DRR Research beyond COTS OCR Software: A Survey[1]

Xiaofan Lin
Imaging Systems Laboratory
HP Laboratories Palo Alto

HPL-2004-167
October 4, 2004*

E-mail: xiaofan.lin@hp.com

document recognition and retrieval. COTS software, OCR

After decades of research, Optical Character Recognition (OCR) has entered into a relatively mature stage. Commercial off-the-shelf (COTS) OCR software packages have become powerful tools in Document Recognition and Retrieval (DRR) applications. One question naturally arises: What areas are left for new DRR research beyond COTS OCR software? There are many discussions around it in recent conferences. This paper attempts to address this question through a systematic survey of recently reported DRR projects as well as our own Digital Content Re-Mastering (DCRM) research at HP Labs. This survey has shown that custom DRR research is still in great need for better accuracy and reliability, complementary contents, or downstream information retrieval. Several concrete observations are also made on the basis of this survey: First, the basic character/word recognition is mostly taken on by COTS software, with a few exceptions. Second, system-level research with regard to reliability and guaranteed accuracy can seldom be replaced by COTS software. Third, document-level structure understanding still has much room to expand. Fourth, post-OCR information retrieval also has many challenging research topics.

* Internal Accession Date Only

[1] To be presented at and published in the SPIE Conference on Document Recognition and Retrieval XII, 16-20 January 2005, San Jose, CA

DRR Research beyond COTS OCR Software: A Survey

Xiaofan Lin
Hewlett-Packard Labs
1501 Page Mill Rd MS 1203, Palo Alto, CA 94087
Email: xiaofan.lin@hp.com

ABSTRACT

After decades of research, Optical Character Recognition (OCR) has entered into a relatively mature stage. Commercial off-the-shelf (COTS) OCR software packages have become powerful tools in Document Recognition and Retrieval (DRR) applications. One question naturally arises: What areas are left for new DRR research beyond COTS OCR software? There are many discussions around it in recent conferences. This paper attempts to address this question through a systematic survey of recently reported DRR projects as well as our own Digital Content Re-Mastering (DCRM) research at HP Labs. This survey has shown that custom DRR research is still in great need for better accuracy and reliability, complementary contents, or downstream information retrieval. Several concrete observations are also made on the basis of this survey: First, the basic character/word recognition is mostly taken on by COTS software, with a few exceptions. Second, system-level research with regard to reliability and guaranteed accuracy can seldom be replaced by COTS software. Third, document-level structure understanding still has much room to expand. Fourth, post-OCR information retrieval also has many challenging research topics.

Keywords: document recognition and retrieval, COTS software, OCR

1. INTRODUCTION TO COTS OCR

Through the Digital Content Re-Mastering (DCRM) project [1] we have the opportunities to experiment with the SDKs of several leading COTS OCR packages:

- ABBYY FineReader SDK [3]
- ScanSoft TextBridge SDK [4]
- IRIS Readiris Toolkit [5]

According to our experience, COTS OCR software usually provides the following functions:

- Retrieve document images from scanners as well as a wide range of image formats such as TIFF, JPEG, BMP, and recently PDF.
- Conduct preprocessing on the input image, including binarization, noise removal, de-skewing, etc.
- Analyze page layout structure and segment the image into regions of different types: text, image, graphics, tables, etc.
- Recognize the characters in the text regions. They usually support multiple languages such as English, French, German, and many other European languages. There are also COTS OCR packages targeting Asian languages like Chinese, Japanese, and Korean [6].
- Render the recognition results into various output formats such as plain text, Microsoft Word, RTF, and PDF.

Besides, some versions of COTS OCR SDKs provide more advanced capabilities:

- Train on special characters or fonts and post-process on custom lexicon. This is very useful when recognizing documents in specific domains, such as history, law, and literature.
- Provide confidence values of OCR results. This enables more possibilities in multi-engine combination.

