Performance Metrics for Document Understanding Systems

J. Kanai, T. A. Nartker, S. V. Rice

Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, NV 89154-4021, USA

G. Nagy

ECSE Department
Rensselaer Polytechnic Institute
Troy, NY 12180-3590, USA

Abstract

Requirements for the objective evaluation of automated data-entry systems are presented. Because the cost of correcting errors dominates the document conversion process, the most important characteristic of an OCR device is accuracy. However, different measures of accuracy (error metrics) are appropriate for different applications, and at the character, word, text-line, text-block, and document levels. For wholly objective assessment, OCR devices must be tested under programmed, rather than interactive, control.

1 Introduction

In every area of science and technology there are accepted standard measures. These metrics provide the basis for progress in the field. Well-defined and unambiguous measures, combined with reproducible experiments, allow knowledge in the field to accumulate and permit the use of quantitative methods.

The customary performance evaluation paradigm requires comparing an observed variable with a reference variable under controlled conditions. In Optical Character Recognition (OCR), the observed variable is the output of an OCR device, the reference variable is its desired ("Truth") output, while the controlled variables are selected attributes of the input images over a sample document domain.

The combinatorial explosion of relations between these variables contributes to the number of possible metrics of interest, to the need for large quantities of input data and to the necessity of automating the measurement task. Although setting up automated test systems is both costly and technically challenging, important advantages include:

- Aside from eliminating human error, the inherent consistency of automated systems tends to avoid bias toward algorithms by "excusing" certain types of errors.
- Experiments are reproducible, which is a basic requirement in all scientific experimentation.
- Large, statistically-significant experiments can be conducted with little additional effort.

We suggest that the situation demands a large number of new metrics which each provide an interesting measure of a small part of the OCR performance domain. From the user's point of view, the appropriate measure of an OCR system is the total cost of document conversion, typically dominated by the cost of correcting residual errors in the output. However, such a measure is not necessarily appropriate for a researcher who seeks to measure progress in developing a new algorithm. A vendor might use a research metric at an early stage of development, an overall cost metric to evaluate the final system and yet another class of metric to tune components before final assembly. Different performance metrics may be important at different stages of OCR system development.

Our notion of "accuracy" may also depend upon the specific application involved. OCR algorithms have typically been compared by measuring their accuracy in recognizing isolated characters. However, users of the information produced are interested in words and their correct reading order and almost never in individual characters. Applications involving complex layouts, unusual typefaces, or chemical or mathematical formulas require different measures of accuracy. Examples of metrics and proposed metrics are described in Sections 3 and 4, respectively.

2 Input data

There are different approaches to providing appropriate test data. Example sets of real world document images with the associated "Truth" representation are one ideal form of input test data. Such data are extremely expensive to prepare.

It is also possible to use simulated data. In this case, it is customary to perturb ideal images by adding "noise." Pseudo-random noise generators range from simple salt-and-pepper bit switching to sophisticated emulations of the distortions of the imaging process [1]. In spite of the appeal of generating large test databases this way, their value for predicting performance in field conditions has not been established. In fact, the evaluation and comparison of real world distortions and simulated distortions is a good example of needed new metrics.

A unique aspect of measurement in OCR is the need for both standard benchmark (public) datasets and for private datasets for independent evaluation. Public datasets can provide a standard against which researchers can measure and compare their progress. Since it is always possible to design or train a system to recognize a given amount of data with high accuracy, independent test data are required for objective final assessment.

0-8186-4960-7/93 $3.00

Before an OCR device can be tested, it is necessary to verify that it will accept values of the input variable of interest, i.e., digitized document images or parts of document images. Also, it must produce output in the form chosen for the "Truth" representation of the input data.

The most important component of a program to automate performance measures is an algorithm to match the output representation of OCR devices with the "Truth" representation. Performance measures are then computed by routines which examine the difference between the matched representations. Because different devices produce different but equivalent output representations of some inputs, it is necessary to normalize representations of all device output to compare their performance.

Routines to compute output errors alone are not sufficient to support large-scale experiments. Statistical and display tools are also needed to help users categorize errors and to analyze the causes of these errors.

