Information Retrieval and OCR: From Converting Content to Grasping Meaning

Jamie Callan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15241, USA

callan@cmu.edu

Paul Kantor
School of Communication,
Information and Library Studies
Rutgers University
New Brunswick, NJ 08901, USA
kantor@scils.rutgers.edu

David Grossman
Computer Science Dept
Illinois Institute of Technology
Chicago, IL 60616, USA

grossman@iit.edu

Introduction

IR and OCR have largely developed independent standards and metrics, with OCR focused on literal accuracy, and IR focused on essential "content/meaning". With more and more media not only paper, but in multiple image formats, the opportunities and challenges for OCR on new formats – video and still images – are enormous. While OCR is assessed in metrics that emphasize words and characters, IR has learned to apply end-to-end metrics that ask whether the needs of the users can be met by existing systems. The same considerations apply also to the problem of providing permanent worldwide access to millions of pages of legacy print documents, representing the shared human record as it existed until just a few years ago.

The International Society for Optical Engineering (SPIE) has held a series of Document Recognition and Retrieval (DRR) conferences. The tenth, DRR X will be held in January 2003, in Santa Clara California. In 2001, Dan LoPresti of Bell Labs decided that the area would benefit from more intense collaboration between those who specialize in finding the words on a page image, and those researchers who know how to find the right documents, given the words. He invited Paul Kantor (Rutgers) to join the DRR Chairs, and together they invited Dave Lewis (Consultant) to give a keynote address at DRR VIII. Dan then stepped down. Paul chaired DRR IX (2002) and then handed the reins to Tapas Kanungo (IBM, Almaden) and together they invited Jamie Callan (CMU), David Grossman (IIT) and Alex Hauptmann (CMU) to join the conference committee for DRR X.

To improve communication between SIGIR and DRR, this group proposed a SIGIR workshop on this area. The workshop on "Information Retrieval and OCR: From Converting Content to Grasping Meaning" was intended to stimulate cross-fertilization between OCR and IR, in hopes that better use of IR will enable the OCR community to avoid expensive hand processing, and to demonstrate that the combination of present static and dynamic image processing and present state-of-the-art robust information retrieval can generate substantial advances in both extraction of messages from image streams and conversion of existing paper variants. It solicited papers dealing with future applications, such as the indexing and retrieval of text embedded in static or video graphic images, with problems of skew, distortion, and obscuration, as well as state-of-the-art discussions of the storage and retrieval of handwritten or print legacy materials.

The workshop was held on August 15, 2002 in Tampere, Finland, immediately following the SIGIR 2002 conference. Although the workshop was intended to appeal to a wide range of IR and OCR researchers (and indeed was proposed at the request of colleagues from the OCR community), it primarily drew people with a background in IR. About a dozen people participated. The small size allowed a very interactive, seminar-style format and very vigorous discussion between and during presentations. Most presentations ran 30% to 50% longer than planned, and our impression is that most of the participants found it very productive.

Presentations

Below we provide very brief descriptions of the workshop presentations, to give a sense of the range of themes and topics covered. The complete workshop proceedings are available in electronic form at http://www.cs.cmu.edu/~callan/Workshops/IR-OCR-02/.

Paul Kantor's talk "*Introduction: Links to Other Groups, Especially Document Recognition and Retrieval Conference X*" opened the workshop and identified key themes and discussion points for the remainder for the day. IR research tends to focus on identifying organizational structure, content, meaning, and relevance, whereas OCR research tends to focus on layout structure, characters, and clean and non-traditional text images. Both IR and OCR can be viewed as a type of "noisy channel" problem, and language models are a point of overlap between the two fields. Both fields study how to smooth, augment, or correct a document representation using a variety of dictionary-based and corpus-based techniques. The two communities collaborated in the early 1990s on search of OCR collections, and results were generally good. There is a perception among some in the IR community that retrieval from OCR collections is now a solved problem, but for non-English text, environments requiring high accuracy, and tasks such as question answering, the problem is far from solved. A combination of more accurate OCR and more robust IR is required.

