Performance Evaluation for Document Analysis

Jonathan J. Hull

Ricoh California Research Center, 2882 Sand Hill Road, Suite 115, Menlo Park, CA 94025

ABSTRACT

A framework for evaluating the performance of a document analysis system is presented. This framework takes into account the task definition for the document analysis system, a data base on which that system is evaluated, the metrics used to evaluate performance, and the generalization of the results achieved beyond the confines of the test. Several recent significant efforts in evaluating document analysis systems are surveyed. How these efforts fit the general framework is discussed. The specific task that was evaluated, the data base used for the evaluation, and the generalization of the derived performance is presented. Most of these projects were designed for limited applications in which the translation of images of text into ASCII was the primary consideration. However, this is only part of what a document analysis system must often calculate. Other, less easily measured tasks, such as the subdivision of a document image into zones that represent regions of graphics, photographs, and text, must also be performed. Generally accepted solutions for measuring the performance of such tasks often do not exist. Several of them are mentioned as areas for future research. © 1996 John Wiley & Sons, Inc.

I. INTRODUCTION

A document analysis system is given an image of a document as input and it translates that image into an ASCII text description. The output description can be a flat file that contains just the ASCII in the document, and it can also contain other descriptive information about the layout of the document such as the fonts used to render the text, the paragraph segmentation of the document, its section layout, and its division into chapters.

The objective of a document analysis system is to help automate other processes that depend on the information contained in a document image. Examples include forms recognition, in which the handwritten text on a form is translated into ASCII. Another important process is information retrieval (IR), in which a user poses queries that are matched to stored documents. An IR system could return the recognized ASCII as well as the images of the original documents that matched the query.

There are often different constraints imposed on the quality of document analysis system output depending on the process that will be applied to it. In a forms recognition task it is often required that individual digits be recognized with high accuracy (<0.5% error rate). In an IR application the accuracy requirements may not be this high. Individual character accuracies as low as 80%–90% may be acceptable, since they may allow the correct documents to be retrieved.

Revised manuscript received 29 May 1996

Performance evaluation is an important part of the development of computer vision systems [1]. Issues in the evaluation of document image analysis systems have been addressed by other authors [2] and metrics have been proposed for the evaluation of OCR systems [3].

The evaluation of the performance of a document analysis system should be performed in the context of application processes that will be applied to its output. That is, users who depend on the extrapolation of performance figures derived in isolation outside the context of their specific application could be disappointed later when they install the working software and discover that the achieved performance does not match the expected performance.

The rest of this article presents a framework in which document analysis systems should be evaluated. This takes into account the application that will be applied to the output data. A survey of several significant efforts that have recently been performed to evaluate document analysis systems is presented and the degree to which they fit this framework is discussed. Several open research problems are summarized.

II. EVALUATION METHODOLOGY

A framework in which a document analysis system can be evaluated is shown in Figure 1. Each step in the evaluation process is discussed below in the order in which it should be considered.

First, the application process (Fig. 1a) should be defined and the performance that the user would like the document analysis system to achieve should be determined. For example, a system might be proposed to recognize the courtesy amounts (strings of digits) on images of bank checks. Because of the high potential cost in lost money and customer confidence, the performance requirement might be stated as a per-digit error rate of <0.01% and an accept rate (percent of images assigned a string of digits) of at least 40%.

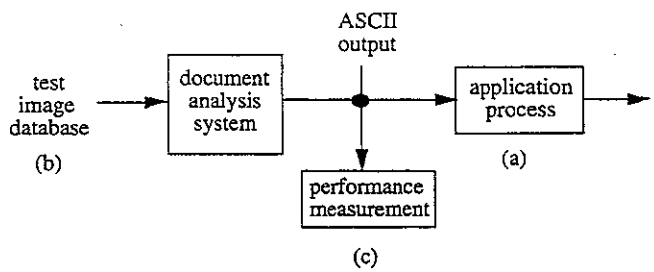

Figure 1. Performance evaluation framework. (a) Application definition; (b) test image data base specification; and (c) performance measurement.

