumption (under-color removal) [23, p. 588]. In some cases, there may be complicated interactions between the black ink and any one of the other inks.

The development of color printer models follows closely that of B&W models. As we saw in [1] and [2], printers tend to produce circular rather than square dots. The same is true for color printers, as illustrated in Fig. 2. The radius of the dots must be at least $T/\sqrt{2}$ so that they are capable of covering a page entirely. This means that the colored dots cover portions of adjacent spaces, thus increasing the amount of light absorbed by the ink. As we saw earlier, we call this *dot overlap*.[3] For many printers, the radius of the dots is larger than the minimal size. Dot overlap is the most elementary characteristic of many printers. A variety of other phenomena affect the color levels produced by a printer. For example, internal light scattering reduces the amount of reflected light. In addition to the phenomena encountered in B&W printing, in color printing one has to account for the spectral characteristics of the inks. As a result, the color level produced by the printer in the vicinity of site (i,j) depends in some complicated way on $\boldsymbol{b}_{i,j}$ and neighboring dots. Due to the close spacing of dots and the limited spatial resolution of the eye, the color level can be modeled as having a constant value $\boldsymbol{p}_{i,j}$ within the area of the square pixel at site (i,j), equal to the average color of the pixel. The color printer model takes the form

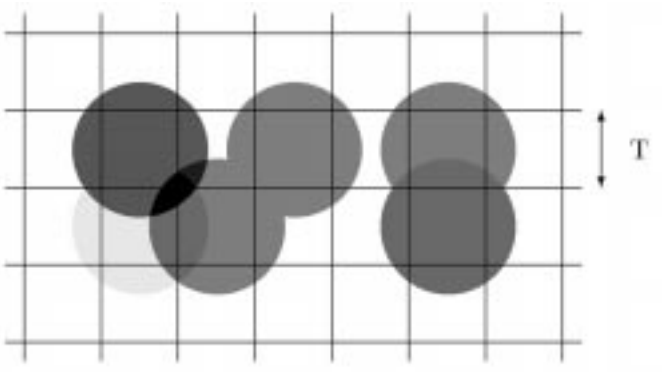


Fig. 2. Color-dot overlap.

$$\boldsymbol{p}_{i,j} = (p_{i,j}^1, p_{i,j}^2, p_{i,j}^3) = \boldsymbol{P}(\boldsymbol{W}_{i,j}) \quad 1 \le i \le N_W, \quad 1 \le j \le N_H \tag{2}$$

where $\boldsymbol{W}_{i,j}$ consists of $\boldsymbol{b}_{i,j}$ and its neighbors, and $\boldsymbol{P}(\cdot)$ is some function thereof. This function could be deterministic or probabilistic. The value of $\boldsymbol{p}_{i,j}$ can be specified in any color space (an RGB space, CIE XYZ, CIE $L^*a^*b^*$, etc. [11]).[4] Thus, as illustrated in Fig. 3, our model generates an array $[\boldsymbol{p}_{i,j}]$ of color levels that has the same dimensions as the array $[\boldsymbol{b}_{i,j}]$ that specifies the dot pattern to be printed.

For the methods we study here, it is essential that the value of $\boldsymbol{p}_{i,j}$ be determined by the dots in a finite window around $\boldsymbol{b}_{i,j}$. In principle, the function $\boldsymbol{P}$ can be specified in a table. The number of elements in the table, however, is 8^9 even for a 3×3 window. This makes it practically impossible to derive the individual elements of this table from measurements of the color of various printed dot patterns, as was done in [24] and [25] for B&W printers. Thus, a simpler model must be derived, based on a physical understanding of the printing mechanism. An example of such a model is described below.

[2] This printer resolution should not be confused with the frequency of classical halftoning screens which is also expressed in dots per inch; using our terminology it should be expressed in macrodots per inch.

[3] The term *dot overlap* should not be confused with the the term *dot gain*, which is used to denote the darkening of the halftone images due to a combination of factors, mainly dot-overlap (mechanical dot gain) and internal light scattering (optical dot gain) [6].

[4] CIE is the Commission Internationale d'Eclairage.

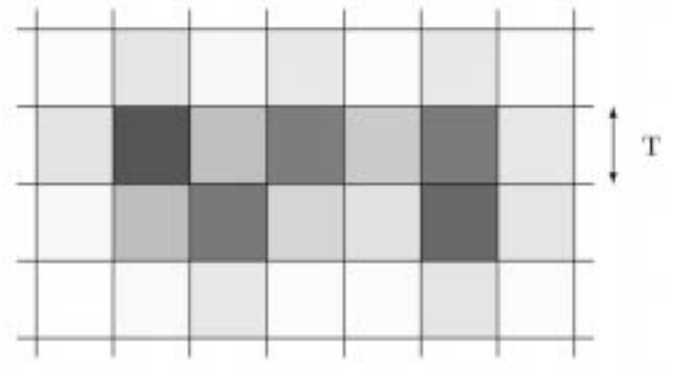


Fig. 3. Printer model output.

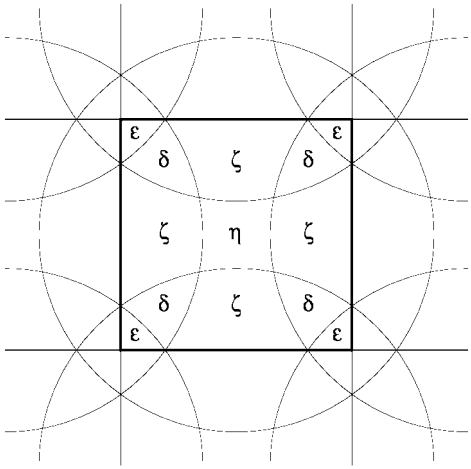


Fig. 4. Overlapping segments of color.

