# Prediction of OCR Accuracy

Luis R. Blando,[1] Junichi Kanai, Thomas A. Nartker, and Juan Gonzalez

## 1 Introduction

The accuracy of all contemporary OCR technologies varies drastically as a function of input image quality [Rice 92, Rice 93, Chen 93, Rice 94]. Given high quality images, many devices consistently deliver output text in excess of 99% correct. For low quality images, even images which are easily read by a human, output accuracy is frequently below 90%. This extreme sensitivity to quality is well known in the document analysis field and is the subject of much current research.

In this ongoing project, we have been interested in developing measures of image quality. We are especially interested in learning to predict OCR accuracy using some combination of image quality measures independent of OCR devices themselves.

Reliable algorithms for measuring print quality and predicting OCR accuracy would be valuable in several ways. First, in large scale document conversion operations, they could be used to automatically filter out pages that are more economically recovered via manual entry. Second, they might be employed iteratively as part of an adaptive image enhancement system. At the same time, studies into the nature of image quality can contribute to our overall understanding of the effect of noise on algorithms for classification.

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

- Pages are printed in black and white (no color).
- Page images have been segmented, and text regions have been correctly identified. The image-based prediction system extracts information 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 Method

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

1. Blando is now a member of Harmonix Corp. in Woburn, MA.

- **Observation 1:** Pages with characters whose strokes are thick tend to have many touching characters. Another by-product of these characters 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 as shown in Figure 1. Thus, a metric that could capture the existence of these "minimally open" loops would detect problems related to touching characters.


Figure 1: Filled loops

- **Observation 2:** Broken characters are usually fragmented into small pieces, and character fragments could have almost any shape as shown in Figure 2. A metric that could weight the existence of these character fragments would detect problems related to broken characters.

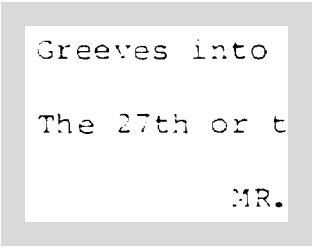


Figure 2: Broken characters

- **Observation 3:** pages with "inverse video" (white letters on black background) or artistic typesetting tend to produce more OCR errors.

Based on these observations, three simple classification rules were developed. The training data utilized included 12 clean pages and 12 degraded pages. These pages were extracted from the DOE data sets. For 21 pages, the median accuracy was computed from the output of eight OCR systems [Rice 92]. For the remaining 3 pages, the median accuracy was computed from the output of six newer systems [Rice 94]. Since the highest character accuracy obtained from the degraded pages was 88.76%, *good* corresponds to an expected accuracy above 90%. On the other hand, *poor* corresponds to an expected accuracy below 90%.

To detect minimally open loops (Observation 1), a *White Speckle Factor* (WSF) was defined. A white speckle is any 8-connected white component whose size is less than or equal to 3 pixels in height and width. It is defined as:

$$\text{WSF} = \frac{\text{\# white bounding rectangle smaller than 3 by 3}}{\text{\# white bounding rectangles}}$$

This metric weighs the amount of white-speckle present. It is expected that image quality goes down as this ratio goes up. Because of the small quantity of training data, the threshold value for identifying *poor* quality pages was manually determined to be 0.1. Thus, the first rule is as follows:

**Rule 1:** if **WSF** $\geq$ 0.1 then *quality_is_poor*

To measure the amount of broken characters in a given page (Observation 2), a *Broken Character Factor* 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, such as Figure 3, the bounding rectangles of character fragments appear near the origin. Thus, this region was defined as the *broken character zone* as shown in Figure 4.

It is important to note that the broken character zone is designed to collect all small black connected components. These small components are mostly character fragments but can also be "dots", such as the dot of "i" and a period, and other small legal characters in small type sizes. Since these dots will fall inside of the *broken character zone*, a density-based measurement is sensitive to the distribution of characters in the page. Therefore, the *Broken Character Factor* (BCF) uses area coverage rather than density to make a decision.

To measure the degree of covering of this zone, it is divided into square cells, at a rate of one per square 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 *Broken Character Factor* is computed as:

