Robust Document Image Understanding Technologies

Henry S. Baird
baird@cse.lehigh.edu

Daniel Lopresti
lopresti@cse.lehigh.edu

Brian D. Davison
davison@cse.lehigh.edu

William M. Pottenger
billp@cse.lehigh.edu

Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015, USA

ABSTRACT

No existing document image understanding technology, whether experimental or commercially available, can guarantee high accuracy across the full range of documents of interest to industrial and government agency users. Ideally, users should be able to search, access, examine, and navigate among document images as effectively as they can among encoded data files, using familiar interfaces and tools as fully as possible. We are investigating novel algorithms and software tools at the frontiers of document image analysis, information retrieval, text mining, and visualization that will assist in the full integration of such documents into collections of textual document images as well as "born digital" documents. Our approaches emphasize *versatility first*: that is, methods which work reliably across the broadest possible range of documents.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture; H.3.6 [**Information Storage and Retrieval**]: Digital Libraries: Collection, Systems issues

General Terms

Algorithms, Design, Experimentation, Management

Keywords

Document analysis, information retrieval, OCR error management

1. INTRODUCTION

The challenges faced by many industries and US government agencies in automating the capture, understanding, and reuse of scanned hardcopy documents include extremely high volumes of documents and a dauntingly wide variety of document types. High-accuracy OCR systems do not exist for many languages and writing systems due to the lack of commercial incentives to develop them. Also, many documents, when scanned, yield images of such low quality that conventional OCR systems fail almost completely. Moreover, later-stage processes, including retrieval and data mining, may be severely impacted by document analysis and OCR errors.

In the Department of Computer Science and Engineering at Lehigh University, we are studying many of these key issues. For example, in the past we have performed work on:

- high–accuracy OCR on low–quality document images [10, 47],
- robust retrieval from noisy text corpora by combining approximate string matching techniques with fuzzy logic [33, 31],
- document image quality modeling and applications of such models to the construction of high–performance OCR systems [1, 22],
- duplicate detection for scanned documents that have been subjected either to OCR [34] or character shape coding [32],
- the impact of recognition errors on document summarization [26],
- reduction of the knowledge engineering cost of textual information extraction [52], and
- building search engines and identifying topical locality within hyperlinked Web pages [16, 15]

Recently, we identified some of the most pressing issues confronting government agencies attempting to build and manage large collections of scanned document images [2].

2. RESEARCH DIRECTIONS

In this section, we discuss some of the research topics that we believe would help solve these problems and that we are capable of addressing.

2.1 "Versatility-First" DIA Research

One promising strategy for improving the performance of image understanding systems by the orders of magnitude that are needed is, we believe, to aim for *versatility first*. For decades the machine vision R&D community has optimized for high speed, and for high accuracy on some (often only a small) fraction of the input images,

HDP'04, November 12, 2004, Washington, DC, USA.

but only later — if at all — for versatility, by which we mean *guaranteed competence over a broad and precisely specified class of images.* As a result, vision technologies still fall far short of both human abilities and users' needs: they are overspecialized, brittle, unreliable, and improving only with painful slowness.

A versatility-first vision research program begins when we select a broad, challenging family of images: *e.g.*, all printed documents potentially containing any of many languages, scripts, page layout styles, and image qualities. Then, we investigate ways to:

- capture as much as possible of these images' variety in a formal generative (often stochastic) model that combines several submodels, *e.g.*, of image quality, layout, and language (this requires both analytical rigor and sophisticated statistical modeling, for significant progress towards this, cf. [27, 2]);
- develop methods for inferring the parameters of such models from labeled training data (which can be difficult even though there is a large relevant literature, *e.g.*, [28, 40]);
- design provably optimal recognition algorithms, for each submodel, and for the system as a whole, for best possible results with respect to the models (an intellectual challenge but sometimes doable, *e.g.*, [41, 42]);
- (only then) reduce run times to practical levels, carefully without loss of generality (this may require inventions but is almost always possible, *e.g.*, [37, 8, 10]);
- organize the system to adapt its model parameters to unlabeled test data, on the fly, and so retrain itself with a minimum of manual assistance (progress has been reported, in recent years, at RPI [46], Bell Labs [5], and PARC [9]); and
- construct 'anytime' recognition systems which, when allowed to run indefinitely, are guaranteed to improve accuracy monotonically to the best achievable, *i.e.*, consistent with the Bayes error of the problem (a daunting, exciting, but as yet almost untouched research domain).

