The mathematics of halftoning

R. L. Adler
B. P. Kitchens
M. Martens
C. P. Tresser
C. W. Wu

This paper describes some mathematical aspects of halftoning in digital printing. ***Halftoning*** *is the technique of rendering a continuous range of colors using only a few discrete ones. There are two major classes of methods:* ***dithering*** *and* ***error diffusion****. Some discussion is presented concerning the method of dithering, but the main emphasis is on error diffusion.*

Introduction

The problem of halftoning in digital printing has provided a splendid new example of a mathematical opportunity in modern technology. *Digital halftoning* is the technique used to display an image with a few immiscible colors discretely applied to paper. This problem involves various fields of science such as the physics of light, the biology of the human visual system, and the mathematics of analog-to-digital conversion. Digital halftoning can be considered a huge optimization problem, but we shall see that by sacrificing optimality one can construct efficient algorithms which achieve very satisfactory results. Halftoning is above all an art, yet mathematics, particularly the theory of dynamical systems, helps to improve this art, while the art suggests a rich collection of mathematical problems and insights.

Specifically, we address the question of how full-color images should be imitated by printers which print at positions on a lattice with colors limited to a few available choices. This question subsumes the simpler, more restrictive one of getting continuous-tone gray-scale images from black and white dots. Black and white digital halftoning was initially created for television well before digital printers! We call the available colors *pixel colors* and the locations on the lattice *pixel locations*. Purists reserve the term *pixels* for screens and instead use *pels* for printing, but we shall stick to pixels. *Pixel colors* are for us the colors of inks or toners. In some printers the choice of pixel colors can be augmented by superposition of these substances, but this is not allowed in highlight printers. The two images of Brian Wu (see **Figure 1**) illustrate halftoning for gray scale: Figure 1(b) is a coarse-grain halftone image of the fine-grain image in Figure 1(a), which is approximately a full gray-scale image.

The halftone image in Figure 1 was made by a clustered dither mask, which is discussed later. It was modified to accommodate artificially large pixels. The mathematical problem of digital halftoning is to construct an algorithm whose output is a sequence of pixel colors that creates the full-color visual experience of an input image sequence. For gray-scale images, one wants discrete black and white dots to appear as continuous tones.

In discussing the mathematics of printing, one cannot avoid using the language of geometry. It has long been known that our perception of color can be modeled by vector addition in a color space, the dimension of which we call the *color dimension*. For rendering gray-scale images with black and white dots, the color dimension is 1; for the color space of full color, it is 3.

Colors and gamuts

Standard ink and toner colors are cyan (C), magenta (M), yellow (Y), and black (K). The letters in parentheses represent color vectors. Added to this set is white (W), taken for the color of paper. The primary colors for light are red (R), green (G), and blue (B). In color space we have the following relations:

$$W = R + G + B,$$

$$M = W - G = R + B,$$

$$C = W - R = G + B,$$

$$Y = W - B = R + G.$$

Since inks and toners act as light absorbers, only blue light is reflected from a combination of magenta and cyan on white paper. Similarly, only red light is reflected from magenta and yellow, and green from cyan and yellow.

We assume that the coefficient of the color vector of inks is either 0 or 1, although the use of more levels is on the

0018-8646/03/$5.00 © 2003 IBM

(a) (b)

Figure 1

(a) Gray-scale image. (b) halftone image.

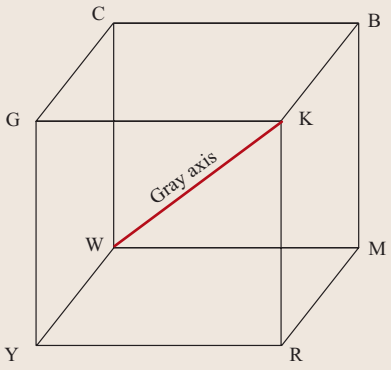

Figure 2

Cubic color gamut.

rise, and 16 (4 bits of information per pixel) is typically used in the new high-end machines.

Vector addition is used to describe the perceived color generated by a combination of light rays emanating from a single source. However, a printed image is made up of many sources, each emitting its own pixel color. The eye averages received light over areas consisting of many pixels. This means that we do not describe the perceived color in a printed image by taking positive linear combinations of color vectors at individual pixels, but rather by convex combinations.

Given the set of pixel colors available to a printer, the set of convex combinations of these pixel color vectors is called the *printer gamut*, or simply *gamut*. The geometrical significance of this is that the gamut is a polytope in color space (defined in the subsection on color error diffusion). The printer gamut is a subset of the gamut of the human eye (see the Commission Internationale de l'Eclairage (CIE) colorimetric system in [1]), which is a much more complicated convex body in color space [2].

When all eight colors C, M, Y, R, G, B, K, and W are available to a printer and arranged in the cube, as in **Figure 2**, the Cartesian product structure of the cube allows for significant mathematical simplification in printing, as follows:

1. Each component of a color vector along the C, M, Y axes is treated separately with W in the same way one would deal with a black and white image. The page is overprinted three times, one for each pixel color, and possibly a fourth time in order to print as much K as possible (black ink or toner is cheaper, and K, W render better grays than C, M, Y, W).
2. When overprinting the sheet multiple times, one must take care of interference (Moiré) patterns due to misalignment in the superposition of colors, a phenomenon which is now well understood (see for instance [3, 4] and the discussion at the end of the subsection on clustered masks).

Actually, the eight color vectors form the vertices of a six-sided figure which is far from cubic in any perceptual color space such as the one given by the CIE colorimetric system. Furthermore, constraints such as the impossibility of superimposing inks to get R, G, or B, or the addition of other shades to augment printer gamuts, such as "light magenta" and "light cyan," also lead to significant departures from the cube.

Index dimension

While printing is done in two dimensions, we define an *index dimension* as the dimension of the coordinate system used to specify pixel locations. If pixel locations are simply ordered, the index dimension is 1, but if locations are given by a pair (i, j) of integers, the index dimension is 2, and so on. From the point of view of dynamical systems, color dimension can be thought of as a phase-space dimension, while index dimension is that of time: This is indeed our mindset in treating the digital halftoning problem by using an algorithm which corresponds to running a nonautonomous dynamical system.

