Spatial Sampling of Printed Patterns

Prateek Sarkar, *Student Member, IEEE,*
George Nagy, *Senior Member, IEEE,*
Jiangying Zhou, *Member, IEEE,* and
Daniel Lopresti, *Member, IEEE,*

Abstract—The bitmap obtained by scanning a printed pattern depends on the exact location of the scanning grid relative to the pattern. We consider ideal sampling with a regular lattice of delta functions. The displacement of the lattice relative to the pattern is random and obeys a uniform probability density function defined over a unit cell of the lattice. Random-phase sampling affects the edge-pixels of sampled patterns. The resulting number of distinct bitmaps and their relative frequencies can be predicted from a mapping of the original pattern boundary to the unit cell (called a modulo-grid diagram). The theory is supported by both simulated and experimental results. The modulo-grid diagram may be useful in helping to understand the effects of edge-pixel variation on Optical Character Recognition.

Index Terms—Spatial sampling, random phase sampling, digitization, optical character recognition, document defect models, scanner models, modulo-grid diagram, locales.

1 Introduction

DIGITIZATION is the process of spatial, intensity, and temporal quantization of an analog pattern. Printed patterns are scanned into digital images by spatial and intensity quantization. High-contrast (essentially black and white) images such as text for Optical Character Recognition, and line drawings are often scanned in binary mode. Binary scanning can be modeled as a sequence of the following steps.

1) The analog intensity pattern is convolved with the point-spread function (PSF) of the sensor.
2) The resulting "smoothed" signal is spatially sampled.
3) Finally, the samples are quantized to 0 or 1 by comparing them against a preset threshold.

The order of the last two steps can be swapped without altering the result. We do this in our model to conveniently isolate the spatial sampling step. Our model for scanning is illustrated in Fig. 1 for a 1D signal. (We shall liberally use one-dimensional illustrations throughout the paper, but all of our results apply to 2D.)

Here, we study only the spatial quantization aspect of digitization, and predict the variability of pixel configurations under uniform random-phase sampling of printed patterns. This variability is the result of the uncontrollable displacement of the sampling grid relative to the page. Unlike other types of noise in scanners, random phase noise is an intrinsic consequence of finite sampling resolution. In OCR (Optical Character Recognition), where the smallest patterns (periods, commas, quotation-marks) are typically represented by only a few pixels at common point-sizes (8-10 pts) and sampling

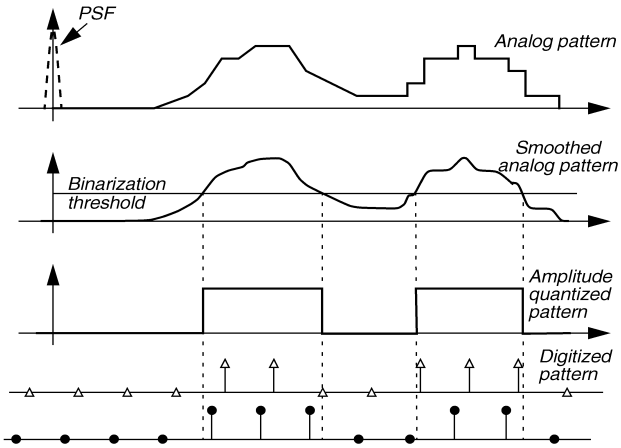

Fig. 1. Our model of digitization of an analog pattern illustrated in 1D. Two different digitizations resulting from displacement of the sampling lattice are shown.

resolutions (200-400 dpi), random-phase noise may be the limiting factor in the recognition of high-quality print. Even in larger patterns, thin strokes, such as the top serif of the numeral "1" or the horizontal bar of a Times-Roman "e," may be adversely affected.

An example of the effect is shown in Figs. 2 and 3. The latter presents the different pixel configurations (bitmaps) obtained by digitizing a black disk (for example, a period) of diameter equalling 1.2 sampling intervals. The disk gives rise to four different digitized configurations. If the location of the sampling grid is uniformly distributed (as is the case in practice), then the various patterns occur with the relative frequencies shown in Fig. 3.

The type of information shown in Fig. 3 bears on the following OCR problems:

1) Minimum sample size for adequate representation of a set of patterns.
2) Minimum sampling resolution for the recognition of a family of patterns, such as a typeface of given point size.
3) The design of OCR-features that are less sensitive to edge-noise.
4) The construction of pseudorandom defect models for OCR.
5) The reconstruction of the original (analog) pattern from a set of digitized samples.
6) Reducing OCR errors by multiple scans of a page.

We hope that the theory presented here eventually opens up ways of addressing these OCR issues.

Our principal tool for obtaining the information shown in Fig. 3 is the modulo-grid diagram shown in Fig. 2. It is obtained from the original pattern by overlaying the pattern boundaries in each cell of the digitizing grid on the unit cell. In the sequel, we demonstrate that the modulo-grid diagram decomposes the spatial displacement space into regions of isomorphic digital representations of a given pattern. Furthermore, the areas of these regions correspond to the relative frequencies of the digital patterns.

A qualitative discussion of this phenomenon, emphasizing the correlated nature of the resulting edge-noise in OCR, appeared in [11]. Nadler mentioned the effects of random-phase sampling on scanned characters in a 1972 survey paper and illustrated it in a recent textbook on Pattern Recognition [12]. Ingold estimated the difference expected between different scans of the same pattern and devised a method of constructing ternary templates that are insensitive to edge noise [6]. Pavlidis investigated conditions under which the connectivity and shape of bilevel patterns are preserved under digitization. To this end, he formulated a "compatibility condition" that restricts the minimum size and maximum boundary curvature of both foreground and background components of the analog pattern as a function of the sampling interval

- *P. Sarkar and G. Nagy are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180. E-mail: sarkap@rpi.edu, nagy@ecse.rpi.edu.*
- *J. Zhou is with Panasonic Information and Networking Technologies Laboratory, Panasonic Technologies, Inc., Two Research Way, Princeton, NJ 08540. E-mail: jz@panasonic.research.com.*
- *D. Lopresti is with Bell Laboratories Research, Lucent Technologies, 700 Mountain Ave., Murray Hill, NJ 07974. E-mail: dlopresti@bell-labs.com.*

Manuscript received 5 May 1997. Recommended for acceptance by J.J. Hull. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 106321.