- Geometrically map OCR results to the images. The coordinates associated with OCR text can be provided at different levels: word or character. Some SDKs can also provide the coordinates based on the original images while others only provide the coordinates on the preprocessed images after de-skewing and rotation.

Obviously, the above functions cover a large portion of the area of document recognition. Over the years, the COTS OCR packages are steadily increasing their accuracy through either refinement of algorithms or internal combination of multiple algorithms (This is only a reasonable guess based on the companies' press releases [3][4] because the OCR vendors seldom publish technical details of their proprietary OCR engines). Moreover, the COTS OCR SDKs usually come with well defined APIs, ready-to-use examples, and detailed documentation on the usage. In addition, OCR software is gradually becoming commodity accompanied by falling prices. Thus, more and more DRR applications are tapping into the functionalities of COTS OCR.

2. DRR APPLICATIONS BASED ON COTS OCR

We have surveyed the recently published DRR projects that involve COTS OCR. Table 1 shows how they split the job between COTS OCR and custom DRR research. From this table, it can be seen that DRR research plays several important roles that cannot be fulfilled by COTS OCR:

2.1. Custom DRR can help to achieve the required accuracy or reliability.

COTS OCR software strives to work reasonably well on the majority of document images. The software vendors usually put more R&D resources to features or improvements that are visible to the "average" applications. As a side effect, its accuracy in a specified domain is not guaranteed. Thus, custom DRR can effectively fill in the white spaces. Special DRR methods can be applied in several ways to achieve the accuracy or reliability required by certain applications:

- Preprocess the image to facilitate later OCR.

Such applications include the recognition of text in video [18][16][25][30] or difficult images [9][12][17][19]. Typically, special algorithms are designed to locate the text regions, which are then binarized using adaptive thresholding techniques. Although COTS OCR packages are becoming better recognizing documents with complex color background, they are still focusing on images scanned from paper documents and may perform poorly on the raw low-quality images taken through other channels such as cameras, video recorders, and microfilms. Doermann et al gave a detailed survey on the technical challenges and existing solutions in this area [34].

- Combine multiple COTS OCR engines.

Classifier combination has proven to be effective in increasing the recognition accuracy [1][9][28][35][46]. Since there are more than one COTS OCR engines on the market, it is natural to combine them to achieve accuracy higher than that of any individual engine. Compared with general classifier combination, the combination of COTS OCR engines has several unique characteristics:

1. Simple methods such as voting are more feasible. Because each COTS OCR engine is a self-contained complex system and only exposes high-level APIs, it is very difficult to couple them closely and apply sophisticated combination scheme that requires detailed information from the individual engines. In our own DCRM work [1], we pointed out that in many cases the label of the top choice is the only information guaranteed by all of the engines. Not all of the COTS OCR engines will provide scores or alternatives in a comparable manner.

2. Simple combination methods work well! Interestingly enough, simple combination methods perform satisfactorily. For example, we observed a reduction of 30% (relative) in error rate by combining threes COTS OCR engines using majority voting [1]. Belaid has achieved the goal of reducing error rate to 1/10000 by selecting a primary OCR engine and improving it with a secondary OCR engine using some heuristics [28]. To some extent, it is due to the fact that different COTS OCR engines are developed by different teams using different techniques in all of the steps from preprocessing, layout analysis, character segmentation, to actual recognition and post processing. Thus, COTS OCR

engines are expected to be uncorrelated to each other: When one engine performs badly on a certain document, the other engines are likely to perform well on that document.

- Post process the OCR results with domain knowledge.

COTS OCR engines will employ general post-processing techniques such as dictionary to improve the raw recognition results. However, many DRR projects utilize more sophisticated domain-specific post processing techniques to push the envelope. For example, Hauser et al correlated the historical data with OCR results to improve bibliographic information recognition for building MEDLINE database [20]. Nartker et al used document-level knowledge to correct OCR errors [21]. If a word appears more than once in a document, the MANICURE system can potentially correct one OCR error with the recognition results on other pages.

