To Appear: ICDAR'95: Third International Conference on Document Analysis and Recognition, Motreal, Canada, August 14-16, 1995.

Prediction of OCR Accuracy Using Simple Image Features

Luis R. Blando[1], Junichi Kanai, and Thomas A. Nartker
Information Science Research Institute
University of Nevada, Las Vegas
USA

Abstract

A classifier for predicting the character accuracy achieved by any Optical Character Recognition (OCR) system on a given page is presented. This classifier is based on measuring the amount of white speckle, the amount of character fragments, and overall size information in the page. No output from the OCR system is used. The given page is classified as either "good" quality (i.e., high OCR accuracy expected) or "poor" (i.e., low OCR accuracy expected). Results of processing 639 pages show a recognition rate of approximately 85%. This performance compares favorably with the ideal-case performance of a prediction method based upon the number of reject-markers in OCR generated text.

1: Introduction

To evaluate the performance of an OCR system, the character accuracy on a given processed page is determined by comparing the OCR output with the ground-truth data (correct text)[1,2]. In the real world, however, the corresponding ground truth data are not available. Therefore, other ways to estimate character accuracy are needed.

It was shown that on average rekeying documents is more cost effective than editing corresponding OCR-generated text unless a minimum character accuracy of 95-98%, depending on document complexity, is achieved [3]. An accuracy estimator algorithm would act as a filter to screen pages for rekeying and save substantial cost.

A study conducted by Information Science Research Institute (ISRI) showed that OCR systems achieved character accuracy rates better than 99% for high quality page-images; however, the rates varied widely (from 84.56% to 94.83%) for low quality page-images [1]. Furthermore, studies in [4, 5] showed that touching and broken characters seem to be the most important source of OCR problems. These results suggest that the character accuracy of a given page could be predicted by measuring its image quality.

Algorithms for estimating the quality of any given page would be beneficial for other applications as well. Such algorithms would be used to automatically determine improvement (or degradation) made by adaptive document image-restoration algorithms, such as [6].

An image quality estimator would also be essential to the operation of adaptive OCR algorithms. The quality information would be used to select an appropriate classifier or a set of weights for combining results obtained from multiple classifiers.

Devices for measuring print quality are presented in [7,8]. These devices use character templates to determine the print quality of a page. Since the font information of a given page is usually not available, this approach cannot be utilized to predict character accuracy without performing font recognition first.

In this paper, we propose a prediction technique based upon measuring the features associated with degraded characters. In order to limit the scope of the research, the following assumptions are made:

1. Pages are printed in black and white (no color).
2. Page images have been segmented, and text regions have been correctly identified. The image-based classifier extracts features from text regions only.

This prediction system simply classifies the input image as either *good* (i.e., high accuracy expected) or *poor* (i.e., low accuracy expected).

2: Prediction Method

Features associated with degraded characters (or character images) that cause OCR errors are used to determine the quality of the input image and to predict OCR accuracy. A small set of sample pages with low character accuracy was carefully inspected, and the following observations were made.

[1] L.Blando is currently a member of Harmonix Corp. in Woburn, MA.

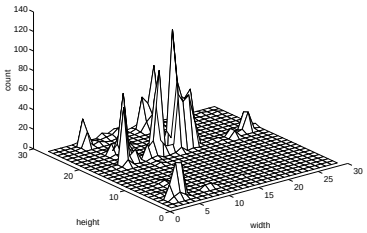

Figure 1: Frequency Distribution of the Bounding Rectangles.

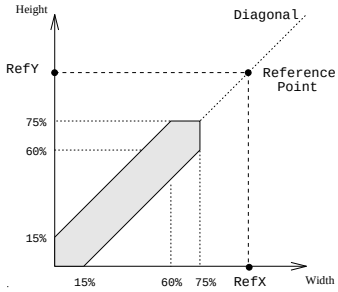

Figure 2: Broken Character Zone Coordinate Definition

Observation 1: When characters in a page are thicken by distortion, it tends to cause many touching characters. Another by-product of characters with thick strokes is that the loops in letters like "a" and "e" often get filled up completely or present only a minimal white portion in the center. Thus, a metric that could capture the existence of these "minimally open" loops would be a good way to detect problems related to touching characters.