A. Circular-Dot-Overlap Color Printer Model

In [1] and [2] we developed a *circular dot-overlap* model for B&W printers that accounts for the dot-overlap characteristic like that illustrated in Fig. 2. In the B&W case, all the dots have the same color. In the color case, the overlapping segments can take different colors depending on the color of the neighboring dots. Fig. 4 shows the segments of different colors that we can get within a pixel. The color of each segment depends on the absorption properties of the overlapping dots of ink and can be specified using RGB or CIE XYZ tristimulus values (linear light). Assuming that the paper saturates with each type of ink, the color of each segment depends on which types of ink overlap the segment. A table can provide the color for each of the 16 different ink combinations. In the "one-minus-RGB" case we will examine below, the elements of the table are trivial to obtain. Otherwise, the elements of the table can be obtained by measuring the color of the different ink combinations using a spectrophotometer. Tables I and II provide the CIE XYZ tristimulus values for the two test printers.

Once the color of each segment is determined, the average color $\boldsymbol{p}_{ij}$ of the pixel (i,j) can be obtained as a weighted sum of the colors of the different segments with the weight being the area of the segment, provided that the color of each segment is specified in linear-light coordinates. This is based on the additive color mixing theory [23, Ch.7] on which the Neugebauer equations [23, p. 535–538] are based. The average value of the pixel can then be converted to any other color system. Thus, the printer model $\boldsymbol{P}$ is completely specified. The area of each segment is easy to calculate in terms of the parameters α, β, and γ, used to specify the B&W circular dot-overlap model

$$\epsilon = \beta, \quad \delta = \gamma - \beta, \quad \zeta = \alpha - 2\gamma, \quad \eta = 1 - 4\alpha + 4\gamma. \tag{3}$$

The parameters α, β, and γ are the ratios of the areas of the shaded regions shown in Fig. 5 to T^2, and can be expressed in terms of the ratio ρ of the actual dot radius to the minimal dot radius $T/\sqrt{2}$ [2]. For the two test printers, we found that the dot size is close to the minimal, i.e. $\rho = 1$, and $\alpha = .143, \beta = 0$, and $\gamma = 0$. However, the proposed model is valid for a wide range of values of the parameter ρ. (The above equations assume $1 \leq \rho \leq \sqrt{2}$, but similar formulas have been

TABLE I
CIE XYZ TRISTIMULUS VALUES FOR DIFFERENT INK COMBINATIONS OF THE CANON PRINTER

$CMYK$ specification	X	Y	Z
(0,0,0,0) - WHITE	84.70	88.70	98.62
(1,0,0,0) - CYAN	17.04	23.29	62.09
(0,1,0,0) - MAGENTA	36.37	19.96	39.13
(0,0,1,0) - YELLOW	65.66	77.19	11.70
(0,1,1,0) - RED	26.38	15.20	4.51
(1,0,1,0) - GREEN	6.02	15.62	8.50
(1,1,0,0) - BLUE	6.42	4.03	22.88
(1,1,1,0) - BLACK (CMY)	2.82	3.14	4.61
(0,0,0,1) - BLACK (K)	6.60	6.87	7.33
(1,0,0,1) - BLACK (CK)	2.39	2.76	4.27
(0,1,0,1) - BLACK (MK)	3.59	2.95	4.02
(0,0,1,1) - BLACK (YK)	5.08	5.64	2.97
(0,1,1,1) - BLACK (MYK)	2.85	2.49	2.22
(1,0,1,1) - BLACK (CYK)	2.12	2.79	2.51
(1,1,0,1) - BLACK (CMK)	1.83	1.76	3.15
(1,1,1,1) - BLACK ($CMYK$)	1.83	1.96	2.24

TABLE II
CIE XYZ TRISTIMULUS VALUES FOR DIFFERENT INK COMBINATIONS OF THE TEKTRONIX PRINTER

CMY specification	X	Y	Z
(0,0,0) - WHITE	83.20	86.76	102.16
(1,0,0) - CYAN	14.41	18.37	63.20
(0,1,0) - MAGENTA	31.32	16.74	22.81
(0,0,1) - YELLOW	64.06	74.89	9.79
(0,1,1) - RED	27.57	15.85	3.35
(1,0,1) - GREEN	5.21	14.10	8.41
(1,1,0) - BLUE	4.76	3.27	18.15
(1,1,1) - BLACK	1.91	2.42	3.14

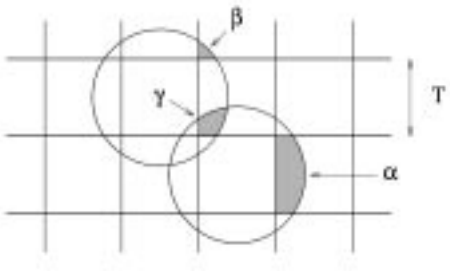


Fig. 5. Definition of α, β, and γ for printer model.

derived for larger values of ρ.) For illustrative purposes, in the examples in the following sections we will consider the value $\rho = 1.25$.

The circular dot-overlap model is a simplification of printer behavior and provides a good first-order approximation for the behavior of many printers. However, there are many other phenomena that it does not account for. As we disussed above, more elaborate models of the form of (2) could account for such phenomena. The advantage of such models is that they can be used with the model-based techniques we will consider in the following sections to optimize the quality of the printed image on a pixel-by-pixel basis. This is in contrast to traditional techniques (e.g. Neugebauer, Murray–Davies, and other equations [6]), whereby various effects are accounted for on an aggregate basis.

The "One-Minus-RGB" Case

When the original images are specified in the RGB domain, one can assume a simple "one-minus-RGB" relationship between the cyan, magenta, and yellow inks and the RGB image values [11, pp. 152–153], as follows:

$$c = 1 - r$$
$$m = 1 - g$$
$$y = 1 - b.$$

In this case, the color of each segment in Fig. 4 can be determined (in RGB coordinates) independently for each of the color components. This is because cyan ink affects only the red component, magenta affects only the green, and yellow affects only the blue. The black is not necessary in this case, since a combination of the other three inks produces a perfect black. Thus, the average color $\boldsymbol{p}_{i,j}$ for the site (i,j) can be determined independently for each color component, and the printer model of (2) becomes *separable*

$$\begin{aligned}\boldsymbol{p}_{i,j} &= (p^R_{i,j}, p^G_{i,j}, p^B_{i,j}) \\ &= (\mathcal{P}^R(W^C_{i,j}), \mathcal{P}^G(W^M_{i,j}), \mathcal{P}^B(W^Y_{i,j}))\end{aligned} \tag{4}$$

where each of the RGB components is specified by a circular-dot-overlap model that is identical to the gray-scale model [1], [2]. As we will see in the following sections, the "one-minus-RGB" assumption is reasonable in many cases and leads to significant computational and performance advantages.

