Document Image Recognition and Retrieval: Where are we?

Michael D. Garris
National Institute of Standards and Technology
225/A216
Gaithersburg, MD 20899 USA

ABSTRACT

This paper discusses survey data collected as a result of planning a project to evaluate document recognition and information retrieval technologies. In the process of establishing the project, a Request for Comment (RFC) was widely distributed throughout the document recognition and information retrieval research and development (R&D) communities, and based on the responses, the project was discontinued. The purpose of this paper is to present "real" data collected from the R&D communities in regards to a "real" project, so that we may all form our own conclusions about where we are, where we are heading, and how we are going to get there. Background on the project is provided and responses to the RFC are summarized.

Keywords: document image recognition information retrieval

1. BACKGROUND

In January of 1997, the National Institute of Standards and Technology (NIST) and the Department of Defense (DoD) entered into a joint project for assessing document image recognition and retrieval technologies. The aim of the project was to study and evaluate the automated production and usability of large-scale, on-line collections of digital documents.

Some of the proposed topics to be explored included:

1. How should these collections be constructed in an automated way using document recognition technology?
2. What impact does document recognition errors have on the accuracy of information retrieval?
3. What types of information should be provided from document recognition (in addition to text) that will facilitate and enhance information retrieval?
4. What types of data (in addition to text) should be indexed and retrieved to improve the usability of these collections?
5. What types of "real-world" integrated applications can be solved today, solved next year, solved in five years?

A key focus was to study the interface between document recognition and information retrieval technologies. The project was forward-looking in terms of evaluating the detection and utilization of non-text information (metadata) in the processes of recognition, indexing, and retrieval. Metadata are the physical and logical elements of a document including fonts, page layout, figures, tables, language, etc. Metadata may be used to construct queries and/or it may be part of the information retrieved. Thus the project was named, Metadata/Text Retrieval Conference (METTREC).

METTREC was intended to bring developers of document recognition technologies and information retrieval technologies together to work on integrated tasks through a series of evaluation conferences similar to the TREC and OCR Systems Conferences run by NIST.[1-3] Researchers and developers would be invited to work on a series of cooperative and measurable tasks that have relevance to real-world applications. Data was to be disseminated, results collected and tabulated, benchmarks measured, and conclusions published. In all this, NIST was chartered to provide program leadership, prepare and disseminate training and testing materials, implement benchmarking software and test-beds, host conferences and related workshops, and publish conference reports.

From January 1997 through January 1998, a team at NIST worked on preparing a large document image database[4] for future evaluations. The team also evaluated the UW Scoring Package[5] and as a result developed word-level alignment and scoring technology of our own. A technical planning committee of recognized experts was also formed during this time. All this was required to lay a foundation from which the first evaluation conference could be proposed.

Other author information: Email: mgarris@nist.gov; Telephone: 301-975-2928; Fax: 301-975-5287.

A break-out meeting was held for the technical planning committee at the SPIE Document Recognition V Conference.[6] At this meeting, the committee was updated on the progress made at NIST, and plans for the first evaluation conference were sketched out. When asked what organizations might be interested and capable of participating, it was suggested (and all the members of the committee in attendance agreed) that a survey of interest in participation should be conducted within the document recognition and information retrieval R&D communities.

This was a logical step to take, so a METTREC RFC was drafted by NIST and was approved by the planning committee, and on February 18, 1998 its distribution commenced.

The following groups of people received the RFC. The approximate number of recipients are listed with each group:

□ METTREC planning committee	17
□ METTREC mailing list	25
□ NIST OCR Systems Conference report recipients [2,3]	250
□ NIST public domain Form-Based Handprint OCR System recipients [7]	525
□ 5th SDAIR attendees [8]	100
□ SDIUT-97 presenters [9]	30
□ SPIE Document Recognition V attendees [6]	61
□ TREC-7 participants [10]	70
	1078 (with some redundancy)

The RFC was also posted to the following digest and news groups:

□ comp.ai.doc-analysis.ocr

□ comp.theory.info-retrieval

□ sci.image.processing

□ comp.ai.neural-nets

2. RFC SECTION 1: PROPOSED CONFERENCE

The RFC was comprised of two main sections. The first section presented a proposed evaluation conference including a cursory description of tasks and time schedules. The second part contained a list of questions to which responses were requested in order to ascertain the level of interest and capabilities within the R&D communities.

