Measurement of printer parameters for model-based halftoning

Thrasyvoulos N. Pappas
AT&T Bell Laboratories
Signal Processing Research Department
Murray Hill, New Jersey 07974

Chen-Koung Dong*
First Boston Corporation
New York, New York 10055

David L. Neuhoff†
University of Michigan
Department of Electrical Engineering and Computer Science
Ann Arbor, Michigan 48109

Abstract. We present a new approach for estimating printer model parameters that can be applied to a wide variety of laser printers. Recently developed “model-based” digital halftoning techniques depend on accurate printer models to produce high-quality images using standard laser printers (typically 300 dpi). Since printer characteristics vary considerably, e.g., write-black versus write-white laser printers, the model parameters must be adapted to each individual printer. Previous approaches for estimating the printer model parameters are based on a physical understanding of the printing mechanism. One such approach uses the “circular dot-overlap model,” which assumes that the laser printer produces circularly shaped dots of ink. The circular dot-overlap model is an accurate model for many printers but cannot describe the behavior of all printers. The new approach is based on measurements of the gray level produced by various test patterns and makes very few assumptions about the laser printer. We use a reflection densitometer to measure the average reflectance of the test patterns and then solve a constrained optimization problem to obtain the printer model parameters. To demonstrate the effectiveness of the approach, the model parameters of two laser printers with very different characteristics were estimated. The printer models were then used with both the modified error diffusion and the least-squares model-based approach to produce printed images with the correct gray-scale rendition. We also derived an iterative version of the modified error diffusion algorithm that improves its performance.

1 Introduction

Digital halftoning is the process of generating a pattern of binary pixels that the eye perceives as a continuous-tone image. Digital halftoning is necessary for display of grayscale images in media in which the direct rendition of gray tones is impossible. Examples of such media include paper and binary cathode-ray tube (CRT) displays. In this paper we examine “model-based” digital halftoning techniques that have been developed recently§ and depend on accurate printer models to produce high-quality images using standard laser printers (typically 300 dpi). The goal of this paper is to develop an experimental procedure for estimating model parameters for a wide variety of printers.

Model-based halftoning can be used to improve the quality of grayscale images transmitted by facsimile. A new approach to gray-scale facsimile is proposed in Refs. 5 and 6, in which the image is transmitted in grayscale form using high-fidelity image coders. This approach allows more efficient encoding of data and, more importantly, permits the halftoner to be tuned to the individual printer on which the document is printed.

Figure 1 shows how halftoning works. A halftoning algorithm generates a binary pattern of pixels that is printed and perceived by the eye. All halftoning techniques rely on the fact that the eye acts as a spatial low-pass filter. The performance of a halftoning technique is also affected by the behavior of the display device. Most halftoning techniques assume that the displayed binary pattern consists of identically shaped dots of two colors, usually on a rectangular grid. This assumption does not hold for most printing devices, which introduce significant distortions. Such distortions make many halftoning techniques unsuitable for printers.
Printer characteristics are known to vary considerably from printer to printer. For example, write-black laser printers have very different characteristics than write-white laser printers. Traditional halftoning techniques are designed to be fairly robust to printer distortions. As a result, they compromise both spatial and gray-scale resolution. Model-based halftoning techniques, on the other hand, exploit the characteristics of each particular printer to maximize the quality of the printed images. Thus, they depend on an accurate printer model, whose parameters must be adapted to each individual printer.

The printer model predicts the “gray” level of the printed pattern as a function of the specified bits. The parameters of the printer model can be derived from a physical understanding of the printing mechanism or from direct measurements of the reflectance of various printed patterns. One example of the first approach uses the circular dot-overlap model, which assumes that the laser printer produces circularly shaped dots of ink. Similar circular dot-overlap models were also used in Ref. 7 to modify the ordered dither thresholds so that they result in a linear gray scale and in Refs. 8 and 9 to correct for printer distortions in error diffusion. The gray level at each pixel of the printed image is proportional to the area of the pixel that is covered by ink. This area can be calculated easily from the radius of the dots. Thus, all we need to obtain the gray values at each location is an estimate of the radius of these circularly shaped ink dots.

The circular dot-overlap model is an accurate model for many printers but cannot describe the behavior of all printers. The black dots produced by actual printers are not perfectly round, they are not perfectly black, and their size and shape may depend on the presence of adjacent dots. For example, as was pointed out in Ref. 1, a white line surrounded by several black lines is brighter than when surrounded by two single lines. For some printers the circular dot approximation may not be valid at all. In this paper we develop an approach for estimating printer model parameters that is based on measurements of the gray level produced by various test patterns; this approach makes very few assumptions about the laser printer. The measurement approach is general and can be applied to any write-white or write-black printer. A reflection densitometer is used to measure the reflectance of a set of test patterns. The measured reflectance can be related to the printer model parameters by means of iterative techniques to solve a constrained optimization problem.

The measurement approach can also be extended to apply to any digital printing device, even though the emphasis of this paper is on laser printers. High-resolution devices* (e.g., phototypesetters) used for high-quality and high-volume printing behave very differently from laser printers. For example, physical limitations make it impossible for such devices to print isolated “black” or “white” dots. The strong and complicated dependence of each printed pixel on neighboring pixels makes it difficult to develop analytical models such as the circular dot-overlap model. The measurement approach may be the only way to obtain models for such devices. Most of these devices use traditional halftoning techniques that, as we saw, are robust to printer distortions. The use of model-based halftoning could either improve their quality further or lower the cost by reducing the resolution requirement.

To demonstrate the effectiveness of the approach, we estimated the model parameters of two laser printers with very different characteristics. The printed models were then used with two model-based halftoning techniques: the modified error diffusion algorithm1 and the least-squares model-based (LSMB) halftoning algorithm.3 We also propose a multipass version of the modified error diffusion algorithm that improves its performance. The most important criterion for the effectiveness of the measurement approach is the accuracy of the gray-scale rendition of the printed images.