However, there are a number of problems with the "one-minus-RGB" assumption [11, pp. 153–154]. In particular, the relationships ignore the specific spectral characteristics of the inks and the RGB primaries. Even in the ideal case of block inks, the "one-minus-RGB" assumption does not hold. The actual inks are far from the ideal. Fig. 6 shows the CIE (x,y) chromaticity diagram with the inks of the two test printers and the secondary colors formed with their combinations. Also shown are the primaries of the Rec. 709 RGB video standard [11. p. 135]; the chromaticities that they can reproduce are inside the solid triangle. The figure also shows the primary and secondary colors of the block inks we discussed earlier in this section.

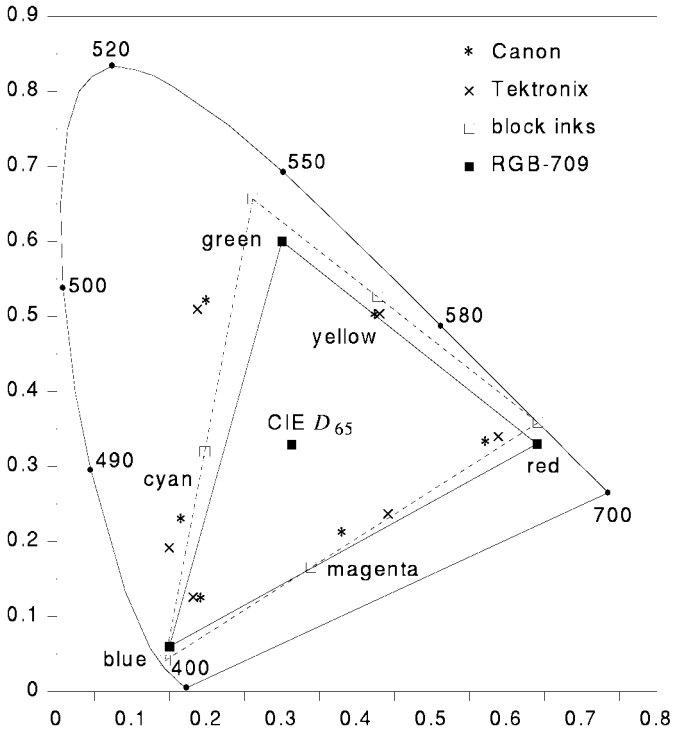


Fig. 6. Inks of the two test printers, the ideal block inks, and the primaries of the Rec. 709 RGB video standard plotted on the CIE (x,y) chromaticity diagram. Also plotted are the two-ink combinations.

In summary, we described a general class of models for laser printers (which, in fact, can be applied to any color display device) as well as a specific model, the circular dot-overlap model, which accounts for the effects of overlap between neighboring dots of the same and different colors and for the absorption characteristics of the inks. We now examine model-based techniques that rely on these color printer models.

III. Screening

In screening, the binary image is generated by comparing each pixel of the continuous-tone image to an array of image-independent thresholds [4]. A dot of ink is placed when the value of the image is greater than the corresponding threshold. The threshold array is usually periodic (ordered dither). The thresholds can be arranged to produce various types of patterns. In clustered dither they produce dots that are clustered together to form bigger dots, which we call *macrodots*. Alternatively, the thresholds can be arranged so that they produce dispersed dots. The main advantage of

screening techniques is the fact that the required amount of computation is minimal and can be carried out in parallel.

The classical screen has been the most popular for printing because it is robust to printer distortions. This is at the expense of limited spatial resolution and visible halftone textures. In classical screening the term *dot gain* is used to denote the darkening of the halftone image due to a combination of factors, mainly dot-overlap (mechanical dot gain) and internal light scattering (optical dot gain) [6]. Dispersed-dot screening techniques produce images with better spatial resolution and better textures, but are more sensitive to printer distortions. An example of such dispersed-dot techniques is blue-noise screening (BNS), which attempts to approximate the performance of error diffusion at much faster execution times. Two approaches to obtaining blue-noise screens are the power-spectrum matching technique [12] and the void-and-cluster method [13].

Color screening is applied to each color independently. Thus, screens of three or four different colors are superimposed. When the screens are periodic (e.g. the classical screen), misregistration of the screens creates a number of problems that are not present in the B&W case. For example, due to the spectral absorption properties of the inks, misregistration of the screens results in significant color shifts. Angular misregistration results in Moire patterns. Misregistration also causes loss of spatial resolution (sharpness). A more detailed discussion of such phenomena can be found in [5] and [6]. The usual way to minimize Moire patterns and color shifts is to rotate the screens relative to each other, but this further constrains the spatial resolution and increases the visibility of the halftone textures. These problems can be virtually eliminated by using error diffusion, which is aperiodic, or blue-noise screens, which are virtually aperiodic (they have very large periods).

In color halftoning, the term *classical screening* usually refers to rotated screens, while the term *dot-on-dot screening* is used when the same screen is used for all color components. In three-color printing, the most common orientations for the cyan, magenta, and yellow components are 15°, 4°5, and 75°. In four-color printing, the usual orientations for the cyan, magenta, yellow, and black components are 15°, 75°, 0°, and 45°. Both cases result in classical rosette patterns [5].

When the screen for each color is applied independently, one can only account for overlap between dots of ink of the same type. This can be done using the B&W techniques for each screen. For example, Roetling and Holladay [26] used a circular dot-overlap model to design classical screens, and Schulze and Pappas [14] used a similar model to design blue-noise screens using the void-and-cluster method. Then, the color depends on the absorption properties of the inks. Since there is no way to order the three-dimensional (3-D) color space, it is impossible to design a 3-D screen. The screens for each color component have to be independent. In traditional screening, the ink characteristics and the dot gain, as well as other printing conditions, can be accounted for using the Neugebauer and other equations to produce the color separations for a given printer and paper type [6, Ch. 10].

