Spatial Sampling Effects in Optical Character Recognition

Daniel Lopresti Jiangying Zhou

Matsushita Information Technology Laboratory
Two Research Way
Princeton, NJ 08540, USA

George Nagy Prateek Sarkar

Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Abstract

In this paper we examine the effects of random-phase spatial sampling on the optical character recognition process. We start by presenting a detailed analysis in the case of 1-dimensional patterns. Empirical data demonstrate that our model is accurate. We then give experimental results for more complex, 2-dimensional patterns (i.e., printed, scanned characters). Spatial sampling seems to account for a significant amount of the variability seen in practice.

1 Introduction

The study of noise in scanned characters, along with the development of synthetic character generators, has attracted a great deal of interest over the past three decades [8, 1, 2, 5]. The issue of evaluating noise models from the standpoint of the OCR error behavior they induce has also recently received attention [6, 9]. However, most previous work along these lines has rarely involved the direct comparison of real and synthetic characters. As a result, it appears that a very significant – and completely unavoidable – source of character variation, random-phase spatial sampling, has been largely overlooked (or at least underemphasized) by the OCR community. The objective of this communication is to demonstrate that random spatial sampling noise gives rise to several readily observable phenomena that bear directly on optical character recognition. Greater appreciation of this type of noise may also lead to more realistic synthetic character generators.

Spatial sampling noise is the consequence of the random, uniformly distributed displacement of the scanner sampling grid with respect to the characters. Its importance lies in the fact that it produces correlated, character-dependent variations that affect OCR systems very differently from independent, identically distributed (i.i.d.) perturbations with the same signal-to-noise ratio. Furthermore, such sampling-phase noise cannot be eliminated by technical scanner refinements, although its effect can be mitigated by increasing the sampling resolution.

In this paper, we examine the effects of spatial sampling from several standpoints. We begin by presenting a brief theoretical analysis for the case of 1-dimensional patterns in Section 2. Our model has been confirmed experimentally – Section 3 gives data to support this. In Sections 4 and 5, we consider the more complex case of 2-dimensional patterns (*i.e.*, characters). Finally, we summarize the paper and offer some conclusions in Section 6.

2 Analysis of 1-Dimensional Case

In this section, we explore the effects of sampling variation in one dimension. We shall assume that the grid shift is a random variable uniformly distributed over a distance equal to one sampling interval. The digitized pattern is generated by sampling an analog pattern at integral coordinate points (see Figure 1).

For a single black stroke of length $L = l$, digitization produces a run-length of N black pixels with the following probability distribution on N:

$$\Pr(N = n | L = l) = \begin{cases} \lceil l \rceil - l & \text{if } n = \lceil l \rceil - 1 \\ 1 - \lceil l \rceil + l & \text{if } n = \lceil l \rceil \\ 0 & \text{otherwise} \end{cases}$$

In the case of an analog pattern made up of multiple strokes, we place the pattern along a coordinate axis. As we slide it along the axis by small amounts, the resulting digitized pattern changes. We see a new pattern as soon as one of the boundaries crosses a sampling point, and we continue to see the same pattern until another boundary crosses a sampling point. Figure 1 shows a binary waveform and the corresponding

$\{0, 1\}$ strings produced at four different displacements of the sampling grid.

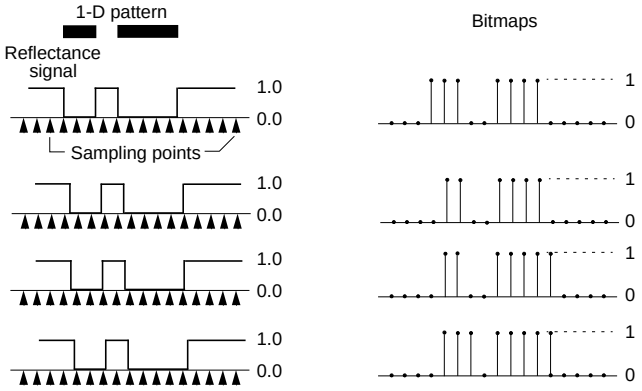

Figure 1: Spatial sampling in 1-dimension.