- Use various system-level quality control measures.

Most real-world applications impose some sort of upper bound on the allowable error rate. They do not expect the computers to achieve error rate lower than the upper bound on every page, but they do hope that problem pages can be automatically spotted so that the human operators can handle them properly. On the other hand, it is critical to keep human intervention as little as possible to reduce the overall cost and to increase the efficiency. Because different applications have different requirements and workflows, COTS OCR software can do little in this regard and significant DRR research is needed. Sarkar and Baird introduced a triage step to predict on which pages COTS OCR cannot achieve the expected accuracy so that those pages will be manually processed [29]. Thoma et al created a whole workflow to build MEDLINE database [27]. They created a component to modify OCR's raw confidence scores and a component to correct institutional affiliations. In our DCRM work [8], we employed a multi-pass solution in which the first several passes are automatic and the final manual pass only process less than 1% of the pages.

2.2. Custom DRR can enable the downstream information retrieval (IR) applications.

In many DRR applications, OCR is just the start point instead of the end point. Much richer downstream information retrieval applications, which are mostly driven by custom DRR components, are the ultimate goals. DRR research along this line has several major topics:

- Analyze and understand the higher-level or deeper document structure.

COTS OCR software only provides the page-level layout analysis. Thus, new DRR components are required to extract document structure at higher level. As a valuable part of digital library effort, book structure recognition has attracted much attention [2][7][14]. The objective is to locate and label the table of the contents and chapters/articles in a book based on the OCR results. Besides the higher level structure analysis, another direction is the deeper semantic structure analysis, such as bibliographic information extraction [11][15], business letter logical element extraction [24], and fax proper noun extraction [32]. Kink and Dengel described a rule-based document logical structure analysis method on top of the raw physical structure from OCR [43]. This type of structure analysis usually involves some models. The challenge is to keep the model as general and adaptive as possible. Liang [24] and Mao [15] introduced automatic adaptation in their systems so that the models can adjust to new documents. Viola et al introduced unified machine learning in fax routing based on OCR results [47]. In our book structure analysis work [8], instead of modeling the layout of title pages, our model only assumes that text strings in the table of content pages will re-appear in the title pages. This model is much more general than the layout models.

- Fuse OCR results with other sources to improve IR performance.

Some IR application can work without OCR. However, integration of OCR results with other sources can significantly boost the IR accuracy because OCR serves as an input channel quite independent of the other sources, such as the raw image information. Ozawa et al combined image-based matching and OCR-based matching to locate presentation slides in video [16]. Aradhye et al also combined the recognized text from video and the map images to recognize the geographic locations featured in the video [18]. In the "improving" module of the smartFix system [44], Dangel and

Klein used constraint solving to post process OCR results. It is very effective in processing table documents because the different fields are usually bound by some business logic rules.

- Deal with the imperfect text input.

The biggest difference between OCR-based IR and conventional IR is that OCR is not a perfect process and errors in the generated text are the rule rather than the exception. So much DRR research is devoted to making IR tolerant to OCR errors. Jin et al [22] introduced content-based probabilistic correction in the post processing so that the IR system will be trained to accept the common error patterns. In addition to the practical solutions to reduce the effects of OCR errors on IR, there is also more theoretic investigation on to what extent OCR errors will affect the IR process. The classic work is done by Taghva et al at UNLV. They first studied how the average recall and precision rates are affected by OCR errors [36]. Their observation is that most of the time the impact is insignificant. Their recent work shows that running headers and footers will not affect proximity searching too much [37]. In our own research, we demonstrated that how OCR errors affect a commonly used component in IR --- Part-of-speech (POS) Tagger [38]. POS tagging error rate rises linearly with the OCR error rate.

2.3. Use COTS OCR feedback to boost other document understanding tasks.

Opposite to the preceding subsection, another interesting topic is to use OCR feedback to improve the upstream processing. Ma et al utilized information (font, words, and text lines) from COTS OCR to segment structured documents such as dictionaries and phone books [23]. In our DCRM research [8], we also used OCR feedback to further distinguish text regions from non-text regions.