We should also point out that because of the absorption properties of the inks, the interactions between the overlapping screens affect the spatial frequency characteristics of the resulting patterns. These interactions are a lot more noticeable when a separate black ink is used $(CMYK)$. In such a case, the component screens must be designed jointly to guarantee the blue-noise characteristic. For example, Yao and Parker propose a technique that minimizes the low-frequency noise that results from the interactions between the overlapping screens [27].

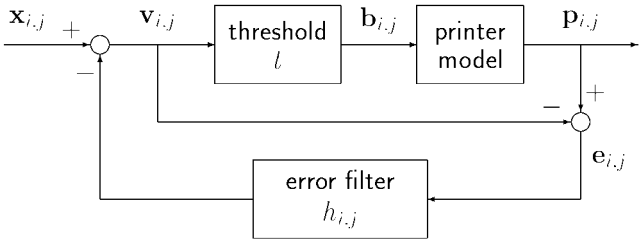


Fig. 7. Color modified error diffusion algorithm.

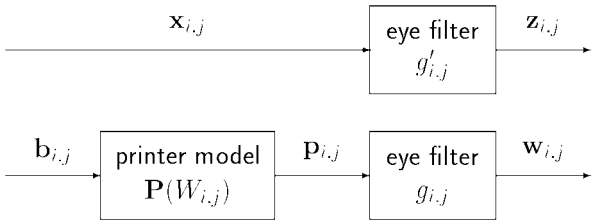
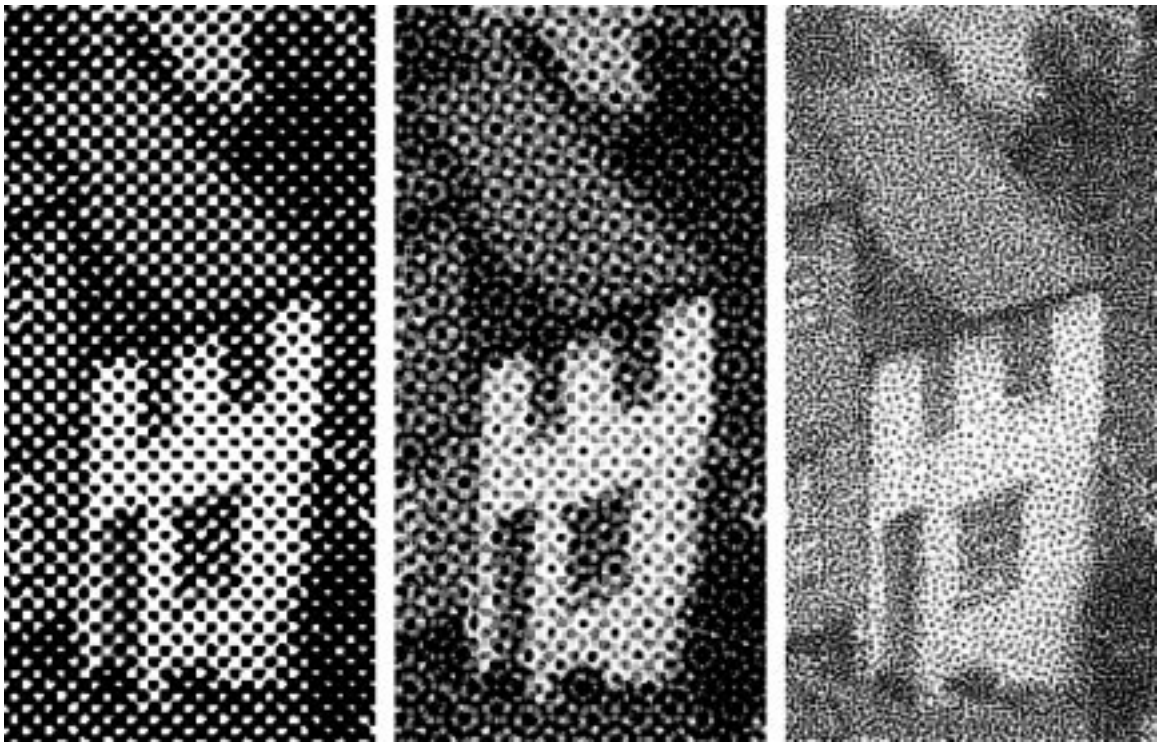


Fig. 8. Color least-squares model-based halftoning.

Fig. 9 shows a magnified detail of an image halftoned with different screening techniques. The resolution of the image is 256 × 128 pixels. Figs. 9–11 simulate a 400 dpi color printer with dot-gain $\rho = 1.25$ and is magnified by a factor of six.[5] The resolution of the magnified image is 66.7 dpi. In Fig. 9(a) the same 45-degree classical screen was used for all color components. As we saw above, this is usually called dot-on-dot screening. In Fig. 9(b), the classical screens for the cyan, magenta, and yellow components were at 15°, 45°, and 75° orientations that produce the classical rosette pattern [5]. No dot-overlap compensation was used with any of these classical screens. This figure demonstrates that the classical screens are still fairly robust to dot-overlap distortions. In Fig. 9(c), we used the model-based blue-noise screens designed by Schulze and Pappas [14] based on the void-and-cluster method [13]. In our simulations, we assumed perfect inks and used uncorrelated screens for the cyan, magenta, and yellow components. The black ink was used only when the other three components were equal to one. This choice results in the best image quality for the electrographic printer we used. Observe that, like in the B&W case, blue-noise screening results in images that are sharper and have better texture. Moreover, as we discussed above, blue-noise screens are fairly robust to shifts and rotations of the three color components.

[5] Alternatively, it simulates a 300 dpi printer with the same dot-gain, magnified by a factor of 4.5.



(a) (b) (c)

Fig. 9. Screening techniques (detail of 400 dpi printout, magnified by six). (a) Dot-on-dot screening. (b) Rotated classical screens. (c) Model-based blue-noise screening.

