

Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty
Publications and Presentations

Department of Electrical and Computer
Engineering

1-1-2003

Estimating Degradation Model Parameters from Character Images

Hok S. Yam
Boise State University

Elisa H. Barney Smith
Boise State University

This document was originally published by IEEE in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*. Copyright restrictions may apply.

Estimating Degradation Model Parameters from Character Images

Hok Sum Yam and Elisa H. Barney Smith
Electrical and Computer Engineering Department
Boise State University, Boise, Idaho 83725
EBarneySmith@boisestate.edu

Abstract

This paper discusses the use of character images to determine the parameters of an image degradation model. The acute angles in character images provide information used to find the model parameters. Three experiments are conducted to evaluate the use of characters. In the first experiment, large quantities of corners from character images are used to investigate how their contribution affects the mean and the standard deviation of the parameter estimators. In the second experiment, we focus on the relationship between the angles of the corners used in estimation and the estimation results. In the last experiment, we examine how likely the text in a common page would offer a reasonable estimation result compared to the results from experiments 1 and 2.

1. Introduction

Degradations in text images affect character recognition. These degradations are typically introduced through the bilevel processes of printing and scanning. It is desirable to have a calibrated model of the document process to predict how a document image will look after being subjected to these printing and scanning processes. This model could be used to generate large training sets of synthetic characters with the model parameters matched to the source document. This would ease the burden of hand segmenting and labeling data. A model will also allow researchers to conduct controlled experiments to improve OCR performance.

The degradation model used for this research is convolution followed by thresholding [1]. Several methods to calibrate this model have been proposed [2, 3, 5, 6]. These methods have included using combinatorial experiments to try every possible parameter value combination and taking measurements from specialized test charts such as a star sector test chart or large wedges. A more convenient way to calibrate the model would be to directly use character images rather than using specialized test charts. To do this we need to be able to use characteristics available in text images. We also need to know how many text images are needed to get high enough model calibration accuracies and whether the characters available in common documents are likely to have these characteris-

tics.

In this paper a model parameter estimation method that uses the corners is applied to corners existing in text images. Experiments are run to evaluate how many characters are needed to get an estimate within a given confidence level. Analysis is also conducted to see which corner angle measures, and thus which characters, are most useful for providing consistent estimators. The system is presented character samples with an occurrence frequency similar to what would be expected for a page of text, in order to simulate estimation from a common text page.

2. Estimation from Corner Images

The convolution and thresholding model is parameterized by two parameters: the width of the point spread function (PSF), w , and the binarization threshold level, Θ . A corner that is deformed by this model is shown in Figure 1. The vertex of the corner will become rounded and the edges of the corner will be displaced from their original positions, but will remain parallel to the original corner edges. Parameter estimation is done by relating the corner erosion distance to the amount of rounding seen after a corner is degraded. The black corner erosion distance, d_b , is [3, 5]

$$d_b(w, \theta; \phi) = \frac{-wESF^{-1}(\theta) + f_b^{-1}(\theta; w, \phi)}{\sin\left(\frac{\phi}{2}\right)} \quad (1)$$

where ESF (Edge Spread Function) is the cumulative marginal of the PSF, and f_b is the greyscale profile of the blurred corner along the bisector, defined as

$$f_b(d_{0b}; w, \phi d) = \int_{x=0}^{\infty} \int_{y=x \tan \frac{\phi}{2}}^{x \tan \frac{\phi}{2}} PSF(x - d_{0b}) dy dx \quad (2)$$

For any one measurement from a corner with an angle of measure ϕ , a locus of (w, Θ) points that could have produced the measurement will be available. The angle of the wedge is needed to determine this locus of points. This angle can be measured from the degraded corner because the edges in the degraded corner will be parallel to the original corner edges.

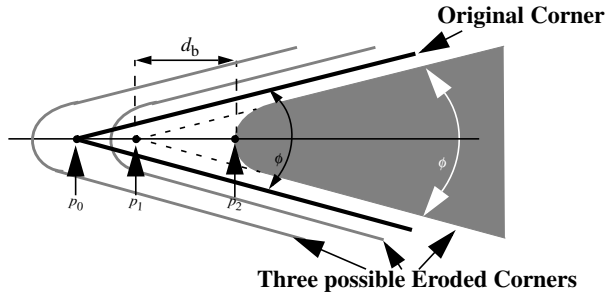


Figure 1: A degraded corner will have a rounded apex and the edges will be parallel to the original stroke edges. This can occur in three possible ways relative to the original corner.