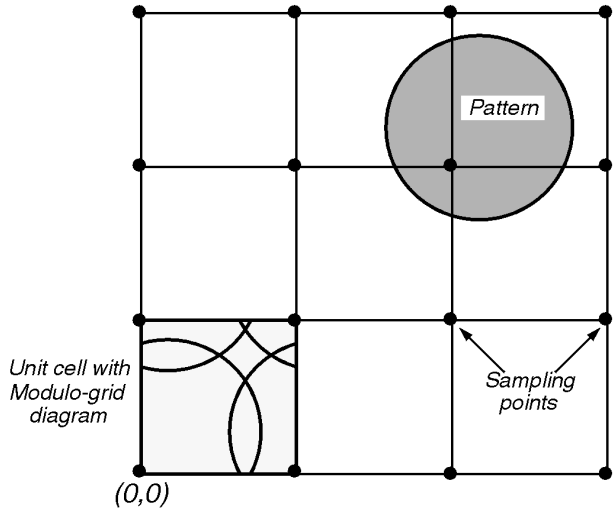

Fig. 2. The modulo-grid diagram for a disk whose diameter equals 1.2 sampling intervals.

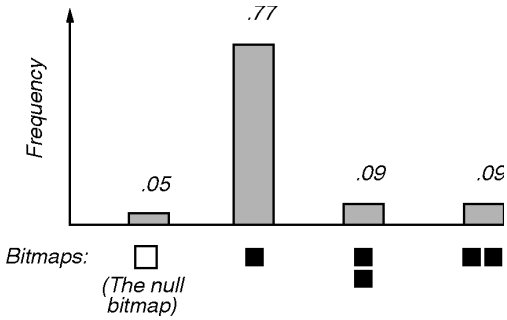

Fig. 3. The four bitmaps that can be obtained by sampling the circular disk shown in Fig. 2 and their corresponding occurrence probabilities.

[16]. The spatial sampling effect is also familiar to digital typographers [17].

Analysis of the effect of sampling in one dimension and simulation results, highlighting possible applications to OCR, were presented in [18] and [13]. The modulo-grid approach to spatial sampling was originated by Havelock [4], [5], who was interested in the precision of localization of terrestrial objects by remote sensing as a function of image resolution, object shape and signal-to-noise ratio. He defined the concept of a "locale," which is precisely a union of regions of the modulo-grid diagram that yield isomorphic bitmaps. Related work on the subpixel localization of edges and circles, and on the design of optimal fiducial marks, was presented by O'Gorman et al. [14], [15].

The contribution of the current paper is the formal proof of the correspondence between the regions of the modulo-grid diagram and the distinct digitizations, and between the area of the regions and the frequency of the digitized patterns.

The modulo-grid diagram is presented in Section 2. Section 3 presents our main results in the form of two theorems. Section 4 contains some examples that illustrate possible applications to OCR, and Section 5 mentions possible generalizations and some unsolved problems.

2 The Modulo-Grid Diagram

When a bilevel spatial pattern is sampled with a lattice (or grid) of delta functions, the result depends on the positioning of the grid relative to the pattern itself. For the sake of simplicity, we consider first sampling in one dimension. If we have a uniform 1D sampling lattice of delta functions, then a simple black stroke of length

Fig. 4. Sampling a black stroke 2.3 sampling-intervals long may yield a bitmap of three or two black pixels.

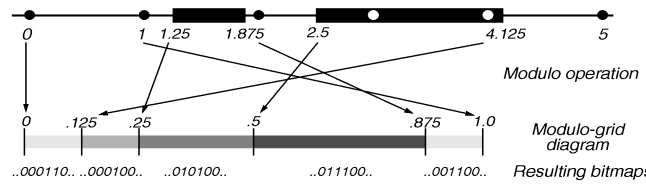

Fig. 5. A 1D pattern of black strokes, and the corresponding modulo-grid diagram. The bitmap associated with each region is also shown.

equalling 2.3 sampling intervals can digitize to a bitmap of either two or three pixels (Fig. 4).[1]

Let us now consider a more complex bilevel 1D pattern, such as the one shown in Fig. 5. When the grid is slid along the pattern, the bitmap changes only when one of the sampling points crosses over an edge of the pattern. Sampling a 1D pattern of black and white strokes could produce as many distinct bitmaps as the number of edges (four in this case) in the pattern [18]. Further, since the sampling lattice repeats with a period of one sampling interval, it should be possible to obtain all the possible bitmaps by sliding the grid by no more than a sampling interval.

This relationship between the analog pattern boundary and the digital configuration can be captured by what we call the modulo-grid diagram. For a 1D pattern, the modulo-grid diagram can be constructed in the following manner. Overlay the grid at arbitrary position over the pattern, and register the pattern in the grid coordinate system. For each edge in the pattern, located at x, insert a mark in the interval [0, 1) at x mod 1.[2] The mark at 0.5 in Fig. 5, for example, is due to the edge at 2.5. The marked interval [0, 1) forms the modulo-grid diagram.

To understand how the modulo-grid diagram works, move the grid with respect to its original location. As the origin of the grid moves in [0, 1) every time it crosses over a mark within the modulo-grid diagram, some sampling point crosses over an edge in the pattern, and vice-versa. Each interval between adjacent marks in the modulo-grid diagram thus represents a *region* of identical pixel configuration. Since the digital representation does not change within a region, the length of a region is proportional to the frequency of occurrence of the corresponding pixel configuration. Owing to the periodicity of the lattice, the modulo-grid diagram wraps around and the two end regions give rise to isomorphic bitmaps.

Extending the concept to two dimensions leads to a *unit cell* (Fig. 6) in place of the unit interval, and the "marks" are replaced by segments of the pattern boundary. An example of a 2D modulo-grid diagram for a circular disk pattern was shown in Fig. 2. The diagram is formed by taking segments of the pattern's boundary contour that lie in each of the "grid cells" and retracing these segments at their coordinates, modulo one, with respect to both axes. The regions are demarcated by the boundary segments.