The experimental results demonstrate that when the measured model parameters are used with the two model-based techniques, they produce the correct gray scale and maintain the overall performance of these techniques. For one of the laser printers we used, the circular dot-overlap model and the measurement approach produce results that are equally good. The circular dot-overlap model does not apply to the second laser printer. Thus, even though the measurement approach is considerably more complicated than the circular dot-overlap model, it can be used to establish the limitations of this and other simple models.

The remainder of this paper is organized as follows. Section 2 discusses printer models. Section 3 reviews model-based halftoning techniques. Section 4 describes the measurement approach for estimating printer model parameters. The application of the measurement approach to specific laser printers and halftoning techniques is presented in Sec. 5. The conclusions are summarized in Sec. 6.

2 Printer Models

The purpose of a printer model is to predict the “gray” levels produced by a printer. Accurate and objective predictions of gray level make it possible not only to correct for the effects of printer distortions, but also to take advantage of them to produce more gray levels. A printer is a device that can generate dots of ink on a piece of paper in designated areas, usually on a Cartesian grid. We will refer to grid locations with an as “black” dots and grid locations without ink as “white” dots. Two types of printers perform this function: write-black printers and write-white printers. The difference between the two is that write-black printers print dots of black ink against a white background, while write-white printers, effectively, “print” dots of white ink against a black background. In reality, both printer types use black ink for printing. The actual difference lies in the way the drum of the printer and the particles of carbon or ink are charged. The nature of the distortions in these two types of printers is quite different. There are many causes for these distortions. They include the spreading of the laser beam, interactions of the laser and the charge applied to the drum, the type of toner particles used, and the heat finishing. The printer distortions

*The typical resolution of such devices is more than 1000 dpi.
have a significant effect on the actual darkness of a printed image.

In the remainder of this section, we review the general formulation for the printer model that was proposed in Refs. 1 and 2 and discuss approaches for obtaining specific models. In our description of printer models, we use the following notation. The printer is controlled by an $N_w \times N_H$ binary array $\{b_{ij}\}$, where $b_{ij} = 1$ indicates that an ink dot is to be printed at pixel $(i, j)$ and $b_{ij} = 0$ means that no ink dot is to be printed. The pixel $(i, j)$ is located $iT$ inches from the left and $jT$ inches from the top of the image.

As a result of printer distortions, the gray level produced by the printer at any point in the image depends in some complicated way on the surrounding bits. Let $u(s, t)$ be the gray level produced by the printer at point $(s, t)$ located $s$ inches from the left and $t$ inches from the top of the image. Then,

$$u(s, t) = f(s, t; B_{s,t})$$

$$\frac{T}{2} \leq s \leq N_w T + \frac{T}{2}, \frac{T}{2} \leq t \leq N_H T + \frac{T}{2},$$

where $B_{s,t}$ denotes the set of bits in a neighborhood of the point $(s, t)$ and $f$ is some function. However, due to the close spacing of the dots and the limited spatial resolution of the eye, the gray level $u(s, t)$ of the printed image can be modeled as having a constant value $p_{i,j}$ in the vicinity of site $(i, j)$ as follows

$$\bar{u}(s, t) = p_{i,j}$$

$$\frac{iT - T/2}{T} \leq s \leq \frac{iT + T/2}{T}, \frac{jT - T/2}{T} \leq t \leq \frac{jT + T/2}{T},$$

for all $1 \leq i \leq N_w$ and $1 \leq j \leq N_H$. Although the gray level is not actually constant, the eye responds, essentially, only to the average gray level over the site. It is this average gray level that $p_{i,j}$ represents; namely,

$$p_{i,j} = \frac{1}{T^2} \int_{iT - T/2}^{iT + T/2} \int_{jT - T/2}^{jT + T/2} f(s, t; B_{s,t}) \, ds \, dr$$

$$1 \leq i \leq N_w, 1 \leq j \leq N_H.$$

(3)

Thus, the printer model becomes

$$\bar{u}(s, t) = p_{i,j}$$

$$\frac{iT - T/2}{T} \leq s \leq \frac{iT + T/2}{T}, \frac{jT - T/2}{T} \leq t \leq \frac{jT + T/2}{T},$$

where

$$q(s, t) = \begin{cases} 1, & \text{if } |s| \leq T/2, |t| \leq T/2 \\ 0, & \text{otherwise} \end{cases}$$

(5)

It follows from Eq. (3) that the average level $p_{i,j}$ depends on the neighboring bits. Thus the printer model takes the form

$$p_{i,j} = \Phi(W_{i,j}) \begin{cases} 1, & \text{if } b_{i,j} = 1 \\ f_1 x + f_2 y - f_3 a, & \text{if } b_{i,j} = 0 \end{cases},$$

(6)

where $W_{i,j}$ is a window that consists of the bits in some neighborhood of $b_{i,j}$, and $\Phi$ denotes some function thereof. Note that both the binary array $\{b_{i,j}\}$ that specifies the dot pattern to be printed and the array of gray levels $\{p_{i,j}\}$ specified by the printer model have the same dimensions.

Our task is to find the function $\Phi$. For the model-based techniques, it is essential that the window $W_{i,j}$ be finite. To obtain such a printer model, two approaches can be taken. The first approach is to model the physical behavior of a printer, that is, to specify the function $f$ of Eq. (1) and then obtain the function $\Phi$ using Eq. (3). An example of this approach was presented in Refs. 1 and 2 and is reviewed in the following subsection. The second approach is to obtain the function $\Phi$ directly from measurements, without worrying about the exact form of the function $f$. This second approach is the main goal of this paper.

### 2.1 Circular Dot-Overlap Model

The circular dot-overlap model proposed in Ref. 1 is a simple and very effective model of printer behavior. It assumes that each printed dot is circular with a uniform distribution of ink. This idealization of printer behavior was also considered in Refs. 7 and 8. The radius of the dots produced by a printer must be at least $T/\sqrt{T^2}$, where $T$ is the spacing of the Cartesian grid, so that a page can be blackened entirely. We refer to a printer that produces dots of minimal size as the ideal printer. Actual printers produce dots that are larger than the minimal size. We will use $\rho$ to denote the ratio of the actual dot radius to the radius of the dot of the ideal printer. The effective gray level of a printed pixel is assumed to be the percentage of the area of the pixel that is covered by ink. If an appropriate value of $\rho$ is chosen, then the effective gray level of all the pixels of a 2-D pattern can be calculated using simple trigonometric formulas.