Suppose over a region R an image has input color vectors $\gamma(\iota)$ at pixel locations ι, where ι specifies an index of some given index dimension; and we print pixel color vectors $\rho(\iota)$. The total error $\varepsilon(R)$ made over the region R is

$$\varepsilon(R) = \sum_{\iota \in R} [\gamma(\iota) - \rho(\iota)] . \quad (1)$$

The question as to whether a uniform bound on $\varepsilon(R)$ exists has a negative answer for two-dimensional regions R. We recall that the uniform distribution theory developed by T. van Aardenne-Ehrenfest, H. Davenport, K. F. Roth, and others (see for example [5]) shows that it is not possible to bound the error over all possible two-dimensional rectangles using black and white pixels. This is in strong contrast to one-dimensional results, which show that these errors can be very well bounded for black and white printing [6–12]. Thus, mathematics seems to imply that, while there is no inherent difficulty in digital printing on a line, there may be one in doing it on a page.

Although we wish to make $\varepsilon(R)$ as small as possible, we do not require a uniform bound on this quantity. All that is really needed is that the average error $(1/N)\varepsilon(R)$, where N is the number of pixels in R, be small for all regions of reasonable size and shape. The eye's averaging is somewhat different, but the average given above is sufficient for our purposes. For more realistic models of the human visual system and the role they play in digital halftoning, see for instance [6, 7, 13–17].

One gets a sense of the size of the combinatorial optimization problem associated with digital halftoning from the fact that typical printers print 600 pixels per linear inch. For a modest size 4×4 image, this gives 5 760 000 pixels! A *sine qua non* for industrial applications is fast and cheap methods of deciding at each pixel location an output pixel color which gives good imitations of the input. Two classes of techniques are primarily used: *dithering* and *error diffusion*. Dithering makes decisions pixel by pixel on the basis of a threshold mask. Error diffusion is a dynamical system which makes decisions on the basis of a running error.

Dithering

Dithering is based on a completely local determination which is both simple and fast. We first consider the easier case of black and white (BW) printing, for which the color dimension is 1. Furthermore, we assume that the index dimension is 2. A dithering mask (mask for short) is specified by an $n \times m$ matrix M of threshold coefficients $M(i, j)$. The numbers $M(i, j)$ range over the unit interval in a uniform fashion. The image to be halftoned is given by an $h \times v$ matrix Γ of input gray levels $\Gamma(i, j)$, also numbers in the unit interval. Typically, n and m are much smaller than either v or h. The output image (or halftone image) is given by an $h \times v$ zero-one matrix, $\Phi = [\Phi(i, j)]$, which controls the printing of black as follows: At pixel location (i, j), black is printed if and only if $\Phi(i, j) = 1$. The matrix $\Phi(i, j)$ is defined by

$$\Phi(i,j) = \begin{cases} 1 & \text{if } \Gamma(i,j) > M(i \mod n, j \mod m), \\ 0 & \text{otherwise.} \end{cases}$$

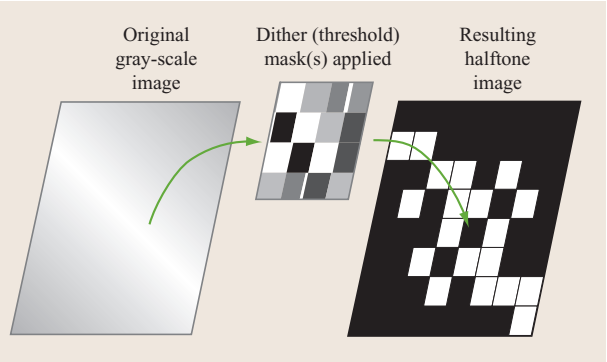

Figure 3

Dither mask.

The method of dithering for black and white printing is illustrated in **Figure 3**.

The problem of dithering is to arrange the numbers $M(i, j)$ in the mask so that good image quality results. This can be quite printer-dependent, at least when it comes to fine tuning. The more thresholds in the mask, the more continuity in the gray spectrum and the smoother the picture. A commonly used upper limit to the number of gray levels is 256. Masks are constructed so that constant gray images look good, the idea being that natural images, which are nearly constant in small areas, will then also look good. There is a price to be paid for the instant decision-making ability of a mask: Namely, it is always possible to construct improbable images on which it will fail badly. However, this does not happen for the usual pictures for which masks are designed.

Two main classes of masks

One usually distinguishes two classes of masks:

1. The *dispersed masks*, mostly used by ink-jet printers, which are slow but reliably print single ink dots.
2. The *clustered masks*, used by laser printers, which do not reliably print isolated dots.

Dispersed masks

One might think that numbers $M(i, j)$ should be placed in the mask as randomly as possible. However, this produces undesirable irregular clusters and other unpleasant defects. To avoid such shortcomings, one tries to place numbers in the mask so that $\Phi(i, j) = 1$ will be well dispersed for each gray level. Masks by their nature (dispersed or clustered) impose an *increasing gray-level constraint* (or *stacking constraint*); i.e., if some pixel is printed for a gray level, it is also printed for any darker

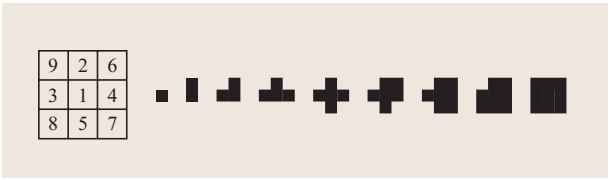

Figure 4

Dither mask and resulting halftone patterns.

one (see **Figure 4**). This constraint presents a serious difficulty in designing masks.

Another disturbing feature of an improperly constructed dispersed mask is that periods n and m of the mask dimensions can sometimes be detected. To avoid this, masks are often constructed so that the product $n \times m$ is much larger than the number of distinct threshold levels. Particularly popular in this respect are *blue noise* masks, so called because low frequencies in the power spectrum of the patterns turn out to be attenuated [18]. The most pleasing pattern for gray level 0.5 is the maximally dispersed checkerboard pattern, but too much regularity in the presence of noise is undesirable, so checkerboard patterns are avoided in most blue noise masks. Adler, Thompson, Tresser, and Wu based an invention [19] on the aperiodic Thue–Morse sequence[1] in order to tile a large mask by two much smaller ones. The memory required for this is considerably less than that for one large mask alone. This gives an alternative (or a complement) to blue noise masks, and can be used in the clustered mask case also.