CCC 0899–9457/96/040357–06

Second, a data base should be defined which accurately reflects the application process and the environment in which it will be tested (Fig. 1b). The original document images that are scanned for inclusion in the data base should be chosen from a stream of "live" data that reflects how the system will be used in practice. Also, an image capture device should be used that is equivalent to that which will be used in practice. It should be operated at the same sampling rate and resolution to help ensure that the same performance is obtained on the test and in practice. Several data bases have recently become available that can be used to develop and test OCR systems. These include the CEDAR CD-ROM of gray-scale images of handwritten city names, state names, and ZIP codes [4], as well as the University of Washington CD-ROM of machine-printed text [5].

In the bank check recognition example, a data base of over 100,000 bank checks scanned at 200 pixels/inch (ppi) binary (one bit/pixel) might be used. The original paper checks should be selected from actual data before it is presented to human operators for keying. The large number of images required are determined by the low error rate and the need to have a suitable confidence interval around the error rate. The sampling rate is typical for the type of high-speed image capture device that might be used in this application.

The third step in the evaluation framework is running the document analysis system on the image data base gathered in the second step. This process should be performed by a third party in an environment as close as possible to that which will be used in practice. If possible, the organization that performs the test should have no self-interest in its outcome. This will help guarantee that the results are not tampered with. However, in performing the test the third party should be provided reasonable access to the developers of the document analysis system so that any question about how to operate the software can be resolved.

The fourth step in the evaluation framework is measuring the performance of the system on the test images (Fig. 1c). This should be a simple implementation of the test statistics defined in the first step. Some additional consideration might also be given to measuring parameters that allow implementation tradeoffs to be determined.

For example, an error-reject analysis of the performance of a bank check recognition systems might be a valuable comparison tool. One system (system A) might have a 0.01% digit recognition error rate with a 40% accept rate and a 0.5% digit recognition error rate with a 50% accept rate. Another system (system B) might have a 0.01% digit recognition error rate at 40% and a 0.02% digit recognition error rate at 50% accept. In this case, system B might be the preferred choice, since its performance seems to be more stable that that of system A, even though both systems might have met the basic performance criteria that were set before the test was performed.

III. OBJECTIVES AND SAMPLING DESIGN

The objective of analyzing the performance of a system is an important part of designing how the data are sampled. The basic approach outlined above of randomly sampling the data which the system will encounter in practice is effective in giving users an idea of the performance they can expect. However, this can obscure problems that could be significant in actual practice.

An example in automated bank check processing is the recognition of European handwriting. A character recognition algorithm developed in the United States might not have been trained on ones that look like sevens, and sevens that have horizontal bars through the main stroke. A data base gathered with simple random sampling might contain a small fraction of characters written like this. Indeed, it might be impossible to identify this as a potential performance problem until the system that contains this recognition algorithm is deployed in general-purpose use. Such a system would probably perform to specification until it encountered a long sequence of check images from a customer with a European handwriting style. System performance would then drop significantly. The overall negative impact of this drop would be proportional to the number of images that have this handwriting style. In the worst case, automatic recognition performance could drop to 0% and all the images would have to be hand-keyed.

The sampling design used to choose the images in the data base can be varied to give a better overall picture of system performance. Stratified and sequential sampling methods are frequently used in developing performance models [6]. In a stratified sampling procedure, known characteristics are used to divide the sample space into strata. Data can then be sampled systematically within each strata. The effect of this procedure is to reduce the variance within strata and produce an overall standard error that is less than would have occurred in a simple random sample.

An important part of constructing a stratified sample is choosing how to divide the input data into strata. This requires knowledge of characteristics of the data that will effect system performance. In the previous example this might be the style of handwriting (American versus European).

An additional example of the use of stratified sampling was a study of mail piece image characteristics conducted for the United States Postal Service (USPS) [7]. The USPS wanted to learn about physical characteristics of mail pieces that might affect automated mail processing. Based on expert knowledge of the mail stream it was known that the mail "quality" varied by the number of pieces processed at a given post office. Also, the percentage of mail that was gathered from collection boxes and the percentage that was generated by large volume mailers (e.g., credit card statements from a bank) were known to be important determinants of the physical characteristics of the mail.