At this point in the project, the team at NIST had invested a year in preparing data and technology. It became clear that an evaluation conference was needed to bolster project visibility and ensure continued program support. Experience gained and progress made from the initial year of preparation were instrumental in defining the scope of the first proposed conference.

For example, issues of metadata were determined forward-looking to the point that they had to be postponed for the first round of evaluations. Based on experience gained with commercial OCR packages, it was concluded that there was virtually no commercial technology that could assist NIST in automatically detecting and labeling metadata objects; a capability needed to cut the cost of preparing metadata-based ground truth in document images. If commercial technology was not available, then there was certainly not going to be time to develop the necessary technology in-house at NIST and still achieve an evaluation within a reasonable time-frame.

The lack of commercial metadata-oriented technology also cast doubt on the existence of core technology that could be evaluated in a conference context. For a technology evaluation conference to work, there needs to be a number of valid alternatives to a common task, so that the alternatives can be tested and compared. If only one or two organizations provide technical solutions to a given task, there is really little to be gained from administering an organized evaluation.

In light of this, a proposal was put together that basically continued where the Confusion Track in TREC-5[11] left off with metadata-related issues relegated to a secondary track. This way interested parties could discuss future metadata evaluations in more of a workshop setting. Within the primary track, optical character recognition (OCR) would suffice as the enabling recognition technology, and retrieval would be text-based. Unlike TREC-5, the proposed METTREC evaluation would involve real OCR results from multiple participants, and the evaluation would be conducted on a larger set of document pages. Under this scenario, the aim would be to study the impact of OCR errors on the quality of information retrieval.[12,13]

In this paper, and for the METTREC project in general, a reasonable distinction is made between optical character recognition (OCR) and document recognition (DR). For the sake of clarity, a working definition for both is provided below.

OCR is the process of locating all the characters within a document image and, applying whatever pattern recognition technology available, so that each plausible character segment is assigned a class with an associated confidence value. Upon completion, an OCR system provides the hypothesized location and content of each word in the image. This technology has a long history of R&D,[14] it has been commercially available for a number of years, and it has been the focus of the past UNLV evaluations.[8]

DR on the other hand builds upon OCR technology as its foundation, extracting higher-level information and understanding about the contents of a document image. Categories of higher-level information include analyzing and identifying physical, logical, and linguistic attributes of a page. These include detection of fonts, layout analysis, inferring reading order, identifying languages, building dynamic thesauruses, and much more. While OCR is rather fixed in purpose and scope (lending itself to carefully engineered solutions), areas of interest in DR are diverse and much more open-ended. Also, while OCR has a long R&D history and has been aggressively commercialized, DR is less mature and has found its way into very few general-purpose products.

These definitions, while overly-simplistic, are important to the discussion in this paper. The conference proposal can be found in the first part of the RFC listed in the Appendix A.

3. RFC SECTION 2: QUESTIONS

The purpose of the RFC was to put a solid evaluation proposal before the DR and IR R&D communities, to solicit feedback on the level of interest of organizations to participate, and to receive back concerns and recommendations. Therefore, the second part of the RFC was comprised of a list of 22 probing questions carefully constructed to collect this kind of feedback.

The distribution of the RFC commenced on February 18, and by April 1, 22 responses had been received. Fourteen of the respondents stated at least potential interest in participating with 10 indicating a willingness to participate. Of the 10 parties, 5 were likely to participate and 5 were questionable. Their responses have been compiled and summarized in Appendix B along with the corresponding questions.