A white corner on a black background will suffer a degradation similar to the degradation for a black corner. The white corner erosion distance, d_w , can be related to the black corner erosion distance, d_b , by

$$d_w(w, \Theta; \phi) = d_b(w, 1-\Theta; \phi). \quad (3)$$

These two erosion distances can be combined and the intersection of the loci of (w, Θ) points will yield a unique estimate of the unknown parameters to the model. The noise that is introduced into the image through spatial quantization and sensor noise will yield errors in measurements of the erosion distances. Also measurements of the original corner angle may be in error. This yields the need to use erosion distances from many corners and to average the intersection locations to decide on a final estimate of the model parameters.

3. Estimation from Character Images

Three different experiments were conducted to evaluate the feasibility of using text images to estimate model parameters. The first experiment considers how the number of corners used in the estimation process affects the final parameter estimate. The second experiment considers how the angle measure of the corner affects the estimates. The third experiment considers the natural frequency distribution of the characters on the page to evaluate whether enough of each corner is likely to be available to use in estimation.

We restrict our analysis to sans-serif fonts so that straight edges meet at the corners. The exact font details are not necessary because the angle of the corner is measured from the resulting image. The sans-serif fonts give some *a priori* information on the general shape of the character and the existence of 'simple' corners. The characters *kvwxyzAKVWXYZ* were chosen for use in these experiments because they contain acute angles for both exterior (black) and interior (white) corners. Samples from the 12 point font Arial are used in these experiments.

We use photo typeset images to separate the effects of printing from the effects of scanning. Each page consists of between 21 and 26 samples of each character. The characters were well spaced on the page to aid in segmentation. We scanned a page of text 10 times. Each time, we slightly moved the page (no more than 5 degree skew range) before it was scanned to make more variation in the sample images. Results in this paper focus on using two different thresholds to convert the grey-scale images to bilevel images. Some characters were not considered because they contained missing pixels inside the character strokes or broken strokes. For the threshold 64, 5438 white corners and 5288 black corners were generated. For the threshold 125, 5990 white corners and 6132 black corners were generated. Fewer corners were found for threshold 64 due to a larger amount of noise resulting in rejecting some character images. For all thresholds there were nearly equal numbers of black and white wedges.

To measure the corner erosion, each corner in the character image was located and lines were fit to the stroke edges. The lines associated with each corner are parallel to the original stroke if they are sufficiently far from the corner apex, allowing measurement of ϕ . Sufficiently far is defined by the width of the point spread function, which is not known. The PSF was assumed to be a bivariate Gaussian density function with standard deviation w . Edge points at or close to the corner apex were not used to fit the lines. These lines were extrapolated to find their intersection and the distance from that intersection point to the nearest pixel in the image is used to measure the erosion distance d_b or d_w .

3.1 Experiment 1: Number of Corners

The first experiment was intended to see how the number of corners used in estimation affects the mean and standard deviation of the estimation results. Because character images have more noise from measurement errors (pixel drop out, errors in edge finding, etc.) than test charts, using character images requires more samples than would be needed if specialized test charts were used. N_b corners were randomly chosen from the black corner population, and N_w from the white corner population. The mean of the intersections between N_b black and N_w white loci were used to determine the parameter estimate. This estimation was repeated 100 times and the average and sample standard deviation of the 100 resulting estimates were recorded. The values of N that were chosen were $N=N_b=N_w = \{5, 10, 20, 50, 100, 200\}$. Because of the relationship in Equation 3, the loci from white and black corners have different orientations.