$$\text{BCF} = \frac{\text{\# of cells occupied}}{\text{\# of cells}}$$

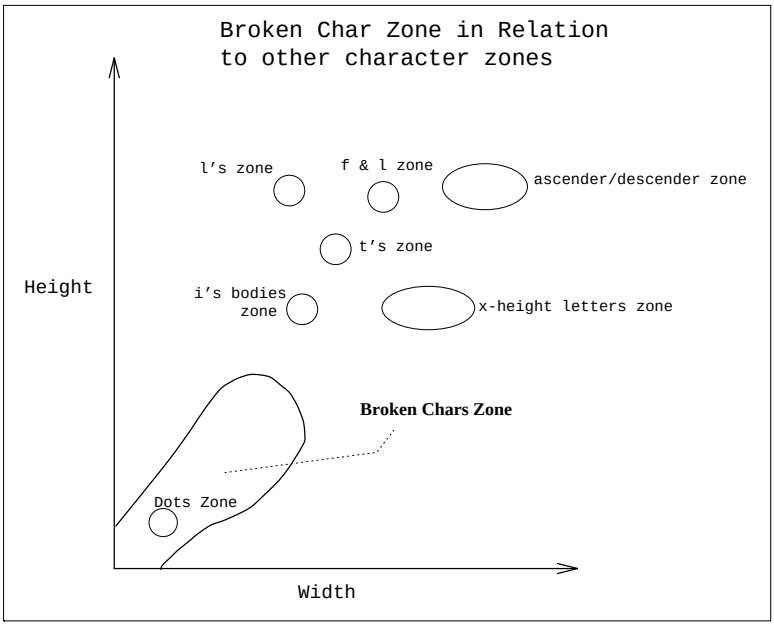


Figure 3: Broken Character Zone and Other Character Zones

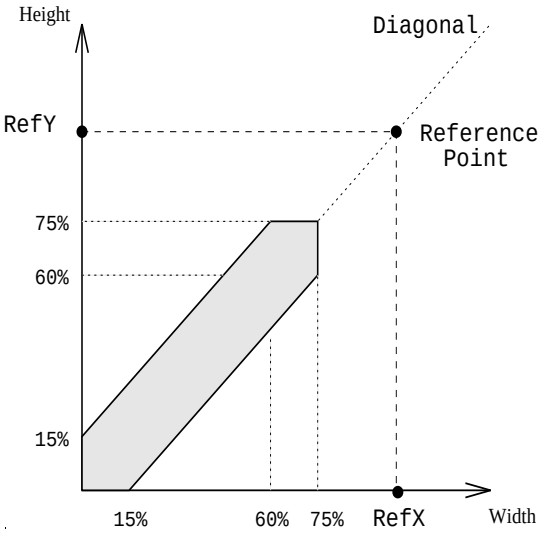

Figure 4: Definition of Broken Character Zone

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 5. From observations on the training data, an area coverage of 70% or more is a very strong indicator of the presence of many broken characters on the page. Thus, the second rule is defined as

**Rule 2:** if **BCF** > 0.7 then *quality_is_poor*

A 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 points), they are highly likely to be characters. 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. The third rule is defined as:

**Rule 3**: if max(Avg. White Width, Avg. White Height) > 30 pixels

and $\frac{\text{\# black connectedcomponents}}{\text{\# white connected components}} < 1.5$

then *quality_is_poor*

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

# 3 Experimental Setup

Two sets of test data were utilized. The first set is a subset of DOE samples. The complete set 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 to produce a binary image [Rice 94]. Since 21 pages in this data set were used to train the image-based prediction system, the remaining 439 pages were used as the DOE test set.

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 [AdAge 93]. 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. Thus, 21 pages were used as a training set, and two sets containing 439 pages and 200 pages, respectively, were used as the test data sets.

The performance of this prediction system was measured using the median character accuracy of six systems[1] obtained in the *Third Annual Test of OCR Accuracy* [Rice 94].

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:

$$\text{Character accuracy} = \frac{(n - \#\,\text{errors})}{n}$$

where $n$ is the total number of characters in the ground truth data.

# 4 Results and analysis

Table 1 shows the confusion matrix for the DOE test set. The classifier misclassfied 15 *poor* quality pages as *good* and 53 *good* quality pages as *poor*. Thus, its error rate was 15%.

The 15 *"poor → good"* misclassifications were carefully examined. The page images did not present substantial degradation. All but four of these problematic pages contained numerical tables in them. The OCR output for many of these tables were illegible and generally useless. Yet, from an image defects point of view, it could find no evidence justifying a *poor* label. Table 2 shows the confusion matrix without pages containing tables.

Table 3 summarizes the relationship between the number of black connected components in a page and the classification result. The table shows that this system has problems in classifying pages containing less than 200 connected components.

Table 4 shows the confusion matrix for the magazine data. The error rate was 13.5%, and similar performance was observed.

1. These six systems, which will not be identified with experimental results, are Caere OCR (Ver.132), Calera WordScan (Ver. 4), EDT ImageReader (Ver 2.0), EperVision RTK (Ver. 3.0), Reccognita Plus DTK (Ver. 2.00D12), and XIS OCR Engine (Ver. 10).

