LA-UR 97-214

CONF-970231--24

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: **AN AUTOMATED SYSTEM FOR NUMERICALLY RATING DOCUMENT IMAGE QUALITY**

AUTHOR(S): Michael Cannon, P. Kelly, S. Sitharama, N. Brener

SUBMITTED TO: SPIE Conference on Electronic Imaging
San Jose, CA
Feb. 9-14, 1997

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

DISCLAIMER

An automated system for numerically rating document image quality

Michael Cannon* and Patrick Kelly,
Los Alamos National Laboratory
S. Sitharama Iyengar and Nathan Brener,
Louisiana State University

ABSTRACT

As part of the Department of Energy document declassification program, we have developed a numerical rating system to predict the OCR error rate that we expect to encounter when processing a particular document. The rating algorithm produces a vector containing scores for different document image attributes such as speckle and touching characters. The OCR error rate for a document is computed from a weighted sum of the elements of the corresponding quality vector. The predicted OCR error rate will be used to screen documents that would not be handled properly with existing document processing products.

Keywords: image quality, speckle, touching characters, broken characters, OCR error rate

1. INTRODUCTION

The Department of Energy has undertaken the classification review of 230 million pages of legacy documents dating from the present back to the Manhattan Project. Present plans call for the scanning of each document into digital image form. Some documents are of good quality, but the quality of others is degraded by repeated photocopying, FAXing, carbon-copying, and aging fibrous paper. The DOE intends to convert documents of sufficient quality to text using OCR, while those of lower quality will be stored as images. In order to determine which documents can be successfully OCR'd, we are developing an algorithm to predict the OCR error rate that can be expected from a particular document or page of a document. Our prediction is based on the values of three quality features that are extracted from the document image. Our preliminary results are based on a corpus of degraded documents that we produced ourselves by repeatedly photocopying selected pages from a book.

2. SAMPLE DOCUMENT CORPUS

In order to create an algorithm under controlled conditions, we created a corpus of documents spanning a range of gradually decreasing quality. We did this by repeatedly photocopying a page from a book, so that we had nine versions of it, the original and 8 following generations. Each successive generation is increasingly plagued with commonly-encountered attributes of lower quality document images: speckle, fattened brush strokes, shrinking white connected components, and touching characters. We also created a second set of degraded documents by repeatedly photocopying a second page from the book in the same manner, producing a total corpus of 18 documents.

We computed histograms of the sizes of white and black connected component for each of the nine photocopy generations; sample plots of the zeroth and fifth generation histograms are shown in Figures 1 and 2.

* Address correspondence to tmc@lanl.gov

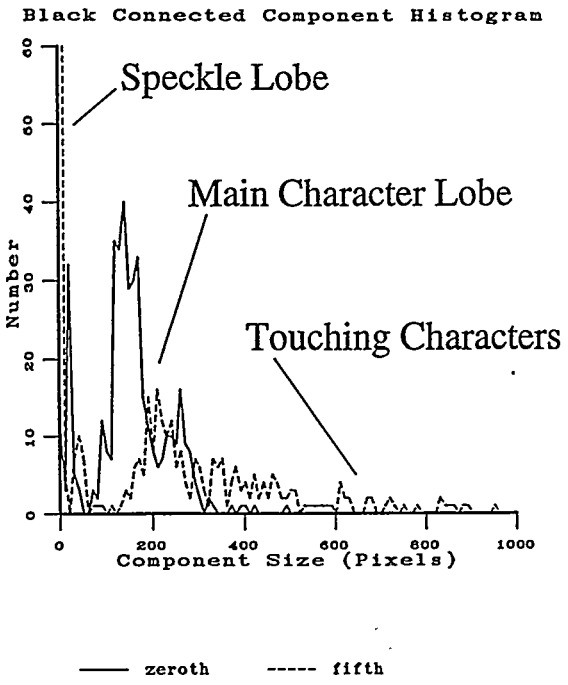

Figure 1. Histograms of black connected component sizes for the zeroth and fifth generation photocopies of one of our sample documents. Annotation indicates portions of the histogram that yield clues to the quality of the documents.