As shown in Figure 2a, the average of the estimates of the width of the PSF is nearly constant as N increases. The

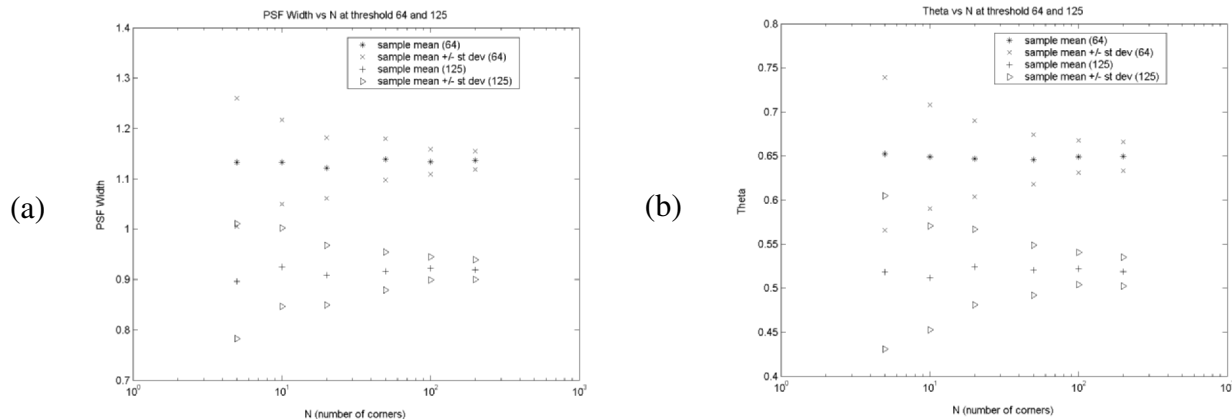


Figure 2: Results as N varies. Sample mean and sample standard deviation of (a) PSF width at thresholds 64 and 125, (b) binarization threshold at thresholds 64 and 125.

sample standard deviation decreases rapidly as N increases. These phenomena can also be found for the binarization threshold in Figure 2b. A minimum of $N=100$ samples are required to have the standard deviations for estimates of both w and Θ less than 4% of the mean.

The width of PSF supposedly remains the same with all thresholds, since the characters were scanned from the same scanner. However, the estimates at these thresholds differ by about 0.2 pixels. The theory expects some threshold dependent bias to exist if the assumed PSF shape does not match the actual PSF shape of the scanner [3]. This small error is negligible. These estimation results will be used to compare with the next experiment.

3.2 Experiment 2: Angle Measure

A study was done to see how using erosion measurements from corners of different angle measures would affect the estimation results. The angles were grouped into 10 degree ranges. Histograms showing the frequency of occurrence of angles in these ranges are shown in Figure 3. The frequency for a threshold of 125 is shown, but the histogram looks similar for other thresholds.

Estimates were calculated based on selecting data from these ranges. For each pair of ranges, all the black corners in the range were compared with all the white corners from the other range. The estimation results over all the angle combinations at threshold 125 are shown in Figure 4. The cross symbol represents the best guess of the actual PSF width and binarization threshold based on results from experiment 1. No single angle combination was found to be the best indicator of that estimate. Also there was no predominant bias evident. Use of multiple angle widths is needed to reach that estimate. Some angle combinations contribute more than the others. Table 1 shows the sample variance in the $N_b \times N_w$ intersections calculated for each pair of angle ranges. The sample vari-

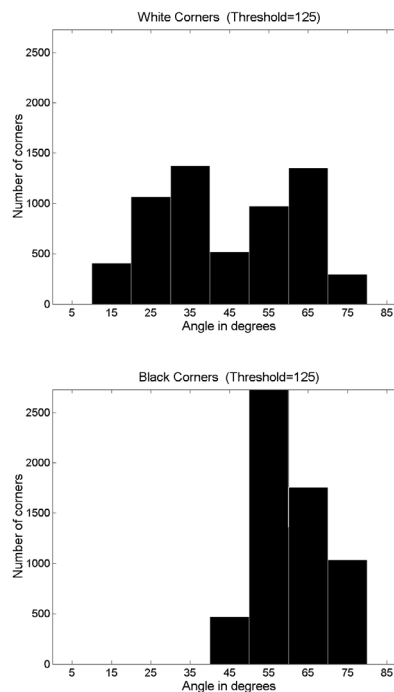


Figure 3: Histograms of corner angle occurrence.

ance on the estimates is low for small white and for large black corners.