3 Examples of metrics

We present a range of increasingly more complicated performance metrics that we have implemented. (See [9].) Our primary concerns are the following issues:

- the collection and preparation of input test data (including the truth data);
- normalization of outputs produced by devices;
- algorithms for comparing the output to the corresponding truth data;
- algorithms for analyzing deviations of the output from the corresponding truth data (i.e., computing the measure of interest.)

3.1 Isolated character accuracy

The overwhelming majority of OCR performance results are reported for isolated characters. Such data may take either the form of test alphabets with an equal number of samples of each class (digits, upper case, lower case, punctuation, special symbols), or of characters extracted from a number of sample documents roughly corresponding to their frequency of usage.

When a system classifies an individual character, its output is typically either a character label or a reject marker that corresponds to an unrecognized character. By comparing outputs with the class labels of the corresponding inputs, the numbers of recognized characters, misrecognized characters, and rejects are determined. The standard display of the results of classifying individual characters is the Confusion Matrix, such as Table 1.

The Error/Reject Curve is a plot of the fraction of substitution errors against the fraction of characters that are rejected outright. The interesting properties of this curve are derived in [3], which also shows how to estimate the error rate from the reject rate.

True ID	Recognized as a	b	c	d	R	E
a	9	0	0	1	0	1
b	0	8	0	0	2	0
c	2	0	6	1	1	3
d	1	0	0	9	0	1
	3	0	0	2	3	5

Table 1: Sample Confusion Matrix

3.2 Character accuracy using text-lines and text-blocks

When text-line databases are used to evaluate devices, such databases permit measures beyond isolated characters in two important aspects. First, they allow the testing of character (and word) level segmentation. Touching and broken characters are difficult problems [2] and are responsible for a significant fraction of all OCR errors [6]. Second, they bring linguistic context into play. Since not only individual characters but also words are available, morphological (n-gram) and lexical techniques can be used to improve accuracy.

Text-blocks form the lowest level of input accepted by most commercial OCR devices. Text-block databases can also test text-line extraction methods, which tend to be vulnerable to tight spacing and to image distortions such as skew and curvature.

Current commercial OCR devices typically produce a stream of character labels and meta-characters from these types of inputs but do not return the location of each character. They recognize the ASCII characters and label them with the usual ASCII code. The ability of different devices to recognize non-ASCII characters varies widely, and so do their labeling schemata.

Although character locations are not given, the difference between the correct character stream and the generated character stream can be automatically measured, and has been termed "edit distance" [10]. The edit distance is the minimum number of character insertions, substitutions, and deletions required to correct the generated text. The edit distance between two character streams is usually determined in two steps. The first step, synchronization, divides the two strings into sets of matching and non-matching segments [4, 7]. The second step examines the non-matching segments to determine the editing operations required to convert the generated string into the correct string.

With isolated characters, substitution errors are easy to count. On the other hand, text-line based data may introduce one-to-many (m $\rightarrow$ rn) and many-to-one (rn $\rightarrow$ m) substitution errors. When consecutive segmentation errors occur, it is practically impossible to determine automatically the corresponding pairs of input and output labels. Therefore, a list of substitution error strings rather than a confusion matrix is used to categorize substitution errors.

A string synchronization tool also works as an analysis tool. When a text-line is not extracted, the character string corresponding to the text-line matches with nothing (or a null character). Therefore, missed text-lines can be easily identified.

3.3 Confidence metric

In the U.S., most commercial OCR devices return a single candidate rather than a set of candidates for each decision. When characters are classified with low confidence, these devices typically attach a suspect marker to the characters. These marked characters are usually highlighted during post-editing processes. A special (reject) character is usually reserved for completely unrecognizable characters.