Two strata were chosen for this experiment (volume processed and percentage of collection mail) and each stratum was divided into five parts such that about 20% of the 473 major post offices in the United States fell into each part. One post office was then chosen from each of the 25 cells in this design. A sequential sample of mail pieces was then collected by dividing the mail processed in a given day into 500 sequences that each contained the same number of mail pieces and choosing a single mail piece to represent each sequence. About 100 physical characteristics of each sample (e.g., color, size, weight, etc.) were recorded. This process was carried out at each office for a period of eight days to obtain a sample of 4000 pieces. This procedure provided information about the variation in mail piece characteristics at the offices represented by the 25 cells in the table. Statistical projections from this sample to the national mail stream were also developed.

IV. CASE STUDIES

Two significant series of competitive tests have been performed of various document analysis systems over the past 5 years. The National Institute of Standards and Technology (NIST), under sponsorship from the Bureau of the Census, conducted tests of

hand-printed character recognition in 1992 and handwritten phrase recognition in 1994. The objective of these tests was to determine the feasibility and cost effectiveness of automating the entry of handwritten data from forms.

Competitive tests of commercial OCR systems for machine-print recognition have also been conducted on a yearly basis (since 1992) by the Information Science Research Institute (ISRI) at the University of Nevada at Las Vegas (UNLV). The objective of these tests was to measure the overall performance of commercial OCR systems.

The rest of this section discusses these projects and how they fit the performance evaluation framework presented above.

A. NIST1: Isolated Handwritten Character Recognition. The first NIST test of handwritten OCR accuracy was conducted in May 1992 [8]. The task was defined to be the recognition of isolated handwritten characters scanned from forms. This was a simple OCR task that required no contextual analysis.

A data base was defined on which this test was performed. A sample form was designed that contained 30 fields. Each field was associated with a random string of digits or alphabetic characters that were printed above an empty box. The forms were distributed to 2100 subjects who handwrote the character strings in the corresponding boxes. These forms were scanned at 300 pixels/inch in binary mode (one bit/pixel). Each field was subsequently segmented into individual characters and the string of digits above each box was used to assign a truth value to the individual images. The assignment of truth values to images was subsequently confirmed by manual inspection. This resulted in over 250,000 isolated numerals and over 100,000 isolated alphabetic characters that could be used as training data. An additional set of 500 forms were used to generate over 50,000 isolated numerals and over 25,000 isolated alphabetic characters for a test data set.

The test was conducted using a two-phase process. In the first phase, the training data were distributed to 26 voluntary participants. After a period of familiarization during which they adapted their systems to the format of the data, the test data were also distributed. A short, fixed amount of time was designated during which the test was to be conducted and the results returned to NIST. A detailed specification was provided for the results file so that the evaluation methodology could be simplified.

The evaluation of performance was performed by NIST. The data in the results file from each participant were used to plot error versus reject performance for each system in isolated digit recognition, isolated lower case recognition, and isolated upper case recognition. Figure 2 shows the error versus reject-rate curves for the systems that provided confidence values on the digit images in the test data. It should be noted that the error-rate axis is plotted on a log scale. It is seen that the best system had about a 1.5% error rate when a decision was made for every digit in the test set. The maximum error rate was about 20% at the zero reject level. The range of performance narrowed significantly to between 0.03% and 3.0% error when 50% of the data were rejected (i.e., decisions were made for half the digits).

This test fit the proposed evaluation framework fairly well. The objective was clearly defined, a data base was gathered that reflected that task, a test was conducted, and the results of that test were analyzed.

However, there were a few points that deserve discussion. The training data were collected from a population of professional census data collectors. These are people who spend a lot of time working with handwritten forms and understand the implications of automating the recognition and data entry from forms. The testing data were gathered from a population of high school students. They may have been less well motivated but maybe more representative of the general population. Perhaps not unexpectedly, as a general rule, the results showed a statistically significant decrease in performance on the test data as compared to the training data. This illustrates the point that the data base should be as reflective of the actual application as possible.