The amount of dot overlap at each pixel can be expressed in terms of the parameters $\alpha, \beta$, and $\gamma$, shown in Fig. 2. These parameters are the ratios of the areas of the shaded regions shown in the figure to $T^2$, the area of the pixel. They can easily be expressed in terms of the radius $\rho$. In terms of these parameters, the circular dot-overlap model becomes

$$p_{i,j} = \Phi(W_{i,j}) = \begin{cases} 1, & \text{if } b_{i,j} = 1 \\ f_1 x + f_2 y - f_3 a, & \text{if } b_{i,j} = 0 \end{cases},$$

(7)

where $f_1$ is the number of horizontally and vertically neighboring dots that are black, $f_2$ is the number of diagonally neighboring dots that are black and not adjacent to any horizontally or vertically neighboring black dot, and $f_3$ is the number of pairs of neighboring black dots in which one is a horizontal neighbor and the other is a vertical neighbor.

The circular dot-overlap model is an idealization of printer behavior. It is very simple and provides a good first-order approximation for the behavior of many printers. However, it does not adequately account for all of the printer distortions. Significant discrepancies often exist between the predictions of the model and the measured values. Some idea of such discrepancies may be obtained by considering a set of horizontally invariant, vertically periodic patterns that were printed on a 300-dpi write-black printer. The different patterns, represented by one period, are listed in Table 1. The table also lists the frequency of ones in each pattern, the gray

1Note that the function $f$ could be deterministic or probabilistic, as suggested in Ref. 9.
level predicted by the circular dot-overlap model with \( \alpha = 0.33 (p = 1.25) \), and the measured reflectance density of each printed pattern. The higher the measured density of the pattern, the darker it appears to the eye. As expected, patterns with the same frequency of ones differ substantially in their reflectance densities. Notice that there are pairs of patterns, such as 101010 and 110110, for which the printer model prediction is opposite from the measured density.

More importantly, however, for some printers the circular dot-overlap model is not valid at all. To predict accurately the behavior of any printer, we consider a direct measurement approach.

### 2.2 Measurement Approach

An alternative approach for predicting printer behavior is by direct measurement of the printer parameters. The parameters of the printer model can be obtained directly from measurement of the gray level of various printed test patterns by means of a reflection densitometer. The measurement approach makes minimal assumptions about the printing process and can be applied to any write-black or write-white printer. The development of such a measurement approach is the main focus of this paper and will be discussed in detail in Sec. 4.

The measured printer model parameters can be used by various halftoning techniques to enhance their accuracy of gray-scale rendition. Two such model-based techniques are the modified error diffusion algorithm and the least-squares model-based method. We now consider these two techniques.

### 3 Model-Based Halftoning

In this section we review model-based halftoning techniques that depend on accurate printer models to account for printer distortions and produce high-quality printed images. We are primarily interested in laser printers, which generate distortions such as dot overlap but, as we discussed in the introduction, the techniques apply to any printing device. Conventional techniques, such as clustered-dot ordered dither, resist distortions at the expense of spatial and gray-scale resolution. Model-based techniques, on the other hand, rely on printer models to exploit printer distortions in order to increase both gray-scale and spatial resolution. We consider two model-based techniques, the modified error diffusion algorithm\(^{1,2} \) and the least-squares model-based algorithm.\(^{3,4} \)

We use \( [x_{ij}] \) to denote a gray-scale image, where \( x_{ij} \) denotes the pixel located at the \( i \)th column and the \( j \)th row of a Cartesian grid. The typical gray-scale resolution is \( 2^8 = 256 \) levels; we use a normalized scale in the interval from 0 = white to 1 = black. We assume that the image has been sampled so there is one pixel per dot to be generated.\(^6 \) Thus the gray-scale image array \( [x_{ij}] \) and the binary image array \( [b_{ij}] \) have the same dimensions.

#### 3.1 Modified Error Diffusion

Error diffusion\(^{10} \) is a halftoning technique that produces sharper images than conventional screening techniques. The standard error diffusion algorithm is very sensitive to printer distortions, however. Stucki\(^8 \) was the first to suggest the use of a dot-overlap model to account for printer distortions. In Refs. 1 and 2 we showed that, by incorporating a printer model into error diffusion, it is possible to not only correct for the effects of printer distortions, but also to take advantage of them to produce more gray levels. We refer to the resulting algorithm as modified error diffusion. In Ref. 2 we show that, while Stucki’s algorithm is more efficient computationally, the modified error diffusion has better performance.

A block diagram of the modified error diffusion algorithm\(^{1,2} \) is shown in Fig. 3. Without loss of generality, we assume that the image is scanned left to right, top to bottom. The binary image \( [b_{ij}] \) is obtained by thresholding a “corrected” value \( v_{ij} \) of the gray-scale image. The modified error diffusion algorithm uses a printer model to estimate the gray level \( p_{ij} \) of the printed pixels. The difference between this gray level and the corrected gray-scale image is defined as the error \( e_{ij} \) at the “instant” \( (i,j) \).\(^7 \) “Past” errors are low-pass filtered and subtracted from the current image value \( x_{ij} \) to obtain the corrected value of the gray-scale image. The

---

\(^6\)When the samples of a given image are fewer than the number of dots to be generated, interpolation is necessary.

\(^7\)In the standard error diffusion algorithm, the errors are calculated under the assumption that the printed pixels are either perfectly black or perfectly white, and thus their gray value is equal to the binary value assigned to the output pixels.
threshold \( t \) is typically fixed at 0.5, the middle of the grayscale range.