Table 1: How the job is split between COTS OCR and novel DRR research?

No	Source	Functions from COTS OCR	Unique DRR research
1	Aradhye et al [18]	Recognize text in video	Image preprocessing and text location, information fusion after OCR
2	METAe project group [7]	Modified ABBYY OCR is used to recognize Fraktur text on old books.	Book structure element recognition
3	Lin and Simske 1][2][8]	Recognize text on books and journals	Document structure analysis, multi-pass image processing for the image part of PDFs (Multiple COTS OCR engines are used for high reliability)
4	Barrett et al [9]	Recognize printed text on microfilms	Image preprocessing, compression (Multiple OCR engines are used)
5	Besagni et al [11]	Recognize the text in the reference section	Generate bibliographic reference structure
6	Esposito et al [12]	Recognize the text in historical documents	Image preprocessing and text location (Binarized text images in specified locations is recognized)
7	Taghva et al [13]	Recognize the text in the Licensing Support Network collection	Proximity query search
8	He et al [14]	Recognize Chinese books	Hierarchical logical structure extraction
9	Mao et al [15]	Recognize medical journals	Logical labeling of the title pages
10	Ozawa et al [16]	Recognize text in slide images and video	Integration of image-based and OCR-based matching
11	Shin et al [17]	Recognize text on business cards	Preprocessing the images to compensate for various lighting conditions
12	Maderlechner et al [19]	Recognize text in legal registers	Preprocessing the images and information extraction after OCR

13	Hauser et al [20]	Recognize medical journals	Correlating OCR results with historical data to improve the accuracy
14	Nartker et al [21]	Recognize UNLV database	OCR correction using document-level knowledge
15	Jin et al [22]	Recognize TREC corpus	Post process using content-based probabilistic correction
16	Ma et al [23]	Recognize text on structured documents such as dictionaries	Use OCR to bootstrap page segmentation
17	Liang et al [24]	Recognize business letters	Logical labeling on top of OCR
18	Du et al [25]	Recognize text in video	Preprocessing and text location
19	Doermann et al [26]	Recognize bilingual dictionaries	Post processing to build the bilingual lexicon
20	Thoma et al [27]	Recognize medical journals	Whole workflow from scanning to QA
21	Belaid et al [28]	Industrial document capture application with required accuracy	Use multiple OCR engines to reduce error rate
22	Sarkar et al [29]	Recognize Patent Literature	Introduce triage step to decide on which documents OCR can achieve desired accuracy
23	Hull et al [30]	Recognize text in video	Adaptive thresholding before OCR, and novel Video Paper system
24	Souza et al [31]	Recognize degraded text	Selection of the best image preprocessing filter to reduce error rate
25	Likforman-Sulem et al [32]	Recognize fax images	Combination of OCR and text analysis to extract proper nouns
26	Zhang et al [33]	Recognize thick bound books	Preprocessing images to straighten text lines
27	Stefan and Klink [43]	Recognize business letters and University of Washington corpus	Document logical structure on top of the raw physical layout from OCR
28	Dangel and Klein [44]	Recognize mainly business form documents such as invoices	Task-adaptive document analysis and understanding system
29	Nakano et al [46]	Recognize Japanese text	Combination of multiple commercial Japanese OCR packages to reduce error rate, with focus text alignment
30	Viol et al [47]	Recognize text on fax images	Fax routing based on OCR results

3. COMMON ISSUES USING COTS OCR

There are a couple of common issues when using COTS OCR:

- Text-text alignment

When multiple COTS OCR engines are used, it is critical to align the results before combination. That is because different OCR engines can locate slightly different text or the same text in different orders. Variants of dynamic programming techniques are effective to solve this problem [1][28][46].

