DOCUMENT IMAGE QUALITY: MAKING FINE DISCRIMINATIONS

HENRY S. BAIRD
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304 USA

E-mail: baird@parc.xerox.com
Tel.: +1-650-812-4481
FAX: +1-650-812-4374

We estimate, using synthetically generated images, the smallest changes in document image quality that can be distinguished reliably and fully automatically by Kanungo's bootstrapping test [Kan96]. Six parameters of a physics–based document–image degradation model [Bai92] are varied, one at a time: for each, over a range of parameter-value differences, two sets of synthetic images are generated pseudorandomly and the two sets tested for statistical equivalence using Kanungo's method. The rate at which Kanungo's method rejects the hypothesis that the two sets are drawn from the same distribution is analyzed as a function of parameter difference (a specialized "power function"). The finest discriminations afforded by the method are given by the width of the power function at a low fixed reject threshold. The data show that remarkably fine discriminations are possible – often subtler than are evident to visual inspection – for all six parameters. As few as 25 reference images are sufficient. These results suggest that Kanungo's method is sufficiently sensitive to a wide range of physics–based image degradations to serve as an engineering foundation for many image–quality estimation and OCR engineering purposes.

Keywords: document image quality, image degradation, bootstrapping, just noticeable differences

1 Introduction

1.1 Document Image Quality

Low–quality images of documents pose serious technical challenges to current recognition technologies [ISRI95]. This affects the majority of FAXes, microfiche and microfilm images, a substantial fraction of the vast legacy document collections in libraries and corporate archives, many newspapers, and even, due to the high cost of systematic quality control in batch scanning, a surprisingly large fraction of images of carefully printed documents acquired by high-quality scanners. As yet, we have no proven methodologies for measuring image quality or for systematically improving recognition accuracy on low-quality images. This is one in a series of studies with the ultimate goal of supplying such methodologies. The aim of this study is to characterize critical engineering properties — sensitivity to small changes in degradation and uniformity across many degradation parameters — of a statistical technique that has shown potential to be fundamental to this research program. (This paper is an expanded version, with new results, of [Bai98].)

1.2 Kanungo's Bootstrapping Method

Kanungo's method [Kan96] for the estimation of parameters of image degradation models applies a statistical bootstrapping technique to the task of distinguishing between two sets of images which may be drawn from

This is the fully fleshed–out text from which the 4-page camera–ready copy was excerpted for publication in *Proc., Int'l Conf. on Document Analysis and Recognition*, Bangalore, INDIA, September 20–22, 1999.

the same or different distributions (more precisely, but equivalently, their degradation–model parameters are drawn from these distributions). Such a test can decide to *reject* the null hypothesis that the distributions are the same, at a given confidence level (but it cannot, for fundamental theoretical reasons, *accept* the hypothesis with any given confidence). The method requires that the user pick a false reject ('misdetection') rate, in order to design the test: we set this to 5%. Kanungo has shown [KHBSM94] how to use the 'power function' (reject rate as a function of a model parameter) to estimate parameters of a morphologically motivated local degradation model. We now extend the use of power functions to the measurement of small changes in image quality.

1.3 A Physics-Based Image Degradation Model

A ten-parameter model was proposed in [Bai92] that approximates some aspects of the known physics of machine-printing and imaging of text, including symbol size, spatial sampling rate and error, affine spatial deformations, jitter, speckle, blurring, and thresholding. [Bai93] reviewed the then state of the art of methods for the calibration of such models, which did not include bootstrapping. We now apply bootstrapping and power-function analysis to this physics-based model.

2 Experiments

We have run two sets of experiments, using pseudo–randomly generated synthetic images, to provide a baseline reject threshold and to assess the sensitivity of Kanungo's method to small changes in image quality.

In each test, two sets of images to be compared contained samples drawn from distributions in which all variation among the images was due to spatial sampling error (uniform in X and Y over the range [0,1] pixel) and two kinds of per-pixel randomization, sensitivity and jitter (cf. [Bai93]). The other six parameters of the model — skew angle, size of blurring kernel, binarization threshold, pixel sensor sensitivity error, horizontal scaling, and vertical scaling — were fixed, with known means and zero variance. The images were of the letter 'A' in the Adobe Times–Roman typeface, at a typesize of 8 point, rendered as a bilevel image at 300 pixels/inch (Figures 1 and 2 illustrate this for two values of the blurring parameter).