The impulse response of the low-pass filter is \([h_{ij}]\). It has nonsymmetric half-plane support, so that only the past errors are filtered. This is the 2-D equivalent of "causality," and enables the algorithm to make instantaneous decisions at each point. Thus, the modified error diffusion algorithm, like standard error diffusion, requires only one pass through the data. The filter coefficients are positive and their sum is equal to one. This guarantees stability. In the examples of this paper we use the error diffusion filter proposed by Jarvis, Judice, and Ninke, shown in Fig. 4.

The modified error diffusion equations are\(^1\)

\[
v_{i,j} = x_{i,j} - \sum_{m,n} h_{m,n}e^{i,j}_{-m,j-n}, \quad (8)
\]

\[
b_{i,j} = \begin{cases} 
1, & \text{if } v_{i,j} > t \\
0, & \text{otherwise}
\end{cases}, \quad (9)
\]

\[
e^{i,j}_{m,n} = p^{i,j}_{m,n} - v_{m,n} \quad \text{for } (m,n) < (i,j), \quad (10)
\]

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

\[
p^{i,j}_{m,n} = \mathcal{P}(W^{i,j}_{m,n}) \quad \text{for } (m,n) < (i,j), \quad (11)
\]

where \(W^{i,j}_{m,n}\) consists of \(b_{m,n}\) and its neighbors as in Eq. (6). Notice, however, that here the neighbors \(b_{k,l}\) have been determined only for \((k,l) < (i,j)\); they are assumed to be zero (i.e., white) for \((k,l) \geq (i,j)\). Since only the dot-overlap contributions of the past pixels can be used in Eq. (11), the past errors keep getting updated as more binary values are computed. Hence the dependence of the error and the printer model output on the "instant" \((i,j)\). This is illustrated in Fig. 5, which shows the binary values necessary for the calculation of the error for all the points in the error diffusion filter mask, assuming a 3 \(\times\) 3 window \(W^{i,j}_{m,n}\). For the points on the edge of the filter mask, the error depends on "future" points, which have not yet been determined. These points, shown as question marks in Fig. 5, are assumed to be white (or some other predetermined pattern). We will refer to the value of these points as the initial background.

The initial background assumption produces a bias in the gray scale of the printed images. This bias is toward darker images when the initial background is assumed to be white because the contribution to the gray level due to the surrounding pixels is underestimated; so the algorithm tends to produce too many black dots. As we will see in Sec. 5, the extent of this bias depends on the printer model and the initial background. For the circular dot-overlap model of the previous section, the bias is small and is almost imperceptible. For other printer models, this bias can be large and can cause the algorithm to diverge. Thus, the addition of the printer model can affect the stability of the algorithm. To eliminate the bias, we propose a multipass version of the modified error diffusion algorithm, whereby, in each iteration the future pixels (initial background) are assumed to have the value given by the previous iteration. The multipass algorithm is discussed further in Sec. 5.

### 3.2 Least-Squares Model-Based Halftoning

LSMB halftoning\(^3,4\) takes advantage of both an eye model and a printer model. It finds an optimal halftoned reproduction by minimizing the squared error between the perceived intensity of the original gray-scale image \([x_{i,j}]\) and the perceived intensity of the printed halftoned image \([b_{i,j}]\).

Given a gray-scale image \([x_{i,j}]\), the LSMB method finds the binary image \([b_{i,j}]\) that minimizes the sum of the squares of the differences between the two perceived images \([z_{i,j}]\) and \([w_{i,j}]\):

\[
E = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (z_{i,j} - w_{i,j})^2, \quad (12)
\]

where, as illustrated in Fig. 6,

\[
z_{i,j} = x_{i,j} \ast g_{i,j}', \quad (13)
\]

\[
w_{i,j} = p_{i,j} \ast g_{i,j} = \mathcal{P}(W_{i,j}) \ast g_{i,j}, \quad (14)
\]

and \(\ast\) indicates convolution. As in Ref. 3, we have allowed the impulse responses \([g_{i,j}]\) and \([g'_{i,j}]\) for the eye filters corresponding to the continuous-tone and halftone images to be different. In fact, for the remainder of this paper, we drop
independently. One-dimensional halftoning is simpler than 2-D halftoning and easier to analyze, but is seldom used in practice because it does not exploit the 2-D properties of the eye. One-dimensional least-squares halftoning can be implemented with the Viterbi algorithm.\(^4\) The Viterbi algorithm provides an efficient way to search the solution space and leads to a global optimum in a finite number of steps. Unfortunately, the Viterbi algorithm cannot be used in two dimensions. Therefore, as we saw earlier, iterative techniques are necessary to obtain an approximate solution of the 2-D least-squares problem. Thus, in addition to the theoretical interest, the 1-D problem makes it possible to isolate the performance of the printer model from that of the optimization algorithm. Moreover, as we will see in the next section, the 1-D printer models are considerably easier to obtain and understand.

4 Measurement of Printer Model Parameters

The proposed method is based on direct measurement of the reflectance of a set of printed test images. The measured reflectance can be related to the printer model parameters by a set of linear equations. We can then formulate a constrained optimization problem that incorporates various constraints on the model parameters. This optimization problem can be solved by standard iterative techniques. We will refer to the resulting printer model as the *measurement model*. The approach is general and makes very few assumptions about the laser printer.

As our test vehicles, we used an HP LaserJet II, which is a 300-dpi write-black printer, and a Data Products LZR 2665, which is a 300-dpi write-white printer. As a measuring device, we used a Macbeth RD922 (Answer II) reflection densitometer. This densitometer measures the average reflectance over an area with a diameter of ~4 mm. The reason for choosing an instrument that measures the average reflectance of the test patterns, instead of one that measures the gray level of individual pixels, is that we do not have to estimate the transfer function of the instrument. Also, measurement of average reflectance does not require precise alignment of the measuring device with the printed patterns.