4. CONCLUDING REMARKS

The responses to the RFC demonstrate a general lack of interest and motivation for participation in a joint DR and IR evaluation conference. From the 10 parties who responded as being potentially able to participate, 5 were OCR respondents (3 had questionable participation), 4 were IR respondents (2 had questionable participation), and there was 1 possible metadata participant. A number of conclusions can be drawn from these results.

The forward-looking goal of METTREC was to pursue the use of automatically recognized metadata and measure its impact on information retrieval. We conclude that metadata cannot be readily detected with existing OCR technology (remember the distinction made earlier between DR and OCR). In fact, since the discontinuation of the UNLV evaluation conferences, it appears that a well-organized OCR research community no longer exists. On the other hand, the DR community is really in its infancy. As a result, very little research has been developed into technology tools for automatically detecting metadata in legacy paper documents that can be used for IR.

It is also apparent from the RFC responses that the IR community is not prepared to address the use of metadata even if such information could be automatically detected. IR researchers, in general, acknowledge that metadata is interesting and might be useful, but no one seems to be actually trying to exploit it. This conclusion is supported by a recent experience where, Donna Harman (the manager of the Natural Language Processing and Information Retrieval Group at NIST), recently was a key note speaker at an IR workshop in Pittsburgh.[15] While speaking, she explicitly raised the issue of what can/should be done with metadata. No one from the audience responded, and upon seeking individuals out after the talk regarding this topic, no one had any response. Her group is now beginning to do research in these areas.

One of the lessons learned from this project is that there is a window of opportunity within which technology evaluation conferences can play a significant role. A technology must be mature enough so that multiple alternative solutions exist for comparison, and yet the technology cannot be so mature and commercialized that the market has determined what solutions are viable. Both of these factors were significant in this project. On one hand, we have determined that metadata detection and utilization is very much a new research topic within both the DR and IR communities, so very few parties can be expected to have mature enough systems to participate. On the other hand, OCR technology, which was proposed for the initial evaluation conference, has been aggressively commercialized, so that very few parties from the OCR community were sufficiently motivated to participate.

This leads to another conclusion. Technology evaluation conferences run by NIST are voluntary, and organizations are expected to cover their own costs of participating. In order for an organization (especially a non-academic one) to participate, there needs to be a significant potential reward for participating on which a business case can be made. Otherwise, motivation will be left to rely solely on investing in R&D for science sake alone, an academic pursuit which will not attract participation from commercial entities.

In closing, NIST is mandated with the task of developing and evaluating technology that will impact private industry. Without a scientific focus on metadata or any significant commercial activity, NIST management has decided to discontinue long term effort in this area. NIST will, in the short term, publish and document the data and tools that have been developed on this project. These resource will be published and distributed as a NIST Special Database for research purposes.[4] There has been some discussion of using some of this data in a future TREC track.