Character labels generated by an OCR device can be classified as Correct, Marked Error, Unmarked Error, and False Mark [8]. Not all recognition errors should be considered equal. Marked Errors are easily corrected due to the presence of reject characters and/or suspect makers, whereas Unmarked Errors require costly proofing techniques. False marks are correctly-generated output labels that have been marked as suspect. The cost of correcting a string can be described by the cost model

$$\begin{aligned} Cost &= A(\#Marked\_Errors) \\ &\quad +B(\#Unmarked\_Errors) \\ &\quad +C(\#False\_Marks) \end{aligned}$$

where A, B, and C are appropriate weights. This model charges a fixed cost for correcting each error. To reflect the cost of post-editing more accurately, the extent to which errors are "bunched" may also be taken into account.

3.4 Word accuracy

When word accuracy is measured, the smallest acceptable input is an isolated word. Yet most commercial OCR devices do not accept isolated words. Thus, text-blocks or pages must be used. The output from a device is a stream of characters and not a stream of words.

To automatically compare outputs from devices with truth data, the correct demarcation of words is required. In the US OCR research community, "word" is often defined as a string of characters between interword spaces. However, this definition is different from the ordinary sense of "English word," as the following four examples illustrate:

[word word, word. "word"

By this definition, [word becomes an error if the character "[" is misclassified. An important topic for investigation is an algorithmically sound and intuitively acceptable definition of "word." Development of a comparison tool depends on the postulated definition of "word."

The concept of a marked character can be extended to analyze word errors. In this case, a marked word is defined as a word that contains at least one marked character. Some unmarked errors are spell checker-resistant (e.g., fall → fail). Such errors are extremely difficult to detect even by humans, and they should be assigned a high penalty.

For text retrieval applications, discriminating between stopwords and non-stopwords is important. Stopwords are common words, such as *the, and, but,* which are normally not indexed because they have essentially no retrieval value. Thus, non-stopword accuracy is an even more important metric for these applications [8].

3.5 Zoning

Before text on a page can be recognized, it must first be located. Commercial OCR devices support automatic zoning capabilities, whereby the device identifies the text regions and their reading order. The device finds columns of text, and if they are not part of a table, defines a separate zone for each column so that the generated text will be de-columnized. In addition, the device identifies graphic regions in order to exclude them. Zone representation schemes are not standardized, and the following methods are used by commercial OCR devices: rectangles, piece-wise rectangles, nested rectangles, and polygons. Therefore, geometrical comparison of zones is not feasible.

Since OCR devices attempt to locate all text on the page, an automatic zoning metric based on the number of character string (text-line) moves required to transform an OCR output to the correct reading order can be defined [5]. For example, when a page is not correctly de-columnized as shown in Figure 1, Text-Line-3 must be moved between Text-Line-2 and Text-Line-4. In this case, a single move operation corrects the reading order. Better zoning programs need fewer moves to correct the reading order.

Text-Line-1	Text-Line-3
Text-Line-2	Text-Line-4

(a) Two Column Format

Text-Line-1	Text-Line-1
Text-Line-2	Text-Line-3
Text-Line-3	Text-Line-2
Text-Line-4	Text-Line-4
(b) Correct Order	(c) OCR-Generated

Figure 1: Zoning Error

The number of moves needed to correct the reading order of an OCR output is calculated in two steps. First, an OCR-generated character stream is matched with the correctly ordered character stream, and transposed matches are identified. Second, the minimum number of moves required to correct the reading order is calculated. To our knowledge, this is the only metric that evaluates automatic zoning performance independently of zone representation schemes.

4 Other proposed metrics

There are many other potentially useful metrics for OCR research and the evaluation of OCR devices. In the following sections, we present examples of needed OCR metrics and suggest areas where other types of metrics are required.

4.1 Inverse formatting

How well a device extracts from page images the format information such as font, horizontal and vertical spacing, and text-line alignments, should be measured. Some important applications of OCR technology, such as file generation for word processors, require this information. The major problem in this domain is the equivalent representation of format information, such as horizontal spaces generated by tabs or

space characters. To automatically measure the performance, the ground truth data format and methods for normalizing equivalent representations must be developed.

4.2 Tables and formulas

The accurate conversion of printed material containing tabular data and equations is of paramount interest, if only because such material is so onerous to enter and verify. Although there has been considerable research on automating this type of symbol recognition, we do not know of any effort to evaluate the accuracy of the process.