The purpose of a printer model is to predict the “gray” level of printed pixels based on the binary values of the surrounding pixels. The gray level of each printed pixel depends in a complicated way on the pixels in its neighborhood. According to the printer model of Eq. (6), the gray level produced at site \( (i,j) \) depends on the pixels in a window \( W_{ij} \) surrounding pixel \((i,j)\). Let \( W^1, W^2, \ldots, W^n \) denote all of the possible patterns for the window \( W_{ij} \). Then, the mapping \( \mathcal{P} \) assigns a gray level \( p^k \) to each window pattern \( W^k \), an example of which is shown in Fig. 8. Thus, the printer model is specified by the set of parameters \( \{ p^1, p^2, \ldots, p^p \} \).

The first step in the procedure for obtaining the printer model parameters is to determine the size and shape of the

---

**Figure 6** LSMB halftoning.

**Figure 7** Frequency and impulse response of the selected eye model.

**Figure 8** Printer model.
window $W_{ij}$. In the following discussion, we consider a $3 \times 3$ square window. Once the window size and shape are selected, we must determine the function $\mathcal{P}$ for all possible window patterns $W^k$. Thus, for each window pattern $W^k$, we must find the corresponding gray level $p^k$. For a $3 \times 3$ window, the total number of window patterns that can be obtained is $n = 2^9 = 512$.

To determine the mapping $\mathcal{P}$, we must first relate the unknown printer parameters $p^k$ to reflectance measurements of various test patterns. Let $m^1$, $m^2$, ..., $m^l$ denote the gray levels that correspond to the measured reflectances of all the test patterns. Figure 9 shows an example of a test pattern. It is obtained by periodically repeating a $4 \times 4$ pattern. The basic pattern is repeated both vertically and horizontally. Since the densitometer measures the average reflectance over an area that contains several periods of the test pattern, the measured reflectance of the printed test pattern is equal to the reflectance of one period of the pattern. This, in turn, is equal to the average of the reflectances of the 16 pixels contained in one period of the pattern. Thus, according to the printer model we have

$$m^1 = \frac{1}{16} \sum_{j=1}^{4} \sum_{i=1}^{4} p_{kj}.$$  \hspace{2cm} (15)

However, each of the gray levels $p_{kj}$ corresponds to one of the unknown parameters $p^k$. The window patterns $W^k$ corresponding to two of those parameters are indicated in the figure by solid-line squares. Thus, we can obtain an equation relating the unknown parameters to a measurement.

With the preceding procedure, additional test patterns with the same and different periods can be generated, printed, and measured to provide more equations. Note that the period of the printed test patterns does not have to equal the size of the window of the printer model. In principle, the period of the test patterns can take any value. In practice, however, we choose test patterns with short periods. The main reason for avoiding test patterns with long periods is that the reflection densitometer can only measure reflectance over a finite area. If the period is too long, then relationships between the measurements and the unknowns cannot be established.

The unknown printer parameters can be obtained from the equations relating them to the measurements. However, to make the solution computationally tractable, we first consider various possibilities for reducing the number of unknown parameters. First, we can use symmetry. We assume that window patterns that are reflected or rotated are mapped to the same gray level $p^k$. This reduces the number of unknown parameters from 512 to 102. We can also assume that the gray level of a white dot surrounded by eight white dots is 0, and the gray level of a black dot surrounded by eight black dots is 1. This eliminates two more unknown parameters.

Further reduction in the number of unknown parameters can be based on some additional understanding of printer behavior. In write-black printers, black dots of ink tend to spread beyond the pixel boundaries. It is thus reasonable to assume that the gray value corresponding to all window patterns with a "1" at the center is 1. Conversely, in write-white printers, white dots tend to spread beyond the pixel boundaries. Thus, we can assume that the gray value corresponding to all window patterns with a "0" at the center is 0. This reduces the number of unknown parameters to 50.

All of these assumptions are based on some understanding of the printing mechanism. They help reduce the number of unknowns and thus make the problem tractable. However, they also limit the applicability of the approach, e.g., to write-black and write-white printers. Fortunately, most laser printers fall in one of these two categories and also satisfy the symmetry constraints. Also, for reasons we will see later, these assumptions also help to maintain stability in the modified error diffusion algorithm.

The $3 \times 3$ window we considered earlier captures the distortions caused by the pixels in the immediate neighborhood of each printed pixel. However, more distant pixels may also have an effect. Thus, we considered printer models with different window sizes and shapes. The alternative window shapes we considered are shown in Fig. 10. The procedure for generating test patterns and equations for these windows is similar to the procedure outlined earlier for the $3 \times 3$ square window. The only difference is in the number of unknowns and the number of equations involved. For the $3 \times 3$ square window, there are $n = 2^9 = 512$ possible patterns. For the cross-shaped windows with lengths 3 and 5, there are $n = 2^5 = 32$ and $n = 2^9 = 512$ possible patterns, respectively. Finally, for the $5 \times 5$ square window, the number of unknowns is $n = 2^{25} = 33,554,432$. In principle, this window should give the best results. Unfortunately, however, the number of unknowns is too large and makes the problem intractable.
For each window size and shape, we selected test patterns with different periods to generate equations relating the unknown parameters to the reflectance measurements. For the $3 \times 3$ square window, we used test patterns with periods $2 \times 2$, $2 \times 3$, $3 \times 3$, and a subset of the test patterns with period $4 \times 4$. We generated all the possible test patterns for each period and eliminated redundant equations by comparing the coefficients of the equations. This resulted in a total of 200 test patterns for the $3 \times 3$ window. For the $5 \times 5$ cross-shaped window, test patterns with periods $2 \times 2$, $2 \times 3$, $3 \times 3$, and a subset of the test patterns with period $5 \times 5$ were used. The total number of equations generated was 244.

The equations can be placed in matrix form $A^*P = M$, where $M$ is the vector of measurements and $P$ is the vector of unknown printer model parameters that correspond to the set of window patterns. This linear system cannot be solved directly, because of two potential problems: a possible rank deficiency of the matrix $A$ and discrepancies in the measurements.

The first problem, where the matrix $A$ is not full rank, occurs when not enough independent equations can be generated. It is, in fact, difficult to obtain enough independent equations to guarantee that the matrix $A$ has full rank. For the $3 \times 3$ square window, 3 of the singular values of the matrix are zero. It is our speculation that this problem is caused by the fact that some sets of unknown parameters always appear together in the printed test patterns.