It can be shown that the number of possible digital variations equals the number of boundaries (color transitions) in the analog pattern [11]. Furthermore, given the coordinate locations of the transitions in the analog pattern, the distance the pattern can be moved between two successive boundary crossings equals the relative probability we will see the corresponding digital pattern. Figure 2 plots frequencies-of-appearance for the patterns of Figure 1 at two sampling resolutions.

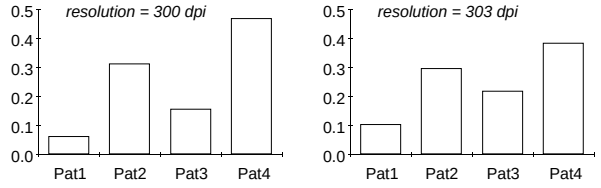

Figure 2: Digitized pattern probabilities.

As a direct consequence of spatial sampling variation, we note that stroke length is not unimodally distributed. The histogram for a single black stroke has a sidelobe on the right if its length is just under an integral number of sampling intervals, and on the left if its length is just over (see Figure 3).

3 Evaluation of 1-Dimensional Case

To test the validity of our model, we conducted an experiment comparing the predicted and observed probabilities for a simple pattern. For input, we used a linear bar-code from [10]. The pattern was placed with the bars slightly tilted ($\approx 2^{\circ}$) to simulate a uniform

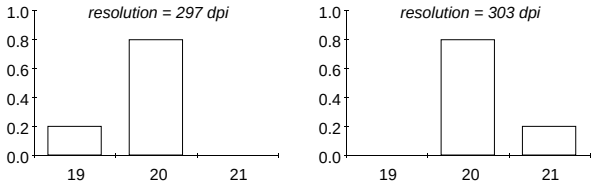

Figure 3: Stroke-length histograms.

Quantized Pattern	*Probability*			
	Predicted	*Observed*		
		Test	*Training*	*Overall*
4,7,4,4,3,5,3	0.01	0.066	0.025	0.041
4,7,3,5,3,5,3	0.19	0.170	0.180	0.176
3,8,3,5,3,5,3	0.40	0.290	0.390	0.350
3,7,4,5,3,5,3	0.02	0	0.031	0.019
3,7,4,4,4,4,4	0.24	0.283	0.220	0.243
3,7,4,4,3,5,4	0.02	0.066	0.031	0.045
3,7,4,4,3,5,3	0.12	0.085	0.087	0.086
Partial Sum	1.0	0.962	0.957	0.959
3,7,4,5,3,4,4	0	0.012	0	0.007
3,7,4,4,4,5,3	0	0.006	0.019	0.011
3,8,3,4,4,5,3	0	0.018	0.009	0.015
3,7,4,4,4,4,3	0	0.006	0	0.004
3,8,3,5,3,4,4	0	0	0.009	0.004

Table 1: Experimental evaluation of a 1-dimensional pattern consisting of four black strokes.

shift-rate in the longitudinal direction. The scanning spot-size was set at 1/75 of an inch. The rows of the resulting scanned pattern were used as 1-dimensional samples. The total number of patterns so obtained was 267.

Individual stroke widths were estimated from a portion of the sample set (161 of the patterns). These were used to predict the possible digitized patterns and their respective probabilities of appearance. The results were then compared to the observed frequencies for the remainder of the sample set (106 patterns). Table 1 presents this data.

Similar experiments were carried out with other 1-dimensional patterns [11]. In all cases we found that our spatial sampling model could predict the most common digital patterns, and also provide a relatively accurate estimate of their frequencies-of-appearance. Less than 10% of the patterns in each case did not belong to the predicted set of variations. It is evident there are other factors that affect the digitization process, many related to the scanning hardware and difficult to analyze, but clearly random-phase noise in sampling is itself a major factor worth studying.

4 Models for 2-Dimensional Case

While 1-dimensional patterns are amenable to rigorous analysis, more complex 2-dimensional shapes (*e.g.*, printed characters) are not. In this section, we describe an empirical model for characterizing digitization noise in 2-dimensional patterns.