IV. Modified Error Diffusion

Error diffusion [4], [28] is a popular method for generating halftone images for displays, such as some CRT's that do not suffer from substantial dot overlap or other distortions. The standard color error diffusion algorithm is applied to each color component independently. It offers the same advantages as the B&W algorithm, namely sharp images with visually pleasant textures. Of course, it is equally sensitive to printer distortions as the B&W algorithm. This problem can be corrected by using the color printer models we presented in Section II. In the B&W case, Stucki [29] was the first to suggest using a dot-overlap model to account for printer distortions. In [1] and [2], Pappas and Neuhoff showed that, by incorporating a printer model into error diffusion, it is possible not only to correct for the effects of printer distortions but also to exploit them to produce more gray levels than could be obtained without them. We refer to their algorithm as modified error diffusion. In [2], it was shown that while Stucki's algorithm is more efficient computationally, the MED has better performance. At first glance, the extension of the MED algorithm to color appears to be straightforward. As we will see below, however, depending on the properties of the color inks, the color MED algorithm is not necessarily separable, i.e., it cannot always be applied to each color component independently. Moreover, it may become unstable.

The block diagram of the color MED algorithm is similar to that of the B&W case [1], [2] and is shown in Fig. 7. Let $[\boldsymbol{b}_{i,j}]$ be the halftone image produced by error diffusion with $\boldsymbol{b}_{i,j}$ defined as in (1). The intensity of each color component is assumed to vary between zero and one. We assume that the image has been sampled so there is one pixel per dot to be generated. Without loss of generality, we assume that the image is scanned left to right, top to bottom. The MED algorithm uses a printer model to estimate the color value of the printed pixels. The error $\boldsymbol{e}_{i,j} = (e^R_{i,j}, e^G_{i,j}, e^B_{i,j})$ is defined as the difference between the output of the printer model $\boldsymbol{p}_{i,j}$ and the "corrected" color image $\boldsymbol{v}_{i,j} = (v^R_{i,j}, v^G_{i,j}, v^B_{i,j})$. Previous errors are filtered and subtracted from the current image value $\boldsymbol{x}_{i,j}$ to obtain the "corrected" value of the gray scale image. The color MED equations are

$$\boldsymbol{v}_{i,j} = \boldsymbol{x}_{i,j} - \sum_{m,n} h_{m,n} \boldsymbol{e}^{i,j}_{i-m,j-n} \tag{5}$$

$$b^C_{i,j} = \begin{cases} 0, & \text{if } v^R_{i,j} > t \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

$$b^M_{i,j} = \begin{cases} 0, & \text{if } v^G_{i,j} > t \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

$$b^Y_{i,j} = \begin{cases} 0, & \text{if } v^B_{i,j} > t \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

$$b^K_{i,j} = \begin{cases} 1, & \text{if } b^C_{i,j} = 1, b^M_{i,j} = 1, \text{ and } b^Y_{i,j} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$\boldsymbol{e}^{i,j}_{m,n} = \boldsymbol{p}^{i,j}_{m,n} - \boldsymbol{v}_{m,n} \text{ for } (m,n) < (i,j) \tag{10}$$

where $(m,n) < (i,j)$ means (m,n) precedes (i,j) in the scanning order and

$$\boldsymbol{p}^{i,j}_{m,n} = \boldsymbol{P}(\boldsymbol{W}^{i,j}_{m,n}) \quad \text{for } (m,n) < (i,j) \tag{11}$$

where $\boldsymbol{W}^{i,j}_{m,n}$ consists of $\boldsymbol{b}_{m,n}$ and its neighbors as in (2), but here the neighbors $\boldsymbol{b}_{k,l}$ have been determined only for

(a) (b) (c)

Fig. 10. Halftoning techniques (detail of 400 dpi printout, magnified by six). (a) Modified error diffusion. (b) Least-squares model-based halftoning. (c) Least-squares model-based halftoning with sharpening.

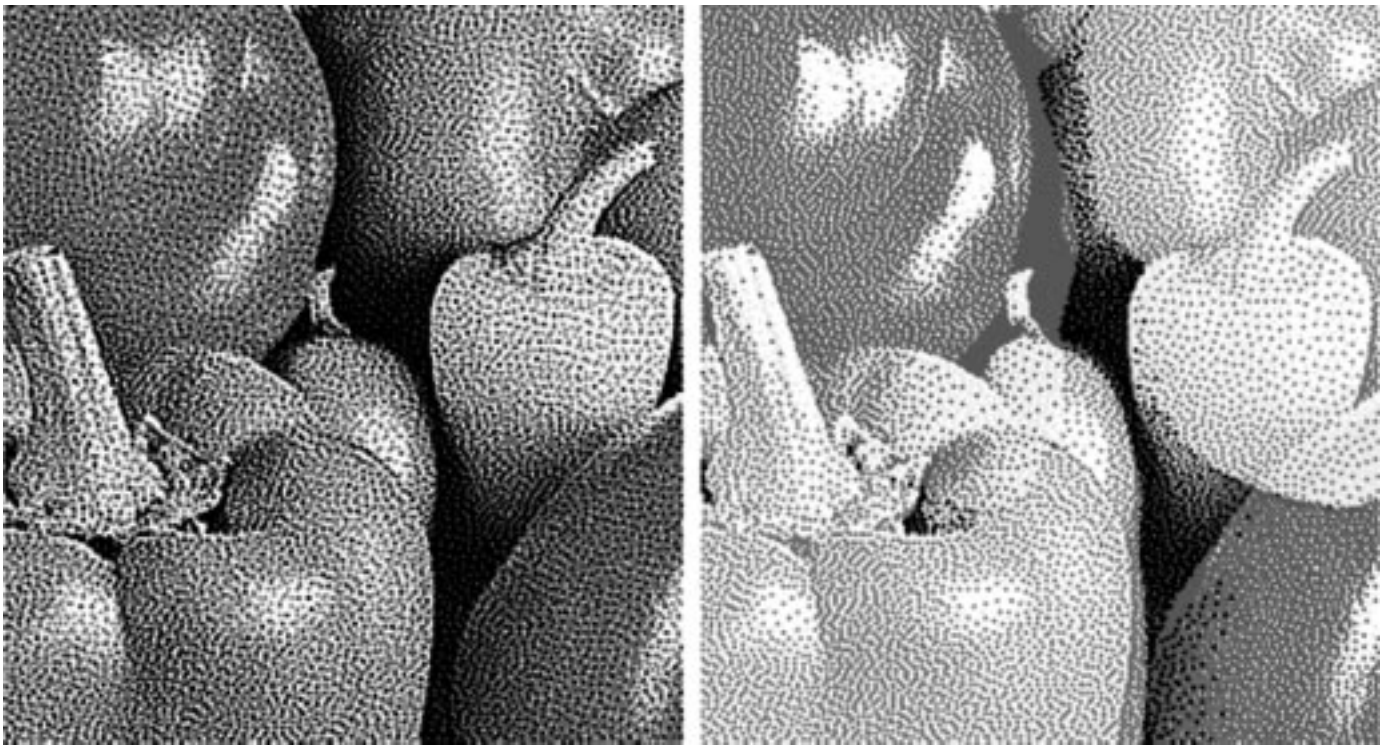
$(k,l) < (i,j)$; they are assumed to be zero (white) for $(k,l) \geq (i,j)$. Since only the dot-overlap contributions of the "past" pixels can be used in (11), the "past" errors keep getting updated as more pixel values are computed. Hence the dependence of the error and the printer model output on (i,j).[6] Notice that in the MED equations above, we assume that the black component $b^K_{i,j}$ cannot be specified independently, and is equal to one if and only if all of the other three components are equal to one. This choice resulted in the best image quality for the electrographic printer we used in our experiments. A different choice may be preferable for other printer technologies. As in the B&W case, we fix the threshold t at .5, the middle of the range of each color, and use the same filter $h_{i,j}$ for all color components. Note that the simple component by component threshold we chose above, can be replaced by more elaborate schemes, e.g., minimum Euclidean distance. In addition, different color spaces can be used. For example, Kim *et al.* [19] adopted our circular dot-overlap printer model to implement a variation of error diffusion whereby the quantization occurs in $L^*a^*b^*$ space, which is perceptually uniform.