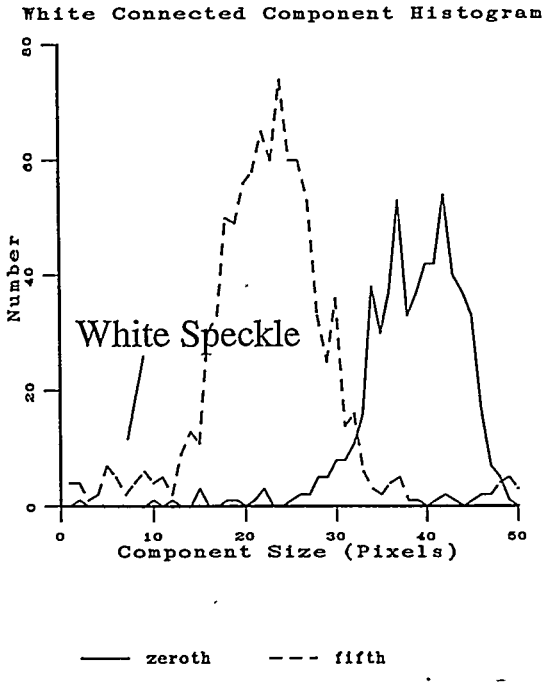

Figure 2. Histograms of white connected component sizes for the zeroth and fifth generation photocopies of the same document as that in Figure 1. Very small white connected components are labeled as White Speckle.

We observed several interesting features of these histograms.

- The number of very small black connected components increases with photocopy generation due to the increasing incidence of speckle.
- The main lobe of black connected components, initially centered at approximately 150 pixels in size, broadens and shifts to the right as brush strokes fatten.
- Black connected components greater than 600 pixels in size occur as characters begin to touch.
- The number of very small white connected components increases as white connected components shrink in size due to the fattening of black brush strokes, and as neighboring characters begin to touch and create still more small white connected components.

Some of the corresponding degradations in the images from which these histograms were computed are shown in Figure 3.

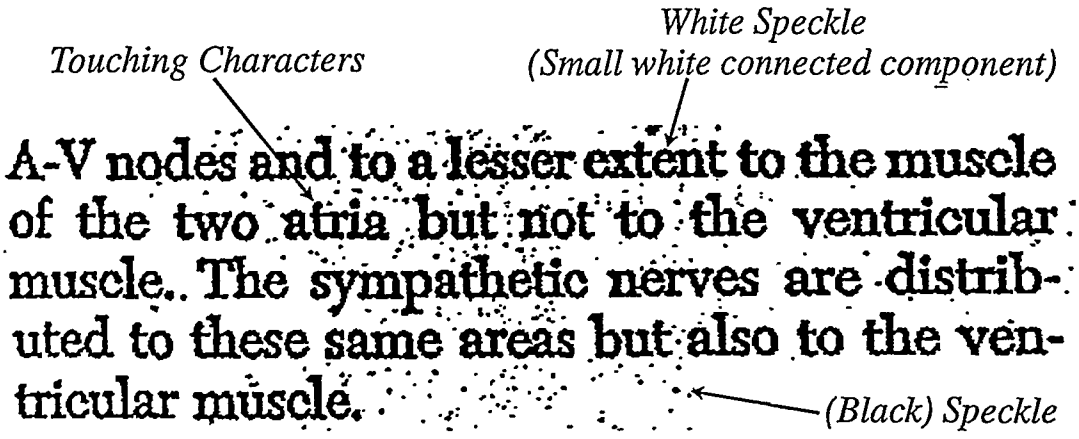

Figure 3. A sample chunk of degraded text, showing black speckle, white speckle due to shrinking white connected components, and touching characters.

We computed the OCR error rate, the percentage of incorrect words, for each photocopy generation of our two sets of documents. This was done using the Caere OmniPage Pro OCR package and comparing its output with the original page. We averaged the error rates of the two document sets together, generation by generation. As expected, the error rate increased with each generation, and is shown in Figure 4.

Based on these observations of the histograms and OCR error rate, we have derived three easily computed features that quantitate the increasing degradation of each photocopy generation. We are also able to predict the OCR error rate based on these derived features.