Clustered masks

It is difficult to construct a good clustered mask when its size is not much larger than the number of gray levels to be rendered. Furthermore, there is a tradeoff between the number of gray levels to be rendered and the size of clusters: The bigger the cluster, the more levels the mask can render, but the more noticeable the cluster. Thompson, Tresser, and Wu, using ideas inspired by celestial and statistical mechanics, have addressed this difficulty in patents[2] [24–26]. The main idea behind the new masks is to start with a large mask consisting of several copies of a small mask with 256 levels. There are certain gray levels for which there is an ideal pattern: for example, the checkerboard for the 0.5 gray level. One attempts to keep these patterns and obtain the intermediate threshold values by interpolation. This involves a trial-and-error procedure in which one repetitively resets thresholds and judges print quality. This method is very versatile, and it can be used for a variety of situations: for example, to create several masks for machines whose behavior degrades between maintenances, to adapt a mask to a new printer that has been tuned for another, etc. Other inventions mentioned above concern variations on this theme. These ideas can also be used for the simpler problem of building dispersed masks.

For standard color dithering, four different masks are used, one for each nonwhite color (including K, which is applied first, since it is usually the cheapest ink). Using the same mask four times is usually avoided because of the possibility of misregistration that causes disturbing Moiré patterns. In the past, color printing was done with screens that performed like masks (indeed, the word *screen* is still used in lieu of *dithering mask*). To minimize Moiré patterns, the screens for separate colors were set at different angles. For digital printing, there is a technique in designing the different masks that is equivalent to the effect of rotating a screen. For a description of this method, see [4]. A pending patent by Rao, Thompson, Tresser, and Wu proposes to build what they call *semi-digital printers*: printers intermediate to digital and analog ones, in which each color plane is treated as in the usual digital printing, but the screens are rotated as in traditional analog printing.[3]

Calibration of dither masks

In constructing dither masks, it is usually assumed that the number of black pixels in a halftone pattern is proportional to the gray level. We refer to such dither masks as *linear dither masks*. This assumption is not true when nonlinear effects such as *dot gain* or *dot overlap* are introduced. Another nonlinear effect occurs when a quantity called *lightness* [4] is used to measure gray levels.

One way to compensate for nonlinear effects is to first apply a tone reproduction curve (TRC) to the input data and then use a linear dither mask, as shown in **Figure 5**. The TRC describes the relationship between input gray levels and output ones. Using 8-bit numbers to represent gray levels, the TRC as used in Figure 5 maps an 8-bit number to another one. Since the TRC is generally not the identity map, this means that the output of the TRC has less than 256 distinct values (see **Figure 6** for a 2-bit input/output example). Thus, after the application of such a dither mask, there will be fewer than 256 distinct halftone patterns, which reduces the number of renderable gray levels, especially when the nonlinearity is strong. However, for small nonlinearities it is the method of choice.

[1] The Thue–Morse sequence is a sequence of 0s and 1s formed by a series of concatenations. In each step, a new block is concatenated to an old one, the new block being formed by interchanging 0s and 1s in the old. The first four blocks of the sequence are 0, 01, 0110, 01101001, $\cdots$. The total sequence has the property that no subblock is repeated more than twice in succession [20–23].

[2] M. J. Stanich, G. Thompson, C. Tresser, and C. W. Wu, patent pending.

[3] R. Rao, G. R. Thompson, C. P. Tresser, and C. W. Wu, patent pending.

A way to compensate for strong nonlinearities is to directly modify thresholds in the mask, which has the advantage of preserving the number of renderable gray levels. Furthermore, since it obviates the need for a TRC, it can increase processing speed. The construction of such a nonlinear dither mask proceeds as follows. First, a series of good halftone patterns are chosen to be reproduced by the dither mask; this provides a set of data points. A curve, serving as a tone reproduction curve, is then obtained by interpolating, say, a spline curve between these data points. However, instead of using the curve as in Figure 5, the inverse of the curve is used to determine the number of black pixels needed for each gray level. This information is then used to generate a dither mask with correct output. In contrast to the linear dither mask, the difference in the number of black pixels between halftone patterns of successive gray levels is not constant, but is determined by the inverse curve. The stacking constraint requires that interpolation generate a monotonic curve. See [26] for more details. This method can be used in color printers in which multiple masks are used [27]. For related matter, see also [28].

Several of the new techniques for dithering that we have discussed have been employed in the halftone design of the IBM line of laser printers, ranging from the desktop models to production print machines.

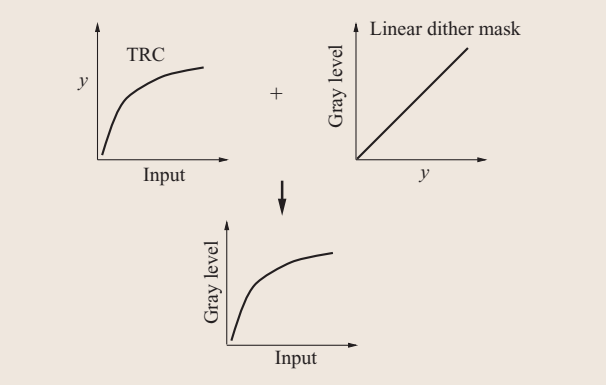

Figure 5

Use of tone reproduction curve (TRC) to compensate for nonlinear effects.

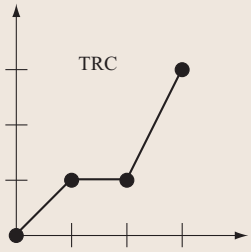

Figure 6

TRC with 2-bit input and output.

Error diffusion

For error diffusion, as before, we first consider the easier case of color dimension $d = 1$, e.g., black and white printing. In addition, we take the index dimension to be 1. Instead of Equation (1), we define the running error $\varepsilon(n)$ by

$$\varepsilon(n) = \sum_{k=1}^{n} [\gamma(k) - \rho(k)], \tag{2}$$

where, at pixel location k, $\gamma(k)$ is a gray defined proportionally to its darkness by a number from 0 for white to 1 for black, and $\rho(k)$ the printed pixel color, with $\rho(k) = 0$ for white and 1 for black. The running error satisfies the recursion

$$\varepsilon(n + 1) = \varepsilon(n) + \gamma(n + 1) - \rho(n + 1), \tag{3}$$

where $n = 0, 1, 2, \cdots$ and $\varepsilon(0) = 0$.