Observation 2: Broken characters are usually fragmented into smaller pieces, and character fragments could have almost any shape. A metric that could weigh the existence of these character fragments would be a good way to detect problems related to broken characters.

Observation 3: Pages with "inverse video" (white letters on black background) or unusual typesetting tend to produce more OCR errors.

Based on these observations, a prediction system consisting of three simple rules was developed. The training data utilized included 12 clean pages and 12 degraded pages. These pages were extracted from scientific and technical documents, and did not contain small fonts or numerical tables. For 21 pages, the median accuracy was computed from the output of eight OCR systems. For the remaining 3 pages, the median accuracy was computed from the output of six newer OCR systems. Since the highest character accuracy obtained from these degraded pages was 88.76%, *good* corresponds to the expected accuracy above 90%. On the other hand, *poor* corresponds to the expected character accuracy below or equal to 90%.

To detect "minimally open loops" (Observation 1), a *White Speckle Factor* (WSF) was defined. White speckle is any white 8-connected component whose size is less than or equal to 3 pixels high and wide. To measure the level of white speckle, the WSF is defined as:

$$WSF = \frac{\text{Num. of White Bounding Rectangles } \leq 3 \times 3}{\text{Total Number of White Bounding Rectangles}}$$

It is expected for image quality to degrade as this ratio increases. Due to the small size of the training data, the threshold value for identifying *poor* quality pages was manually determined to be 0.1. Thus, the first rule is

Rule 1: IF $WSF \geq 0.1$ THEN *poor*

To measure the amount of broken characters in a given page (Observation 2), a *Broken Character Factor* (BCF) was defined. In general, the sizes and shapes of character fragments vary widely. Thus, their bounding rectangles will have many different widths and heights. When a frequency distribution of the bounding rectangles is plotted as a 3-D histogram, as in Figure 1, the bounding rectangles of character fragments appear near the origin in a particular region. Thus, this region was defined as the *broken character zone*.

It is important to note that this zone will enclose *all* small black connected components. Since some of these small components are valid, such as the dot of "i" or a period, a density-based measurement, which is sensitive to the distribution of characters in the page, cannot be used. Therefore, the area covered rather than density is used. To compute the area coverage, the zone is divided both vertically and horizontally into cells of one pixel by one pixel. The bounding rectangles of black connected components are allocated to these cells according to their width and height. After these cells are filled by the connected components, the BCF is computed as:

$$BCF = \frac{\text{Number of Cells Occupied}}{\text{Number of Cells}}$$

To eliminate the effects of font sizes, the average height and the average width of bounding rectangles are used as a reference point, and the shape of the *broken character zone* is defined as shown in Figure 2. The training data showed that an area coverage of 70% or more is a very strong indicator of the presence of many broken characters in the page. Therefore, the second rule is defined as:

Rule 2: IF $BCF \geq 0.7$ THEN *poor*

The third rule that uses the number of white connected components and their sizes as features was also defined to detect "inverse video" regions (Observation 3). In an inverse video region, white connected components correspond to characters. If either the average height or the average width of white connected components exceeds 30 pixels (approximately 7 pts), they are likely to be characters rather than the background Moreover, the number of black connected components in the region should be small because of the background. Thus, the ratio of the number of black connected components to the number of white connected components also provides useful information, and a

threshold value of 1.5 was manually chosen. The third rule is defined as:

Rule 3:

IF $(max(\overline{WhiteWidth}, \overline{WhiteHeight}) \geq 30$ pixels)

and $\left(\frac{\text{Num. of Black CCs}}{\text{Num. of White CCs}} < 1.5\right)$ THEN $poor$

This prediction algorithm classifies a given page as *poor* if at least one of these three rules is activated.