3.3 Experiment 3: Typical Pages

If model parameters are to be estimated from characters in documents, the availability of suitable data from these documents must be analyzed. The letter occurrence frequency in English language documents was determined from a sample of 2,194,661 characters gathered from scientific and technical documents at ISRI [10]. Character

frequencies from our dataset are shown in the Table 2. In the same table, the expected number of each character on a page of 1000 characters is estimated.

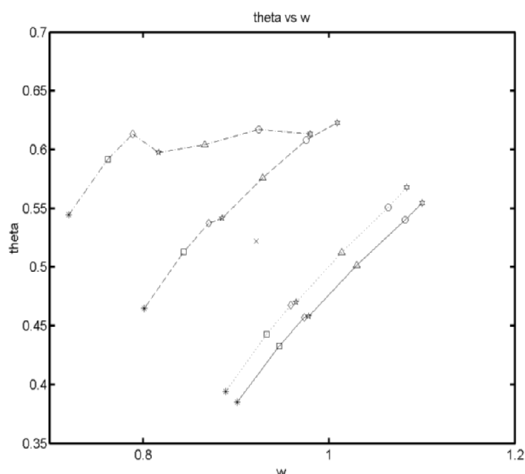
The next experiment was to sample $P=\{1, 2, 5\}$ sets of characters to estimate the expected number of suitable characters on P pages of 1000 characters. Each set randomly selects 56 characters from our dataset with the distribution in Table 2. All the corners present in these $56 \times P$ characters were used to estimate the model parameters (w, Θ). As shown in Table 2, some characters have higher frequencies of occurrence than the others. Some characters such as K, X, and Z do not occur frequently enough so their expected number was set to 1 to assure their presence in the sample.

This experiment was repeated 100 times. Each set of 56 characters contains a total of 211 corners; 109 corners (51.7%) are white and 102 corners (48.3%) are black. The results are shown in Figure 5. Again the sample mean does

not change significantly, and the standard deviation decreases gradually. The averages of the estimates are similar to the averages in experiment 1. The width estimates differ by 5% and the threshold estimates by less than 1%, even though some of the corners that were needed to get the original results come from characters that occur infrequently.

4. Conclusions

Three experiments to estimate scanner parameters using character images were presented. From these experiments the number of corners needed for a 'good' estimate of the model parameters were estimated. A minimum of $N=100$ corners is required to return a good estimation result for this scanner, noise level and printing process. The effect of the angle measure on the estimation value and variance has also been studied. The corners available from characters on a typical page of text serve as a potential estimation



Line:

- dashdot 40-50 degree B with each W
- dashed 50-60 degree B with each W
- solid 60-70 degree B with each W
- dotted 70-80 degree B with each W

Symbols:

- Star 10-20 degree W with each B
- Square 20-30 degree W with each B
- Diamond 30-40 degree W with each B
- Circle 40-50 degree W with each B
- Triangle 50-60 degree W with each B
- Pentagon 60-70 degree W with each B
- Hexagon 70-80 degree W with each B

Figure 4: Estimation results obtained by combinations of angle ranges.

Table 1: Sample standard deviation of estimates as a function of corner angle.

		$\hat{\sigma}_w$				$\hat{\sigma}_\Theta$			
		Black Corners				Black Corners			
		40-50	50-60	60-70	70-80	40-50	50-60	60-70	70-80
White Corners	10-20	0.256	0.204	0.215	0.211	0.2281	0.1766	0.1731	0.1747
	20-30	0.246	0.184	0.193	0.191	0.2176	0.1576	0.1527	0.1537
	30-40	0.253	0.191	0.200	0.198	0.2182	0.1607	0.1576	0.1586
	40-50	0.336	0.263	0.272	0.266	0.2044	0.1903	0.1929	0.1942
	50-60	0.306	0.239	0.249	0.244	0.2195	0.1979	0.1981	0.1999
	60-70	0.266	0.211	0.219	0.216	0.2350	0.1941	0.1937	0.1956
	70-80	0.321	0.251	0.256	0.251	0.2037	0.1969	0.2079	0.2096

mechanism. Calibration of the degradation model from the source image opens the way for using the degradation model to improve character recognition and document analysis.