Our experience inventing, building, testing, patenting, and applying systems of this type has convinced us of their promise — successes so far include:

- a world record in accuracy (99.995% characters correct) achieved by exploiting semantic as well as syntactic models of image content [6];
- a page reader that is quickly and easily 'retargetable' to new languages including Japanese, Bulgarian, and Tibetan [3];
- an automatically self-correcting classifier that cuts its own error rate by large factors without retraining, given merely a single hint [5];
- a high-accuracy tabular-data reader that, with only 15 minutes of clerical effort, can be trained to a new table-type, applied to over 400 different forms [48];
- a printed-text recognition technology, trainable with low manual effort, that maintains uniformly high accuracy over an unprecedentedly broad range of image qualities; and
- world-class web security technology (CAPTCHAs) able to block programs ('bots, spiders, etc) from abusing web services, by means of automated Turing tests that exploit the gap in ability between humans and machines in reading degraded images of text [4, 13].

2.2 Retrieving from Noisy Sources

Most published methods for retrieval of document images first attempt recognition and transcription followed by indexing and search operating on the resulting (in general, erroneous) encoded text using, *e.g.*, standard "bag-of-words" information retrieval (IR) methods. Early papers by Taghva, *et al.* show that moderate error rates have little impact on the effectiveness of traditional information retrieval measures for relatively long documents [49, 50]. The excellent survey by Doermann [19] summarized the state of the art (in 1997) of retrieval of entire multi-page articles as follows:

1. at OCR character error rates below 5%, these IR methods suffer little loss of either recall or precision; and
2. at error rates above 20%, both recall and precision degrade significantly.

A crucial open problem, which we are studying, is the effectiveness of "first OCR, then IR" methods on short passages such as, in an extreme but practically important case, fields containing key metadata (title, author, etc.). Within such passages, dictionary solutions may not help interpretation of arcane or unusual words (such as names of people and places). Approximate string matching techniques offer some promise for improving recall, as we have shown in earlier papers [33, 31].

To compare the behavior of traditional Boolean retrieval versus our proposed "fuzzy" methods, we performed a large-scale experiment involving a total of 59.6 million query evaluations and a database of 1,000 news articles gathered from the Internet. To simulate the output of an OCR process, we coded in C a Unix filter for generating errors based on a confusion matrix derived from analyzing a large corpus of real OCR output, yielding error patterns that appear authentic.

Curves for this experiment are presented in Figure 1. Approximate string matching exhibits an impressive degree of robustness in terms of recall. By the time the noise level approaches 20%, the traditional retrieval model is returning fewer than 50% of the true hits, while the fuzzy algorithm captures 95%. For documents that have suffered severe damage (noise levels of 60% or greater), the traditional approach misses over 90% of the hits, whereas the fuzzy method still returns over half.

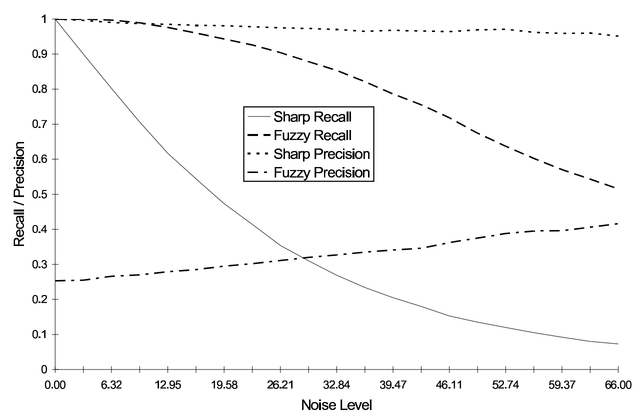

Figure 1: Results for Boolean retrieval under confusion matrix noise.

With regards to precision, the fuzzy model is understandably less selective: only about 30% of the hits it returns are "true." Still, it is often preferable to return a little too much data than to miss something important.

2.3 Summarizing Noisy Documents

In a recent paper [26], we examined some of the challenges in summarizing noisy documents. In particular, we broke down the summarization process into four steps: sentence boundary detection, preprocessing (part-of-speech tagging [35] and syntactic parsing), extraction, and post-editing [25]. We tested each step on noisy documents and analyzed the errors that arose, finding that these modules suffered significant degradation as the noise level in the document increased. We also studied how the overall quality of

Table 1: OCR performance relative to ground-truth (average precision and recall).