Simple error diffusion

Simple error diffusion is defined by taking $\rho(n + 1)$ to satisfy the greedy algorithm: Namely, $\rho(n + 1)$ takes the value 0 or 1, whichever minimizes $\varepsilon(n + 1)$ with a tie-breaking rule. It is easy to show that $\varepsilon(n)$ lies in the interval $[-\frac{1}{2}, \frac{1}{2}]$ under (3) for any choice of the sequence $\{\gamma(n)\}_{n\in\mathbb{N}}$. Thus, $\varepsilon(n)$ is bounded, and $\varepsilon(n)/n \to 0$.

Setting $x(n + 1) = \varepsilon(n) + \gamma(n + 1)$, we can rewrite Equation (3) as

$$\varepsilon(n + 1) = \varepsilon(n) + \gamma(n + 1) - \rho^*[x(n + 1)], \tag{4}$$

where $\rho^*[x(n + 1)]$ is the closest (with a tie-breaking rule) endpoint of the interval [0, 1] to $x(n + 1)$. (Since their domains of definition are different, we chose not to abuse notation by using ρ for ρ^*.) It is easy to see that $x(n)$ is trapped in the interval $[-\frac{1}{2}, \frac{3}{2}]$. Bounding $x(n)$ is equivalent to bounding $\varepsilon(n)$. Adding $\gamma(n + 2)$ to both sides of Equation (4) and reducing indices, we get

$$x(n + 1) = x(n) + \gamma(n + 1) - \rho^*[x(n)]. \tag{5}$$

The advantage of Equation (5) over Equation (4) is due to the fact that (5) can be interpreted as a time-dependent dynamical system, as follows. The orbits $x(n)$ of this system can be expressed recursively by $x(n) = f_{\gamma(n)}[x(n - 1)]$, where $\{\gamma(n)\}$ is a sequence of points in the unit interval

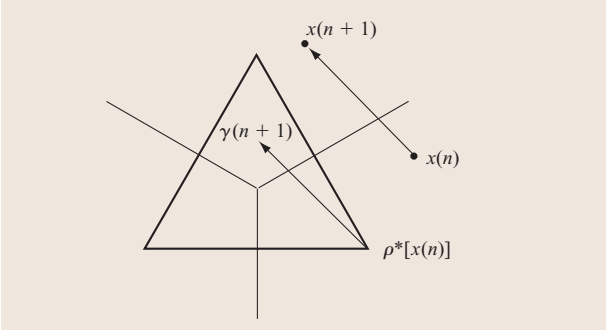

Figure 7

Two-dimensional error diffusion in action.

and $\{f_{\gamma(n)}\}$ is a sequence of mappings selected from a family of mappings $\{f_\gamma | \gamma \in [0, 1]\}$, each defined by

$$f_\gamma(x) = x + \gamma - \rho^*(x), \tag{6}$$

$\rho^*(x)$ being the closest endpoint (with a tie-breaking rule) of the unit interval to x. Trapping the orbits $x(n)$, and hence bounding the errors $\varepsilon(n)$, is equivalent to the following.

Theorem 1

There exists a bounded interval ($J = [-1/2, 3/2]$) *containing* $[0, 1]$ *which is invariant under all members* f_γ *of the family.*

Sturmian sequences

When the sequence $\{\gamma(n)\}_{n\in\mathbb{N}}$ is constant, $\gamma(n) \equiv \gamma_0$, Equation (6) taken mod 1 is a classical dynamical system, i.e., rotation on the circle by angle γ_0. It is easy to check that the sequence $\{\rho^*[x(n)]\}_{n\in\mathbb{N}}$ is a (non-exceptional) Sturmian sequence with the proportion of 1s equal to γ_0. The non-exceptional Sturmian sequences have the property that a given proportion of 0s and 1s are distributed as evenly as possible (the exceptional ones are those that are not generated by rotations) [29, 30]. This is remarkable considering that these well-distributed sequences are all generated by a single simple greedy algorithm. We are currently investigating a higher-dimensional generalization of this striking fact.

Color error diffusion

We now turn our attention to color printing. In color space, the printer gamut is a polytope P with pixel color vectors as vertices. A *polytope* is defined as the convex hull of a finite set of points: Equivalently, it is a compact intersection of a finite number of half spaces.

As we have mentioned, when all eight colors M, C, Y, R, G, B, K, W are used, replacing P with a cube gives reasonable results even though P is far from a cube. A generalization to simple error diffusion is *vector error diffusion*, in which $\varepsilon(n)$, $\gamma(n + 1)$, $x(n + 1) = \varepsilon(n) + \gamma(n + 1)$, and $\rho^*[x(n + 1)]$ of Equation (4) are vectors in 3-space (see also [31]).

The boundedness of $\|x(n)\|$ and $\|\varepsilon(n)\|$ is a corollary to the following d-dimensional generalization to Theorem 1, although we give a proof here for only $d = 2$.

Theorem 2

Let P *be an arbitrary polytope and* $\{f_\gamma | \gamma \in P\}$ *a family of mappings defined as in* (6) *by*

$$f_\gamma(x) = x + \gamma - \rho^*(x), \tag{7}$$

where now x *and* γ *are vectors in n-space,* $\gamma \in P$, *and* $\rho^*(x)$ *is the closest vertex of* P *to* x. *Then there exists a bounded convex set* $P' \supseteq P$ *which is invariant under any member* f_γ *of the family. Furthermore,* P' *can be constructed to be arbitrarily large*.

It is easy to give examples of sequences $\{\gamma(n)\}_{n\in\mathbb{N}}$ in which the $\gamma(n)$ are not in P such that $\{x(n)\}$ and hence $\{\varepsilon(n)\}$ diverge. **Figure 7** illustrates the geometrical nature in two dimensions of Equation (7) with triangular P.

By taking flatter and flatter triangles, one can check that the ratio of the diameters of P' and P can be arbitrarily large. This can be taken as an indication of the non-triviality of the general problem.

Notice that we have considered Euclidean distance only in Theorem 2, so that only the shape of P counts. For some important choices of P such as a cube with faces parallel to the coordinate planes, other norms such as the max or the sum of the absolute values of the coordinates might be useful as well.