David Grossman set the stage for discussion with his talk "*Retrieving OCR Text: A Survey of Current Approaches*". He reviewed past academic research, and some of the workshop participants with commercial OCR experience contributed insights from the commercial world. A key point to emerge from the discussion is that there is a mismatch in the experimental methodologies of the two communities. The IR community is skeptical of any collection smaller than a hundred thousand documents; for an OCR task, the IR community would also want "ground truth" versions of most documents (at least the relevant documents). The OCR community is skeptical of "simulated" collections, for example collections taken by "corrupting" ASCII text using an OCR error model; the OCR community prefers real paper scanned by real machines, which makes creating large-scale collections, and generating "ground truth" equivalents, very expensive. Documents created electronically are a potential solution, because ground truth is available, and paper can be printed (if necessary) and scanned back in.

Kevyn Collins-Thompson's talk "*A Clustering-Based Algorithm for Automatic Document Separation*" described work done at Microsoft Research on identifying document boundaries in a stream of page images produced by a bulk OCR device. A variety of problems at the intersection of OCR and IR arise in such an application, and the software must compensate to the extent possible. For example, people make characteristic errors such as accidentally rescanning a stack of pages (resulting in duplicates), putting a stack of pages in backwards (resulting in a reversed order), or scanning the "odd" pages of a double-sided document followed by the "even" pages. Page headers and footers (e.g., dates, page numbers, running headings) play an important role in this application, so OCR errors in these regions of the document can have a significant effect on how accurately document boundaries are found. Kevyn described a three-stage architecture that delivered reasonably accurate results in preliminary tests.

Rong Jin presented "*A Content-based Probabilistic Correction Model for OCR Document Retrieval*" in which characteristic OCR errors are incorporated directly into a statistical language model used for ad-hoc document retrieval. The language model for a scanned document is adjusted using OCR output and an automatic spelling-correction algorithm. The adjusting procedure forces the document towards self-consistency in selecting among spelling correction candidates, i.e., candidates that already appear in the document are more likely to appear than candidates that do not. The result is a document language model (*not* a spelling-corrected document) that can be used for document retrieval. Experiments with TREC OCR track data ("clean" documents that were "corrupted" according to an OCR error model) showed that the

method works well. The method can be extended to a variety of other "noisy document" tasks such as cross-lingual information retrieval, multi-media information retrieval, and spoken-document retrieval.

Tomi Klein's talk "*A Voting System for Automatic OCR Correction*" argued for the use of multiple OCR algorithms in recognizing text because different algorithms assess probabilities differently and tend to make different errors. "Local" (OCR-specific) and "global" (database-specific) dictionaries are formed, providing multiple forms of evidence for a classification procedure that selects the "right" word. One novel feature is that the algorithm can recognize as "correct" words, such as personal names, that don't appear in any of its dictionaries. This approach to OCR reduces the error rate on Hebrew characters from 12-14% to 3-4%. Improvements for English and French text are less dramatic, in part because the baseline accuracy for these languages is already very high. One conclusion is that this approach may be most effective on "difficult" languages such as Hebrew and Arabic.

Adenike Lam-Adesina presented "*Examining the Effectiveness of IR Techniques for Document Image Retrieval*" which studied the use of automatic relevance feedback on OCR documents. Experiments were done on the TREC SDR collection; an OCR collection was created by formatting the ASCII text using a variety of formats, printing it, and scanning it back. The initial word error rate was 15-20%, making it a very difficult collection. The initial results with automatic relevance feedback were weak because noisy corpus statistics produced poor term weights for expansion terms. Collection enrichment with a comparable corpus "smoothed" the corpus statistics from the OCR collection, making automatic relevance feedback on an OCR collection almost as accurate as it was on the "clean" baseline.