The second problem is that there are discrepancies in the measurements. These are caused by measurement noise and modeling errors. Modeling errors occur when the window size is too small to include all of the pixels that affect the reflectance of a printed pixel. For instance, for the $3 \times 3$ square window, it is quite possible that dots that are two pixels away from a pixel affect its gray value.

The first problem suggests that we need some additional constraints, and the second problem suggests that we need to find a best fit to the measurements. A reasonable set of constraints on the parameter vector $P$ is that its components $p^i$, which are gray-scale values, must be between 0 and 1. We can then minimize the error between the reflectance measurements of the test patterns and the reflectance predicted by the parameter vector $P$. The minimization is done in the least-squares sense. Thus, we arrive at the following constrained optimization problem:

\[
\text{minimize: } ||A^*P - M||^2_2
\]

subject to:

\[
\begin{align*}
0 & \leq p^1 \leq 1 \\
0 & \leq p^2 \leq 1 \\
& \vdots \\
0 & \leq p^n \leq 1
\end{align*}
\]

To solve the constrained optimization problem, we used an optimization routine by Fletcher and Harwell from the Harwell Subroutine Library. It is based on Davidon’s method, which uses an approximation to the inverse Hessian matrix. The linear inequality constraints are dealt with by projection techniques. An initial estimate of the solution that satisfies the constraints must be provided. At each iteration, a new estimate is found that satisfies the constraints and produces a lower error value. The algorithm continues for several iterations until it cannot move toward a point with a lower error value.

If there is a unique minimum, then the algorithm converges to that point. Unfortunately, several local minima are often present in the solution space; in such cases, it is not always possible to determine the global minimum. However, if the matrix $A$ is full rank, then a unique minimum exists in the solution space, and the constrained optimization routine returns the optimal solution. When $A$ is not full rank, there are two possibilities. If enough of the constraints become active, then a unique solution vector exists and the algorithm converges to that vector. Otherwise, there may be several local minima, and the starting point determines which solution vector is returned by the algorithm. Even though this can happen in practice, we have found that the performance of the printer model is not strongly affected by which solution is used to construct our printer model.

For the printer model with window size $3 \times 3$, the matrix $A$ is not full rank, even when we make all the assumptions we made earlier to reduce the number of unknown parameters. For the cross-shaped window with length 3, the matrix $A$ is full rank, while for the cross-shaped window with length 5 it is not.

As we will see in the next section, the best results were obtained with the $3 \times 3$ measurement model. A reasonable starting point is the vector of parameters given by the circular dot-overlap model. Typically, the solution vector satisfies all the constraints and the residuals are very small.

### 4.1 One-Dimensional Models

We also considered 1-D printer models. Such models can be used with 1-D halftoning techniques, whereby each row or column of the image is halftoned independently. One-dimensional halftoning is seldom used in practice because it does not exploit the 2-D properties of the eye.

There are many reasons for studying the one-dimensional problem, even though it is of limited practical significance. First, the problem is simpler in one dimension in terms of the number of unknown parameters and measurements involved. It is thus easier to test the various assumptions that we make and to look for potential problem areas. Second, the model can be tested using the Viterbi algorithm to obtain the optimal solution to the 1-D LSMB halftoning problem.

The solution of the 2-D least-squares problem is obtained by iterative techniques and is not guaranteed to be optimal. We thus have to rely on approximate solutions to the least-squares problem and on the modified error diffusion algorithm in order to test the 2-D model.

We considered 1-D models with windows of sizes 3, 5, and 7. For the 1-D printer models, the matrix $A$ becomes full rank if we assume that the gray value of all window patterns with “1” at the center is 1. This assumption is valid for write-black printers only. For write-white printers, the matrix $A$ becomes full rank when the gray value of all window patterns with “0” at the center is assumed to be 0.

### 5 Application to Model-Based Halftoning

In this section, we evaluate the performance of the measurement approach. In our experiments, we used the HP LaserJet II write-black printer and the Data Products LZR 2665 write-white printer as our test vehicles. The characteristics of these
Fig. 13: CD, MED (1 iteration)
Fig. 14: MM, MED (1 iterations)

Fig. 15: CD, MED (5 iterations)
Fig. 16: MM, MED (5 iterations)

Fig. 17: CD, LSMB (5 iterations)
Fig. 18: MM, LSMB (5 iterations)

Fig. 19: 2D halftoning

Fig. 20: 1D halftoning
two printers are very different. As our measurement device, we used the Macbeth RD922 (Answer II) reflection densitometer. Based on the procedures outlined in Sec. 4, we estimated the parameters of each printer model and then used them with the modified error diffusion and the LSMB halftoning algorithms.

Our initial experiments produced printed images that did not have the right gray level. This is because the measured reflectance of the printed images does not necessarily correspond to the perceived gray level. On the other hand, our experience indicates that the perceived gray level is proportional to the amount of ink on the paper. For example, a gray-scale ramp halftoned using the “classical” screen, which is robust to printer distortions, is perceived as having a linear variation of gray level. Accordingly, we need to calibrate the measured reflectances. To establish a relationship between measured reflectance and perceived gray level, we printed a gray-scale ramp on the HP LaserJet, which is a 300-dpi printer with a fair amount of dot overlap. To minimize the effects of the overlap, we used a 4 × 8 classical screen at a resolution of 100 dpi. The 100-dpi resolution was simulated by pixel repetition. The 4 × 8 classical screen produces 33 different patterns with gray levels that increase approximately linearly. Figure 11(a) shows a comparison of the reflectance values obtained from the densitometer versus the expected reflectance (assuming that 1 — perceived gray level is proportional to reflectance). Observe that the measured and expected values are different. Note that the measured reflectance of the solid black areas is not zero, yet they are still perceived as black. The measured reflectance of most of the patterns is lower than expected. One of the reasons for this discrepancy could be the effects of multiple internal reflections. The amount of multiple internal reflections depends, among other factors, on the intensity of the incident light. The incident light used by the densitometer differs substantially from the light used for normal viewing of the printed images.