Notation

We adhere to the convention of writing the inner product of vectors as a dot product.

Let v_i, $i = 0, \cdots, n - 1$, $n \geq 2$, be the vertices of a convex polygon P and n_j, $j = 1, \cdots, n$, the unit normal vectors to edges $v_{j+1}v_j$ of P. Let P_t denote the polygon generated by moving the edges of P perpendicularly outward a distance t: i.e.,

$$P_t = \bigcap_j \{x : x \cdot n_j \leq d_j + t\},$$

where $P = P_0$. The *Voronoi region* R_{v_i} of vertex v_i is the set of points closer to v_i than to any other vertex. It is defined as

$$R_{v_i} = \bigcap_j \{x : \|x - v_i\| \leq \|x - v_j\|\}.$$

For $\gamma \in P$ we define the mapping $f_\gamma : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$f_\gamma(x) = x + \gamma - v(x), \tag{8}$$

where $v(x) = v_i$, the index i being the smallest for which $x \in R_{v_i}$.

Lemma 1

If $n_i \neq n_0$ and $n_i \neq n_1$, then either $(v_2 - v_1) \cdot n_i > 0$ or $(v_0 - v_1) \cdot n_i > 0$. Furthermore, if $(v_2 - v_1) \cdot n_i > 0$, then (n_1, n_i) is a basis, and if $(v_0 - v_1) \cdot n_i > 0$, then (n_0, n_i) is a basis of $\mathbb{R}^2$.

Proof

It is clear from **Figure 8** that n_i can be written as $n_i = an_0 + bn_1$, where either $a < 0$ or $b < 0$. Suppose $a < 0$; then $(v_2 - v_1) \cdot n_i = a(v_2 - v_1) \cdot n_0$. Since $v_1 \cdot n_0 = d_0$ and $v_2 \cdot n_0 < d_0$, it follows that $(v_2 - v_1) \cdot n_0 < 0$, and thus $(v_2 - v_1) \cdot n_i > 0$. Furthermore, since $a \neq 0$, n_i is not parallel to n_1, and thus (n_1, n_i) form a basis. The case in which $b < 0$ is similar. □

Proof [Proof of Theorem 2]

It suffices to show that for t large enough,

$$f_\gamma(P_t \cap R_{v_i}) \subset P_t$$

for all $\gamma \in P$.

Let $x \in P_t \cap R_{v_1}$. Then $f_\gamma(x) \cdot n_0 = (x + \gamma - v_1) \cdot n_0 = x \cdot n_0 + \gamma \cdot n_0 - v_1 \cdot n_0$. Since $v_1 \cdot n_0 = d_0$ and $\gamma \cdot n_0 \leq d_0$, it follows that $(x + \gamma - v_1) \cdot n_0 \leq x \cdot n_0 \leq d_0 + t$. Similarly, $(x + \gamma - v_1) \cdot n_1 \leq x \cdot n_1 \leq d_1 + t$. Let $n_i \neq n_0$, $n_i \neq n_1$. Without loss of generality, we assume that by Lemma 1 $(v_2 - v_1) \cdot n_i > 0$ and (n_1, n_i) is a basis. Clearly, $(n_1 \cdot n_i)^2 < 1$. Let $x = c_1 n_1 + c_2 n_i$. Then $c_1 + c_2 n_1 \cdot n_i = x \cdot n_1$ and $c_1 n_1 \cdot n_i + c_2 = x \cdot n_i$. Solving for c_2, we obtain

$$c_2 = \frac{1}{1 - (n_1 \cdot n_i)^2} [x \cdot n_i - (n_1 \cdot n_i) x \cdot n_1].$$

Therefore, $(v_2 - v_1) \cdot x = c_2(v_2 - v_1) \cdot n_i = \alpha[x \cdot n_i - (n_1 \cdot n_i) x \cdot n_1]$, where

$$\alpha = \frac{(v_2 - v_1) \cdot n_i}{1 - (n_1 \cdot n_i)^2} > 0.$$

Since $x \in R_{v_1}$, $|x - v_1|^2 \leq |x - v_2|^2$. This implies that $(v_2 - v_1) \cdot x \leq (1/2)(|v_2|^2 - |v_1|^2)$, and therefore $x \cdot n_i - (n_1 \cdot n_i) x \cdot n_1 \leq (1/2\alpha)(|v_2|^2 - |v_1|^2)$, which implies that

$$x \cdot n_i \leq |n_1 \cdot n_i|(d_1 + t) + \frac{1}{2\alpha}(|v_2|^2 - |v_1|^2),$$

provided $d_1 + t > 0$. Since $(1/2\alpha)(|v_2|^2 - |v_1|^2)$ is independent of t and $|n_1 \cdot n_i| < 1$, it is clear that $x \cdot n_i \leq d_i + t$ for large enough t. Therefore, $x \in P_t$, and the proof is complete. □

Remark

The above method works for $d = 1, 2$ but will not work for $d \geq 3$. Moving all faces outwardly a fixed perpendicular distance is doomed to failure. A simple tetrahedral counterexample can be constructed in which one of the vertices passes out of its Voronoi region. A new idea for an invariant set is needed. We deal with this in a subsequent work. However, the theorem can easily be proved for special polytopes P such as the cube or the regular simplex in any dimension. Also, Adler et al. [6, 7] have proved the existence of an error bound in vector error diffusion in all dimensions under a rule different from the greedy algorithm.

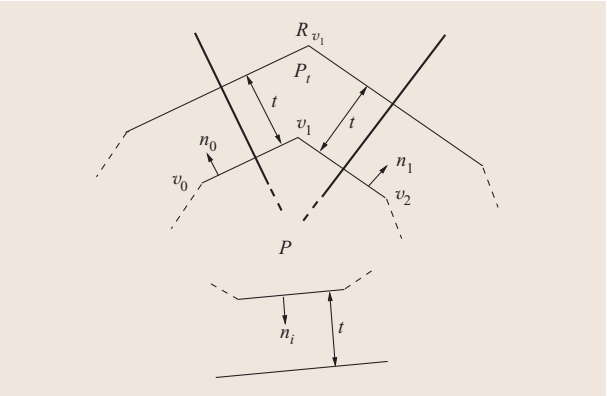

Figure 8

Polygons P, P_t and Voronoi region R_{v_1}.