	Per-Character					
	All Symbols		Punctuation		Per-Word	
	Prec.	Recall	Prec.	Recall	Prec.	Recall
OCR.clean	0.990	0.882	0.869	0.506	0.963	0.874
OCR.light	0.897	0.829	0.556	0.668	0.731	0.679
OCR.dark	0.934	0.739	0.607	0.539	0.776	0.608
OCR.fax	0.969	0.939	0.781	0.561	0.888	0.879
OCR.skew	0.991	0.879	0.961	0.496	0.963	0.869

summarization was affected by the noise level and the errors made at each stage of processing.

The summarization pipeline was evaluated using documents exhibiting both synthetic and real noise. In the latter category were pages that had been printed, possibly degraded in some way, scanned at 300 dpi using a UMAX Astra 1200S scanner, and then OCR'ed with Caere OmniPage Limited Edition. These included:

clean The page as printed.

fax A faxed version of the page.

dark An excessively dark (but legible) photocopy.

light An excessively light (but legible) photocopy.

skew The clean page skewed on the scanner glass.

Note that because the faxed and photocopied documents were processed by running them through automatic page feeders, these pages can also exhibit noticeable skew.

In examining the accuracy of the OCR process using edit distance techniques [21], we determined that OCR performance varied widely depending on the type of degradation, as shown in Table 1. Punctuation symbols were particularly hard-hit due to their small size, which is critical because of their importance in delimiting sentence boundaries. For clean text, sentence boundary detection is not a big problem; the reported accuracy is usually above 95% [39, 44, 45]. However, since such systems typically depend on punctuation, capitalization, and words immediately preceding and following punctuation to make judgments about potential sentence boundaries, detecting sentence boundaries in noisy documents is a challenge due to the unreliability of such features.

We also found that syntactic parsers may be very vulnerable to noise in a document (Table 2). Even low levels of noise tended to lead to a significant drop in performance. For documents with high levels of noise, it may be better not to rely on syntactic parsing at all since it will likely fail on a large portion of the text, and even when results are returned, they will be unreliable.

Employing three measures used in the Document Understanding Conference [20] for assessing the quality of generated summaries, unigram overlap between the automatic summary and the human-created summary, bigram overlap, and the simple cosine, we evaluated the overall performance of our test summarization system. Not surprisingly, summaries of noisier documents generally had a lower overlap with human-created summaries (for full details, see [26]).

As our results showed, the methods we tested at every step were fragile, susceptible to failures and errors even with slight increases in the noise level of a document. Clearly, much work needs to be done to achieve acceptable performance in noisy document summarization. We need to develop summarization algorithms that do not suffer significant degradation when used on noisy documents. We also need to develop the robust natural language processing techniques that are required by summarization. These would include, for example, sentence boundary detection systems that can reliably identify sentence breaks in noisy documents.

Table 2: Percentage of sentences with incomplete parse trees from the ESG parser [35]. Sentence boundaries were first detected using two different tokenizers and individual sentences were given to ESG as input.

	Tokenizer 1	Tokenizer 2
Original	10%	5%
OCR.clean	2%	3%
OCR.light	46%	53%
OCR.dark	37%	43%
OCR.fax	37%	30%
OCR.skew	5%	6%

2.4 Automating Metadata Creation

A large portion of the expense and effort in bringing document images online is the extraction, correction and creation of metadata. In previous work we have developed techniques for classification of and automatic assignment of keywords to documents [14]; this work can be merged and extended with our work on information extraction techniques (*e.g.,* [52]) to aid in the automatic assignment of various types of textual metadata to document images.

In terms of correcting OCR errors in metadata, in the case of Lehigh University's "Digital Bridges" digital library [17], the librarians involved in the project estimate that complete correction of OCR errors in metadata took approximately 10 minutes per page, or six pages an hour; so for a 300 page book, a total of 50 hours was required [36]. Based on feedback we have received, there is no doubt that the need for extensive manual post-processing is regarded as a major hurdle in the construction of large collections from scanned document images.