The graph of Fig. 11(a) can be used to “correct” the measurements of the densitometer, that is, to account for the discrepancy between the measured reflectance and the perceived gray level of the printed images. In fact, one can use the reflectance density measurements directly for this calibration. The measured reflectance density, shown in Fig. 11(b), is defined as the logarithm to the base 10 of the reciprocal of the reflectance factor. The results of the model-based halftoning techniques indicate that this calibration eliminates any bias in the printed images.

We now examine the performance of our measurement models in model-based halftoning. The model-based halftoning algorithms were tested on several images. The test images included “Lena” and a gray-scale ramp. The resolution of “Lena” is 512 × 512 pixels. We used an interpolation scheme consisting of an expander (Ref. 18, pp. 105—109) and an equiripple low-pass FIR filter (Ref. 18, pp. 462—480) to obtain a 1024 × 1024 image. The resolution of the gray-scale ramp is 1200 × 90 pixels. Figures 13 through 20 were printed on the write-black HP LaserJet II. Only a section of the halftoned “Lena” image is shown in these figures.

5.1 Modified Error Diffusion

As we mentioned in Sec. 3, the addition of the printer model can affect the stability of the algorithm. For example, for a write-black printer, if the assumption that all printed ones take on the gray level of 1 is relaxed, then the algorithm becomes unstable and diverges, as can be seen in Fig. 12. This figure shows the result of the modified error diffusion algorithm applied to a gray-scale ramp when the above assumption is satisfied in (a) and when it is not in (b). The figure has been magnified by a factor of 6. Similarly, for a write-white printer, if the assumption that all printed zeros take on the gray value of 0 is relaxed, then the algorithm becomes unstable.

To prevent instability, tight constraints on the measurement model are necessary. One set of constraints is the assumption discussed above, namely, that all ones have a gray value of 1 for write-black printers and all zeros have a gray value of 0 for write-white printers. As we saw in the previous section these assumptions also help reduce the number of unknowns and the degrees of freedom. In addition to these assumptions, the initial background of the output image must
be fixed to specified values. For write-black printers, we
found that the initial background should be white. For write-
white printers, the initial background should be black. Both
of these assumptions were true in Ref. 1 and, thus, no insta-
bilities were observed. However, the initial background as-
sumption causes some bias in the algorithm, as we see in the
following.

As an alternative approach to eliminate the instability
problem, we considered the multipass modified error diffu-
sion algorithm, which was introduced in Sec. 3. As we saw
in Sec. 3, the modified error diffusion algorithm produces a
bias in the gray scale of the printed images. When the mea-
surement model constraints of the previous paragraph are
satisfied, this bias is small and almost imperceptible. Other-
wise, the bias may cause the modified error diffusion algo-
rithm to diverge. The multipass algorithm can (but is not
guaranteed to) eliminate the instability problem.

Assuming no instability, the multipass algorithm produces
the correct gray scale. The difference with the one-pass al-
gorithm is very small and appears to be insignificant. How-
ever, when the multipass error diffusion result is used as an
initial estimate for the least-squares approach, the resulting
error is much lower than that obtained when the one-pass
result is used as an initial estimate. More importantly, the
image with the lower error preserves the (visually pleasant)
error diffusion texture, while the image with higher error is
very grainy. This is because of the bias of the one-pass mod-
ified error diffusion algorithm. As the iterative least-squares
algorithm tries to modify the binary image to restore the
correct gray level, it destroys the error diffusion patterns.
The multipass algorithm also eliminates some of the low-
frequency artifacts that the one-pass algorithm produces.
However, sometimes, the images produced by the multipass
algorithm appear to be slightly grainier than those produced
by the one-pass algorithm.

Figures 13 and 14 show the results of modified error dif-
fusion (MED) with the circular dot-overlap (CD) and mea-
surement models (MM), respectively. The best results were
obtained with the 3 x 3 measurement model. Figures 15 and
16 show the result of multipass modified error diffusion (five
iterations) with the two models. Figure 19 compares the per-
formance of the two printer models on a gray-scale ramp.
The first ramp in Figure 19 was halftoned with a 4 x 8 clas-
sical screen (CL). Since the classical screening technique is
fairly robust to printer distortion, the 2-D classical ramp is
used as a reference for the performance of the model-based
techniques. The second and third ramps were halftoned using
the five-pass modified error diffusion (MED-5) with the cir-
cular dot-overlap and measurement models, respectively. Fi-
ally, the fourth and fifth ramps were halftoned using the
LSMB approach with the two models.

The preceding figures, printed on the write-black HP
LaserJet II, show that the performance of MED with our
measurement model is quite good. The rendition of gray scale
is relatively accurate, and the performance of error diffusion
(sharpness and few low-frequency artifacts) is maintained.
In fact, the performance of the circular dot-overlap model
and the measurement model on the write-black printer is
comparable. We also used a Data Products LZR 2665 write-
white printer. The circular dot-overlap model does not apply
to this printer and, thus, the measurement model must be
used. Our experiments showed that our procedure for esti-
mating printer model parameters works on this printer as well
as it does on the write-black printer.

Our experiments also indicated that the performance of
the measurement model is not strongly influenced by which
starting point is used in the optimization algorithm to obtain
the vector of printer model parameters. The difference in
performance of MED when the models corresponding to two
different starting points are used is small. In the examples
of this section, we used as a starting point the vector of parameters given by the circular dot-overlap model with $\rho = 1.25$.

5.2 Least-Squares Model-Based Halftoning

The 2-D LSMB algorithm is iterative and produces images that may be only local optima of the least-squares problem. Here we demonstrate the approach that updates only one pixel at a time, using the multipass MED result as an initial estimate. In our experiments, we found that this version gives the best results.

Examples made with this technique are shown in Figs. 17 and 18 using the circular dot-overlap and measurement models, respectively. In these examples, the initial estimate for the least-squares algorithm is the output image from five iterations of error diffusion, shown in Figs. 15 and 16. The number of iterations for the least-squares method is 5. Similarly, the fourth and fifth ramps in Fig. 19 were halftoned using the LSMB technique with the circular dot-overlap and measurement models, respectively.