Beyond simple error diffusion

In simple error diffusion, only the error accumulated up to the last pixel is explicitly taken into consideration. Instead, one can take into account several past errors and weight them. In practice, the system of weights is often taken as a probability vector (non-negative entries that sum to one), which is not necessarily constant, as follows:

$$\varepsilon(n+1) = \left[\sum_{i=0}^{m} w_i(n) \varepsilon(n-i) \right] + \gamma(n+1) - \rho(n+1), \tag{9}$$

where $w_i(n) \geq 0$, $\Sigma\, w_i = 1$, and $\rho(n+1)$ is the greedy algorithm which minimizes $\varepsilon(n+1)$. Setting

$$x(n+1) = \left[\sum_{i=0}^{m} w_i(n) \varepsilon(n-i) \right] + \gamma(n+1), \tag{10}$$

we have

$$\varepsilon(n) = x(n) - \rho(n); \tag{11}$$

thus,

1	3	5	3	1
3	5	7	5	3
5	7	pixel(k)		

(a)

1	2	4	2	1
2	4	8	4	2
4	8	pixel(k)		

(b)

Figure 9

(a) Jarvis scheme; (b) Stucki scheme.

Figure 10

Image by error diffusion.

$$x(n+1) = \sum_{i=0}^{m} w_i(n)[x(n-i) - \rho^*[x(n-i)] + \gamma(n+1)]$$

$$= \sum_{i=0}^{m} w_i(n) f_{\gamma(n+1)}[x(n-i)], \qquad (12)$$

where $w_i(n) \geq 0$, $\Sigma\, w_i = 1$, and $\rho^*[x(n)] = \rho(n)$. A corollary to Theorem (2), by virtue of the convexity of P', is that $x(n) \in P'$ for all n, hence the error $\varepsilon(n)$ of Equation (9) is bounded.

For printing, the natural index dimension is 2. Trials of space filing curves to overcome the fundamental two-dimensionality of a page, as in [32], give reasonable but largely suboptimal results. Floyd and Steinberg made a crucial invention in the field of digital halftoning [33] which allows index dimension 1 to resemble in some sense index dimension 2. They used a scheme of weights $w_i(k)$ which involved the neighbors to the left and above the current pixel. Soon afterward, Jarvis, Judice, and Ninke [34] introduced a larger 12-neighbor error scheme. Stucki [35] also constructed a 12-neighbor scheme (see also [36, 37]). **Figure 9** shows the two schemes (to obtain individual weights, one normalizes the numbers in a table by dividing by their sum).

For reasons which have not been completely understood, error diffusion creates artifacts in the form of small-scale worms (see **Figure 10**). More recently, Tresser and Wu [38] made improvements to reduce these artifacts using systems of weights chosen to produce predetermined patterns judged to be good for certain gray levels in BW printing, and then interpolating sets of weights for the remaining gray levels. In their scheme there are gray levels with negative weights. In this case, Theorem 2 cannot be applied to prove boundedness of $\varepsilon(n)$.

Calibration for error diffusion

As in the case of the dither mask, one has to take account of the dot gain and dot overlap. A method patented by Tresser and Wu [39] improves on former methods for correction [40, 41].

Constrained error diffusion

Late in the last century, all major manufacturers of high-end printers competed to bring to market their own models based on a common new print engine. The only pixel colors available to a preliminary version of the IBM InfoColor* 100 printer were black, yellow, magenta, cyan, and white. No overlap of toners was allowed. This constraint is shared by certain types of printers such as the Kodak ImageSource** 70cp Series II. Printers with such constraints are often used as highlight printers, to color simple objects such as pie charts. We developed algorithms to transform the preliminary IBM version into a near-full-color printer[4] by using error diffusion on the restricted printer gamut P = convex hull of $\{K, Y, M, C, W\}$, in combination with a gamut-reduction map from the unit three-dimensional color cube to P. Even though no saturated red, green, or blue was printable, the results were quite acceptable. IBM was consequently able to present the first color high-speed Advanced Function Presentation* (AFP*) printer at the 1998 XPLOR trade show, enabling the corporation to enter a market it had previously not penetrated. In subsequent, more advanced printers, in which red, green, and blue are available by superimposing colors, our method developed for the IBM InfoColor 100 printers can be used in a draft mode as a toner or ink saver.

[4] R. L. Adler, M. Martens, J. L. Mitchell, R. Risch, N. Rijavec, C. P. Tresser, and C. W. Wu, patent pending.

Finally, a comprehensive survey of error diffusion and other techniques can be found in the books of R. Ulichney [42] and H. Kang [4].

Acknowledgments

This paper has described both methods and contributions of the IBM Research Mathematical Sciences Department in collaboration with other IBM divisions and departments. In particular we acknowledge the contributions of Jean Aschenbrenner, Gordon Braudaway, Danielle Dittrich, Joan Mitchell, Ravi Rao, Nenad Rijavec, Robert Risch, Mikel Stanich, Gerry Thompson, Fred Mintzer, and Howard Sachar.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of Eastman Kodak Company.