We assume a sampling grid of square cells arranged in vertical columns and horizontal rows. Each cell covers an area equal to one unit, which is also the size of a pixel. Figure 4 illustrates the variability caused by different placements of the sampling grid. Clearly, if the grid is perfectly aligned with the bitmap being sampled, each cell will cover exactly one pixel (*e.g.*, the upper portion of Figure 4). However, in the case the grid is shifted, cells may overlap several pixels of different colors (*e.g.*, the lower portion of Figure 4). Under our model, we decide a cell is white if the area it covers is more white than black.

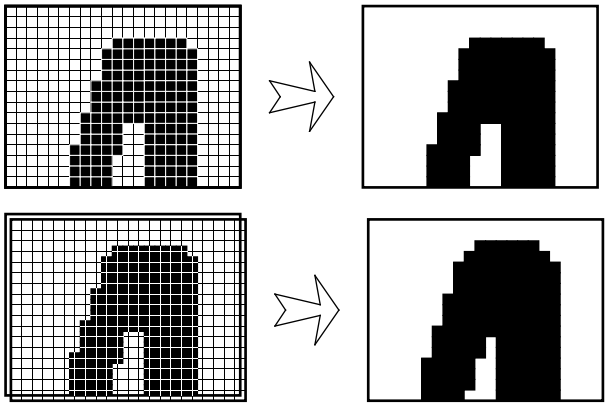

Figure 4: Variation due to sampling grid placement.

If the integral of the point spread function is approximately linear near the threshold, then the effects of threshold variation can be approximated by a random displacement of the sampling delta function (sampling spot) with respect to the edge in the input image [11]. However, such an approximation is valid only on at those edge points in the vicinity of which the edge is roughly straight. The approximation does not reflect accurately, for example, the round-up phenomena at corners induced by the sensor's point spread function. We therefore apply an additional blurring process at the corners of a contour prior to threshold variation.

Figure 5(a) simulates random-phase spatial sampling on three instances of a 12-point Helvetica 'M'. Figure 5(b) is the same reference character digitized with a small amount of independent, identically distributed threshold noise. The effects of combining random-phase sampling with threshold noise are illustrated in Figure 5(c). For comparison, Figure 5(d) shows three scan-digitized samples printed using a 600 dpi laserprinter.

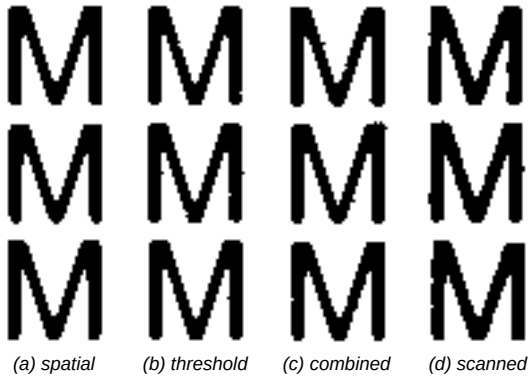

Figure 5: Synthesized and real character bitmaps.

5 Evaluation of 2-Dimensional Case

In this section, we present a study of the effects of spatial-sampling on the shape of 2-dimensional patterns.

We begin by considering some simple statistical measures. Figure 6 shows the distinct effects of the sampling phenomenon on the measured area for 900 instances of a 12-point Computer Modern Sans Serif 'e'. Histogram (a) is for spatial sampling, (b) for i.i.d. threshold noise, (c) for the combined effects of sampling and threshold variation, and (d) for real scanned characters. Similarly, Figure 7 shows stroke-width histograms for the same sets of characters. It is quite clear, for example, that average stroke width is more affected by sampling than by threshold noise.

Next, we consider the effects of digitization on a chain-code representation of the character's boundary. A chain-code consists of a string of symbols representing the vectors joining two neighboring pixels along the boundary of a shape [4]. We measure the difference between two shapes by comparing their chain-codes using well-known string matching techniques [12, 3]. Figure 8 shows, for example, two instances of the letter '-', the chain-code for each, and the difference between the two images in terms of string edit distance. From the results of the string matching, we compute an error distribution profile along the chain-code of a canonical reference character.