- Text-image registration

As discussed earlier, sometimes the recognized text will be fused with the image-based method. For this purpose, we have to associate the text with the corresponding image region. The positions reported by the OCR engines may be based on the preprocessed images (for example, after de-skewing). Thus it is necessary to map those positions to the original coordinates [16][39].

- System-level fault tolerance

Although COTS OCR has already gone through the standard testing process for commercial software, it can still fail (We mean run-time exceptions instead of recognition errors here) for various reasons. This is especially noticeable in high-volume batch processing such as that of a digital library project. Since COTS OCR is provided as-is without source code, it is almost impossible to correct the particular failures. The best strategy is then to build system-level fault tolerance [40] so that the overall DRR system can always work in a predictable fashion even when the COTS OCR components fail occasionally.

- Programming APIs

One recurring problem with using COTS OCR SDKs is the lack of common APIs for OCR. Usually each vendor has its own proprietary APIs, which can also change from one version to another. So application developers have to learn the new APIs whenever switching to another SDK or even a new version. Besides, different APIs make it almost impossible to integrate low-level components from different OCR vendors (for example, use the layout analysis of one SDK and then recognize the text in the text regions with another SDK) to form a pipeline. We may borrow successful experience from the speech recognition community, in which a couple of open API standards, such as Microsoft SAPI [48] and Java JSAPI [49], are widely adopted.

4. CONCLUSIONS

Figure 1 summarizes how custom DRR and COTS OCR can work together in various applications. Based on this survey, we have the following observations:

- Novel DRR research still plays a critical role in DRR applications even as the basic OCR components become more and more commoditized.
- In most DRR applications, the basic character/word recognition is taken on by COTS software with the following exceptions:

In this paper, we have focused on applications dependent on the recognition of printed documents in Roman languages. If offline handwriting or languages not supported by COTS OCR is involved, research on character/word recognition is still needed. Besides, some promising paradigms such as document degradation models [41][42] and Character Shape Code [45] may lead to big leaps in the recognition rate, especially on low-quality documents and historical literature.

- System-level research with regard to reliability and guaranteed accuracy is in great need and can seldom be replaced by COTS software.
- Document-level structure understanding still has much room to expand.
- Document information retrieval also has many challenging research topics.

This survey has shown that custom DRR research is still in great need for better accuracy and reliability, complementary contents, or downstream information retrieval. Besides, the adoption of a common set of OCR APIs can greatly benefit the research and development based on COTS OCR packages.