References

1. W. S. Stiles and G. Wyszecki, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Second Edition, John Wiley & Sons, Inc., New York, 1982.
2. D. B. Judd and G. Wyszecki, *Color in Business, Science, and Industry*, John Wiley & Sons, Inc., New York, 1975.
3. I. Amidror, *The Theory of the Moiré Phenomenon*, Kluwer Academic Publishers, Dordrecht, Netherlands, 2000.
4. H. Kang, *Color Technology for Electronic Imaging Devices*, SPIE Optical Engineering Press, Bellingham, WA, 1997.
5. J. Beck and W. W. L. Chen, *Irregularities of Distribution*, Cambridge University Press, Cambridge, England, 1987.
6. R. L. Adler, B. Kitchens, M. Martens, A. Nogueira, C. Tresser, and C. W. Wu, "Error Bounds for Error Diffusion and Other Mathematical Problems Arising in Digital Halftoning," *Proc. SPIE* **3963,** 437–443 (2000).
7. R. L. Adler, B. Kitchens, M. Martens, A. Nogueira, C. Tresser, and C. W. Wu, "Error Bounds for Error Diffusion and Related Halftoning Algorithms," *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2001, pp. II-513–II-516.
8. H. G. Meijer, "On a Distribution Problem in Finite Sets," *Nederl. Akad. Wetensch. Indag.* **35,** 9–17 (1973).
9. H. G. Meijer and H. Niederreiter, "On a Distribution Problem in Finite Sets," *Compositio Math.* **25,** 153–160 (1972).
10. H. Niederreiter, "On the Existence of Uniformly Distributed Sequences in Compact Spaces," *Compositio Math.* **25,** 93–99 (1972).
11. R. Tijdeman, "On a Distribution Problem in Finite and Countable Sets," *J. Combin. Th. (A)* **15,** 129–137 (1973).
12. R. Tijdeman, "The Chairman Assignment Problem," *Discrete Math.* **32,** 323–330 (1980).
13. R. Näsänen, "Visibility of Halftone Dot Textures," *IEEE Trans. Syst., Man, Cybernet.* **14,** 920–924 (1984).
14. J. Sullivan, L. Ray, and R. Miller, "Design of Minimal Visual Modulation Halftone Patterns," *IEEE Trans. Syst., Man, Cybernet.* **21,** 33–38 (1991).
15. J. Sullivan, R. Miller, and G. Pios, "Image Halftoning Using a Visual Model in Error Diffusion," *J. Opt. Soc. Amer.* **10A,** 1714–1724 (1993).
16. R. A. Ulichney, "The Void-and-Cluster Method for Dither Array Generation," *Proc. SPIE* **1913,** 332–343 (1993).
17. M. Yao and K. J. Parker, "Modified Approach to the Construction of the Blue Noise Mask," *J. Electron. Imaging* **3,** 92–97 (1994).
18. R. A. Ulichney, "Dithering with Blue Noise," *Proc. IEEE* **76,** 56–79 (1988).
19. R. L. Adler, G. R. Thompson, C. P. Tresser, and C. W. Wu, "Dithering Masks with Very Large Periods," U.S. Patent 6,088,123, 2000.
20. A. Thue, "Über Unendishle Zeichenreihen," *Videnskapsselskapets Skrifter. I. Mat.-naturv.*, Klasse, Kristiania, 1906, pp. 1–22.
21. A. Thue, "Über die Gegenseitige Lage Gleicher Teile Gewissser Zeichenreinen," *Videnskapsselskapets Skrifter. I. Mat.-naturv.*, Klasse, Kristiania, 1912, pp. 1–67.
22. M. Morse, "Recurrent Geodesics on a Surface of Constant Negative Curvature," *Trans. Amer. Math. Soc.* **22,** 84100 (1921).
23. M. Morse and G. A. Hedlund, "Unending Chess, Symbolic Dynamics, and a Problem in Semigroups," *Duke Math. J.* **11,** 1–7 (1944).
24. G. Thompson, C. Tresser, and C. W. Wu, "Clustered Aperiodic Mask," U.S. Patent 5,917,951, 1999.
25. G. Thompson, C. Tresser, and C. W. Wu, "Multicell Clustered Mask with Blue Noise," U.S. Patent 6,025,930, 2000.
26. C. W. Wu, C. Tresser, G. R. Thompson, and M. Stanich, "Supercell Dither Masks with Constrained Blue Noise Interpolation," *Proceedings of NIP17, Imaging Science and Technology International Conference on Digital Printing Technologies*, 2001, pp. 487–490.
27. J. L. Mitchell, G. R. Thompson, C. W. Wu, T. J. Trenary, and Y. Qiao, "Multilevel Color Halftoning," *Proceedings of the Imaging Science and Technology 9th Color Imaging Conference*, 2001, pp. 189–193.
28. R. Rao, G. R. Thompson, C. P. Tresser, and C. W. Wu, "Microlocal Calibration of Digital Printers," U.S. Patent 5,943,477, 1999.
29. M. Morse and G. A. Hedlund, "Symbolic Dynamics II. Sturmian Trajectories," *Amer. J. Math.* **62,** 1–42 (1940).
30. R. M. Siegel, C. Tresser, and G. Zettler, "A Decoding Problem in Dynamics and in Number Theory," *Chaos* **2,** 473–493 (1992).
31. J. R. Sullivan, R. L. Miller, and T. J. Wetzel, "Color Digital Halftoning with Vector Error Diffusion," U.S. Patent 5,070,413, 1991.
32. L. Velho and J. de Miranda Gomes, "Digital Halftoning with Space Filling Curves," *Computer Graph.* (SIGGRAPH '91 Proceedings, T. W. Sederberg, Ed.) **25,** 81–90 (1991).
33. R. Floyd and L. Steinberg, "An Adaptive Algorithm for Spatial Grey Scale," *Society for Information Display Symposium, Digest of Technical Papers*, 1975, pp. 36–37.
34. J. F. Jarvis, C. N. Judice, and W. H. Ninke, "A Survey of Techniques for the Display of Continuous Tone Pictures on Bilevel Displays," *Computer Graph. & Image Process.* **5,** 13–40 (1976).
35. P. Stucki, "MECCA—A Multiple-Error Correcting Computation Algorithm for Bilevel Image Hard-Copy Reproduction," *Research Report RZ-1060*, IBM Research Laboratory, Zurich, Switzerland, 1981.
36. Z. Fan, "Error Diffusion with a More Symmetric Error Distribution," *Proc. SPIE* **2179,** 150–158 (1994).
37. K. T. Knox, "Error Diffusion: A Theoretical View," *Proc. SPIE* **1913,** 326–331 (1993).
38. C. P. Tresser and C. W. Wu, "Target Patterns Controlled Error Management," U.S. Patent 6,101,001, 2000.
39. C. P. Tresser and C. W. Wu, "Model Based Error Diffusion with Correction Propagation," U.S. Patent 5,946,455, 1999.
40. Y. Lin and T. C. Ko, "A Modified Error-Based Error Diffusion," *IEEE Signal Process. Lett.* **4,** 3638 (1997).

41. T. N. Pappas and D. L. Neuhoff, "Printer Models and Error Diffusion," *IEEE Trans. Image Process.* **4,** 66–80 (1995).
42. R. Ulichney, *Digital Halftoning*, MIT Press, Cambridge, MA, 1987.