5. Acknowledgement

Portions of this work were completed with funding from NSF-EPSCoR grant #9720634 and the BSU Faculty Research Grant program.

Table 2: Character occurrence frequencies in ISRI dataset.

Character	Frequency of occurrence in ISRI sample (%)	Number of members in sample page of 1000 characters
k	0.276	3
w	0.888	9
v	0.891	9
x	0.232	2
y	1.122	11
z	0.107	1
A	0.616	6
K	0.031	1(force)
M	0.562	6
W	0.290	3
V	0.158	2
X	0.006	1(force)
Y	0.095	1
Z	0.035	1(force)

6. References

- [1] Henry S. Baird, "Document image defect models," in H.S. Baird, H. Bunke and K. Yamamoto (eds.), Structured Document Image Analysis, Springer-Verlag, June 1992.
- [2] Henry S. Baird, "Calibration of document image defect models," Proc. of Second Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, April 1993, pp. 1-16.
- [3] Elisa H. Barney Smith, "Optical Scanner Characterization Methods using Bilevel Scans," Ph.D. thesis, Rensselaer Polytechnic Institute, December 1998.
- [4] Elisa H. Barney Smith, "Characterization of Image Degradation Caused by Scanning," Pattern Recognition Letters, Volume 19, Number 13, 1998, pp. 1191-1197.
- [5] Elisa H. Barney Smith, "Estimating scanning characteristics from corners in bilevel images," Proceedings SPIE Document Recognition and Retrieval VIII, Volume 4307, San Jose, CA 2001, pp. 176-183.
- [6] Elisa H. Barney Smith, "Scanner parameter estimation using bilevel scans of star charts," Proceedings International conference on Document Analysis and Recognition 2001, Seattle, WA, 2001, pp. 1164-1168.
- [7] Tin Kam Ho and Henry S. Baird, "Large-Scale Simulation Studies in Image Pattern Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 10, October 1997, pp. 1067-1079.
- [8] Theo Pavlidis, Minghua Chen, and Eugene Joseph, "Sampling and Quantization of Bilevel Signals," Pattern Recognition Letters, Vol. 14, July 1993, pp. 559-562.
- [9] Prateek Sarkar, George Nagy, Jianguyng Zhou, Daniel Lopresti, "Spatial Sampling of Printed Patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998, pp. 344-351.
- [10] ISRI data complements of Tomas Nartkar and Ron Young.

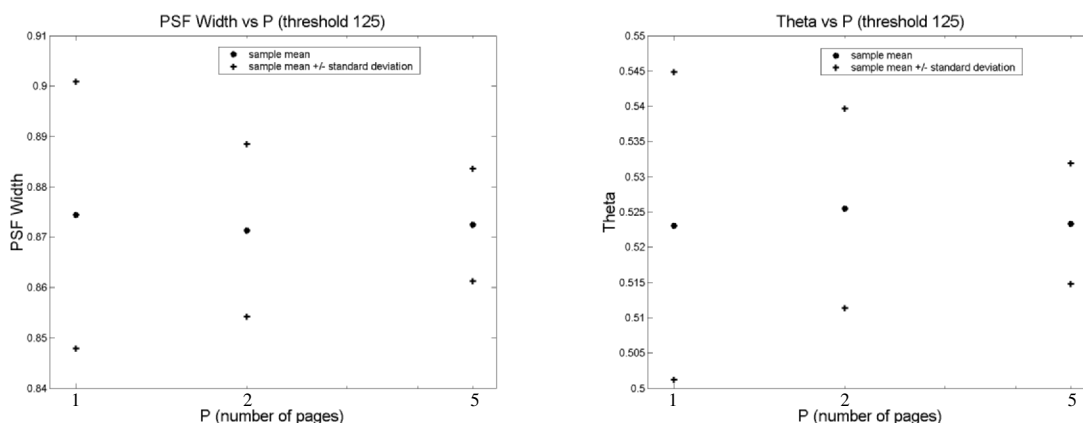


Figure 5: Estimation results as P increases with ISRI character occurrence frequencies.