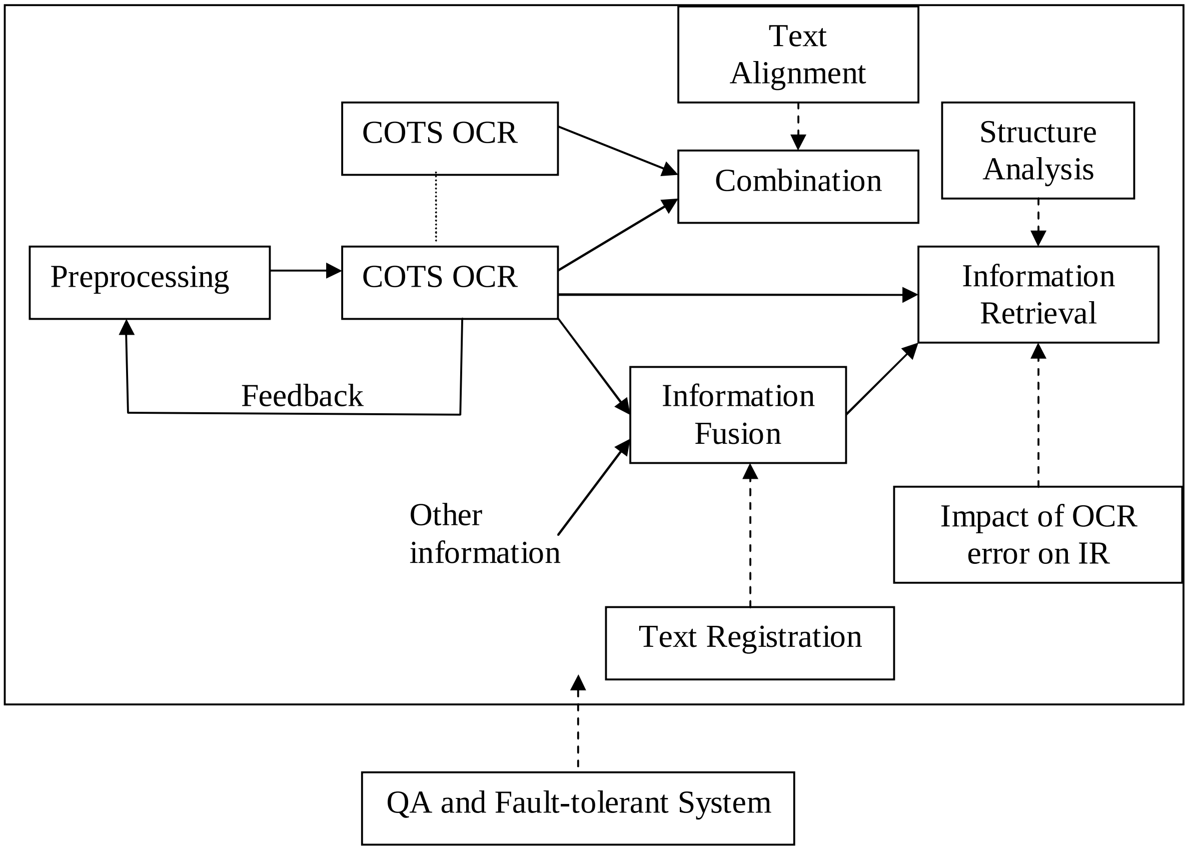

Figure 1: DRR research in connection with COTS OCR

ACKNOWLEGEMENTS

This survey is largely motivated by the author's research experience in HP Labs' DCRM project, in which many colleagues supported the author through technical collaboration, management, and logistics. The author is particularly grateful to the following DCRM team members: Joann Becker, Igor Boyko, John Burns, Douglas Heins, Sheelagh Hudleston, Steven Simske, Yan Xiong, and Sherif Yacoub. The author would also thank anonymous reviewers for their valuable comments.

REFERENCES

1. X. Lin, "Reliable OCR for Digital Content Re-mastering, Document Recognition and Retrieval IX," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 223-231, San Jose, 2002.
2. X. Lin, "Text-mining Based Journal Splitting," *Proc. International Conference on Document Analysis and Recognition 2003*, pp. 1075-1079, Edinburgh, UK, August 2003.
3. ABBYY Web Site, www.abbyy.com
4. ScanSoft Web Site, www.scansoft.com
5. I.R.I.S. Web Site, www.irisusa.com

6. X. Ding, D. Wen, et al., "Document Digitization Technology and Its Application for Digital Library in China," *Proc. of Document Image Analysis for Libraries*, pp. 46-53, Palo Alto, 2004.

7. European Metadata Engine Project, http://meta-e.aib.uni-linz.ac.at/ocr/ocr.html

8. S. Simske and X. Lin, "Creating Digital Libraries: Content Generation and Re-Mastering," *Proc. of Document Image Analysis for Libraries,* pp. 33-45, Palo Alto, 2004.

9. W. Barrett, L. Hutchison, et al, "Digital Mountain: From Granite Archive to Global Access," *Proc. of Document Image Analysis for Libraries*, pp. 104-121, Palo Alto, 2004.

10. S. Prateek, H. Baird, et al, "Triage of OCR Output Using 'Confidence' Scores," *Proc. SPIE Conference on Document Recognition and Retrieval IX,* pp. 20-25, San Jose, 2002.