If we now slide the grid around, with its origin confined to the unit cell of the modulo-grid diagram, each time the origin crosses over an edge from one region to another, a sampling point at some location moves in or out of the pattern, and vice versa. Different bitmap configurations, therefore, correspond to different regions in

1. We are using the image processing terminology of "bitmap" and "pixel" in anticipation of the transition to two dimensions.
2. x *mod* 1 is defined as $x - \lfloor x \rfloor$.

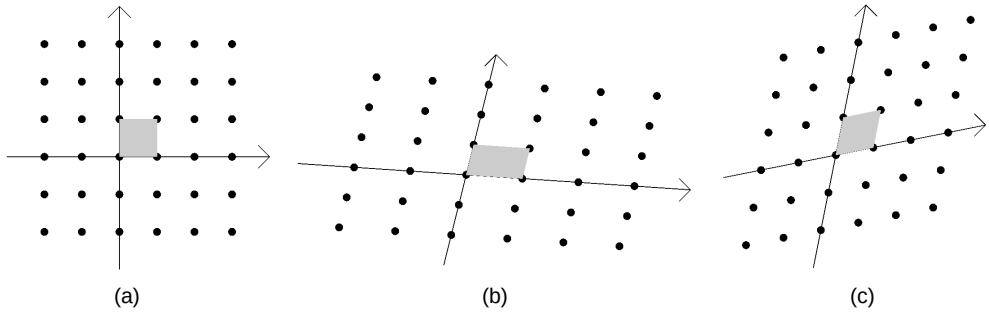

Fig. 6. Examples of sampling grids.

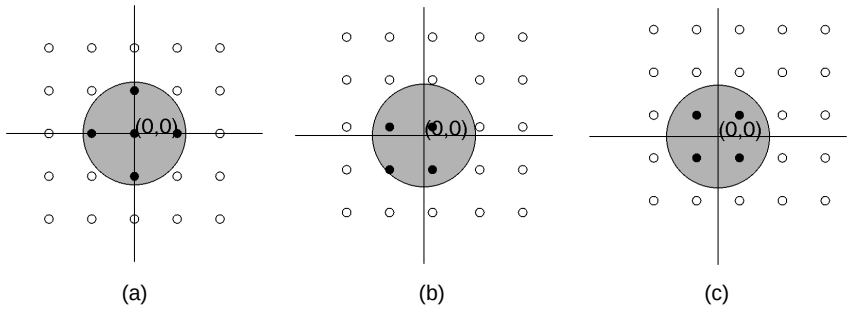

Fig. 7. Digitizations of a circular disk resulting from different grid positions.

the modulo-grid diagram. The converse is not necessarily true as shown at the end of Section 3. Thus the modulo-grid diagram yields only an upper bound on the number of possible bitmap configurations for a given pattern.

In the following section, we develop a mathematical platform for discussing the various concepts presented here and present two theorems that summarize our contribution.

3 2D Spatial Sampling Theory

Let $\Im$ be the set of all subsets of Euclidean space R^2 that are closed, bounded, and regularized,[3] and whose boundary consists of only a finite number of disjoint simple continuous closed curves. The "regularization" condition excludes sets that contain isolated and dangling lines and points from $\Im$ [10], so that the formalization is not complicated by abstractions.

We define an intensity pattern as a function $f(\mathbf{x})$, $\mathbf{x} \in R^2$, that is zero everywhere in R^2 except in a region $X \in \Im$. X is called the *support* of f and represents the pattern foreground where $f(\mathbf{x}) \neq 0$, and $R^2 - X$ represents the background where $f(\mathbf{x}) = 0$. We shall consider only bilevel intensity patterns where the $f(\mathbf{x}) = 1$ in X. In our scanning model (Fig. 1), the thresholded intensity is a bilevel pattern.

In some applications, the original analog pattern is bilevel. If we can digitize without convolution (as in rasterizing an outline computer font) or if the point spread function is essentially a delta function, then $f(\mathbf{x})$ is the same as the original bilevel pattern. Convolution by a PSF, followed by thresholding causes $f(\mathbf{x})$ to be a thickened or thinned version of the printed pattern.

Spatial quantization of an analog pattern is a mapping of $f(\mathbf{x})$ from its support X to a discrete set of ordered pairs $\{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in$ *a discrete set of points in* $R^2\}$. In practice, a *sampling grid* or *lattice* is overlaid on R^2 space and $f(\mathbf{x})$ is measured at the grid intersections. In the case of bilevel patterns, however, specifying the set of grid intersections that lie within the support X of the pattern completely defines the spatial quantization. Disregarding $f(\mathbf{x})$ underscores the separation of spatial quantization effects from amplitude or intensity quantization, which we do not discuss in this paper.[4] Henceforth we shall refer to X as the spatially-analog pattern or simply *analog pattern*.

Let $\mathbf{u}$, $\mathbf{v}$ be a pair of basis vectors for R^2. A sampling grid, $G(\mathbf{p})$, "positioned" at a point $\mathbf{p}$, is defined as

$$G(\mathbf{p}) = \{\mathbf{g} : \mathbf{g} = k_1\mathbf{u} + k_2\mathbf{v} + \mathbf{p}, \quad k_1, k_2 \in Z\}$$

where Z is the set of all integers. Fig. 6 shows examples of grid configurations. The grid positioned at the origin, $G(\mathbf{0})$, from now on is simply denoted as G. Associated with G is a unit cell, C.

$$C = \{\mathbf{c} : \mathbf{c} = w_1\mathbf{u} + w_2\mathbf{v}, \quad 0 \leq w_1 < 1, 0 \leq w_2 < 1\}.$$

The shaded areas in Fig. 6 show the unit cells for the corresponding grids.

Digitization of X by grid G at position $\mathbf{p}$ is then defined as the set

$$D_{\mathbf{p}} = \boldsymbol{X} \cap G(\mathbf{p})$$

Since any grid, as defined above, can be converted to a square lattice and back by linear transformations on R^2, we restrict our discussion to square grids.