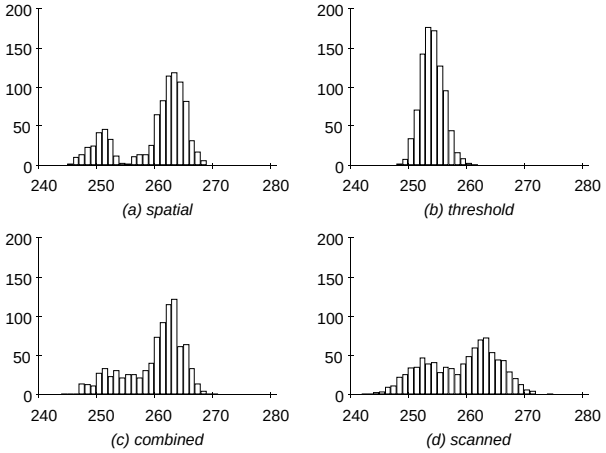

Figure 6: Area histograms for CM 'e'.

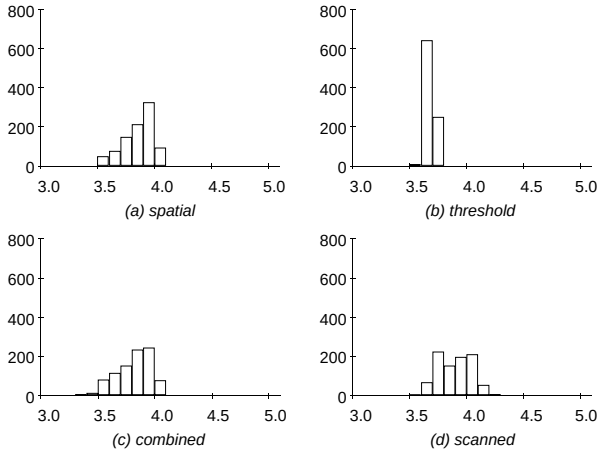

Figure 7: Stroke-width histograms for CM 'e'.

For our tests, we generated two sets of experimental data, both based on characters typeset in 12-point Helvetica:

1. Chain-codes for synthesized characters produced by sampling the reference character bitmap.

2. Chain-codes for real characters that had been printed and scanned.

The experiments were run with three different character patterns: '-', 'e', and 'M'. 1,000 identical characters were printed on a page at 600 dpi and scanned at 300 dpi. Two sets of synthetic data were generated, each containing 1,000 samples. In the first set, we applied only random threshold variation. In the second set, the samples were the result of the combined effects of random grid shifts and threshold variation

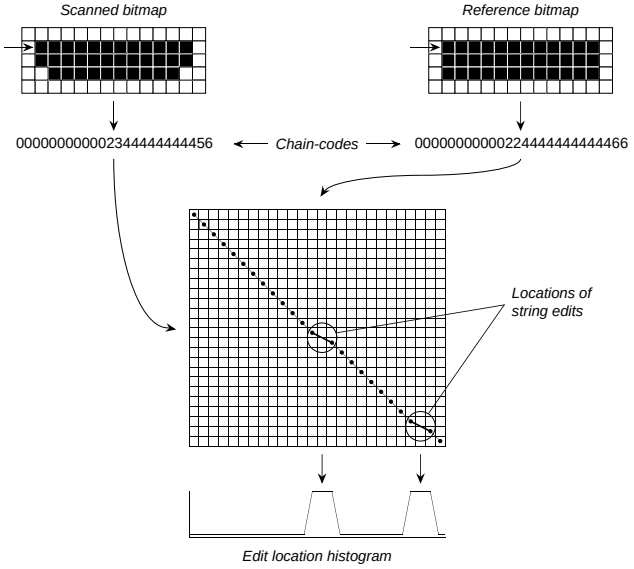

Figure 8: Using string matching to describe the differences between character shapes.

Figures 9 – 11 show histograms of the number of edits performed at each chain-code position for the 1,000 copies of '-', 'e', and 'M', respectively. The upper half of each figure represents the synthetic data and the lower half the real data. As can be seen, the profiles for the real data are quite different from those for the synthetic. It is evident that threshold variation alone accounts for little of the distortion we observed in the digitized patterns.

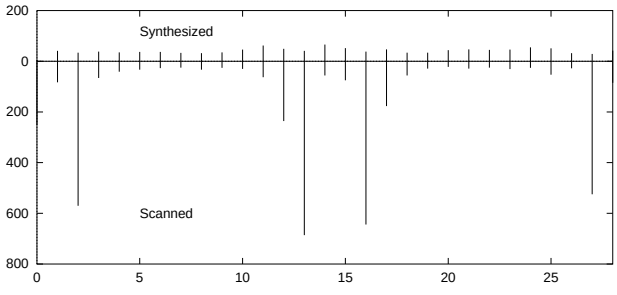

Figure 9: String edit histogram for '-' (threshold only).