3. FEATURE SET

We compute three document quality features that reflect the characteristics of the histograms (Figures 1 and 2) of white and black connected component sizes . We will show

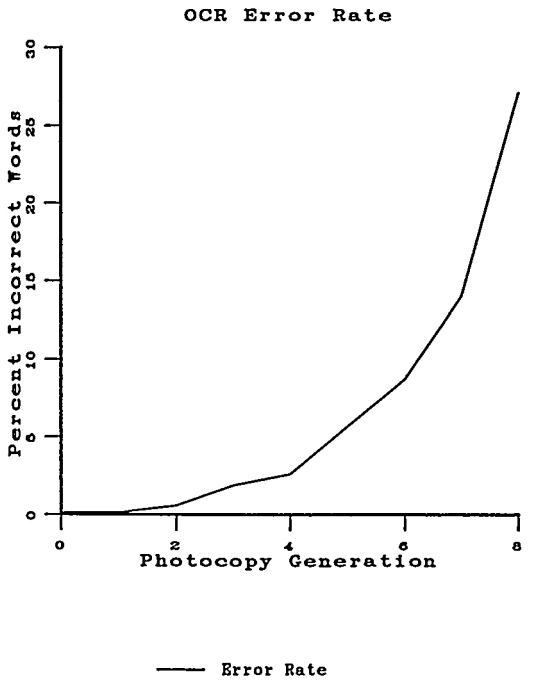

Figure 4. OCR error rates for the nine generations of documents in our test corpus.

that these features are correlated with the increase in OCR error rate that results from a decrease in document image quality. These three features are listed below.

1. *Speckle*. We count the number of black connected components that are less than or equal to ten pixels in size as a measure of the amount of (black) speckle in the document image. This count includes periods and dots over the letters *i* and *j*, a minimal perturbation. We normalize the Speckle count by the total number of black connected components in the image.
2. *Touching characters*. We count the number of black connected components that are larger than 600 pixels in size as a measure of the incidence of touching characters in the document image. We normalize the count by the total number of black connected components in the image. We refer to this feature as the Touching Character Factor (TCF).
3. *Small white connected components*. We measure the incidence of white connected components (actually the increase in the number of very small white connected components) using a method put forth by Blando, et. al.[1]. We count the number of white connected components that are less than or equal to 3x3 pixels in size, normalized by the total number of white connected components in the image. Blando refers to this feature as the White Speckle Factor (WSF), a term we have retained in this paper. We limited our count of white connected components to those less than 300 pixels in size.

These three features are plotted in Figure 5 as a function of photocopy generation. The plotted numbers are derived from our two sets of degraded documents, averaged together generation by generation.

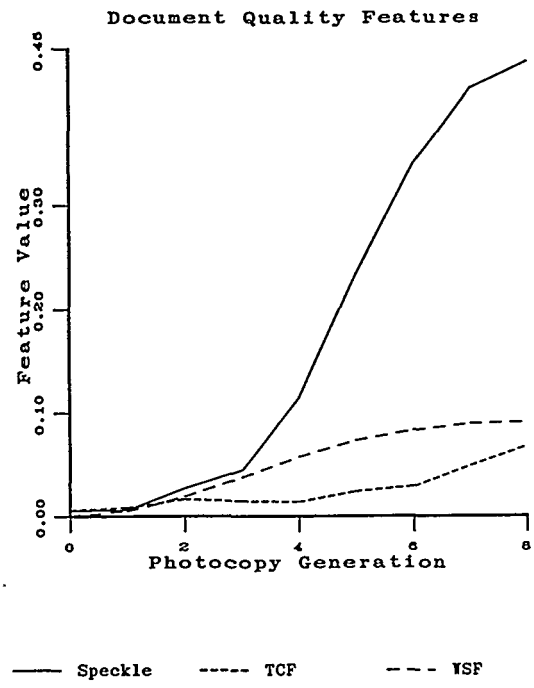

Figure 5. The values of our three document image quality features, Speckle, White Speckle Factor (WSF), and Touching Character Factor (TCF), plotted over the nine generations of our document corpus.