Another interesting characteristic was that the testing was performed by the developers of the systems, not by a disinterested third party. The potential for submitting false results was diminished by the use of a short, fixed time period for conducting the test. However, this aspect of the procedure should be kept in mind when the results are evaluated.

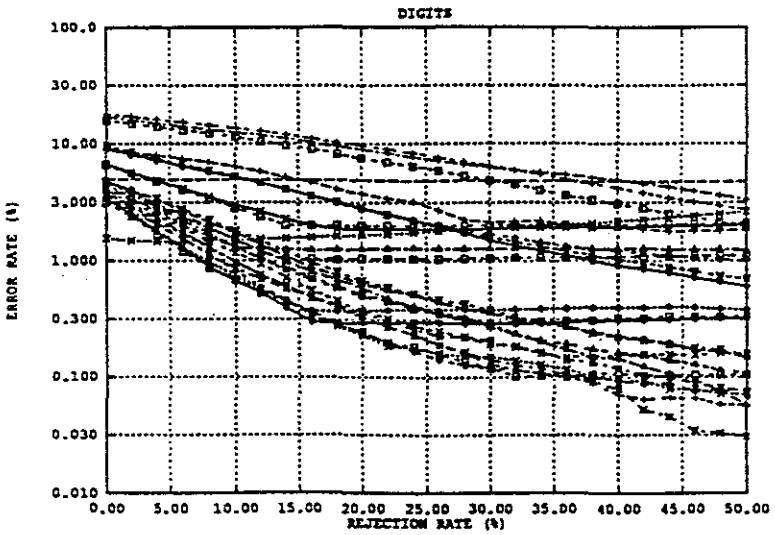

Figure 2. Error rate versus reject rate on the digit recognition portion of the test [8].

B. NIST2: Handwritten Phrase Recognition. The second NIST test of handwritten OCR performance was conducted in late 1993 and early 1994 [9]. The task was defined to be the recognition of handwritten phrases scanned from the Industry and Occupation miniform from the 1990 Census. Specifically, the task was defined as the character-by-character transcription of the handwritten data. This task was designed after the first NIST test was completed as a more realistic test of how a handwritten OCR would be used in practice by the Census Bureau.

The data base used for this test was gathered by scanning both microfilm copies as well as paper copies of census miniforms at 200 dpi in binary mode. A 75×90-mm section of each form was scanned that contained three answer boxes in which the handwritten data had been entered as well as the surrounding machine-printed text that described the questions (see Fig. 3 for an example miniform image).

Two CD-ROMs were produced that contained training data, and a third was produced that contained testing data. The first CD-ROM contained 12,500 miniforms scanned exclusively from microfilm. The second CD-ROM contained 6000 miniforms scanned from microfilm and 6000 miniforms scanned from paper. The third CD-ROM (the test data) contained 3000 miniforms scanned from microfilm and 3000 miniforms scanned from paper.

This test also used a two-phase process in which the participants were provided with the training data and were given an opportunity to familiarize themselves with the data format and train their systems. A 2-week testing phase was used in which the 10 organizations that participated were to run their systems on the test data and return the results to NIST.

The performance evaluation was also conducted by NIST. The confidence value for each answer returned by the recognizer was first compared to a threshold. If it fell below the threshold, the answer was rejected. Otherwise, the answer was compared to the character string typed in by a human operator. If there were any differences between the two strings, the answer was said to contain a field error. A problem was that this measurement can overestimate errors. For example, when the strings "WAITER" and "WAITOR" are compared, they are field errors. However, this counts the same as "WAITER" and "LAWYER" even though the second two alternatives are obviously much further apart.

The field distance was defined to compensate for this. This measures the number of characters that are misclassified when the two strings are aligned so that the number of characters in common are maximized. The alignment was performed with the Levenshtein distance metric. Thus, "WAITER" and "WAITOR" would be assigned a much lower field distance value than "WAITER" and "LAWYER".

The results shown in Figure 4 show a wide range of performance. At a zero reject rate (all answers were accepted), the field distance rate varied from 20% to 70%. For one of the systems, the field distance rate was reduced to 1% when 55% of the fields were accepted. That is, answers were accepted for 45% of the fields and those answers were correct in 99% of the cases. In fact, several of the systems performed well enough to automate a portion of the data collection process. In the case mentioned above, since the 45% that were recognized with high accuracy can be identified, the other 55% could be manually typed in. This would be a significant saving, since otherwise all the data would have to be manually entered.