A pattern digitized by a sampling grid at different locations may give rise to isomorphic digital patterns. For example, patterns (b) and (c) in Fig. 7 correspond to sets:

$$\{(-0.8, -0.8), (-0.8, 0.2), (0.2, -0.8), (0.2, 0.2)\} \text{ and}$$

$$\{(-0.5, -0.5), (-0.5, 0.5), (0.5, -0.5), (0.5, 0.5)\},$$

respectively. However, they are equivalent in terms of pixel configuration.

Let $T(D, \mathbf{t}) = \{\mathbf{d} + \mathbf{t} : \mathbf{d} \in D\}$ denote set D translated by $\mathbf{t}$. Two digital patterns D and D' are *isomorphic* if $\exists \mathbf{t} \in R^2$ such that $T(D, \mathbf{t}) = D'$. The equivalence relation is denoted as $D \equiv D'$.

3. A regularized set is a set that equals the closure of its interior.

4. The effects of spatial and intensity sampling are sometimes related. Fine intensity quantization can compensate for the perceived effects of coarse spatial quantization. This is exploited in "antialiasing" on computer displays.

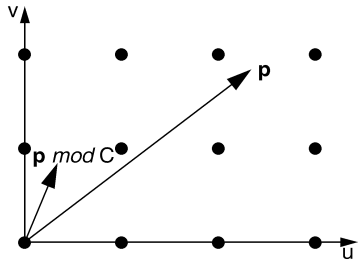

Fig. 8. Modulo-grid operation in 2D.

Note that digitizations, by definition, carry information regarding the location of the grid-origin, unlike *bitmaps* that we obtain from scanners. This makes it necessary to keep in mind the relationship between bitmaps and digitizations: A bitmap is a translation-invariant representation of a digitization, and all isomorphic digitizations correspond to the same bitmap.

On the other hand, a pattern may be digitized into distinguishable bitmaps by different translations of the grid. Fig. 7 shows two different bitmaps resulting from sampling the same analog pattern. As we have mentioned earlier, our tool in analyzing these variations is the modulo-grid diagram. We can now define the modulo-grid operation for a point $\mathbf{p}$ with respect to a unit cell C, "$\mathbf{p}$ mod C" (Fig. 8), as:

$$\mathbf{p} \bmod C = (p_1 - \lfloor p_1 \rfloor)\mathbf{u} + (p_2 - \lfloor p_2 \rfloor)\mathbf{v}, \text{ where } \mathbf{p} = p_1\mathbf{u} + p_2\mathbf{v}.$$

It is easily shown that $D_{\mathbf{p}} \equiv D_{\mathbf{p} \bmod C}$. So we only need to consider grid locations within the unit cell C.

Let X^b denote the boundary of a given analog pattern X.[5] The modulo-grid diagram, introduced in our earlier discussion, is formally defined as the result of superimposing the boundary X^b over the unit cell C: $X_C^b = \{\mathbf{p}: \mathbf{p} = \mathbf{x} \bmod C, \mathbf{x} \in X^b\}$. Such a superimposition creates a planar map of boundary segments in C, which partitions C into nonoverlapping, open, connected sets π_i, bounded by points in X_C^b. Each such set is called a region. We have for $i \neq j$, $\pi_i \cap \pi_j = \emptyset$, and $C - \bigcup \pi_i = X_C^b$. In Fig. 9, X^b is a circle, and X_C^b consists of circular arcs in the unit cell. Here, there are nine regions π_i.

Two points, $\mathbf{p}$ and $\mathbf{p'} \in C$, are said to be in the same *locale* if and only if digitizations by $G(\mathbf{p})$ and $G(\mathbf{p'})$ yield the same bitmap. Formally, a locale is a maximal set of points Λ such that

$$\mathbf{p}, \mathbf{p'} \in \Lambda \Leftrightarrow D_{\mathbf{p}} \equiv D_{\mathbf{p'}}.$$

For the disk pattern and the grid shown in Figs. 2 and 9, each region enclosed by the curve-segments in the modulo-grid diagram belongs to a single locale, as Theorem 1 will show.

In the following, we formally prove that each region belongs to a single locale. First, we introduce a lemma, the notation for which is illustrated in Fig. 9.

LEMMA 1. *If* $\mathbf{p}, \mathbf{p'} \in \pi_i$, *then* $\forall \mathbf{x}^* \in G(\mathbf{0})$, *either* $(\mathbf{x}^* + \mathbf{p})$ *and* $(\mathbf{x}^* + \mathbf{p'})$ *are both elements of* X *or are both elements of* $\overline{X} = R^2 - X$.

PROOF. Since $\mathbf{p}, \mathbf{p'} \in \pi_i$ and π_i is a connected open set by definition, there exists a continuous path L that starts at $\mathbf{p}$ and ends at $\mathbf{p'}$ such that $L \subset \pi_i$ [1]. Consequently,

$$L \cap X_C^b = \emptyset. \tag{1}$$

Let us define the curve $L^* = \{\mathbf{x}^* + \mathbf{x} : \mathbf{x} \in L\}$ for any $\mathbf{x}^* \in G(\mathbf{0})$. L^* is a continuous path connecting $(\mathbf{x}^* + \mathbf{p})$ and $(\mathbf{x}^* + \mathbf{p'})$. Fig. 9 illustrates the proof for a particular value of $\mathbf{x}^*$. It follows that $\mathbf{y} \in L^* \Rightarrow (\mathbf{y} - \mathbf{x}^*) \in L$. We claim that $L^* \cap X^b = \emptyset$, i.e., L^* does not include any point on the pattern boundary. Indeed if this were not the case, then there must exist some $\mathbf{y} \in L^* \cap X^b$.

But then, $(\mathbf{y} \bmod C) = (\mathbf{y} - \mathbf{x}^*) \in X_C^b$ (since $\mathbf{y} \in X^b$) and $(\mathbf{y} - \mathbf{x}^*) \in L$ (since $\mathbf{y} \in L^*$). Consequently, $(\mathbf{y} - \mathbf{x}^*) \in L \cap X_C^b$. This contradicts (1).