3: Reject Marker Based Approach

When an OCR system does not recognize a character, it generates a special symbol known as a "reject" marker. Since the character accuracy of a page is expected to degrade as the number of reject markers generated by an OCR system increases, the ratio of the number of reject characters to the total number of characters in OCR generated text can be used to predict character accuracy. This approach is often used as a quality assurance method in large-scale document conversion environments.

This approach uses a single feature to classify a page as either *good* or *poor*. To determine an optimal threshold value in this one-dimensional feature space, a cost (risk) model is required. The following simple cost model is used in this paper:

$$Cost = \alpha \times \#\text{Correct} + \beta \times \#\text{GoodAsPoor} + \gamma \times \#\text{PoorAsGood}$$

The weights α and β are set to 0 and 1, respectively. Since any misclassification of poor quality pages as good quality pages forces the user to correct errors in pages with accuracy below 90%, the weight γ should be relatively high. To study the effects of the weight γ, three values 20, 50 and 100 are used to compare the performance of this approach and the image-based method.

4: Experimental Setup

Two sets of test data were utilized. The first set is a subset of ISRI's Sample 2 data base [1]. It consists of 460 pages that were selected at random from a collection of approximately 2,500 scientific and technical documents (approximately 100,000 pages). Each page was digitized at 300 dpi using a Fujitsu M3096M+ scanner. Since 21 pages in this data set were used to train the image-based prediction system, the remaining 439 pages were used to test it.

The second set consists of 200 pages selected from 100 magazines that had the largest paid circulation in the U.S. in 1992 as reported by *Advertising Age* magazine [9]. For each magazine, two pages were selected at random. Each page was digitized at 300 dpi using a Fujitsu M3096G scanner. The binary images were generated using a fixed threshold of 127 out of 255 by this gray scale scanner.

Six OCR systems, which participated in the *Third Annual Test of OCR Accuracy* sponsored by ISRI [1], processed these data sets and character accuracy data were collected. Each character insertion, deletion, or substitution needed to correct the OCR generated text was counted as an error. Each reject character was also counted as a substitution error in this calculation. Character accuracy is defined as:

Character Accuracy = (n - Number of Errors) / n

where n is the total number of characters in the ground-truth text [1].

5: Results and Analysis

Table 1 and 2 summarize the decisions made by the image-based prediction system processing the Sample 2 data set and the magazine data set. For all systems, this classifier was able to correctly recognize approximately 85% of the pages in each data set.

The number of pages with accuracy above 90% indicates the performance of OCR systems, and better systems are expected to recognize more degraded pages. The tables show that, in general, the number of misclassifying good quality pages as *poor* by the system increased as the performance of OCR systems improved. Similarly, the number of misclassifying poor quality pages as *good* decreased as the performance of OCR systems improved.

Poor quality pages misclassified as *good* in Sample 2 were visually examined by the authors. The images did not show much degradation. A common feature of these pages was numerical tables. The corresponding OCR generated text contained many substitution errors of "0" by "O" and "1" by "l". This observation suggests that numerical data rather than image degradation caused OCR difficulty. This problem cannot be detected by the image-based approach proposed here. To solve the problem, zone attributes from a page segmentation module would be required.

Several pages containing fewer than 200 connected components were also misclassified by the prediction system. This result suggests that a reliable decision cannot be made from a small number of characters in a given page because of the lack of sufficient information.

For each combination of an OCR system, a data set, and a weight, the minimum cost achieved by the reject maker based approach was determined. In practice, it is highly likely that this kind of optimal performance cannot be achieved by a classifier based upon this approach. Similarly, costs were calculated from the results obtained by the image based method.