When the image is specified in RGB space, and the printer model is separable (the simple "one-minus-RGB" relationship holds), there is no coupling between the three color components, and the MED algorithm can be applied to each color component independently. When the printer model is not separable, the color MED equations are coupled. This affects not only the ease of implementation, but also the stability of the algorithm. We adopt a BIBS (bounded input bounded state) definition of stability [30]: The error diffusion is stable if and only if the quantization error $\boldsymbol{e}_{i,j}$ is bounded for any input in the given dynamic range. As we will see below, when we use the circular-dot-overlap model that corresponds to the actual inks of the two test printers (see Tables I and II), the algorithm is unstable. This is true even for the case where we assume no dot overlap.

Now we consider the separable case. We tried the algorithm on a large number of natural images scanned from color prints and color slides, as well as synthetic images and graphics. Fig. 10(a) shows a magnified detail of an image halftoned using MED with a Jarvis–Judice–Ninke filter. As can be seen in the figure, the separable MED algorithm produces images that are sharper and have better texture than both the classical and blue-noise screening techniques. In particular, the blue-noise screen produces considerably grainier textures than MED. The difference is striking at 300 dpi and a viewing distance of two feet. (At 66.7 dpi, the reader may have to squint the eyes to see the difference.) This is contrary to some people's belief that blue-noise screening produces better textures than error diffusion. As in the B&W case [2], it can also be shown that MED produces images with richer and better color tones than screening techniques.

The separable printer models offer the advantages of computational simplicity, robustness to errors in color-plane registration, and most importantly, guaranteed stability of the modified error diffusion algorithm. Moreover, the few artifacts and asymmetries that the B&W error diffusion algorithm is

[6]The assumption that the undetermined pixels are white leads to a very small and difficult to detect bias in the color of the printed image. A multipass version of the MED algorithm eliminates this bias [25].



(a) (b)

Fig. 11. Modified error diffusion (400 dpi printout, magnified by six). (a) Separable printer model. (b) Nonseparable printer model.

known to produce are a lot less visible. This is because there are more degrees of freedom in coloring a pixel (eight colors instead of two). In fact, at 400 dpi, our experiments show that the halftoning artifacts of the MED are almost invisible and the printed MED images look like continuous-tone images.

However, as we discussed in Section II, the "one-minus-RGB" assumption does not hold for any real printer. Thus, the separable algorithm leads to inaccuracies in the overall color reproduction, even though the relative colors still look fine, as indicated by extensive experiments. Thus, the colors of an image look different when it is printed on different printers, and are also significantly different from the continuous-tone image displayed on a CRT. The color of each printout looks fine on its own, but a comparison shows significant color differences.

Unfortunately, the nonseparable algorithm may be unstable. The typical effects of instability can be seen in Fig. 11, which shows an image halftoned with the separable MED in (a) and the nonseparable MED in (b). The resolution of the image is 256 $\times$ 235 pixels. In Fig. 11(b), we used the circular-dot-overlap model that corresponds to the inks of the electrographic Canon printer (Table I). Notice how color smears across object boundaries. As we mentioned above, MED is unstable even for the case where there is no dot overlap. The instability of error diffusion has also been reported elsewhere, e.g. in [19] and [20]. The manifestation of instability artifacts depends on both the ink parameters and the specific image. A typical way to remove instabilities is by forcing the error term to be bounded by clipping. However, this is not guaranteed to produce the correct color (according to the LSMB error criterion defined in the next section), even though it appears to work reasonably well for many images. As we will see in the next section, the LSMB algorithm has no stability problems.

The nonseparable algorithm requires an accurate device for estimating the ink parameters, and more importantly, accurate knowledge of the characteristics of the image acquisition device, in order to reproduce the colors of the original image. Further study of the nonseparable algorithm is left to future work.