4. OCR ERROR RATE PREDICTION METHOD

Note that the features appear to be correlated with the OCR error rates plotted in Figure 4. One may wonder at this point if the Speckle, TCF, and WSF features can be used to predict the OCR error rate of a document image. This point is central to the DOE document quality assessment effort, and has a great impact on how automated the declassification project can be made. We hypothesize that a linear combination of the three features can be used as an error rate predictor. That is,

$$\text{OCR error rate} = \alpha \cdot \text{Speckle} + \beta \cdot \text{TCF} + \delta \cdot \text{WSF} + \gamma, \qquad (1)$$

where α, β, δ, and γ are determined from a training set of data, and the Speckle, TCF, and WSF features are computed from the document image in question. To determine α, β, δ, and γ from our corpus of nine generations of increasingly degraded photocopies, we form a linear system of equations

$$z_i = \alpha \cdot w_i + \beta \cdot x_i + \delta \cdot y_i + \gamma, \qquad (2)$$

where z, w, x, and y represent error rate, Speckle, TCF, and WSF, and i represents the photocopy generation, 0 through 9. We solve the system to obtain the α, β, δ, and γ that minimize the mean square error in predicting z. After we have obtain α, β, δ, and γ, we can use them to "predict" the OCR error rate for the nine generations of photocopied documents in the sample document corpus based on the three quality features derived from each document. These results are plotted in Figure 6, overlaid on the actual OCR error rates.

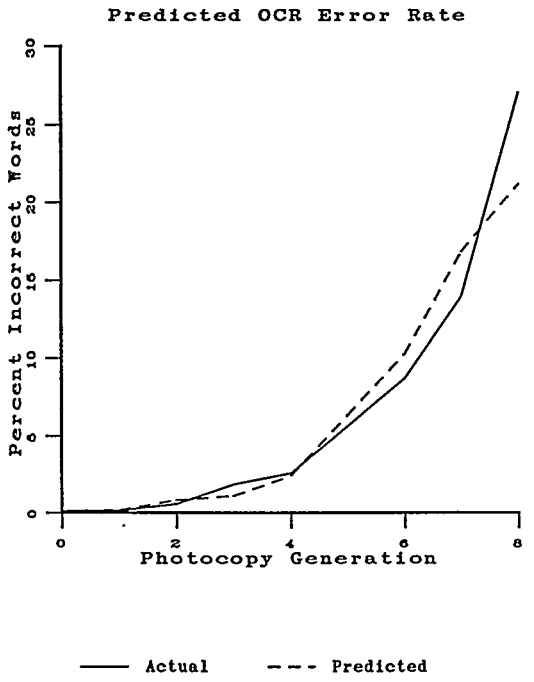

Figure 6. The actual OCR error rate is plotted here, just as in Figure 4. Overlaid on it is the "predicted" error rate based on our three quality measures.

5. FUTURE WORK

We are encouraged with the predictive capability of our three document quality features. Our next step is to refine the Speckle feature calculation so it is limited to areas of the document that are covered with a line of text. We feel that speckle outside a line of text has little bearing on the OCR error rate. We also hope that limiting the Speckle measure to text lines will also reflect the state of broken characters in the document, a quality which at present we do not measure. After we have refined our three features, we will apply our OCR error rate prediction algorithm to a set of actual DOE documents. We hope to train it on a portion of the set, then use it to predict the error rate of documents not included in the training set.

6. CONCLUSIONS

We have developed an algorithm that shows promise in predicting the OCR error rate of a document image. It is based on three features that are easily extracted from histograms of the sizes of black and white connected components in the image.

ACKNOWLEDGMENTS

Judy Hochberg suggested that we develop our algorithm on a set of carefully controlled degraded data, an invaluable piece of insight. Tom Burr has given us assistance in using proper statistical methods for our data. Clint Scovel and Jim White have contributed many meaningful ideas. Tom Curtis, US Department of Energy, has kindly funded our project.

REFERENCE

1. Luis R. Blando, Junichi Kanai, Thomas A. Nartker, *Prediction of OCR Accuracy Using Simple Image Features*, Proceedings ICDAR '95, Montreal, Canada, p319.