2.5 Preserving Uncertainty

The creation of a large-scale repository from document images is likely to introduce many errors into the recognized text, which can degrade retrieval quality (as described in Section 2.3). Instead of enforcing the traditional boundary between OCR and IR, we would, in fact, like to preserve uncertainty throughout our system as much as possible. Doing so allows us to recognize where the system knows about possible errors, permitting better debugging, and possible incorporation of end-user correction and training. Given appropriate feedback about new content, recognition systems can be trained, thus improving their performance on similar future tasks. In general, manual correction of OCR'ed text is infeasible for large-scale efforts — the time and expense are too high. Instead, we believe that what is needed is to design and build a collaborative tool for editing and correction, providing valuable feedback to the underlying recognition model, both to train the system for future recognition tasks, but also to re-evaluate past uncertainty. Thus, the correction of one image from one page of a document could have a ripple effect throughout the document, and perhaps to other documents which had similar uncertainties.

Such a system will require work in a number of areas.

- A collaborative editing scheme. One possibility includes a community approval process, *a la* Slashdot [38]. A good editor will make corrections that are approved by others, increasing the editor's authority, thus decreasing the amount of confirmation required by others in the future. Another direction is simple redundancy in proofreading, as in Project Gutenberg's Distributed Proofreaders effort [18]. The types of editing/correction selected for a particular project will depend heavily on the type of collection and the audience to which it is presented.

- A strong dependency error model, so that when corrections are made, other scenarios with the same uncertainty can be quickly identified. In practice, it may be necessary to go further – to incorporate dependency information into all recognized text, not just uncertain text.

This aspect of our work requires a comprehensive end-to-end model in which document information is managed from initial imaging through OCR, indexing, and end-user presentation.

2.6 Automated Creation of Hypertext Links for Document Images

The presentation of imaged documents from a digital repository should provide at least the functionality of the original documents, and where appropriate, provide improvements that digital representation makes possible.

Text that references or discusses figures or other documents is an excellent candidate for innovative linking and hypertext navigation. Akin to previous work in information extraction from captions in document images (*e.g.*, [23, 24]), we plan to identify, extract, and index such text, in addition to recognizing and indexing *within-image* text and explicit captions. This allows us to make non-text images (*e.g.*, figures, plates) retrievable using text queries (in contrast to most content-based retrieval techniques [51, 53]).

Recognizing textual content that discusses a figure or image would also be useful in deciding to include an image for automated summarization purposes, and finding the first such reference can assist in re-flowing a document for better presentation (see, for example, [11]).

Prior work has focused on the automatic recognition and extraction of scholarly citations (*e.g.*, CiteSeer/ResearchIndex [29, 30]) but has not incorporated the discussion text as part of the cited document. On the Web, in contrast, search engines routinely associate the content in links both to the source document and to the cited Web page [12], since such text is a good descriptor of the target document [15]. This is exploited by motivated Web authors for search engine manipulation and for what is known as Google-bombing [7] — creating enough links with common anchor text to a particular site to place that site at or near the top of the rankings when the anchor text is used as the query.

When indexing images, the major Web search engines use some available text. They all use text within the URL of the image, but some go further. Google's image search is presently capable of using image captions; it also uses page form text (*e.g.*, pull-down menus), but not general text (or titles, etc.). AltaVista's image search, in contrast, apparently uses text from the citing page, which allows for many more matches, but also includes many poor matches.

The accurate selection of relevant text will make an otherwise irretrievable figure accessible via a search engine. We plan to make use of text mining techniques; in particular, given labeled examples of the kinds of references we wish to find, information extraction algorithms can be trained to recognize new occurrences (*e.g.*, as in [52]).

3. CONCLUSIONS

We have touched on several key areas of our research agenda. Collectively we are fortunate to have the experience to span the hardcopy document processing domain from document capture through information retrieval and text mining. As is often the case, it is at the boundaries of the various stages of hardcopy document processing that work in end-to-end systems development is needed. Given the nature of our collective experience, it is precisely in these 'transitional' areas of end-to-end processing systems that we feel most capable of making significant contributions.

Acknowledgements

Portions of the work described in this paper were performed at Bell Labs, Palo Alto Research Center, Matsushita Information Technology Laboratory, Rutgers University, and the National Center for Supercomputing Applications. We gratefully acknowledge the assistance of our co-authors and other colleagues in accomplishing this work. Co-author William M. Pottenger also gratefully acknowledges His Lord and Savior, Yeshua (Jesus) the Messiah for His work of salvation in his life.