Figure 1: Example of a set of 20 synthetic pseudo–randomly generated character images, slightly blurred (blurring parameter set at 0.7 pixels; other parameters are set to nominal mean values, given in Section 2.1).

Figure 2: Example of a set of 20 synthetic images, grossly blurred (blurring parameter set at 2.0 pixels). Effects due to other degradation parameters are more apparent, but they are in fact set to the same values used in Figure 1.

Each test was rerun 100 times (freshly pseudo-randomized each time) in order to estimate the reject rate.

The pseudorandom seeds used were the thirteen least significant bits of the computer's CPU clock. The set-to-set distance used was nearest neighbor; the image-to-image distance was Hamming distance (pixel-wise exclusive OR) after superimposing the centers of the bounding rectangles. Each test involved 100 bootstraps applied to 100 randomized partitions each (for explanations of these, see [Kan96].)

2.1 Baseline Trials

We began with "baseline" trials to test the uniformity of Kanungo's method across all parameters of the model and to estimate a reliable reject threshold for use later. In these trials, we set the number of samples per set (N) to 100.

For each of the six model parameters, we picked a set of eleven parameter values, covering a wide range important in practice, suggested by prior experience.

Parameter	**Range of values**
skew angle	-3 -2 -1 -0.5 -0.25 0* 0.25 0.5 1 2 3
blur kernel	0.3 0.4 0.5 0.6 0.7* 0.8 0.9 1.0 1.1
threshold	0.15 0.175 0.20 0.225 0.25* 0.275 0.30 0.325 0.35
sensitivity	.000 .033 .067 1.000 1.125* 1.158 1.192 1.225 1.250
X scaling	0.6 0.7 0.8 0.9 1.0* 1.1 1.2 1.3 1.4
Y scaling	0.6 0.7 0.8 0.9 1.0* 1.1 1.2 1.3 1.4

While each parameter ranged over the values shown, the other five were fixed at their "nominal mean" values, indicated by '*' above. Thus there were 66 trials: in each, two image sets were generated pseudo–randomly, both drawn from the *same distribution*. The mean reject rate observed over all 66 trials was 5.2%, with a standard error of 3.2%. Note that this observed mean is, as expected, close to the designed misdetection rate of 5%.

Non–zero variance is, of course, to be expected since the number of trials, bootstraps, partitions, and samples per set are finite. Ideally, the variance should be determined only by these choices (and the character's artwork), not by details of the bootstrap implementation. Any implementation of Kanungo's method requires several choices for which the underlying statistical theory doesn't offer any guidance. These include the choice of set-to-set distance function and image-to-image metric. Kanungo [KHB95] compared three set-to-set distance functions and found empirical grounds for preferring the trimmed-mean nearest-neighbor distance. But it has been an open question whether or not these or other choices inadvertently bias the test in ways that matter in practice. The danger is that a particular image metric can conceivably turn out to be more or less sensitive to specific types of image degradation. As a somewhat artificial example of this, suppose that the image–to–image metric normalized aspect ratios before comparing images: then Kanungo's method would be oblivious to many commonly occurring affine geometric distortions.

Happily, we observed no systematic differences in reject rate mean or variance from one parameter to another, and no significant correlation of reject rate with the changing values of any single parameter. This uniformity of reject behavior across all of the parameters of a physics-based model raises our confidence that the bootstrapping test, as implemented, is not biased for or against any particular degradation parameter.

These baseline trials also provide a single, statistically conservative, reject threshold (T) — chosen to be three standard errors above the mean — which is applicable uniformly to all model parameters, for estimating the sensitivity of the method. Here, with N = 100, T is set to 0.15. In similar trials, for N=50 we estimate T=0.20, and for N=25, T=0.25.

2.2 Sensitivity Trials

In the second series of trials, for each of the six parameters we compared pairs of sets of images drawn from *different distributions*: the first using a "reference" value (here set equal to the nominal mean value); and the second using different "probe" values. As above, only one model parameter was varied at a time, and the variances of all six parameters were set at zero. The goal of these trials was to estimate the finest discriminations that Kanungo's method can make: *i.e.* its threshold of sensitivity to changes in degradation.