V. Least-Squares Model-Based Halftoning

The color LSMB halftoning algorithm is a direct extension of the B&W algorithm [3]. It minimizes the squared error between the perceived intensity of the continuous-tone original image and the perceived intensity of the printed halftone image. The original continuous-tone color image is denoted by $[\boldsymbol{x}_{i,j}]$. Again, we assume that the image has been sampled so there is one pixel per dot to be generated. We assume a printer model with the form of (2), and an eye model that has the form of a finite impulse response (FIR) filter with impulse response $g_{i,j}$. We seek the halftone image $[\boldsymbol{b}_{i,j}]$ that minimizes the squared error

$$E = \sum_{i,j} \|\boldsymbol{z}_{i,j} - \boldsymbol{w}_{i,j}\|^2 \tag{12}$$

where, as illustrated in Fig. 8

$$\boldsymbol{z}_{i,j} = \boldsymbol{x}_{i,j} * g'_{i,j} \tag{13}$$

$$\boldsymbol{w}_{i,j} = \boldsymbol{p}_{i,j} * g_{i,j} = \boldsymbol{P}(\boldsymbol{W}_{i,j}) * g_{i,j}. \tag{14}$$

Here, $\boldsymbol{W}_{i,j}$ consists of $\boldsymbol{b}_{i,j}$ and its neighbors as in (2), and "*" indicates convolution. The boundary conditions assume that no ink is placed outside the image borders. As in the B&W case, we allow different impulse responses $g_{i,j}, g'_{i,j}$ for the eye filters corresponding to the continuous-tone and halftone images. As in the B&W case, we used a simple eye model that consists of a two-dimensional (2-D) FIR filter. We used the same filter for all the color components. No attempt was made here to exploit the different sensitivity of the human visual system to the luminance and the chromatic dimensions, as suggested in [16] and [17].

When the printer model is separable as in (4), the least-squares problem can be solved independently for each color component. When the printer model is not separable, the least-squares problem is also not separable. In the nonseparable case, the selection of the black component $b^K_{i,j}$ is not constrained, and thus we can use it to produce an even richer variety of colors (the black ink produces a different color than the combination of the other three or all four inks). However, for some printers the black ink does not mix well with any one of the other inks, and it may be necessary to restrict its use only in combination with the other three inks to produce solid black dots.

The color least-squares solution is obtained by iterative techniques similar to those used in the B&W case. Such techniques find a solution that is only a local optimum; the visual quality of the resulting halftone images depends on the initial estimate and the optimization strategy. The initial estimate of the binary image $[\boldsymbol{b}_{i,j}]$ could be a constant or random image, or the output of a halftoning algorithm, e.g. the MED described in Section IV. In a simple iterative scheme [3], [10] the binary image is updated one pixel at a time, in a raster scan. In a similar scheme [31], each pixel of the binary image is either toggled or swapped with one of its eight neighbors in a given scanning order. Other variations are possible ([9] considered swaps only). More sophisticated (and computationally intensive) schemes use simulated annealing [10], [14]. More details about the various schemes can be found in [32]. In the color case, three (or four) bits must be selected jointly for each pixel. Thus, there are eight (or sixteen) choices for the toggle only case; the complexity increases significantly in the other cases. Note that, unlike error diffusion, there are no stability problems. The separable case, whereby the minimization is done independently for each color component, requires significantly less computation. Finally, as in the B&W case, the color LSMB error criterion provides a metric for halftone image quality.

We now consider the separable case. Fig. 10(b) shows a magnified detail of an image halftoned with LSMB using Gaussian eye filters $g = g'$ with standard deviation 1.2, which corresponds to a viewing distance of 2 ft at 300 dpi, 1.5 ft at 400 dpi, and 9 ft at the displayed resolution. The initial estimate was provided by the multipass MED algorithm. Notice that the MED result of Fig. 10(b) is sharper than the LSMB result. This is because error diffusion introduces some amount of sharpening [33], [34]. As we show in [32], the LSMB approach allows us to control the amount of sharpening by varying the eye filter corresponding to the continuous-tone image. Fig. 10(c) shows an image halftoned with LSMB using the same filter g as in (b) and g' equal to a unit impulse (this results in the maximum amount of sharpening).

In addition to the computational advantage, the separable algorithm is robust to errors in color registration. In practice, however, the "one-minus-RGB" assumption does not hold, and the overall color of the reproduction depends on the ink characteristics, even though the relative colors look fine. The LSMB algorithm with nonseparable printer models offers the possibility of achieving device-independent color within the common range of colors of the different devices. Colors outside the gamut of the printer will be mapped to the closest color in the gamut. The nonseparable algorithm is expected to be more sensitive to color registration. As we saw earlier, it requires an accurate device for estimating the ink parameters, and more importantly, accurate knowledge of the characteristics of the image acquisition device, in order to reproduce the colors of the original image. A systematic study of the performance of the nonseparable algorithm is left to future work.

VI. Conclusion

We presented a new class of models for color printers and a specific model that accounts for dot overlap and the light absorption properties of the inks. We also presented two model-based techniques for color halftoning, the MED and the LSMB algorithm. For images specified in the RGB domain, we showed that if we assume a simple "one-minus-RGB" relationship between the red, green, and blue image specification and the corresponding cyan, magenta, and yellow inks, the algorithms are separable. The separable algorithms produce halftone images that are sharper, have better texture, and have richer and better color tones than both the classical and blue-noise screening techniques. They also offer computational simplicity and robustness to errors in color registration. Since they can exploit three different types of ink, color model-based halftoning techniques produce higher quality images than the corresponding techniques on similar B&W printers of the same resolution. The main drawback of the separable algorithms is the fact that they do not accomplish exact colorimetric reproduction of color. The colors of the resulting images are device dependent. The relative colors, however, look fine, and this may be all that is necessary in some applications. For example, different types of photographic film produce images with different colors, all of which are acceptable. The systematic study of the nonseparable algorithms, which offer the possibility of device-independent color reproduction, will be considered in future work.

Acknowledgment

The author wishes to thank Dr. V. Ostromoukhov of the Swiss Federal Institute of Technology for providing examples of rotated color screens.