The presence of a continuous curve L^* between $(\mathbf{x}^* + \mathbf{p})$ and $(\mathbf{x}^* + \mathbf{p'})$, which never cuts the boundary X^b, implies by the Jordan Curve Theorem that $(\mathbf{x}^* + \mathbf{p})$ and $(\mathbf{x}^* + \mathbf{p'})$ are either both in X or both in $\overline{X}$. □

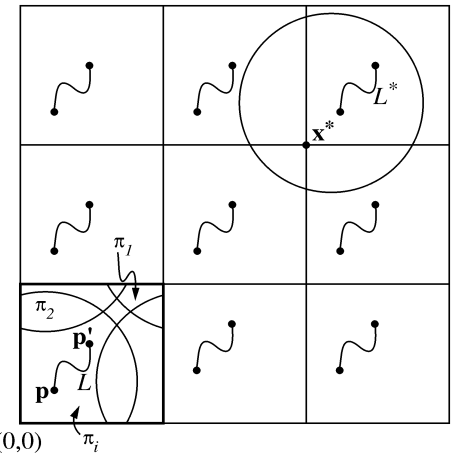

Fig. 9. Illustration for Lemma 1.

Theorem 1 states formally that all grid positions in a region are in the same equivalence class or locale, as suggested in Section 2. The proof follows the reasoning that in moving from a shift of $\mathbf{p}$ to a shift of $\mathbf{p'}$, none of the points in the digitizing grid G moves in or out of the pattern.

THEOREM 1. $\forall i, \mathbf{p}, \mathbf{p'} \in \pi_i \Rightarrow D_{\mathbf{p}} \equiv D_{\mathbf{p'}}$.

PROOF. Let $\mathbf{x} \in D_{\mathbf{p}}$. This means that $\mathbf{x} \in G(\mathbf{p})$ and $\mathbf{x} \in X$. From $\mathbf{x} \in G(\mathbf{p})$, we have $\mathbf{x} = \mathbf{x}^* + \mathbf{p}$, where $\mathbf{x}^* \in G(\mathbf{0})$.

Let $\mathbf{x'}$ be defined as $(\mathbf{x} + \mathbf{p'} - \mathbf{p})$. Evidently, $\mathbf{x'} = (\mathbf{x}^* + \mathbf{p'}) \in \mathbf{G}(\mathbf{p'})$. Furthermore, $\mathbf{x} = (\mathbf{x}^* + \mathbf{p}) \in \mathbf{X} \Rightarrow (\mathbf{x}^* + \mathbf{p'}) = \mathbf{x'} \in \mathbf{X}$ by the premise of this theorem and Lemma 1. Therefore,

$$\mathbf{x'} \in G(\mathbf{p'}) \cap X = D_{\mathbf{p'}}.$$

Thus, we see that $\mathbf{x} \in D_{\mathbf{p}} \Rightarrow \mathbf{x'} = (\mathbf{x} + \mathbf{p'} - \mathbf{p}) \in \boldsymbol{D}_{\mathbf{p'}}$. Lemma 1 can be similarly used to show that $\mathbf{x} \notin D_{\mathbf{p}} \Rightarrow \mathbf{x'} = (\mathbf{x} + \mathbf{p'} - \mathbf{p}) \notin \mathbf{D}_{\mathbf{p'}}$. Thus, $\boldsymbol{T}(\mathbf{D}_{\mathbf{p}}, \mathbf{p'} - \mathbf{p}) = \mathbf{D}_{\mathbf{p'}}$ and consequently, $\mathbf{D}_{\mathbf{p}} \equiv \mathbf{D}_{\mathbf{p'}}$. □

In Theorem 2, we show that if all grid positions are equally likely, then the number of distinct, probable bitmaps induced by a pattern is bounded above by the number of regions in its modulo-grid diagram.

THEOREM 2. *If the position* $\mathbf{p}$ *of the sampling grid* $G(\mathbf{p})$ *is a random variable that is uniformly distributed in probability over the unit cell* C, *then the number of different bitmaps induced by a pattern* $X \in \Im$ *that can appear with nonzero probability is bounded above by the number of regions* π_i *in the modulo-grid diagram of* X.[6]

5. $X^b = X - X^o$ where X^o is the interior of X.

6. Our assumptions regarding the nature of a pattern are more stringent than is required for the proof of Theorem 2. In fact, it is sufficient if the pattern has a boundary contour of Hausdorff dimension less than two [3].

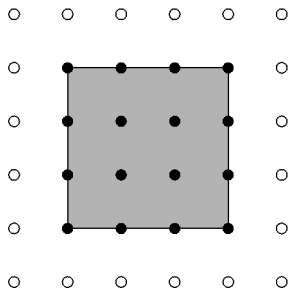

Fig. 10. The grid location at a boundary point produces a new bitmap.

PROOF. Since $X \in \mathfrak{I}$, X^b has a Lebesgue measure of zero. Since X^b maps onto X_C^b, the Lebesgue measure of X_C^b also equals zero. Under the assumption of uniform distribution, the Lebesgue measure of a set equals its probability measure. Thus all digitizations $D_{\mathbf{q}}$ such that $\mathbf{q} \in X_C^b$, put together account for a zero probability of appearance. All *probable* isomorphic digitization classes therefore correspond to grid displacements $\mathbf{p}$ that lie within the regions. By Theorem 1, two different bitmaps, being nonisomorphic, cannot belong to the same region π_i and, hence, the result. □

COROLLARY 1. *The relative frequency of each distinct bitmap, under the assumption of uniformly distributed grid-shifts, is proportional to the area of the corresponding locale.*

The upper bound in Theorem 2 is valid only in a probabilistic sense because grid positions on the boundary of the pattern may actually correspond to new bitmaps. For example, the grid location shown in Fig. 10, which is on the boundary, results in a bitmap which is different from bitmaps generated at any other grid location.

Theorem 2 proves that the number of regions is an upper bound for the number of locales and, therefore, the number of bitmaps, for a given analog pattern. Since the modulo process is periodic in nature, the unit cell can be wrapped around. Topologically, the modulo-grid diagram forms a pattern on a torus: The top and bottom sides, the left and right sides, as well as the four corners, respectively, are equivalent. This is evident in Fig. 11, where the bitmap corresponding to each region is printed on the modulo-grid diagram. For a given pattern, different positions of the grid yield different modulo-grid diagrams. These are however topologically identical when treated as a torus.