Received October 19, 2001; accepted for publication July 24, 2002

Roy L. Adler *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Adler received his Ph.D. degree in mathematics from Yale University in 1961. He had already joined IBM in 1960 as a Research Staff Member in the Research Division Mathematical Sciences Department. From 1982 to 1987 he was Manager of General Mathematical Studies, and from 1987 to 1994, Senior Manager of Computational Mathematics. He has held visiting positions at Stanford University, the University of Warwick, the Hebrew University of Jerusalem, Brown University, Battelle Institute, Columbia University, Yeshiva University, Pratt Institute, and the Mathematical Science Research Institute. Dr. Adler's main area of interest has been classification theorems of ergodic theory and dynamical systems. Theorems in these areas of pure mathematics led to an algorithm to design codes to meet constraints for data storage and transmission channels. One of these codes was used in the IBM 9332 hard disk file. Dr. Adler is the author of 59 research papers, ten patents, and ten patent publications which deal with coding, printing, spine modeling, X-ray data acquisition, and cryptography. He also devised the cryptography system for the IBM Controlled Access System. Dr. Adler was a member of the Editorial Board of the *Journal of Ergodic Theory and Dynamical Systems*. He served two terms as a Trustee of the Mathematical Sciences Research Institute and is currently serving a second term as an elected Trustee of the American Mathematical Society. He has been awarded an IBM Fourth Plateau Invention Achievement Award, two IBM Research Outstanding Innovation Awards, an IBM Outstanding Technical Achievement Award, and an IBM 2000 Research Patent Portfolio Award. An honorary conference was held at Yale University on the occasion of his 60th birthday. He is a Fellow of the New York Academy of Arts and Sciences and the American Academy of Arts and Sciences.

Bruce P. Kitchens *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Kitchens received his Ph.D. degree in mathematics from the University of North Carolina at Chapel Hill in 1981 and has been a Research Staff Member in the Mathematical Sciences Department at the IBM Thomas J. Watson Research Center since 1982. He has been a visitor at the Mathematical Sciences Research Institute, Warwick University, the Université de Paris VI, Northwestern University, the University of Washington, Wesleyan University, and the Institut de Mathematique de Luminy. His mathematical interests are ergodic theory and dynamical systems. In recent years he has been primarily interested in the dynamics of algebraic $\mathbf{Z}^d$ actions on compact topological groups. Dr. Kitchens is the author of the book *Symbolic Dynamics, One Sided, Two-Sided and Countable State Markov Shifts*, which was published by Springer in 1998. His applied interests have been in modulation coding theory and algebraic coding theory, financial mathematics, and mathematics arising in printing.

Marco Martens *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (mmartens@us.ibm.com).* Dr. Martens received his engineer Diploma in 1986 and his Ph.D. degree in 1990 in Delft, The Netherlands. He then spent two years at the Instituto Mathematica Pura e Aplicada (Rio de Janeiro) and six at the State University of New York at Stony Brook, where he proved several major results on rigidity and universality. He joined the Mathematical Sciences Department of the IBM Thomas J. Watson Research Center as Manager of Special

Mathematical Studies in 1998. Dr. Martens held this position until early 2002, when he decided to return full time to basic and applied research. His research work covers both pure mathematics (in particular dynamics and the transition to chaos) and applications to various technologies, from digital printing (where he worked both on halftoning and image compression) to aspects of applied cryptography, and most recently on the foundations of autonomic computing.

Charles P. Tresser *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (tresser@us.ibm.com).* Dr. Tresser received his Ph.D. degree in theoretical physics at the University of Nice in 1981. He joined the Mathematical Sciences Department of the IBM Thomas J. Watson Research Center in 1989, spending most of his time on basic and applied research and as Senior Manager in charge of the Applied Mathematics Department. His research ranged from pure mathematics to applications in various areas of IBM technology. The applications included digital printing, digital watermarks, telephony, medical visualization, broadband spectrum communication, electronic commerce, electronic counterfeiting and tampering protection, and electronic privacy. In 1999 Dr. Tresser left the Research Division to join the Financial Services Sector of the IBM Sales and Distribution Division. He became head of the IBM Financial Services Research Center in 2001, interfacing between Research, the Financial Services Sector, and customers worldwide. Dr. Tresser has been a coauthor of eighteen patents. Several of these inventions have already found their way into IBM technology, earning him an IBM Outstanding Technical Achievement Award. Prior to joining IBM in 1989, he was Directeur de Recherche in the French Centre National de la Recherche Scientifique (CNRS) Physical Sciences for Engineers Division and in the Theoretical Physics Division. Dr. Tresser has also held visiting positions at several institutions, including the Courant Institute of Mathematical Sciences (New York), IMA (Minneapolis), IHES (Bures-sur-Yvette), the Weizmann Institute (Rehovot), The Hebrew University (Jerusalem), the City University of New York Graduate Center, and Columbia University. While at the CNRS, he made key discoveries in chaos theory, a field in which he is a world-recognized leader. He is currently funded by the National Science Foundation for fundamental research in pure mathematics. Dr. Tresser has published more than 120 scientific papers. In addition, he has served as editor of the journals *Nonlinearity* and *Journal of Complexity*, and is currently serving as an editor of *Chaos*. Finally, he was awarded a Vinci of Excellence in the Science pour l'Art Prize by the luxury conglomerate LVMH (Louis Vuiton–Moet–Hennessy) and a Médaille d'Argent from the CNRS.

Chai Wah Wu *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (cwwu@us.ibm.com).* Dr. Wu received his Ph.D. degree in electrical engineering from the University of California at Berkeley in 1995. He held an IBM postdoctoral fellowship in 1996–1997 and since that time has been a Research Staff Member in the Mathematical Sciences Department at the IBM Thomas J. Watson Research Center. He is currently the group leader of the Special Math Studies group at IBM. His research interests include synchronization and control of coupled chaotic systems, circuit theory, digital halftoning, and multimedia security. Dr. Wu has written more than 50 journal papers and is the author of the book *Synchronization in Coupled Chaotic Circuits and Systems*, published by World Scientific in 2002. He holds 19 U.S. patents and has been awarded an IBM Tenth Plateau Invention Achievement Award. He also received an IBM Outstanding Technical Achievement Award for his work on digital halftoning. Dr. Wu was elected a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2001; he served as an associate editor of the *IEEE Transactions on Circuits and Systems*, *Part 1*, during the periods 1997 to 1999 and 2002 to 2004.