5. REFERENCES

1. D. Harman Ed., “The Fourth Text REtrieval Conference (TREC-4),” NIST Special Publication 500-236, November 1995.
2. R.A. Wilkinson, J. Geist, S. Janet, P.J. Grother, C.J.C. Burges, R. Creecy, B. Hammond, J.J. Hull, N.J. Larsen, T.P. Vogel, and C.L. Wilson, “ The First Census Optical Character Recognition Conference,” NIST Internal Report 4912, August 1992.
3. J. Geist, R.A. Wilkinson, S. Janet, P.J. Grother, B. Hammond, N.W. Larsen, R.M. Klear, M.J. Matsko, C.J.C. Burges, R. Creecy, J.J. Hull, T.P. Vogel, and C.L. Wilson, “The Second Census Optical Character Recognition Conference,” NIST Internal Report 5452, May 1994.
4. M.D. Garris, S.A. Janet, and W.W. Klein, “Federal Register Document Image Database,” to be published as a *NIST Special Database 25* on CD-ROM.
5. “UW English Document Image Database I,” CD-ROM, University of Washington, Seattle, WA, 1993.
6. D.P. Lopresti and J. Zhou Eds., “Proceedings of SPIE Document Recognition V,” Vol. 3305, San Jose, CA, January 1998.
7. M.D. Garris, J.L. Blue, G.T. Candela, P.J. Grother, S.A. Janet, and C.L. Wilson, “NIST Form-Based Handprint Recognition System (Release 2.0),” NIST Internal Report 5959 and CD-ROM, January 1997.
8. “Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval,” Information Science Research Institute, University of Nevada, Las Vegas, April 1996.
9. “Proceedings of the 1997 Symposium on Document Image Understanding Technology,” Organized by the Laboratory for Language and Media Processing, University of Maryland, Annapolis, MD, April 1997.
10. D. Harman Ed., “The Seventh Text REtrieval Conference (TREC-7),” to be held at NIST, November 1998.
11. E.M. Voorhees and D.K. Harman Eds., “The Fifth Text REtrieval Conference (TREC-5),” NIST Special Publication 500-238, November 1997.
12. K. Taghva, J. Borsack, and A. Condit, “Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model,” *Information Processing & Management*, Vol. 32, No. 3, pp. 317-327, 1996.
13. Taghva, K., Borsack, J. , and Condit, A., Results of Applying Probabilistic IR to OCR Text, Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 202-212, 1994.
14. G. Nagy, “At the Frontiers of OCR,” Proceedings of the IEEE, Vol. 80, No. 7, July 1992.
15. “Invitational Workshop on Information Retrieval Tools,” The School of Information Sciences at the University of Pittsburgh, Pittsburgh, PA, March 20-21, 1998.
16. M.D. Garris, W.W. Klein, “Creating and Validating a Large Image Database for METTREC,” NIST Internal Report 6090, December 1997.
17. E.M. Voorhees and D.K. Harman Eds., “The Sixth Text REtrieval Conference (TREC-6),” NIST Special Publication, to be published.

APPENDIX A. REQUEST FOR COMMENT

February 18, 1998

PROPOSED METADATA TEXT RETRIEVAL CONFERENCE (METTREC)

This is a request for comments pertaining to possible participation in an upcoming METTREC conference. METTREC (Metadata Text Retrieval Conference) is a technology evaluation project created to examine the interfacing of Document Recognition and Information Retrieval technologies. The project is cosponsored by the National Institute of Standards and Technology (NIST) and the Department of the Defense (DoD).

PART I - PROPOSED METTREC CONFERENCE

This section proposes a set of tasks and the schedule necessary to initiate METTREC evaluations. The conference will be comprised of two tracks:

Track 1.) Text-based recognition and retrieval evaluation

Track 2.) Metadata demonstration

TRACK 1. Text-based Recognition and Retrieval Evaluation

The purpose of this track is to evaluate the impact of Optical Character Recognition (OCR) errors on Information Retrieval (IR). The OCR and IR systems participating in this track are not limited to commercially available products. However, participants must agree to the public release of their resulting scores and performance.

This track can be described in terms of 1. data, 2. evaluation methodology, 3. design of experiment, 4. proposed size of data sets, and 5. schedule.

1. DATA

A portion of the 1994 Federal Register (FR94) will be used in Track 1. FR94 documents are hierarchically structured, and FR94 pages are primarily 3-column text mixed with occasional figures, graphs, tables, maps, etc.

The FR94 has been scanned at 15.75 pixels per millimeter (400 pixels per inch) binary, totaling 249 daily issues, accounting for more than 67,000 pages.[16] Images from one issue are available via anonymous FTP at "sequoyah.nist.gov" under the subdirectory "pub/mettrec." There will be ground truth files associated with each FR94 page. This ground truth will include the page's text in reading order and various metadata tags.

A multi-lingual set of documents is also being prepared for potential use in this and future evaluations.