When combined with spatial sampling noise, however, the results are much closer to the real data. Figures 12 – 14 show histograms of the number of edits performed at each chain-code position for the 1,000 copies of '-', 'e', and 'M', respectively. As before, the upper half of each figure represents the synthetic data and the lower half the real data. The characters generated using the combined spatial/threshold model seem

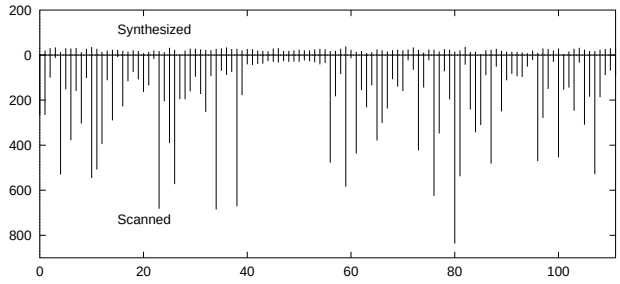

Figure 10: String edit histogram for 'e' (threshold only).

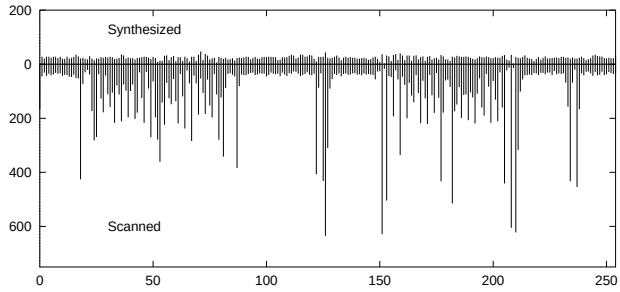

Figure 11: String edit histogram for 'M' (threshold only).

to match the scanned characters quite well.

The effects of sampling noise on chain-code representations are non-uniform; rather, they exhibit certain structural patterns. Each chart has numerous characteristic peaks. The scanned characters seem a bit "richer." This suggests there are other sources of noise which contribute to differences in the chain-codes. Still, the locations of most of the peaks are duplicated precisely in the synthesized characters.

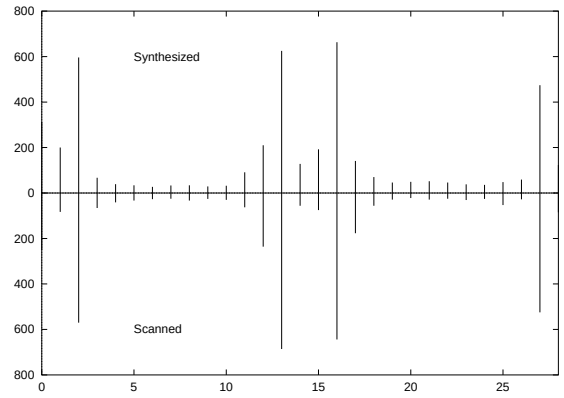

Figure 12: String edit histogram for '-' (combined spatial/threshold).

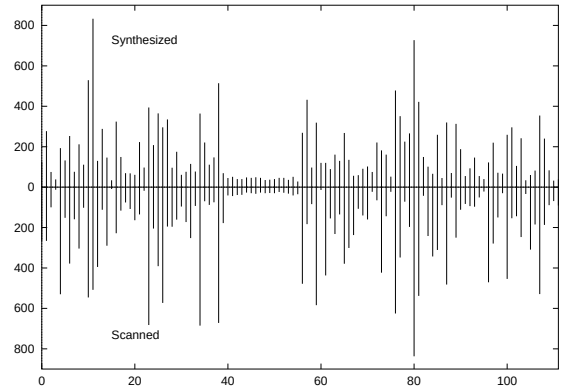

Figure 13: String edit histogram for 'e' (combined spatial/threshold).