4. REFERENCES

[1] H. Baird. Model-directed document image analysis. In *Proceedings of the DOD-sponsored Symposium on Document Image Understanding Technology (SDIUT 1999)*, pages 42–49, Annapolis, Maryland, April 1999.

[2] H. Baird, K. Popat, T. Breuel, P. Sarkar, and D. Lopresti. Assuring high-accuracy document understanding: Retargeting, scaling up, and adapting. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 17–29, Greenbelt, MD, April 2003.

[3] H. S. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80(7):1059–1065, July 1992.

[4] H. S. Baird, A. L. Coates, and R. Fateman. PessimalPrint: a reverse Turing test. *Int'l J. on Document Analysis and Recognition*, 5:158–163, 2003.

[5] H. S. Baird and G. Nagy. A self-correcting 100–font classifier. In *Proceedings, IS&T/SPIE Symposium on Electronic Imaging: Science & Technology*, pages 106–115, San Jose, CA, February 1994.

[6] H. S. Baird and K. Thompson. Reading chess. *PAMI*, 12(6):552–559, 1990.

[7] 'Miserable failure' links to Bush: George W Bush has been Google bombed. BBC News, Dec. 2003. http://news.bbc.co.uk/2/hi/americas/3298443.stm.

[8] D. Bloomberg, T. Minka, and K. Popat. Document image decoding using iterated complete path search with subsampled heuristic scoring. In *Proceedings of the IAPR 2001 International Conference Document Analysis and Recognition (ICDAR 2001)*, Seattle, WA, September 2001.

[9] T. Breuel. Modeling the Sample Distribution for Clustering OCR. In *SPIE Conference on Document Recognition and Retrieval VIII*, 2001.

[10] T. Breuel and K. Popat. Recent work in the document image decoding group at xerox parc. In *Proceedings of the DOD-sponsored Symposium on Document Image Understanding Technology (SDIUT 2001)*, Columbia, Maryland, April 2001.

[11] T. M. Breuel, W. C. Janssen, K. Popat, and H. S. Baird. Paper to PDA. In *Proc., IAPR 16th ICPR*, pages Vol. 4, 476–479, Quebec City, Canada, August 2002.

[12] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, Apr. 1998.

[13] M. Chew and H. S. Baird. Baffletext: a human interactive proof. In *Proc., 10th IS&T/SPIE Document Recognition & Retrieval Conf.*, Santa Clara, CA, January 2003.

[14] Y.-M. Chung, W. M. Pottenger, and B. R. Schatz. Automatic subject indexing using an associative neural network. In *Proceedings of the ACM Digital Libraries 98 Conference*, pages 59–68, Pittsburgh, PA, 1998.

[15] B. D. Davison. Topical locality in the Web. In *Proceedings of the 23rd Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR 2000)*, pages 272–279, Athens, Greece, July 2000.

[16] B. D. Davison, A. Gerasoulis, K. Kleisouris, Y. Lu, H. Seo, W. Wang, and B. Wu. DiscoWeb: Applying link analysis to Web search. In *Poster proceedings of the Eighth International World Wide Web Conference*, pages 148–149, Toronto, Canada, May 1999.

[17] Digital Bridges. Lehigh University Libraries, April 2004. http://bridges.lib.lehigh.edu/.

[18] Distributed Proofreaders. Project gutenberg's distributed proofreaders home page. http://www.pgdp.net/, 2004.

[19] D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3), June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).

[20] Document Understanding Conference (DUC): Workshop on Text Summarization, 2002. http://tides.nist.gov/.

[21] J. Esakov, D. P. Lopresti, and J. S. Sandberg. Classification and distribution of optical character recognition errors. In *Proceedings of Document Recognition I (IS&T/SPIE Electronic Imaging)*, volume 2181, pages 204–216, San Jose, CA, February 1994.

[22] T. K. Ho and H. S. Baird. Large-scale simulation studies in image pattern recognition. *IEEE Trans. on PAMI*, 19(10):1067–1079, October 1997.

[23] Y. Ishitani. Model-based information extraction method tolerant of OCR errors for document images. *Int. J. Comput. Proc. Oriental Lang.*, 15(2):165–186, 2002.