Yuen-Hsien Tseng of Fu Jen Catholic University gave a brief, impromptu presentation on a new OCR test collection for Chinese text. The collection is based on 8,438 pages of news clippings from various sources in Hong Kong, Taiwan, and the People's Republic of China. It contains the 8,438 pages of news clippings, 8,438 OCR images (69% character accuracy), and exhaustive relevance judgements for 30 topics. Each of the 899 relevant documents was also manually transcribed, providing "ground truth" texts for those documents. Yuen-Hsien Tseng can be reached by email at tseng@blue.lins.fju.edu.tw.

Conclusions

Participants agreed that the problem is important, with at least three important application areas: (1) preservation of the world's written cultural heritage, which will increasingly be ignored if it is not machine readable; (2) online access to technical and maintenance materials for hardware (such as airplanes) whose lifetime is measured in decades; and (3) improved interpretation of text in moving and static images, with implications for archiving, indexing, and surveillance.

Several themes emerged in the workshop discussions. Participants agreed that it will be difficult for the IR and OCR research communities to collaborate until a common experimental methodology is adopted. One possibility is the creation of large-scale OCR collections from documents "born digital", as was done by Jones and Lam-Adesina. Such collections can be assembled (relatively) cheaply, and can provide characteristics needed by the IR community, such as well-defined topics, relevance judgements, "ground truth" versions of each document, and a collection that can be studied across many levels of degradation.

Another possible source of data is the degraded images NIST created for the TREC-4 and TREC-5 Confusion tracks. If NIST can make available the confusion matrix, then researchers can evaluate their systems not only in terms of end-to-end document retrieval, but also in terms of the accuracy of the inferred confusion matrix (if the model makes such a matrix explicit).

Another approach to the development of a test collection, on a scale more akin to the usual TREC scale is to use the Reuters collection, interpreting intersections of heading terms as "de facto" topics (this is being tried for the first time in the TREC 2002 Filtering track). Then degraded texts can be produced inexpensively by applying a confusion matrix, or at moderate cost by printing the documents in one or more fonts, scanning them, degrading the images, and applying OCR. A document would be considered relevant only to the topics for which it has the assigned heading terms.

Participants were uncertain whether such collections would be viewed as credible and interesting by OCR researchers.

Good search engines are freely available for anyone wishing to study information retrieval. Examples include MG, Lucene, and Lemur. None of the workshop participants were aware of a comparable OCR resource, i.e., a single OCR engine with an accessible language model and easy access to the important parameters. This deficit was identified as a key obstacle to IR research on OCR collections. Such an engine would make it easier for IR researchers to experiment with a wider range of scanned document types, which would result in more robust solutions.

Multilingual documents are of increasing interest to the IR community, and these would appear to present interesting challenges to the OCR community as well. Klein and Kopel's work suggest that accurate OCR can sometimes benefit from language-specific tuning, raising the possibility of a single document that has differing levels of OCR accuracy in different regions of the document. There is little published IR research on documents in which the "noise" level varies across the document.

Finally, collaboration between the IR and OCR research communities would benefit from a small set of well-defined task scenarios that address issues interesting to both communities. Such scenarios would provide a focal point for developing the common experimental methodologies and common approaches to developing datasets that would link the two communities again.

As mentioned in the Introduction, this workshop was organized in response to encouragement from OCR researchers. The workshop would be well worth repeating in Toronto if more members of the OCR community can be attracted to it. This year the problems of the high-tech sector and the difficulty of justifying international travel may have contributed to the lack of such participation on the part of US researchers, who constitute the majority of the DRR community. Interested IR researchers may also want to attend DRR X, to establish research contacts with that community.

The issues raised at this workshop will be presented to the Document Recognition and Retrieval Conference X in January, 2003. We hope that there will be a response describing the interesting issues as viewed from the OCR and DRR research communities. The papers in this workshop are a starting point, but we hope that they are *just* a starting point.

Acknowledgements

We thank the organizers of the SIGIR conference for their excellent support of the workshop, and the Tampere Chamber of Commerce for spectacular weather.