Tables 3 and 4 compare their performance on the Sample 2 data and the magazine data, respectively. These results show that, in several cases, the image-based prediction technique matches or exceeds the performance

Table 1: Classifying the Sample 2 Data (439 pages) Using the Image-Based Classifier

	# *Good* (# *Poor*)	# Errors		Recognition Accuracy
		G –> P	P –> G	
OCR1	397 (42)	39	19	86.8%
OCR2	416 (23)	50	11	86.1%
OCR3	417 (22)	51	10	86.1%
0CR4	380 (59)	32	26	86.8%
OCR5	381 (58)	37	33	84.1%
OCR6	415 (24)	50	12	85.9%

Table 2: Classifying the Magazine Data (200 pages) Using the Image-Based Classifier

	# *Good* (# *Poor*)	# Errors		Recognition Accuracy
		G –> P	P –> G	
OCR1	183 (17)	25	2	86.5%
OCR2	186 (14)	28	2	85.0%
OCR3	187 (13)	28	1	85.5%
0CR4	181 (19)	25	4	85.5%
OCR5	177 (23)	24	7	84.5%
OCR6	184 (16)	29	5	83.0%

Table 3: Cost-Based Comparison of the Image-Based Classifier and the Reject-Based Method Processing the Sample 2 Data

	γ = 20		γ = 50		γ = 100	
	Image	Reject	Image	Reject	Image	Reject
OCR1	419	419	989	989	1939	1939
OCR2	270	137	600	317	1150	576
OCR3	251	307	551	727	1051	1427
OCR4	552	205	1332	287	2632	387
OCR5	697	356	1687	746	3337	1396
OCR6	290	215	650	425	1250	775

Table 4: Cost-Based Comparison of the Image-Based Classifier and the Reject-Based Method Processing the Magazine Data

	γ = 20		γ = 50		γ = 100	
	Image	Reject	Image	Reject	Image	Reject
OCR1	65	69	125	129	225	229
OCR2	68	84	128	144	228	209
OCR3	48	172	78	326	128	576
OCR4	105	81	225	111	425	161
OCR5	164	86	374	176	724	326
OCR6	129	94	279	181	529	281

of the reject-maker based technique which was optimized for each possible combination. Therefore, the feasibility of the image-based prediction technique has been demonstrated.

6: Summary and Future Work

A classifier for predicting the character accuracy of a given page has been presented. This classifier is based upon measuring the features associated with degraded characters in the page and does not use OCR output at all.

This simple classifier correctly predicted the character accuracy of approximately 85% of the pages in test data sets. In several cases, without any OCR system dependent knowledge, this classifier matched or exceeded the optimal performance of the reject marker based method.

The results also suggest that some OCR errors are not caused by image defects. This method does not detect such errors.

Methods for reporting the degree of image defects are currently being researched. To improve the performance of this classifier, we plan to add more features, increase the quantity of training data, and use statistical pattern recognition techniques to re-design the classifier

Acknowledgments

This research was supported in part by a grant from the US Department of Energy. The authors thank Professor Nagy (RPI) and Professor Bunke (Universtat Bern) for stimulating discussions. We received valuable assistance from J. Gonzalez, S. V. Rice and other ISRI members.

References

[1] S. V. Rice, J. Kanai, and T. A. Nartker, *The Third Annual Test of OCR Accuracy*, TR 94-03, ISRI, University of Nevada, Las Vegas, April, 1994.

[2] J. Esakov, D. P. Lopresti, and J. S. Sandberg, "Classification and Distribution of Optical Character Recognition Errors," *Document Recognition*, SPIE Proc. Series Vol. 1281, pp. 204-212, 1994.

[3] L. A. Dickey, "Operational Factors in the Creation of Large Full-Text Databases," DOE INFOTECH Conference, Oak Ridge, TN, May 1991.

[4] M. Bokser, "Omnidocument Technologies," *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1066-1078, 1992.

[5] T. A. Nartker, et. al., *A Preliminary Report on OCR Problems in LSS Document Conversion*, TR 92-04, ISRI, University of Nevada, Las Vegas, April, 1992.

[6] P. Stubberud, et. al., "Adaptive Image Restoration of Text Images That Contain Touching or Broken Characters," in this proceedings.

[7] W. R. Throssell and P. R. Fryer, "The Measurement of Print Quality for Optical Character Recognition Systems," *Pattern Recognition*, Vol. 6, pp. 141-147, 1974.

[8] M. Bohner, et. al., "An Automatic Measurement Device for the Evaluation of the Print Quality of Printed Character," *Pattern Recognition*, Vol. 9, pp. 11-19, 1977.