This test also fit the proposed framework relatively well. An interesting observation concerned the data format. After the images scanned from microfilm were inspected, it was clear that their quality was much lower than that which would be obtained if the paper itself had been scanned directly. Since this would obviously cause the eventual performance of the system in practice to be estimated unfairly, a second data collection step was needed to gather images directly from paper. This illustrates the point mentioned earlier that the data used to perform the test should be as equivalent as possible to the data the system will encounter in practice; otherwise, the performance in practice could be estimated incorrectly.

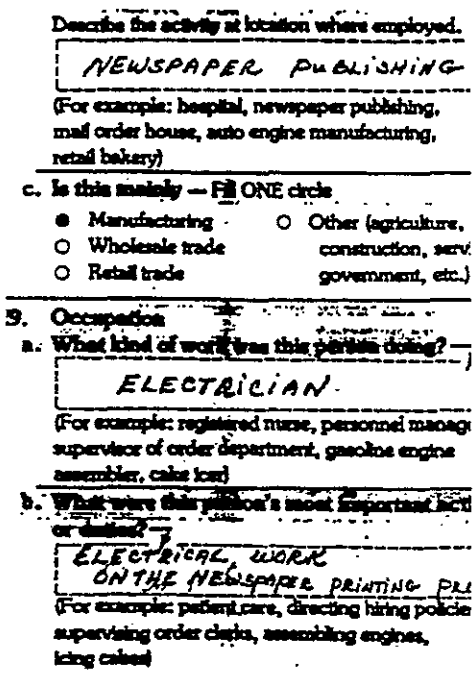

Figure 3. Example miniform image.

C. UNLV Commercial OCR Tests. The Information Science Research Institute at the UNLV has conducted yearly tests of commercial machine print OCR performance since 1992. The purpose of these tests is to provide an unbiased assessment of the performance of these systems and measure improvements in performance on an annual basis. A useful effect of these tests has been to encourage the OCR vendors to improve their performance.

The development of a data base of scanned images of machine print text and an ASCII truth file for those images is an important part of the testing process, since the results of the OCR must be compared with the truth file to derive the performance statistics. The data used in the 1995 test are summarized in Table I. Overall, 1529 page images were processed that contained a total of 6452 zones, 521,069 words, and over 3 million characters. Each page image was scanned and the text zones on each page were located manually. The text in each zone was entered by four independent human operators to ensure high accuracy. Any differences between the four transcriptions were resolved algorithmically.

The tests were conducted entirely by ISRI personnel. The OCR vendors submitted software that ran either on Sun workstations or PCs. Both the 300 ppi binary and grayscale versions of all the

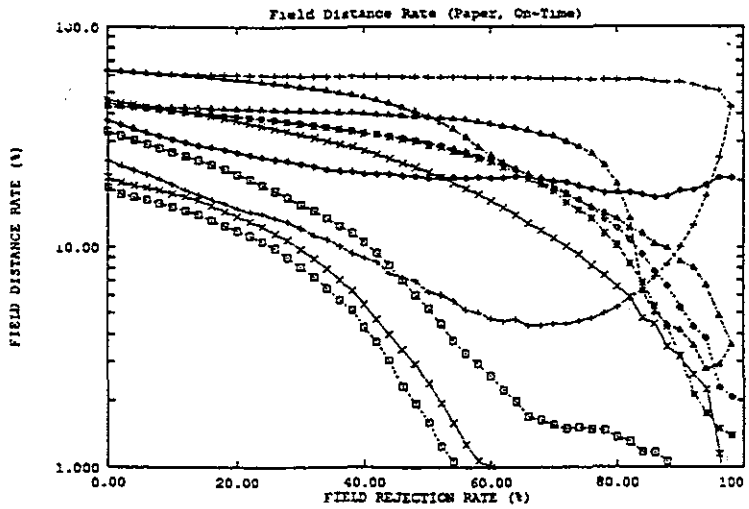

Figure 4. Field distance versus reject rate for the handwritten phrase recognition task. These results were derived from the systems submitted on time that were run on the data scanned from paper (figure excerpted from [9]).

zones in the test data were submitted to each of the participating OCR systems and the ASCII output file of each system was saved for later analysis.