**Table 1: Confusion Matrix for Sample 2 (439 pages)**

| True ID | Recognized | |
|---|---|---|
| | Good | Poor |
| Good | 354 | 48 |
| Poor | 15 | 22 |

**Table 2: Confusion Matrix with Tables as Rejects**

| Tue ID | Recognized | | Pages with Tables |
|---|---|---|---|
| | Good | Poor | |
| Good | 257 | 42 | 103 |
| Poor | 4 | 18 | 15 |

**Table 3: Confusions vs. Number of Connected Components (Tables filtered)**

| IS-AS | Number of Connected Components | | | | | |
|---|---|---|---|---|---|---|
| | [0-100] | (100-200] | (200-300] | (300-400] | (400-500] | (500++ ] |
| Poor-Poor | 6 | 1 | 4 | 0 | 0 | 7 |
| Poor-Good | 2 | 2 | 0 | 0 | 0 | 0 |
| Good-Poor | 6 | 1 | 2 | 2 | 1 | 30 |
| Good-Good | 29 | 17 | 9 | 5 | 4 | 193 |

**Table 4: Confusion Matrix for Magazines (200 pages)**

| True ID | Recognized | |
|---|---|---|
| | Good | Poor |
| Good | 159 | 27 |
| Poor | 1 | 13 |

# 5 The DOE test data as training data

Figure 5 shows a scatter diagram for feature vectors consisting of *White Speckle Factor* and *Broken Character Factor* obtained from Sample 2. The diagram shows three classes: good quality pages, poor quality due to broken characters, and poor quality due to touching characters. To study whether or not a statistical classification method would improve the accuracy, the nearest mean rules using Mahalanobis distance was developed. The distance from the mean vector of a class *i* to a vector $x$ to be classified is defined as:

$$D_i^2(x) = (x-\mu_i)^t K_i^{-1}(x-\mu_i)$$

where *K* is a covariance matrix. The mean vector and covariance matrix of each class are as follows:

$$\text{Good:} \quad \mu_g = \begin{bmatrix} 0.0261 \\ 0.2381 \end{bmatrix}, \; K_g = \begin{bmatrix} 0.0031 & 0.0014 \\ 0.0014 & 0.0275 \end{bmatrix}$$

$$\text{Touching:} \quad \mu_t = \begin{bmatrix} 0.1185 \\ 0.1553 \end{bmatrix}, \; K_t = \begin{bmatrix} 0.0186 & -0.0030 \\ -0.0030 & 0.0110 \end{bmatrix}$$

$$\text{Broken:} \quad \mu_b = \begin{bmatrix} 0.1128 \\ 0.8290 \end{bmatrix}, \; K_b = \begin{bmatrix} 0.0130 & 0.0008 \\ 0.0008 & 0.0053 \end{bmatrix}$$

Figure 6 show the decision boundaries.

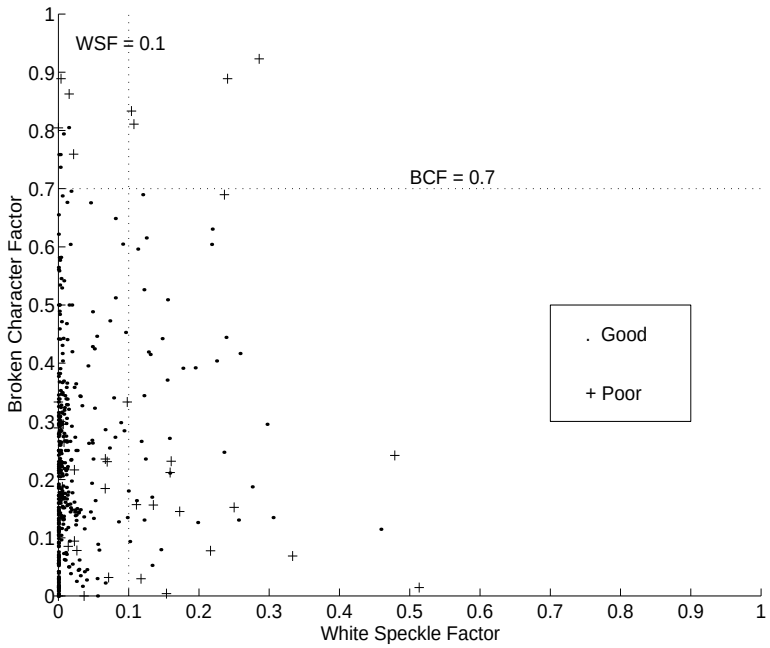


Figure 5: Scatter Diagram for the feature vectors with the heuristic decision boundaries.