We can get a tighter upper-bound by counting as one, the regions that merge together on wrapping around. However, even regions that are not contiguous on the torus can belong to the same locale, as noted by Havelock and illustrated by the example in Fig. 11. In this example, the two regions corresponding to the bitmap of a single black pixel are not contiguous on the torus. On the other hand, the tighter upper bound may be reached (as in the case of the disk in Fig. 9). Fig. 12 shows how the tighter bound increases with increasing size of a pattern, relative to the sampling interval. The dips in the number of locales correspond to the degenerate cases where the disk diameter equals the distance between any two grid points. Degenerate cases are not difficult to construct. A rectangular pattern with sides parallel to the grid axes can yield at the most 4 bitmaps on scanning regardless of its size.

Summary of theoretical results: Given an analog pattern and a sampling grid, the modulo-grid diagram reveals information regarding both the number of different bitmaps and their frequencies when the pattern is digitized. Each region of the diagram corresponds to a specific bitmap (Theorem 1), and its area is proportional to the probability of occurrence of that bitmap (Corollary 1 of Theorem 2) under random-phase sampling.

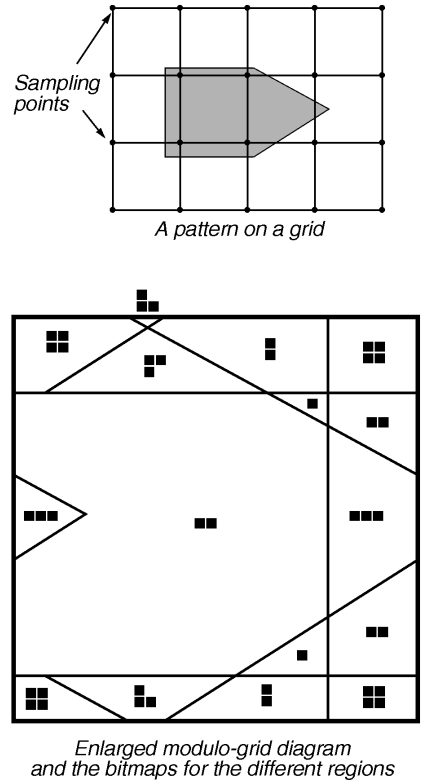

Fig. 11. An analog pattern on a grid, and its modulo-grid diagram showing the regions that yield the same bitmap.

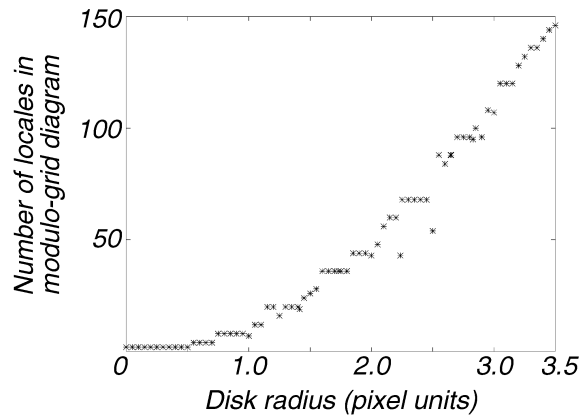

Fig. 12. The number of different bitmaps that an analog pattern (circular disk in this case) can yield on sampling plotted against the radius of the disk. The numbers were obtained by computer simulation. Since the pattern size here is relative to the sampling interval, this also reflects on the rise in the number of possible bitmaps from the same analog pattern as the sampling rate increases.

4 EXPERIMENTAL VERIFICATION

The correlated nature of edge noise in scanned character bitmaps was observed and reported by Nagy as far back as 1968 [11]. Zhou and Lopresti observed that simply scanning a document page thrice with the same scanner and taking a vote among OCR results reduced recognition errors by 30 percent [19]. Since other scanning parameters were the same, this pointed to variable alignment of

Fig. 13. Sample bitmaps of 12 pt. **e**. (a) Printed and scanned. (b) Generated by random phase sampling and some additive noise. (c) Generated by random bit flipping of bitmaps in (a).

the page with respect to the scanner grid as a major factor in OCR performance.

The random phase sampling effect should be considered in the design of pseudo random defect models [18]. Fig. 13 presents a mixture of scanned and simulated bitmaps of **e**s. The simulation procedure for (b) consisted of random phase sampling of an ideal character (as given by the high-resolution outline font of the character), coupled with a very small amount of additive noise.[7] The random bit flipped characters, as in (c) are not representative of scanning degradations and are easily distinguishable from the bitmaps in (a). Independent threshold variations ("sensitivity" in the model described in [2]) and the edge-dependent bit-flipping model of [8] do not induce correlated edge noise. As a consequence, they often produce bitmaps with "hairy" edges and other unusual edge distortions that do not resemble scan-degraded images.

We present two experiments which give more convincing demonstrations of this effect in scanned characters. In each experiment, four sets of 1,000 bitmaps of **e** were used. The four sets of data were obtained by:

1) simulated sampling of an ideal **e** with independent additive noise effect only,
2) simulated sampling of an ideal **e** with the effect of random grid shifts only,
3) simulated sampling combining the above two effects, and
4) actual scanning (using a flat bed scanner) of printed **e**s.

In the first experiment (Fig. 14), the distribution of a simple feature (viz. the number of black pixels in the bitmap) was computed for each set of data, and the results compared against each other. While the distribution for the first set of bitmaps is nearly Gaussian, owing to the (unwarranted) independence assumption, the methods incorporating the random-phase effect can capture the bimodal nature of the distribution seen in real scanned samples (data set 4).

In the second experiment, we computed boundary chain codes for each set of data. We matched each chain code with that of a canonical reference character and plotted the distribution profile of string edits along the chain. Fig. 15 presents the results. As expected, independent additive noise alone results in a uniform error distribution profile. But the model using random grid shifts shows peak error locations that match up well with real scanned characters. This result emphasizes the correlated nature of edge-pixel noise that results from random grid translations.