The images produced using the least-squares method are sharper than those of the MED algorithm. Also, the least-squares approach maintains the texture of error diffusion, which is known to be visually pleasant. This happens only when the multipass MED result is used as the initial estimate. Using an all-white image (or even the one-pass MED) as the initial estimate results in images that are very grainy.

The preceding examples demonstrate that the gray-scale rendition of the halftoned images is relatively accurate. The clipping effect observed at areas where the images are nearly white is due to the fact that the LSMB algorithm avoids placing the halftone dots too far apart, because the LSMB error criterion interprets them as individual dots rather than an average gray level. The clipping point depends on the viewing distance, i.e., the width of the eye filter.

We have thus demonstrated that when our measurement model is used with the two model-based techniques, it produces the correct gray scale and also maintains the overall performance of these techniques.

5.3 One-Dimensional Least-Squares Model-Based Halftoning

In 1-D halftoning, each row or column of the image is halftoned independently. As mentioned previously, one-dimensional LSMB halftoning can be implemented using the Viterbi algorithm to obtain the globally optimum solution. As we saw in Sec. 3, the Viterbi algorithm cannot be used to solve the 2-D problem. Since only approximate solutions to the 2-D problem can be obtained, the 1-D problem offers a unique opportunity to isolate the performance of the printer models from that of the optimization algorithm.

To test the accuracy of gray-level rendition, we used a gray-scale ramp. The first ramp in Fig. 20 was halftoned with a 2-D classical screen. Since the classical screening technique is fairly robust to printer distortion, the 2-D classical ramp is used as a reference for the performance of the 1-D techniques. The second ramp in Fig. 20 was produced using the Viterbi algorithm and the circular dot-overlap model with $\rho = 1.25$. The third, fourth, and fifth ramps were produced using the Viterbi algorithm and measurement models with window sizes 3, 5, and 7, respectively (W3, W5, and W7).

The last ramp shown in Fig. 20 was halftoned using a 1-D equivalent of the classical screen. The resolution of each gray-scale ramp is $1200 \times 74$ pixels.

The gray scale of the ramp that was generated with the circular dot-overlap model is not monotonic, as expected from the discrepancies between the predictions of the model and the measured values demonstrated in Table 1. The problem appears near the top of the ramp. The gray-scale rendition of the ramp that was generated with the measurement model with window size 3 is considerably better than that produced with the circular dot-overlap model. However, it is still not perfect. There is still a minor problem with the nonmonotonicity of the tone scale. In the gray-scale ramps produced by means of the models with window sizes 5 and 7, there is very little improvement over the gray-scale ramp produced with the window of size 3. This indicates that a measurement model with window size 3 captures the most significant printer distortions in one dimension. Thus, there is no need to use larger window sizes to construct our measurement models. This result can also be used to argue that the $3 \times 3$ window size is sufficient for constructing the 2-D measurement models.

The 1-D comparisons demonstrate that the measurement approach produces fairly accurate gray-scale rendition and that LSMB halftoning with a printer model can generate more gray levels than classical screening. Of course, the other advantage of model-based techniques, namely, that they produce sharper images, is not obvious here since we have a smooth ramp image.

6 Conclusions

We presented an approach for estimating printer model parameters based on direct measurements of the reflectance of test patterns. The printer models predict printer distortions and are an essential part of model-based halftoning techniques that depend on accurate printer models to produce high-quality images with standard laser printers. We considered the problem of estimating the parameters of both 1-D and 2-D printer models. We also derived an iterative (multipass) version of the modified error diffusion algorithm that improves its performance.

Even though it is of limited practical significance, the 1-D problem is simpler, and the model can be tested with the Viterbi algorithm to obtain the solution to the 1-D LSMB halftoning problem. Our results indicate that the measurement model performs better than the circular dot-overlap model, which assumes that the printer produces circularly shaped dots of ink.

The performance of 2-D printer models was tested using two model-based halftoning techniques, the modified error diffusion and the least-squares method. Our results using a write-black printer indicate that the performances of the measurement model and the circular dot-overlap model are comparable. The measurement model was the only one that could be applied to a write-white printer, however. We thus demonstrated the necessity of a general approach that can be applied to a wide variety of laser printers.

*The reader is cautioned that the Mach band effect may make the halftoned ramp look more nonmonotonic than it is.
References


20. Thrasyvoulos N. Pappas received the SB, SM, and PhD degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1979, 1982, and 1987, respectively. He has been with the Signal Processing Research Department at AT&T Bell Laboratories, Murray Hill, New Jersey, since 1987. His research interests are in image processing and computer vision. Recently, he has been working on image halftoning and compression based on models of display devices and the human visual system.

Chen-Koung Dong received the SB and SM degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1992. He also received the SB degree from the MIT School of Management in 1992. From June through December 1991 he was a consultant at AT&T Bell Laboratories, Murray Hill, New Jersey, in the Signal Processing Department. He is currently working as a financial analyst at the First Boston Corporation, New York. His fields of interest include image processing, digital halftoning, computer networks, and financial markets.

David L. Neuhoff received a BSE from Cornell University in 1970 and an MS and PhD in electrical engineering from Stanford University in 1972 and 1974, respectively. In 1974, he joined the University of Michigan, Ann Arbor, where he is currently a professor of electrical engineering and computer science. From 1984 to 1989 he was associate chairman of the Systems Science and Engineering Division of the department. From September 1989 through June 1990 he was on leave at AT&T Bell Laboratories, Murray Hill, New Jersey, in the Signal Processing Department. His research and teaching interests are in communications and information theory, especially data compression, quantization, Shannon theory, coding for magnetic recording, halftoning, and constrained coding. Dr. Neuhoff is a member of the Eta Kappa Nu, Tau Beta Pi, and Sigma Xi, and a senior member of the IEEE. He is an associate editor for source coding for the IEEE Transactions on Information Theory.