6 Conclusions

In this paper, we have explored the effects of random-phase spatial sampling on scanned patterns. We verified a prediction that the size of a pattern (*i.e.*, the number of black pixels) has a distinctly multimodal distribution. According to our model, the relative heights of the peaks depend on minute variations in sampling resolution, character-dimension, and stroke width. Variations in the latter are greater than can be accounted for by i.i.d. pixel noise. We demonstrated that there exists a significant amount of structure in the variations in chain-code representations of character boundaries. Furthermore, we showed that it is possible to statistically predict pattern variation: as a function of "ideal" shape, sampling frequency, and threshold noise.

At this early stage, we have only investigated the effects of sampling noise on lower-level features (*i.e.*, pixels). A more complicated, but very important, task is to characterize the effects of sampling on higher-level features and the performance of vision algorithms and systems as a whole. One such preliminary study is presented in [13]. A goal of our future work is to study the relationships between the scanning process, pattern shape, and the mathematical algorithms we use to process them.

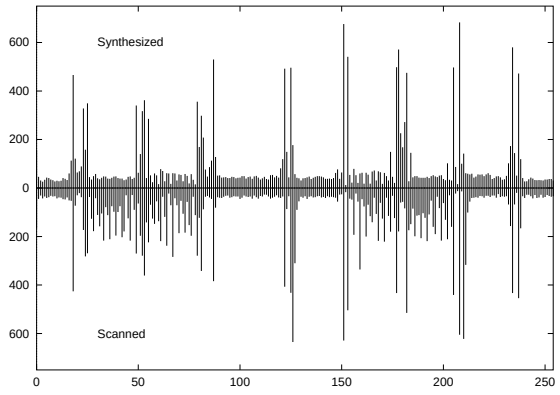

Figure 14: String edit histogram for 'M' (combined spatial/threshold).

Acknowledgments

We would like to thank Dr. Tin Kam Ho for providing the character outlines we used in some of our studies. George Nagy gratefully acknowledges the financial support of the Northern-Telecom/BNR Educational and Research Networking Program.

A portion of this work was conducted at the New York State Center for Advanced Technology (CAT) in Automation and Robotics at Rensselaer Polytechnic Institute. The CAT is partially funded by a block grant from the New York State Science and Technology Foundation.

A longer version of this paper appears as [7].

References

[1] H. S. Baird. Calibration of document image defect models. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, pages 1–16, Las Vegas, NV, April 1993.

[2] M. Buchman. Distortion modeling for document images. In *Proceedings of the DIMUND Workshop on Page Decomposition, Character Recognition, and Data Standards*, Harper's Ferry, WV, August 1993.

[3] J. Esakov, D. P. Lopresti, J. S. Sandberg, and J. Zhou. Issues in automatic OCR error classification. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 401–412, Las Vegas, NV, April 1994.

[4] H. Freeman. Computer processing of line drawing images. *Computer Surveys*, 6:57–98, 1974.

[5] T. Kanungo, R. Haralick, and I. Phillips. Global and local document degradation models. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 730–738, Tsukuba Science City, Japan, 1993.

[6] Y. Li, D. Lopresti, and A. Tomkins. Validation of document image defect models for optical character recognition. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 137–150, Las Vegas, NV, April 1994.

[7] D. Lopresti, G. Nagy, P. Sarkar, and J. Zhou. Spatial sampling effects in optical character recognition. Technical Report #95-6931, Rensselaer Polytechnic Institute, Troy, NY, April 1995.

[8] G. Nagy. On the auto-correlation function of noise in sampled typewritten characters. In *IEEE Region III Convention Record*, New Orleans, LA, 1968.

[9] G. Nagy. Validation of simulated OCR data sets. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 127–135, Las Vegas, NV, April 1994.

[10] T. Pavlidis, J. Swartz, and Y. P. Wang. Fundamentals of bar code information theory. In *IEEE Computer*, pages 74–86, 1990.

[11] P. Sarkar. Random phase spatial sampling effects in digitized patterns. Master's thesis, Rensselaer Polytechnic Institute, Troy, NY, 1994.

[12] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

[13] J. Zhou and D. Lopresti. Repeated sampling to improve classifier accuracy. In *Proceedings of the IAPR Workshop on Machine Vision Applications*, pages 346–351, Kawasaki, Japan, December 1994.