The performance evaluation calculated a number of statistics. The character accuracy estimated the number of edit operations needed to manually correct the errors in the OCR output by the percentage of characters that were correctly recognized. This was determined with a string matching algorithm that determined the minimum number of edit operations needed to transform one string into another [11]. A number of other statistics were also determined. These included accuracy in word and phrase recognition.

The results of this test showed that there was no significant difference among the accuracy rates achieved by three of the participants (Caere, HP, and XIS) on binary images. When gray-scale images were input to the Caere and HP systems they committed from 10% to 40% fewer errors. This was significant, since the use of gray-scale data for character recognition has been discussed in the research community for a number of years. It is useful to see that such methods have met with commercial success.

The ISRI procedures fit the proposed testing framework well. The task was clearly defined and a data base was collected on which the desired performance figures could be determined. Also, the tests were conducted by unbiased third-party personnel (ISRI staff) who ensured that the test results were compiled fairly.

Some interesting observations of the test results concern the testing data sample. It was collected from application-specific sources such as business letters, DOE correspondence, etc. Even though great care was taken to gather images that represent a wide range of images styles and qualities; still, users of the ISRI test results should be careful to properly adapt them to their specific uses. That is, just because one system achieved over a 98% correct rate in character accuracy on fax images, it may not get close to that correct rate when applied to the images output by a particular fax machine. This could be caused by many factors not anticipated by this test.

Table I. Test data used for the 1995 ISRI test of OCR accuracy (table excerpted from [10]).

	Pages	Zones	Words	Characters
Business letters	200	1419	51,460	319,756
DOE sample	785	2280	213,552	1,463,512
Magazine sample	200	1414	114,361	666,134
English newspapers	200	781	84,026	492,080
Spanish newspapers	144	558	57,670	348,091
Total	1529	6452	521,069	3,289,573

V. RESEARCH PROBLEMS

There are several interesting research problems posed by the evaluation of document analysis systems. One such question is determining the quality of zoning output. This is the ability of the OCR to locate paragraphs and groups of paragraphs on a page and determine their correct reading order. An algorithm was proposed in [12] that compared the zones output by a commercial OCR to the zones listed in a truth file. While good performance was reported for this task, it still required a truth file to be available that included such zone data.

The ability of a document analysis system to segment an image correctly into logical components such as "title," "author list," "section titles," etc. was addressed in [13]. These distinctions are important for information retrieval, since they could be used to improve recall and precision performance. Further investigation of the relationship between document analysis systems and retrieval effectiveness is warranted, since this will continue to be one of the main applications for document analysis system output.

Another interesting problem in performance analysis is provided by OCR systems such as the Acrobat Capture system that was recently introduced by Adobe. This system attempts to produce an output file that preserves the physical appearance of the page and includes information about the fonts used, etc. If the approach is not confident in its recognition result, it may substitute a portion of the original image raster. Thus, when systems like this are evaluated their accuracy in preserving appearance should be included in the evaluation criteria. These criteria should include the correct recognition rate as well as the amount of text that is replaced by an image raster.

The metrics used to evaluate a system such as Acrobat Capture should thus be expanded beyond character recognition accuracy. In addition to the usual character recognition correct rate, error rate,

and reject rate, the percentage of characters that were replaced by an image raster and the correctness of this replacement should be considered. There is also a tradeoff between the percentage of characters replaced by their images and impact on information retrieval system effectiveness that should be incorporated in any statistical evaluation. For example, it is trivially possible to substitute image data for all the text in a document and retain 100% effectiveness in image reconstruction. However, this would reduce retrieval effectiveness to zero, since there would be no data available for searching.