11. D. Besagni and A. Belaid, "Citation Recognition for Scientific Publications in Digital Libraries," *Proc. of Document Image Analysis for Libraries*, pp. 244-252, Palo Alto, 2004.

12. F. Esposito, D. Malerba, et al, "Machine Learning Methods for Automatically Processing Documents: From Paper Acquisition to XML Transformation," *Proc. of Document Image Analysis for Libraries*, pp. 328-335, Palo Alto, 2004.

13. K. Taghva, J. Borsack, et al, "The Impact of Running Headers and Footers on Proximity Searching," *Proc. SPIE Conference on Document Recognition and Retrieval XI*, pp. 1-5, San Jose, 2004.

14. F. He, X. Ding, et al, "Hierarchical Logical Structure Extraction of Book Documents by Analyzing Table of Contents," *Proc. SPIE Conference on Document Recognition and Retrieval XI,* pp. 6-13, San Jose, 2004.

15. S. Mao, J. W. Kim, et al, "Style-independent Document Labeling: Design and Performance Evaluation," *Proc. SPIE Conference on Document Recognition and Retrieval XI*, pp. 14-22, San Jose, 2004.

16. N. Ozawa, H. Takebe, et al, "Slide Identification for Lecture Movies by Matching Characters and Images," *Proc. SPIE Conference on Document Recognition and Retrieval XI,* pp. 74-81, San Jose, 2004.

17. K. T. Shin, I. H. Jang, et al, "Block Adaptive Binarization of Business Card Images in PDA Using Modified Quadratic Filter," *Proc. SPIE Conference on Document Recognition and Retrieval XI,* pp. 92-101, San Jose, 2004.

18. H. Aradhye, "Syntax-directed Content Analysis of Videotext --- Application to a Map Detection and Recognition System," *Proc. SPIE Conference on Document Recognition and Retrieval X,* pp. 57-66, Santa Clara, 2003.

19. G. Maderlechner, P. Suda, et al, "Extraction of Valid Data Sets in Registers Using Recognition of Invalidation Lines," *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 67-72, Santa Clara, 2003.

20. S. Hauser, J. Schlaifer, et al, "Correcting OCR Text by Association with Historical Datasets," *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 84-93, Santa Clara, 2003.

21. T. Nartker, K. Taghva, et al, "OCR Correction Based on Document Level Knowledge," *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 103-110, Santa Clara, 2003.

22. R. Jin, C. X Zhai, et al, "Information Retrieval for OCR Documents: A Content-based Probabilistic Correction Model," *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 128-135, Santa Clara, 2003.

23. H. Ma and D. Doermann, "Bootstrapping Structured Page Segmentation," *Proc. SPIE Conference on Document Recognition and Retrieval X*, pp. 179-188, Santa Clara, 2003.

24. J. Liang and D. Doermann, "Content Features for Logical Document Labeling," Proc. SPIE Conference on Document Recognition and Retrieval X, pp. 189-196, Santa Clara, 2003.

25. E. Y. Du, P. D. Thouin, et al, "A Multistage Predictive Coding Approach to Unsupervised Text Detection in Video Images," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 12-19, San Jose, 2002

26. D. Doermann, H. Ma, et al, "Translation Lexicon Acquisition from Bilingual Dictionaries," *Proc. SPIE Conference on Document Recognition and Retrieval IX,* pp. 37-48, San Jose, 2002.

27. G. Thoma, and G. Ford, "Automated Data Entry System: Performance Issues," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 181-190, San Jose, 2002.

28. A. Belaid and L. Pierron, "A Generic Approach for OCR Performance Evaluation," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 203-215, San Jose, 2002.

29. P. Sarkar and H. Baird, "Triage of OCR Results Using 'Confidence' Scores," *Proc. SPIE Conference on Document Recognition and Retrieval IX*, pp. 216-222, San Jose, 2002.