4.3 Logical decomposition

How well a device recognizes logical objects, such as bylines and titles, should be measured. Automated generation of SGML or DSSSL tags from the input pages and automated routing of documents require this information. Therefore, comparison tools and truth representation should be able to deal with hierarchical information. Since some logical objects appear on selected pages only, such as the title on a title page and references on the last page, document databases are required for testing.

4.4 Throughput

The throughput of an OCR device is often advertised as the peak number of characters recognized per second. An alternative is the time required to produce OCR-generated text from a "typical" digitized page. However, the processing time required for a complex page layout or poor quality page is usually much higher. Furthermore, to measure the overall throughput, one cannot ignore the time required to correct zoning and OCR errors [2].

4.5 Page and document quality

Another important topic for investigation is document quality metrics. Paper, print, copier, and scanner quality affect the quality of the relevant document-images and can have a significant impact on the performance of these systems. Thus, it would be ideal to have a measure of the quality of a page. In a large document-conversion environment, such a metric would make possible the automatic determination of which pages could be processed more cheaply manually. For researchers and developers, such a metric would provide an important input to the recognition algorithm. In fact, page quality is a surprisingly elusive concept which deserves more study. A quality metric that enables easy discrimination of high-recognition-cost pages is probably not ideal as an input to a recognition algorithm. Here, measures of the quantity (and severity) of broken and touching characters would be of much greater value.

4.6 Other areas

Printed text documents are only one segment of data-entry. Other segments include typed and handwritten forms, engineering drawings, maps, schematic diagrams, music notation, and graphs. Non-text objects are, of course, often included in mainly-text documents. Although algorithms for recognition of graphic objects do exist, evaluation criteria are not well defined.

5 Conclusion

We have introduced several metrics that we have developed. We also suggested other potentially useful metrics for OCR research and the evaluation of OCR devices.

To develop performance metrics for document understanding systems, the following issues must be considered:

- Preparation of input test data and the representation of truth data.
- Normalization of outputs produced by devices.
- Algorithms for comparing the output to the corresponding truth data.
- Algorithms for computing the measure of interest.

The combined effect of automated experimental environments and new metrics will promote more rapid accumulation of knowledge in the field of Optical Character Recognition.

Acknowledgment

This work was supported by the U.S. Department of Energy.

References

[1] H.S. Baird, "Document Image Defect Models," *Structured Document Image Analysis* (H.S. Baird, H. Bunke, K. Yamamoto, eds), Springer Verlag, 1992, pp. 546-556.

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

[3] C.K. Chow, "On Optimum Recognition Error and Reject Tradeoff," *IEEE Trans. Information Theory 16*, 1970, pp. 41-46.

[4] J. Handley and T. Hickey, "Merging Optical Character Recognition Outputs for Improved Accuracy," *Proc. RIAO 91*, 1991, pp. 160-174.

[5] J. Kanai, S.V. Rice, and T.A. Nartker, A Preliminary Evaluation of Automatic Zoning, ISRI TR-93-02, University of Nevada, Las Vegas, 1993.

[6] T.A. Nartker, J. Kanai, S.V. Rice, A Preliminary Report on OCR Problems in LSS Document Conversion, ISRI TR-92-04, University of Nevada, Las Vegas, 1992.

[7] S.V. Rice, J. Kanai, and T.A. Nartker, "A Difference Algorithm for OCR-Generated Text," in *Advances in Structural and Syntactic Pattern Recognition*, ed. H. Bunke, World Scientific, pp. 333-341, 1992.

[8] S.V. Rice, J. Kanai, and T.A. Nartker, An Evaluation of OCR Accuracy, ISRI TR-93-01, University of Nevada, Las Vegas, 1993.

[9] S.V. Rice, The OCR Experimental Environment, Version 3, ISRI TR-93-04, University of Nevada, Las Vegas, 1993.

[10] R. Wagner and M. Fischer, "The String-to-String Correction Problem," *Journal of the Association for Computing Machinery*, 21:1, 1974, pp. 168-173.