2. EVALUATION METHODOLOGY

In order for these types of experiments to be economically feasible and able to scale up, automated scoring methods will be used.

2a. Evaluating OCR: A Scoring Package developed by NIST for METTREC will be used to align OCR text results with a page's ground truth text and compute word error rates.

2b. Evaluating IR: Information Retrieval performance will be evaluated using "known-item search." A set of queries will be used that are carefully constructed to reference a specific FR94 page. That specific page will be judged to be most relevant (of rank 1). By controlling query composition and which pages should be retrieved, various IR and OCR factors may be isolated for analysis. IR performance will be analyzed by computing statistics on how distant a query's known page is ranked from position 1. This type of analysis was used in the Spoken Document Retrieval Track in the TREC-6 conference.[17]

3. DESIGN OF EXPERIMENT

Proposed tasks do not limit participation to those with both OCR and IR capabilities. Furthermore, the tasks are being designed so as not to require "teaming" between participants. Instead, the interface between OCR and IR will be open for analysis and refinement (a primary aim of the project). OCR results will be reported by multiple participants and subsequently disseminated to multiple IR participants. Participants who can conduct the retrieval tasks by means other than traditional OCR are also encouraged.

A subset of queries will be crafted to represent specific IR-related issues, and another subset will be used to study the impact of document image quality on OCR and subsequently on IR.

A training set (including a defined evaluation subset) containing FR94 page images, corresponding ground truth files, and a number of queries (with top-ranked pages identified) will be disseminated to all participants.

Upon commencement of a testing period, testing material containing a new set of FR94 page images will be disseminated to OCR participants. OCR results reported back by a specified deadline will be disseminated to the IR participants along with a new set of test queries and the ground truth text for the page images. IR results will then be required by a specified deadline.

All OCR and IR results will be scored and analyzed and a conference will be hosted by NIST. During the conference each participant will be given opportunity to discuss their experience and provide a description of their system. Participants are free to determine the level of system disclosure, but a discussion related to approaches, methods, and algorithms is strongly encouraged.

4. PROPOSED SIZE OF DATA SETS

□ Training Set: 2,000 FR94 page images
2,000 FR94 ground truth files
100 of the 2,000 designated for evaluation
5 known-item queries

□ Testing Set: 10,000 FR94 page images
10,000 FR94 ground truth files (IR participants only)
100 known-item queries

5. SCHEDULE

The first METTREC conference is scheduled for the end of September. Working back from there, the following schedule is proposed:

February	Request for Comment distributed
March	Call for participation distributed
April	Training data disseminated
June	Testing data disseminated for OCR
July	Testing data and OCR results disseminated for IR
August	Results reported to NIST
September	Conference hosted by NIST

TRACK 2. Metadata Demonstration

In order to explore the use of metadata in METTREC, a secondary demonstration track is proposed. Track 2 will include a number of additional known item queries along with their top-ranked pages taken from the FR94 pages disseminated for Track 1.

As an example, a query might be:
"Find me the table containing the EPA guidelines on safe drinking water."
where "table" is metadata and "EPA" (being found in an agency heading) is metadata.

Participants, who are able, will be encouraged to demonstrate there ability to detect and/or utilize the metadata included in this track. Participants, who are not able to detect or utilize this metadata, will be asked to comment on how the metadata might be utilized by their systems and what steps would need to be taken to utilize the metadata.

Information gathered from this demonstration track will be used to organize future metadata-based evaluations.