30. J. Hull, B. Erol, et al, "Visualizing Multimedia Content on Paper Documents: Components of Key Frame Selection for Video Paper," *Proc. of International Conference on Document Analysis and Recognition 2003*, pp. 389-392, Edinburgh, 2003

31. A. Souza, M. Cheriet, et al, “Automatic Filter Selection Using Image Quality Assessment,” *Proc. of International Conference on Document Analysis and Recognition 2003*, pp. 508-512, Edinburgh, 2003.

32. L. Likforman-Sulem, P. Vaillant, et al, “Proper Names Extraction from Fax Images Combining Textual and Image Features,” *Proc. of International Conference on Document Analysis and Recognition 2003*, pp. 545-549, Edinburgh, 2003.

33. Z. Zhang and C. L. Tan, “Correcting Document Image Warping Based on Regression of Curved Text Lines,” *Proc. of International Conference on Document Analysis and Recognition 2003,* pp. 589-593, Edinburgh, 2003.

34. D. Doermann, J. Liang, and H. Li, “Progress in Camera-based Document Image Analysis,” *Proc. of International Conference on Document Analysis and Recognition 2003*, pp. 606-616, Edinburgh, 2003.

35. A.F.R. Rahman and M.C. Fairhurst, “Multiple Classifier Decision Combination Strategies for Character Recognition: A Review,” *International Journal of Document Analysis and Recognition*, **Vol 5**, pp.166-194, 2003.

36. K. Taghva, J. Borsack, and A. Condit, “Evaluation of Model-based Retrieval Effectiveness with OCR Text,” *ACM Transactions on Information Systems*, **Vol 14**, pp. 64-93, January 1996.

37. K. Taghva, J. Borsack, et al, “The Impact of Running Headers and Footers on Proximity Searching,” *Proc. SPIE Conference on Document Recognition and Retrieval XI*, pp. 1-5, San Jose, 2004.

38. X. Lin, “Impact of Imperfect OCR on Part-of-speech Tagging,” *Proc. of International Conference on Document Analysis and Recognition 2003*, pp. 284-288, Edinburgh, 2003.

39. X. Lin and J. Burns, “Using OCR in "Run Once, Use Everywhere" Manner to Facilitate Re-use in a Document Processing Workflow,” *Research Disclosure*, pp. 994-995, June 2002.

40. S. Yacoub, “Automated quality assurance for document understanding systems,” *IEEE Software*, May/June 2003, pp. 76-82.

41. H. Baird, “Document Image Defect Models,” *Proc. of IAPR Workshop on Syntactic and Structural Pattern Recognition*, pp. 38-46, June 1990.

42. T. Kanungo, R. Haralick, H. Baird, et al, “A Statistical, Nonparametric Methodology for Document Degradation Model Validation,” *IEEE Trans. Pattern Analysis Machine Intelligence*, **22(11)**, pp. 1209-1223, November 2000.

43. S. Klink, A. Dengel, and T. Kieninger, “Document Structure Analysis Based on Layout and Textual Features,” *Proc. of International Workshop on Document Analysis Systems*, 2000.

44. A. R. Dengel and B. Klein, “smartFIX: A Requirements-Driven System for Document Analysis and Understanding,” *Proc. of International Workshop on Document Analysis Systems*, pp. 433–444, 2002.

45. A. L. Spitz, “Tilting at Windmills: Adventures in Attempting to Reconstruct Don Quixote,” *Proc. of International Workshop on Document Analysis Systems*, pp. 51–62, 2004

46. Y. Nakano, T. Hananoi, et al, “A Document Analysis System Based on Text Line Matching of Multiple OCR Outputs,” *Proc. of International Workshop on Document Analysis Systems*, pp. 463–471, 2004

47. P. Viola, J. Rinker, and M. Law, “Automatic Fax Routing,” *Proc. of International Workshop on Document Analysis Systems*, pp. 484–495, 2004

48. Microsoft Speech SDK Web Site, http://www.microsoft.com/speech/

49. Sun Java Speech API Web Site, http://java.sun.com/products/java-media/speech/