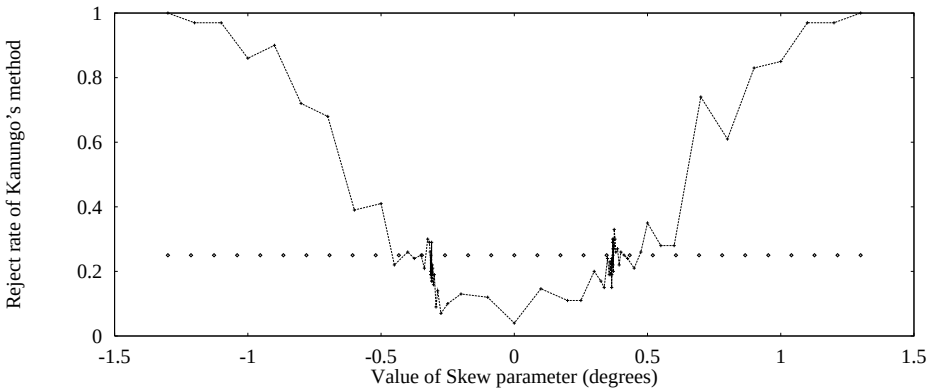

Figure 3: Power function for discrimination trial for the skew parameter. Skew = 0.0 is the reference value. The number of samples per set N = 25. The conservative reject threshold T = 0.25 (indicated by dotted line). Note the iterative refinement of the search for the two critical values where the threshold is crossed.

For this purpose we sampled the power function (Figure 3) — that is, the reject rate of Kanungo's method as a function of the probe value — coarsely at first to locate the two critical values, on either side of the global minimum, where the conservative reject rate T is first exceeded. Then, iteratively, we sampled more finely about these critical values. The average of the absolute distances of the critical values from the reference value — in practice they are roughly equal in magnitude — is identified as the "least discriminable difference" of the method for that model parameter. The results of these measurements for N = 100 and 25 (and T = 0.15 and 0.25 respectively) are given below.

Parameter	**Least discriminable differences**		
	N=100	*N=25*	*Units*
skew angle	±0.02	±0.3	degrees
blur kernel	±0.02	±0.05	pixels
threshold	±0.01	±0.03	intensity
sensitivity	±0.04	±0.075	intensity
X scaling	±0.02	±0.015	dimensionless factor
Y scaling	±0.02	±0.015	dimensionless factor

These data show, in most cases, an expected loss of sensitivity as the number of samples per set is dropped by a factor of four. For the skew angle parameter, the loss is large: a factor of 15 — ! For the rest, the loss is relatively modest: a factor of three at the most. The *increase* in sensitivity for the scaling parameters is unexpected and anomalous, and, although small, is being scrutinized.

In all the trials discussed so far — the baseline trials as well as the discrimination trials immediately above – we generated fresh pseudorandom samples for both sets in each of the 100 executions of Kanungo's method used to estimate the reject rate. So, for N=25, each reject rate estimate required 2*25*100 = 5000 independently generated (and presumably mostly distinct) image samples.

When, in the future, we apply these methods to "real" (unsynthesized) images, we cannot expect often to be able to gather such a large number of distinct images. It is natural to ask how the sensitivity measurements are affected by severely limiting the number of real samples. To answer this question, we modified the above procedure so that, during each reject–rate estimation, we generated only a *single* set of reference images (the set drawn from the distribution determined by the nominal mean value), while continuing to generate 100 fresh sets of probe images. So, for N=25, only 25 reference images are required compared to 2500 probe images. In future applications, real images would be used as the reference set. There would be of course, no obstacle to continuing to use a much larger number of synthetic images. Under this modified, more efficient, and less data-hungry procedure, the results were as follows.

Parameter	**Least discriminable differences**	
	N=25	*Units*
skew angle	±0.3	degrees
blur kernel	±0.05	pixels
threshold	±0.03	intensity
sensitivity	±0.07	intensity
X scaling	±0.015	dimensionless factor
Y scaling	±0.015	dimensionless factor

Remarkably, there is very little deterioration in the sensitivity of the method. These experiments have therefore not revealed any serious obstacle to applying Kanungo's method where only a small number of real images can be expected.