Since variable phase alignment between pattern and grid affects only the edge pixels, the effects of bitmap variability may be

7. Additive noise models the variations in sensor sensitivity and threshold circuitry that come into play when the point-spread function has nonzero diameter. These effects, unlike that of random-phase sampling, result in uncorrelated bit-flipping.

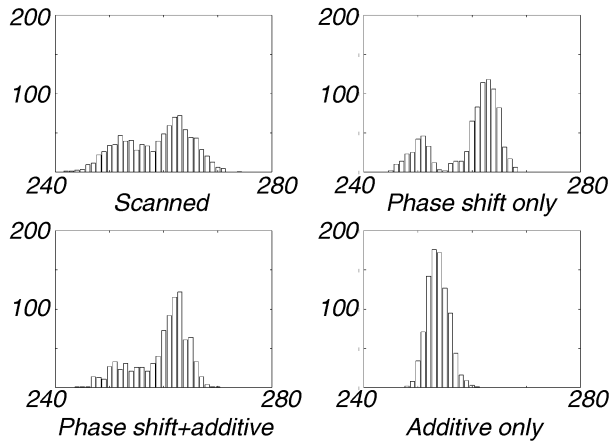

Fig. 14. The distribution of area (the number of black pixels) in a set of 1,000 bitmaps of 12-point Computer Modern Sans Serif **e**. The position of a vertical bar indicates the value of the area, while its height is proportional to the number of bitmaps with that area. Only the random phase-shift model captures the bimodal nature of the distribution shown by the real scanned sample set.

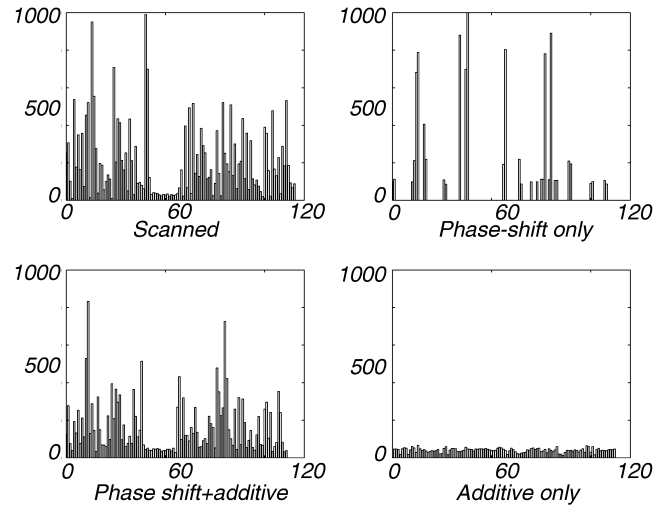

Fig. 15. The random phase-shift model shows a strong correlation with the actual scanning process in terms of the location of string edits in chain codes of random bitmap samples of a given character (12-point Helvetica **e**), when compared against the chain code of a prototype.

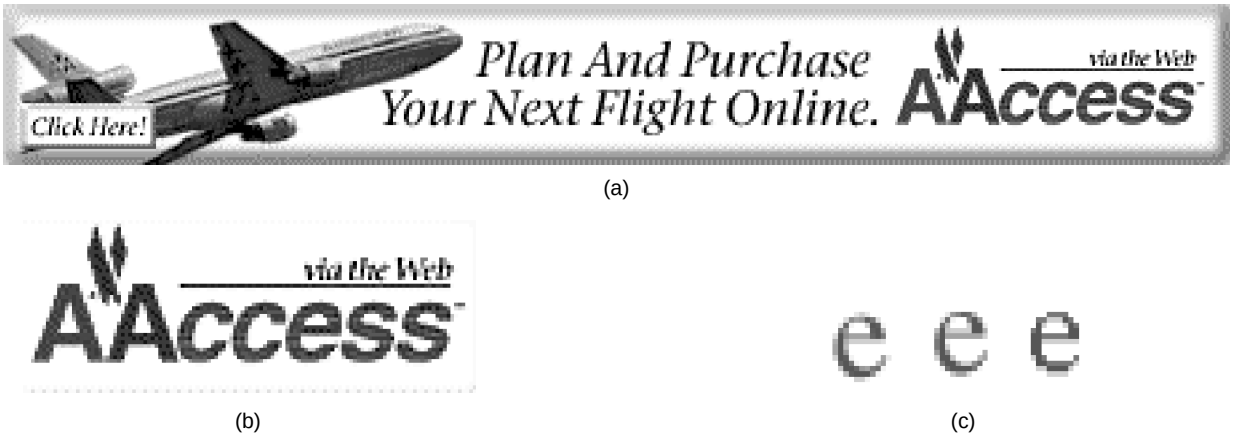

Fig. 16. Spatial sampling variation seen in text images digitized for low resolution displays like computer screens. The image in (a) was picked off the web. A part of it has been magnified in (b) to show the pixel configurations. The character samples in (c) were generated by Adobe Photoshop.

less important for higher sampling rates or, equivalently, bigger type-sizes. However, the experiments show that the effect is prevalent in bitmap samples of standard type-sizes and sampling resolutions, and demonstrate that random-phase sampling noise can account for much of the variation observed in scanned character bitmaps.

We have not attempted to provide a quantitative evaluation of the importance of noise due to spatial quantization relative to other sources of noise. Our primary goal was simply to call attention to an often neglected source of noise. Furthermore, hypothesis tests proposed to evaluate individual parameter settings in pseudo-random defect models, such as that described in [7], cannot be used precisely because the postulated measure of the similarity between two patterns cannot differentiate between correlated and uncorrelated sources of noise. More global tests, such as advocated in [9], are classifier-dependent.