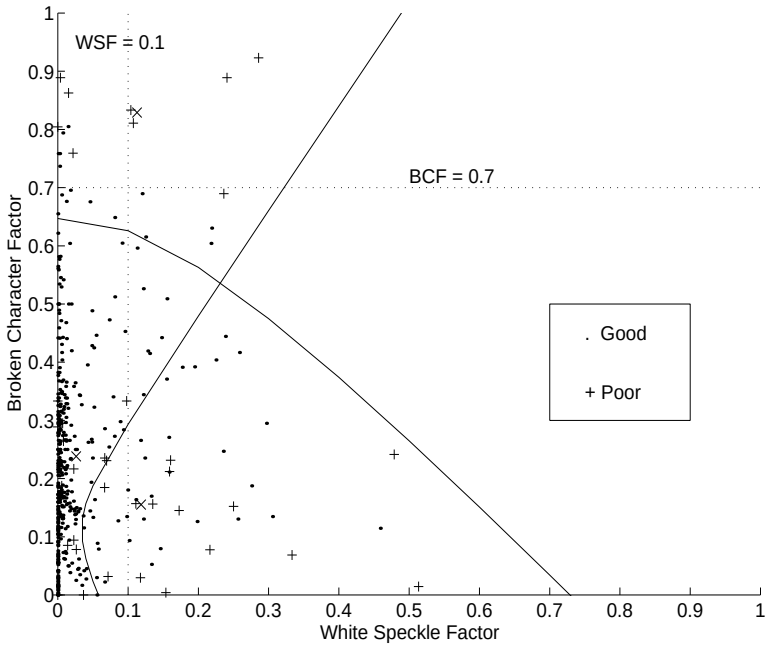


Figure 6: Decision boundaries for the nearest mean rule. The x's show the location of mean vectors.

Two heuristics rules, 1 and 2, were replaced by the nearest mean rules. Table 5 and 6 show that the confusion matrices for the DOE test data and the magazines, respectively, generated by the nearest mean classifier. Although some errors were corrected, some new errors were introduced by this classifier. Hence, no improvement was made. The results showed that better features had to be identified before applying statistical pattern recognition techniques.

**Table 5: Confusion Matrix for Sample Using the Nearest Mean Rule**

| True ID | Recognized | |
|---|---|---|
| | Good | Poor |
| Good | 353 | 49 |
| Poor | 12 | 25 |

**Table 6: Confusion Matrix for Magazines Using the Nearest Mean Rule**

| True ID | Recognized | |
|---|---|---|
| | Good | Poor |
| Good | 157 | 29 |
| Poor | 1 | 13 |

# 6 Summary

A method for predicting the accuracy of OCR generated text has been presented. Our approach measures features associated with degraded text images and does not use any output from an OCR system. Experimental results support the feasibility of this approach. However, some OCR errors, such as recognition of numerical tables, are not caused by image degradation and cannot be detected by this approach. To improve the performance of the prediction system, better features must be identified before applying statistical pattern recognition techniques.

# Acknowledgment

The authors thank Professor Nagy (RPI) for helpful discussions.

# Bibliography

[Blando 94] Blando, Luis R., *Evaluation of Page Quality Using Simple Features,* Master's Thesis, University of Nevada, Las Vegas, 1995.

[Bohner 77] Bohner, M., et. al., "An Automatic Measurement Device for the Evaluation of the Print Quality of Printed Characters," *Pattern Recognition*, Vol. 9, 1977, pp. 11-19.

[Bokser 92] Bokser, Mindy, "Omnidocument Technologies," *Proceedings of the IEEE*, Vol. 80, No. 7, July 1992.

[Chen 94] Chen, Su; Subramaniam, Suresh; Haralick, Robert, M.; and Phillips, Ihsin, T., "Performance Evaluation of Two OCR Systems," In *Poc. of the Third Annual Symposium on Document Analysis and Information Retrieval,* Las Vegas, NV, April 11-13, 1994.

[Esakov 94] Esakov, Jeffery; Lopresti, Daniel P.; and Sandberg, Jonathan S., "Classification and distribution of optical character recognition errors," *Document Recognition*, ed. L. M. Vincent and T. Pavlidis, SPIE Proceedings Series, Vol. 1281, 1994.

[Dickey 91] Dickey, Lois A., "Operational Factors in the Creation of Large Full-Text Databases," *DOE Infrotech Conference*, Oak Ridge, TN, Mary 1991.

[Duda 73] Duda, Richard O., and Hart, Peter, E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973. [Nartker 92]

[Rice 92] Rice, Stephen V.; Kanai, Junichi; and Nartker, Thomas A., *A Report on the Accuracy of OCR Devices*, Technical Report TR-92-02, ISRI, University of Nevada, Las Vegas, 1992.

[Rice 93] Rice, Stephen V.; Kanai, Junichi; and Nartker, Thomas A., *An Evaluation of OCR Accuracy,* Technical Report TR-93-01, ISRI, University of Nevada, Las Vegas, 1993

[Rice 94] Rice, Stephen V.; Kanai, Junichi; and Nartker, Thomas A., *The Third Annual Test of OCR Accuracy*, Technical Report TR-94-03, ISRI, University of Nevada, Las Vegas, 1994