[24] Y. Ishitani. Document transformation system from papers to XML data based on pivot XML document method. In *Proceedings of the 7th International Conference on Document Analysis and Recognition ICDAR 2003*, Edinburgh, Scotland, Aug. 2003.

[25] H. Jing. *Cut-and-paste Text Summarization.* PhD thesis, Department of Computer Science, Columbia University, New York, NY, 2001.

[26] H. Jing, D. Lopresti, and C. Shih. Summarizing noisy documents. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 111–119, Greenbelt, MD, April 2003.

[27] T. Kanungo, H. Baird, and R. Haralick. Estimation and validation of document degradation models. In *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, April 1995.

[28] G. Kopec. An em algorithm for character template estimation. Submitted March 1997; returned for revision, but not revised due to the author's death; available from PARC by request.

[29] S. Lawrence, K. Bollacker, and C. L. Giles. Autonomous citation matching. In O. Etzioni, editor, *Proceedings of the Third International Conference on Autonomous Agents*, New York, 1999. ACM Press.

[30] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[31] D. Lopresti. Robust retrieval of noisy text. In *Proceedings of the Third Forum on Research and Advances in Digital Libraries*, pages 76–85, Washington, DC, May 1996.

[32] D. Lopresti and A. L. Spitz. Comparing the utility of optical character recognition and character shape coding in duplicate document detection. In *Proceedings of the Fourth IAPR International Workshop on Document Analysis Systems*, pages 439–450, Rio de Janeiro, Brazil, December 2000.

[33] D. Lopresti and J. Zhou. Retrieval strategies for noisy text. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, pages 255–269, Las Vegas, NV, April 1996.

[34] D. P. Lopresti. A comparison of text-based methods for detecting duplication in scanned document databases. *Information Retrieval*, 4(2):153–173, July 2001.

[35] M. McCord. *English Slot Grammar*. IBM, 1990.

[36] P. Metzger. Private communication, April 2004.

[37] T. P. Minka, D. S. Bloomberg, and K. Popat. Document image decoding using iterated complete path heuristic. In *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII*, San Jose, CA, January 2001.

[38] Open Source Technology Group, Inc. Slashdot home page. http://slashdot.org/, 2004.

[39] D. Palmer and M. Hearst. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23(2):241–267, June 1997.

[40] K. Popat. Decoding of text lines in grayscale document images. In *Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, May 2001. IEEE. To appear.

[41] K. Popat, D. Bloomberg, and D. Greene. Adding linguistic constraints to document image decoding. In *Proc., 4th International Workshop on Document Analysis Systems*, Rio de Janeiro, Brazil, December 2000. International Association of Pattern Recognition.

[42] K. Popat, D. Greene, J. Romberg, and D. S. Bloomberg. Adding linguistic constraints to document image decoding: Comparing the iterated complete path and stack algorithms. In *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII*, San Jose, CA, January 2001.

[43] T. M. Rath and R. Manmatha. Features for word spotting in historical manuscripts. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 218–222, August 2003.

[44] J. C. Reyner and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., 1997.

[45] M. Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 339–352, Cape Cod, MA, 1989.

[46] P. Sarkar. *Style Consistency in Pattern Fields.* PhD thesis, Rensselaer Polytechnic Institute, May 2000.

[47] P. Sarkar, H. S. Baird, and X. Zhang. Training on severely degraded text–line images. In *Proc., Intl. Conf. on Document Analysis and Recognition*, Edinburgh, Scotland, August 2003.

[48] J. H. Shamilian, H. S. Baird, and T. L. Wood. A retargetable table reader. In *Proceedings of ICDAR*, pages 1–14, Ulm, Germany, August 1997.

[49] K. Taghva, J. Borsack, and A. Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing and Management*, 32(3):317–327, 1996.

[50] K. Taghva, J. Borsack, and A. Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems*, 14:64–93, January 1996.

[51] R. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. Technical Report UU-CS-2000-34, Dept. of Computing Science, Utrecht University, Oct. 2000. http://citeseer.ist.psu.edu/veltkamp00contentbased.html.

[52] T. Wu and W. M. Pottenger. A semi-supervised active learning algorithm for information extraction from textual data. *Journal of the American Society for Information Science and Technology*, 2004. In press.

[53] A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(1), Jan.–Feb. 1999.