Though our motivation has been the analysis of bilevel scanning, random-phase noise does not pertain exclusively to either bitmaps or optical scanning. A new area in which the issue of spatial sampling variation may play a prominent role is the processing of text embedded in in-line images in World Wide Web documents (Fig. 16). Text images embedded in a Web page are often generated through a drawing software (e.g., Adobe Photoshop), by quantizing an abstract, high-level description of an image (e.g., PostScript description) with respect to the origin of the drawing window, at a given resolution (typically 72 dpi). The exact same character in the same font can be quantized differently depending on where it is placed in the image. Fig. 16c shows different ways a letter "e" is quantized by Adobe Photoshop. Because of the low sampling resolution (72 dpi), and because such text is generally free of degradations commonly seen in scanned text such as optical blur and speckle, spatial sampling variation becomes a major source of noise for the text image.

5 Conclusion

We have applied the notion of locales to printed characters, established an upper bound on the number of distinct bitmaps that can be generated by displacements of the scanning grid with respect to an analog pattern under ideal conditions, and computed the frequency of occurrence of each bitmap under uniformly random grid displacements. These extensions of Havelock's ideas open up several interesting areas of investigation. While researchers in remote sensing and computer vision (such as Bruckstein, Havelock, and O'Gorman) are interested primarily in the location of objects, particularly edges, to subpixel accuracy, in OCR the objective is accurate recognition of the character regardless of its location.

The dependence of the number of distinct bitmaps on the size of the patterns, with additional allowances made for additive noise, provides a basis for establishing an acceptable sample size for training OCR classifiers.

Predicting, rather than just counting, all possible bitmap representations of a printed character as a function of the spatial sampling rate leads to the minimum sampling resolution that guarantees a given Hamming distance between bitmaps of characters of different classes. A more difficult problem is determination of the effect of random-phase sampling noise on the types of features used in OCR. Although OCR system designers have for decades attempted to construct features that are resistant to edge noise, there is now hope of introducing quantitative considerations into the design process.

Our work improves the foundation for pseudo-random defect models. Such models are enjoying increasing popularity for generating large sets of identified character bitmaps, but there has been little success so far in estimating their underlying parameters. Random phase sampling appears to reproduce accurately the observed variation among characters scanned from the same source. Conversely, it should be possible to reconstruct an analog pattern optimally from a set of digitized samples that differ primarily with regard to the grid displacement. This might be useful in creating a digital font from a scanned sample of conventional print in a rare typeface or script (e.g., Tibetan).

Finally, we note that the formal treatment can be extended to multilevel amplitude quantization (as in gray-scale OCR) and to higher spatial dimensions (3D voxel models used in range maps and tomography).[8] We conclude that the correspondence between locales and distinct bitmaps offers a rich field of study that extends well beyond the applications proposed in earlier papers.

Acknowledgments

George Nagy and Prateek Sarkar gratefully acknowledge the support of Nortel Educational Research Network. This work has been conducted in the New York State Center for Advanced Technology (CAT) in Automation, Robotics, and Manufacturing at Rensselaer, which is partially funded by a block grant from the New York State Science and Technology Foundation.

8. The generalized modulo-grid diagram for a pattern having a finite number of gray levels can be obtained by considering several supports X_1, X_2, ..., each corresponding to a particular gray level, and inscribing all the boundaries in the unit cell.

REFERENCES

[1] M.A. Armstrong, *Basic Topology*. New York: Springer Verlag, Inc., 1983.

[2] H.S. Baird, "Document Image Defect Models," *Structured Document Image Analysis*, H.S. Baird, H. Bunke, and K. Yamamoto, eds., pp. 546-556. New York: Springer Verlag, 1992.

[3] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, 1990.

[4] D.I. Havelock, "Geometric Precision in Noise-Free Digital Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1,065-1,075, Oct. 1989.

[5] D.I. Havelock, "The Topology of Locales and Its Effect on Position Uncertainty," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 380-386, Apr. 1991.

[6] R. Ingold, "Structure de documents et lecture optique: une nouvelle approche," PhD thesis, Presses Polytechniques Romandes, Lausanne, Switzerland, 1989.

[7] T. Kanungo, H.S. Baird, and R.M. Haralick, "Validation and Estimation of Document Degradation Models," *Proc. Fourth Univ. of Nevada at Las Vegas Symp. Document Analysis and Information Retrieval*, pp. 217-228, Las Vegas, Nev., 1995.

[8] T. Kanungo, R.M. Haralick, and I. Phillips, "Nonlinear Global and Local Document Degradation Models," *Int'l J. Imaging Systems and Technology*, vol. 4, pp. 220-230, 1994.

[9] Y. Li, D. Lopresti, G. Nagy, and A. Tomkins, "Validation of Image Defect Models for Optical Character Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 2, pp. 99-108, Feb. 1996.

[10] M. Mantyla, *An Introduction to Solid Modeling*. Computer Science Press, 1988.

[11] G. Nagy, "On the Auto-Correlation Function of Noise in Samples Typewritten Characters," *IEEE Region III Convention Record*, New Orleans, La., 1968.

[12] M. Nadler and E.P. Smith, *Pattern Recognition Engineering*. New York: John Wiley & Sons, 1993.

[13] G. Nagy, P. Sarkar, D. Lopresti, and J. Zhou, "Spatial Sampling Effects in Optical Character Recognition," *Proc. Third Int'l Conf. Document Analysis and Recognition*, pp. 309-314, 1995.

[14] L. O'Gorman, A.M. Bruckstein, C.B. Bose, and I. Amir, "A Comparison of Fiducial Shapes for Machine Vision Registration," *Proc. IAPR Workshop Machine Vision and Applications* (*MVA '90*), pp. 253-256, Tokyo, Nov. 1990.

[15] L. O'Gorman, "Subpixel Precision of Straight-Edged Shapes for Registration and Measurement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 746-751, July 1991.

[16] T. Pavlidis, *Algorithms for Graphics and Image Processing*. Computer Science Press, 1982.

[17] H.J. Rubinstein, *Digital Typography: An Introduction to Type and Composition for Computer System Design*. Addison-Wesley, 1988.

[18] P. Sarkar, "Random Phase Spatial Sampling Effects in Digitized Patterns," master's thesis, Rensselaer Polytechnic Inst., 1994.

[19] J. Zhou and D. Lopresti, "Repeated Sampling to Improve Classifier Accuracy," *Proc. IAPR Workshop Machine Vision Applications*, pp. 346-351, Kawasaki, Japan, Dec. 1994.