References

[1] T. N. Pappas and D. L. Neuhoff, "Model-based halftoning," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display II*, San Jose,

CA, Feb. 1991, vol. 1453, pp. 244–255.
[2] ———, "Printer models and error diffusion," *IEEE Trans. Image Processing*, vol. 4, pp. 66–80, Jan. 1995.
[3] ———, "Least-squares model-based halftoning," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display III*, vol. 1666, San Jose, CA, Feb. 1992, pp. 165–176.
[4] R. Ulichney, *Digital Halftoning*. Cambridge, MA: MIT Press, 1987.
[5] P. Fink, *Postscipt Screening: Adobe Accurate Screens*. Mountain View, CA: Adobe, 1992.
[6] G. G. Field, *Color and Its Reproduction*. Pittsburgh, PA: Graphic Arts Tech. Foundation, 1992.
[7] T. N. Pappas, "Model-based halftoning of color images," in *IS&T's 8th Int. Congr. Advanced Non-Impact Printing Technology*, Williamsburg, VA, Oct. 25–30, 1992, pp. 270–275.
[8] ———, "Printer models and color halftoning," in *Proc. ICASSP-93*, Minneapolis, MN, Apr. 1993, vol. V, pp. 333–336.
[9] M. Analoui and J. P. Allebach, "Model based halftoning using direct binary search," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display III*, San Jose, CA, Feb. 1992, vol. 1666, pp. 96–108.
[10] J. B. Mulligan and A. J. Ahumada, Jr., "Principled halftoning based on models of human vision," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display III*, San Jose, CA, Feb. 1992, vol. 1666, pp. 109–121.
[11] C. A. Poynton, *A Technical Introduction to Digital Video*. New York: Wiley, 1996.
[12] T. Mitsa and K. J. Parker, "Digital halftoning technique using a blue-noise mask," *J. Opt. Soc. Amer. A*, vol. 9, pp. 1920–1929, Nov. 1992.
[13] R. Ulichney, "The void-and-cluster method for dither array generation," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display IV*, San Jose, CA, Feb. 1993, vol. 1913, pp. 332–343.
[14] M. A. Schulze and T. N. Pappas, "Blue noise and model-based halftoning," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display V*, San Jose, CA, Feb. 1994, vol. 2179, pp. 182–194.
[15] R. Miller and J. R. Sullivan, "Color halftoning using error diffusion and a human visual system model," in *Proc. SPSE 43rd Annu. Meet.*, Rochester, NY, 1990, pp. 149–152.
[16] J. B. Mulligan and A. J. Ahumada, Jr., "Principled methods for color dithering based on models of the human visual system," in *SID Dig. Tech. Papers*, Boston, MA, May 1992, pp. 194–197.
[17] T. J. Flohr *et al.*, "Model based color image quantization," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display IV*, San Jose, CA, Feb. 1993, vol. 1913, pp. 270–281.
[18] B. W. Kolpatzik and C. A. Bouman, "Optimized error diffusion for image display," *J. Electron. Imaging*, vol. 1, pp. 277–292, July 1992.
[19] C. Kim, S. Kim, Y. Seo, and I. Kweono, "Model based color halftoning techniques on perceptually uniform color spaces," in *IS&T's 47th Ann. Conf.*, Rochester, NY, May 15-20, 1994, pp. 494–499.
[20] H. Haneishi, T. Suzuki, N. Shimoyama, and Y. Miyake, "Color digital halftoning taking colorimetric color reproduction into account," *J. Electron. Imaging*, vol. 5, pp. 97–106, Jan. 1996.
[21] M. Rodriguez, "Graphic arts perspective on digital halftoning," in *Proc. SPIE, Human Vision, Visual Processing, and Digital Display V*, San Jose, CA, Feb. 1994, vol. 2179, pp. 144–149.
[22] ———, "Promises and pitfalls of stochastic screening in the graphic arts industry," in *IS&T's 47th Ann. Conf.*, Rochester, NY, May 15–20, 1994, pp. 460–463.
[23] R. W. G. Hunt, *The Reproduction of Colour in Photography, Printing & Television*. Tolworth, U.K.: Fountain, 1987.
[24] C.-K. Dong, "Measurement of printer parameters for model-based halftoning," B.S./M.S. thesis, Mass. Inst. Technol., Cambridge, MA, May 1992.
[25] T. N. Pappas, C.-K. Dong, and D. L. Neuhoff, "Measurement of printer parameters for model-based halftoning," *J. Electron. Imag.*, vol. 2, pp. 193–204, July 1993.
[26] P. G. Roetling and T. M. Holladay, "Tone reproduction and screen design for pictorial electrographic printing," *J. Appl. Photog. Eng.*, vol. 15, pp. 179–182, 1979.
[27] M. Yao and K. J. Parker, "Application of blue noise mask in color halftoning," in *Proc. SPIE, Visual Communications and Image Processing*, Orlando, FL, Mar. 1996, vol. 2727, pp. 876–880.
[28] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial grey scale," in *Proc. SID*, vol. 17/2, pp. 75–77, 1976.
[29] P. Stucki, "MECCA—a multiple-error correcting computation algorithm for bilevel image hardcopy reproduction," Res. Rep. RZ1060, IBM Res. Lab., Zurich, Switzerland, 1981.
[30] Z. Fan, "Stability analysis of error diffusion," in *Proc. ICASSP-93*, Minneapolis, MN, Apr. 1993, vol. V, pp. 321–324.
[31] D. A. Carrara, M. Analoui, and J. P. Allebach, "Recent progress in digital halftoning," in *IS&T's 8th Int. Cong. Advanced Non-Impact Printing Technology*, Williamsburg, VA, Oct. 25–30, 1992, pp. 265–270.
[32] T. N. Pappas, "Model-based techniques for digital halftoning," in *Proc. ICIP-94*, Austin, TX, Nov. 1994, vol. II, pp. 26–30.
[33] R. Eschbach and K. Knox, "Error-diffusion algorithm with edge enhancement," *J. Opt. Soc. Amer. A*, vol. 8, pp. 1844–1850, Dec. 1991.
[34] K. T. Knox, "Error image in error diffusion," in *Proc. SPIE, Image Processing Algorithms and Techniques III*, San Jose, CA, Feb. 1992, vol. 1657, pp. 268–279.

Thrasyvoulos N. Pappas (M'87–SM'95) received the S.B., S.M., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1979, 1982, and 1987, respectively.

He has been a Member of the Technical Staff, Bell Laboratories, Murray Hill, NJ, since 1987. His research interests are in image processing and computer vision. He has worked on image and video segmentation, video/audio integration, and halftoning and compression based on models of display devices and the human visual system.

Dr. Pappas is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is also the Electronic Abstracts Editor for the TRANSACTIONS and a member of the IEEE Image and Multidimensional Signal Processing Technical Committee. He is co-chair for the Conference on Human Vision and Electronic Imaging sponsored by SPIE and IS&T.