3 How fine do these discriminations appear to be?

These discriminations appear to be remarkably fine. For each model parameter, three sets of 20 images were generated and printed: the first grossly degraded, the second at the nominal mean value, and the third whose origin is randomly selected: it may be drawn from the same distribution as the second set, or more or less degraded than it by the least discriminable difference in the table above. A few people were then asked to decide, by visual inspection, which of the three possible origins for the third set is correct. Figures 4–6 are an example; please examine them before reading on.

The correct answer in the cases above is that the third set (in Figure 6) is in fact rotated slightly: by 0.3 degrees, which is the least discriminable difference of Kanungo's method given N=25 reference sample images.

The author, applying a series of such tests to himself, could not make the correct decision reliably; nor could several of his colleagues. For the other five parameters, such tests seem even harder than for skew angle.

This highly informal straw poll cannot support any strong psychophysical claims — but it may help the interested reader grasp intuitively the subtlety of the discriminations that Kanungo's method is capable of making.

4 Discussion

The sensitivity of Kanungo's method to minute changes in image degradation, apparently often subtler than can be distinguished by eye, encourages us to believe that it is suitable for engineering purposes.

These trials have also shown that as few as 25 image samples are sufficient in many cases to make fine discriminations. This raises confidence that Kanungo's method can yield useful results in practical applications where the number of "real" reference images is limited.

Figure 4: **Grossly rotated** synthetic pseudo–randomly generated character images (skew parameter set to 3.0 degrees clockwise). This degradation should be plainly visible to the eye compared with the unrotated images just below.

Figure 5: **Unrotated** synthetic pseudo–randomly generated character images (skew parameter set to 0.0 degrees). Compare them carefully to the images above and below.

Figure 6: These images **may or may not be rotated** – the reader should try to judge which is the case, by inspection, before looking ahead in the text for the correct answer.

It will be interesting, in future work, to assess Kanungo's method's sensitivity to the artwork of other characters and typefaces, to smaller type sizes relative to the spatial sampling rate, to non-zero variance in the model parameters, and to jointly varying parameters.

This success of Kanungo's method on synthetic images clears the way for future, far more effortful, trials on scanning hardware. Ultimately, these engineering studies may give us enough understanding of the behavior of Kanungo's method that we can use it reliably and quickly to fit degradation models to populations of real images.

5 Acknowledgments

The author is indebted to Tapas Kanungo for encouragement, helpful advice, and early versions of his bootstrapping code. Elisa H. Barney Smith carefully critiqued an early draft. Perry Stoll of Xerox ScanSoft kindly made available his implementation of the image degradation model. This project began at Bell Labs, Lucent Technologies and resumed at Xerox PARC. Gary Kopec, Dan Bloomberg, Les Niles, Kris Popat, and Jeanette Figueroa of PARC's Document Image Decoding area patiently assisted the author in countless ways, making a smooth continuation of this work not only possible but a pleasure.

6 Bibliography

[Bai92] H. S. Baird, "Document Image Defect Models," in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer–Verlag: New York, 1992, pp. 546-556.

[Bai93] H. S. Baird, "Calibration of Document Image Defect Models," *Proceedings, 2nd Annual Symposium on Document Analysis and Information Retrieval*, Caesar's Palace Hotel, Las Vegas, Nevada, April 26-28, 1993.

[Bai98] H. S. Baird, "Distinguishing Image Degradations using Bootstrapping," Proceedings, IAPR DAS'98 Workshop, Nagano, Japan, November 4–6, 1998.

[KHBSM94] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, & D. Madigan, "Document Degradation Models: Parameter Estimation and Model Validation," Proc., Int'l Workshop on Machine Vision Applications, Kawasaki, Japan, December 13-15, 1994.

[KHB95] T. Kanungo, R. M. Haralick, & H. S. Baird, "Power Functions and Their Use in Selecting Distance Functions for Document Degradation Model Validation," Proc., IAPR 3rd Int'l Conf. on Document Analysis & Recognition, Montreal, Canada, August 14-16, 1995.

[Kan96] T. Kanungo, *Document Degradation Models and Methodology for Degradation Model Validation*, Ph.D. Dissertation, Dept. EE, Univ. Washington, March 1996 [Supervisor: Prof. R. M. Haralick].

[ISRI95] Information Science Research Institute, 1995 Annual Research Report, Univ. Nevada at Las Vegas, Nevada, pp. 11–50, 1995.