Additional metrics should also be developed that measure the correctness achieved in font recognition. These metrics should incorporate some notion of distance between fonts. For example, incorrectly calculating that a word was printed in Times Roman 10 point from manufacturer *X* rather than the same type face from a different manufacturer should be weighted less than incorrectly calculating that the word was printed in Helvetica.

Other metrics that reflect specialized characteristics such as equation recognition [14] should also be developed. These systems convert the image of an equation into a logical form that can be used as an entry point to lookup in a table of solved integrals. Thus, correct calculation of the final logical form of the equation is what should be measured. Errors in font or point size recognition are irrelevant in this case. However, incorrect recognition of a single character may be sufficient to cause the system to look up the incorrect version of the equation.

VI. DISCUSSION AND CONCLUSIONS

A framework for evaluation of document analysis systems was presented. This included the four steps of system definition, data base collection, test operation, and results evaluation. Each of these steps must be performed carefully so that the results that are achieved can be projected to the performance of the evaluated systems when used in practice.

Three significant evaluations of system performance were also surveyed. These included two tests performed by NIST to determine the performance of handwritten text recognition systems as well as an ongoing series of tests done by ISRI at UNLV that benchmark commercial machine-print OCR systems. It was seen that these tests fit the proposed framework well. However, it was also pointed out that potential users of these results should still be cautious in extrapolating the achieved performance to their application domain.

ACKNOWLEDGMENTS

The anonymous referees contributed several comments that were useful in revising the manuscript.

REFERENCES

1. R. M. Haralick, "Performance characterization protocol in computer vision," in *Proceedings of the Workshop on Performance vs. Methodology in Computer Vision*, 1994, pp. 26–32.
2. G. Nagy, "Document image analysis: automated performance evaluation," in *Document Analysis Systems*, A. L. Spitz and A. Dengel, Eds., World Scientific, 1995, pp. 137–156.
3. J. Kanai, T. A. Nartker, S. V. Rice, and G. Nagy, "Performance metrics for document understanding systems," in *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, October 20–22, 1993, pp. 424–427.
4. J. J. Hull, "A database for handwritten text recognition research," *IEEE Transact. Pattern Anal. Machine Intell.* **16**, 550–554 (1994).
5. I. T. Phillips, S. Chen, and R. M. Haralick, "CD-ROM document database standard," in *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, October 20–22, 1993, pp. 478–483.
6. L. Kish, *Survey Sampling* (Wiley, New York) 1965.
7. *Mail Characteristics Study Findings: Vol. 1: Summary Report.* Kenan Systems Corporation, March 1991. Available from USPS Library, USPS Headquarters, Washington, DC.
8. R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burgess, R. Creecy, R. Hammond, J. J. Hull, N. J. Larsen, T. P. Vogl, and C. L. Wilson, "The first census optical character recognition systems conference," NISTIR 4912, National Institute of Standards and Technology, Gaithersburg, MD, August 1992.
9. J. Geist, R. A. Wilkinson, S. Janet, P. J. Grother, R. Hammond, N. J. Larsen, R. M. Klear, M. J. Matsko, C. J. C. Burgess, R. Creecy, J. J. Hull, T. P. Vogl, and C. L. Wilson, "The second census optical character recognition systems conference," NISTIR 5452, National Institute of Standards and Technology, Gaithersburg, MD, May 1994.
10. *1995 Annual Research Report*, UNLV Information Science Research Institute, 1995.
11. E. Ukkonen, "Algorithms for approximate string matching," *Information Control* **64**, 100–118 (1985).
12. J. Kanai, S. V. Rice, T. A. Nartker, and G. Nagy, "Automatic evaluation of OCR zoning," *IEEE Transact. Pattern Anal. Machine Intell.* **17**, 86–90 (1995).
13. K. Taghva, A. Condit, and J. Borsack, "An evaluation of an automatic markup system," in *Document Recognition II, Proceedings of SPIE – The International Society of Optical Engineering 2422*, February 6–7, 1995, pp. 317–327.
14. R. J. Fateman and T. Tokuyasu, "Progress in recognizing typeset mathematics," in *Document Recognition III, Proceedings of SPIE— The International Society of Optical Engineering 2660*, Jan. 28–2, 1996, pp. 37–50.