APPENDIX B. TABULATION OF RESULTS FOR METTREC REQUEST FOR COMMENT

PART II - QUESTIONS

April 1, 1998

1. REGARDING YOUR PARTICIPATION

Question	Summarized Response
1a. Are you able/interested in participating in this evaluation? (Why or why not?)	5 OCR (2 yes, 3 maybe) 4 IR (2 yes, 2 maybe) 1 Metadata (can't meet schedule)
1b. Can you recommend someone else who may be able/interested participating? (If so, who?)	No specific recommendations
1c. Are you able to participate in the proposed OCR tasks, IR tasks, or both?	10 Respondents: 3 OCR only 2 IR only 5 OCR & IR (2 of 5 use commercial off-the-shelf OCR) 1 Metadata only
1d. If you are capable of OCR tasks, but not IR tasks, do you require "teaming" with a specific IR provider or will you consent to having your results disseminated to many IR participants? (Why, and to what extent?)	2 Respondents require teaming
1e. Are you able to do the IR tasks in Track 1 (especially on very low quality document images) with input from other than traditional text-based OCR? (If so, are you able to demonstrate this in the evaluation?)	2 Respondents answered Yes

2. REGARDING THE PROPOSED EVALUATION SCHEDULE AND SCALE

Question	Summarized Response
2a. Are you willing to comply with the proposed schedule culminating with a conference at the end of September? (Why or why not?)	10 Respondents: 4 Yes 2 Yes, but very tight 2 Possible schedule clashes 2 No
2b. If unable to comply with the proposed schedule, in what time-frame might you be ready to participate?	No response from those able to participate
2c. Will you have difficulty with the scale of the evaluation? (If so, why?)	6 Respondents: 4 No 1 Yes 1 Maybe (other 4 assumed No)
2d. Is 10,000 pages too much for OCR participants to process? (If yes, how many pages are you willing to process?)	5 Respondents: 2 No 2 Yes 1 Maybe
2e. Is 10,000 pages too little for IR experiments? (If yes, how many pages at a minimum should be used?)	4 IR Respondents: 4 OK as a start
2f. Are 50-100 known item test queries sufficient? (Why or why not?)	3 IR Respondents: 3 OK, but more is better
2g. Are the size of proposed training sets sufficient? (If not, how many pages or queries would you prefer to have?)	4 Respondents: 4 Yes
2h. Is 100 pages of the training set sufficient for an evaluation set? (If not, how many pages?)	4 Respondents: 3 Yes 1 No
2i. What factors would you like to see analyzed in these types of evaluations?	1. OmniFont capabilities, variety of page types 2. Proportional credit for a page returned in position 2..10 3. Graph of sorted rank positions of the known items

3. REGARDING THE USE OF METADATA

Question	Summarized Response
3a. What metadata are you able to automatically detect? (What metadata are you able to detect in the FR94 pages?)	1 Respondent: 1 Tables
3b. If metadata were provided, how would you use it in your IR system?	No respondent could use metadata
3c. If you have the capability of using metadata in an IR system, what metadata can you use? (What metadata can you use from the FR94 pages?)	No respondents

4. REGARDING LANGUAGE CAPABILITIES

Question	Summarized Response
4a. Does your OCR technology handle languages other than English? (If so, what languages?)	3 Respondents: 1 Arabic, Chinese 2 French 3 Dutch, Danish, French, German, Italian, Norwegian, Portuguese, Spanish, Swedish, UK&US English
4b. Does your IR technology handle languages other than English? (If so, what languages?)	3 Respondents: 1 French 2 Chinese 3 Dutch, German, English, French

5. REGARDING THE USE OF MULTIMEDIA DOCUMENTS

Question	Summarized Response
5a. What multimedia capabilities does your technology support?	7 Respondents: 1 Speech 2 Image zones 4 None
5b. What multimedia capabilities might your technology support in the future?	1 Respondents: 1 Grayscale or color documents

6. GENERAL COMMENTS AND SUGGESTIONS

Question
6a. Do you have any other general comments or suggestions?
Summarized Response
1. What are some potential sources for funding participation.
2. There is an incompatibility between the amount of data to measure OCR and IR performance.
3. What is the definition of metadata and what are its possible uses.
4. The proposed test is limited in the range of fonts and image types.
5. OCR error rates can potentially be dominated by reading order errors.